Dynamic Hyperparameter Optimization for Bayesian Topical Trend Analysis

Tomonari Masada
Nagasaki University
1-14 Bunkyo-machi
Nagasaki, Japan
masada@cis.nagasaki-u.ac.jp

Daiji Fukagawa
National Institute of Informatics
2-1-2 Hitotsubashi
Chiyoda-ku, Tokyo
daiji@nii.ac.jp

Atsuhiro Takasu
National Institute of Informatics
2-1-2 Hitotsubashi
Chiyoda-ku, Tokyo
takasu@nii.ac.jp

Tsuyoshi Hamada
Nagasaki University
1-14 Bunkyo-machi
Nagasaki, Japan
hamada@cis.nagasaki-u.ac.jp

Yuichiro Shibata
Nagasaki University
1-14 Bunkyo-machi
Nagasaki, Japan
shibata@cis.nagasaki-u.ac.jp

Kiyoshi Oguri
Nagasaki University
1-14 Bunkyo-machi
Nagasaki, Japan
oguri@cis.nagasaki-u.ac.jp

ABSTRACT
This paper presents a new Bayesian topical trend analysis. We regard the parameters of topic Dirichlet priors in latent Dirichlet allocation as a function of document timestamps and optimize the parameters by a gradient-based algorithm. Since our method gives similar hyperparameters to the documents having similar timestamps, topic assignment in collapsed Gibbs sampling is affected by timestamp similarities. We compute TFIDF-based document similarities by using a result of collapsed Gibbs sampling and evaluate our proposal by link detection task of Topic Detection and Tracking.

Categories and Subject Descriptors: I.2.6 [Artificial Intelligence]: Learning; H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval

General Terms: Algorithms, Experimentation

Keywords: Temporal Analysis, Topic Modeling, Topic Detection

1. INTRODUCTION
Bayesian approach is an outstanding trend in data mining. Especially, latent Dirichlet allocation (LDA) [4] leads to challenges in various fields [5][7][8][14]. In this paper, we focus on topical trend analysis and propose a new approach, called Latent dYNamically-parameterized Dirichlet Allocation (LYNDA), by regarding the parameters of topic Dirichlet priors in LDA as a function of document timestamps. In LYNDA, a topic multinomial for each document is drawn from a Dirichlet prior whose parameters are an instantiation of a per-topic function defined over document timestamps. However, if topic Dirichlet priors are markedly different among documents, we cannot fully take advantage of Bayesian approach, because too different per-document topic multinomials result in overfitting to word co-occurrence patterns local to each document. Therefore, we regard the parameters of topic Dirichlet priors as a smooth function of continuous timestamps. Consequently, documents having similar timestamps are generated based on similar Dirichlet priors and exhibit similar topic mixture. LYNDA has an advantage in its simplicity. We can use collapsed Gibbs sampling (CGS) for LDA [6] as is and can estimate hyperparameters in the course of CGS by using an existing gradient-based method such as L-BFGS [11][9]. CGS for LDA provides a topic assignment to all word tokens and induces predictive word probabilities conditional on each document [16]. Since document similarities can be computed based on these probabilities, we compare LYNDA with competing methods by link detection task of Topic Detection and Tracking (TDT), where better document similarity leads to a better result. Note that our approach may be applied to other probabilistic models where Dirichlet priors are prepared for timestamped data.

2. PREVIOUS WORKS, OUR PROPOSAL
In this paper, we focus on probabilistic topical trend analysis. Dynamic Topic Models (DTM) [3] and its continuous time version (cDTM) [13] utilize transitions of the parameters of per-topic word multinomials for modeling document temporality, where the vectors drawn from time-dependent Gaussians are used to obtain multinomial parameters. However, Gaussian is not a conjugate to multinomial. Therefore, the authors of [10] discuss that inference becomes complicated and propose Multiscale Topic Tomography Models (MTTM) based on a completely different idea. In MTTM, the time interval is segmented into two pieces recursively to obtain a binary tree representing the inclusion among subintervals. Further, each leaf node is associated with a Poisson distribution for word generation. However, MTTM cannot be applied to continuous timestamps. When compared with these works, LYNDA is remarkable in two features below.
First, inference is easy to implement. While LYND\(A\) requires additional computations for hyperparameter estimation, we can use CGS for LDA as is and can adopt an existing gradient-based method for hyperparameter estimation. In contrast, the above works require heavily customised implementation. Second, LYND\(A\) can be applied to continuous timestamps. While cDTM has this feature, a special implementation technique is required for efficient memory usage.

More appropriately, we can compare LYND\(A\) with Topics over Time (TOT) [15], because TOT shares both above features. TOT extends LDA just by adding per-topic Beta distributions, which are defined over continuous timestamps. However, these Beta distributions cause the following problem. The full conditional probability that topic \(k\) is assigned to \(i\)th token of document \(j\) in TOT is proportional to 

\[
\frac{(n_{ki} + \alpha_k)(n_{kj} + \beta_j)}{\sum_{i=1}^K (n_{ki} + \alpha_k)} \times \frac{\Gamma(a_k + b_j)}{\Gamma(a_k)\Gamma(b_j)} \left(\frac{\alpha_k}{\beta_j}\right)^{k-1} (1 - \tau_j)^{b_j-1},
\]

where \(\Gamma(.)\) denotes the standard gamma function, and \(\tau_j\) is the timestamp of document \(j\). The first half corresponds to the full conditional in LDA [6]. The latter corresponds to the likelihood of Beta distribution and means that the same topic is likely to be assigned to the word tokens appearing in the documents having similar timestamps. Our preliminary experiments reveal that this likelihood grows unboundedly as CGS proceeds, and that the first half comes to play only a marginal role. To solve this problem, we should keep the likelihood comparable with the first half. In our evaluation experiment, we multiply all Beta parameters \(a_k, b_j\) by the same constant and keep them less than a specific limit.

In contrast, LYND\(A\) controls time-dependency of topic assignment not by augmenting LDA with probability distributions without priors, but by making hyperparameters time-dependent to achieve a moderate fit to timestamp data. In LYND\(A\), we make topic Dirichlet priors time-dependent by defining 

\[
p(\theta_j | f_1, \ldots, f_K) \propto \prod_k \theta_{jk}^{(\tau_j)-1}
\]

for document \(j\), where \(f_k(\cdot)\) are per-topic differentiable functions defined over continuous timestamps. In this paper, we adopt unnormalized Gaussian densities and define 

\[
f_k(\tau) = \xi_k \exp\left(-\frac{(\tau - m_k)^2}{2\xi_k^2}\right)
\]

We think that Gaussian density is appropriate to represent the dynamism of topical trends. Consequently, each topic Dirichlet prior is determined by the three parameters: \(m_k\) specifies the position of the peak popularity of topic \(k\), \(s_k\) cells how large its popularity changes, and \(\xi_k\) represents the intensity of popularity. We introduce \(\xi_k\) to automatically rescale the unnormalized Gaussian density separately for each topic \(k\). These parameters are estimated by empirical Bayes method. The log likelihood of a topic assignment can be obtained as 

\[
\log \left( \prod_k \frac{\Gamma(\sum f_k(\tau_j))}{\Gamma(\alpha_k)\Gamma(f_k(\tau_j))} \right)
\]

We maximize this by optimizing \(\xi_k, m_k, s_k\), and \(\alpha_k\) with L-BFGS [11][9]. However, in the preliminary experiments, \(m_k\) showed no significant change after optimization. Therefore, we fix \(m_k\) to \((k - 1)/(K - 1)\) by normalizing the timestamp domain to \([0, 1]\). As a result, we can put an equal interval between the peaks of neighboring Gaussians and thus can cover the entire time interval impartially.

We compare LYND\(A\) also with LDA where hyperparameters are optimized straightforwardly. The log likelihood of a topic assignment in LDA is obtained as 

\[
\log \left( \prod_k \frac{\Gamma(\sum_{n_{ij} + \alpha_k})}{\Gamma(\sum_{\sum_{n_{ij} + \alpha_k}})} \right)
\]

We compare LYND\(A\) with five methods: TFIDF, LDA with fixed hyperparameters (\(\alpha_k = 0.5\) for all \(k; \beta = 0.01\)), HypLDA, TOT2, and TOT5. The number of topics \(K\) is set to \(100\). While we tested \(K = 50\) and \(200\), we only obtained worse results for \(K = 50\) and comparable results for \(K = 200\). We prepare 30 runs of CGS from different random initializations. The number of iterations of CGS is \(500\) for LDA, TOT2, and TOT5. In the course of CGS for HypLDA and LYND\(A\), we estimate hyperparameters by L-BFGS once for every 10 iterations. HypLDA and LYND\(A\) require 1,000 iterations, because convergence is slow due to the incorporation of hyperparameter estimation.

Our evaluation is conducted as follows. Assume that the
number of on-topic documents for a TDT-topic \( t \) is \( T_t \). By computing the similarities between each on-topic document and all documents, we have similarities for \( T_t \times T_c \) document pairs. Among these pairs, we call \( T_t \times T_c \) pairs of two on-topic documents “correct” and the rest “incorrect.”

We would like to approximate this ideal split by devising a document similarity and then by setting a similarity threshold to obtain a split. We introduce two functions of similarity threshold \( \eta \), \( R_c(\eta) \) is the number of correct pairs whose similarities are larger than \( \eta \), \( A_c(\eta) \) is the number of incorrect pairs whose similarities are larger than \( \eta \). Now we can define two evaluation measures: miss probability \( P^M(\eta) \equiv 1 - R_c(\eta)/(T_t \times T_c) \) and false alarms probability \( P^A(\eta) \equiv A_c(\eta)/(T_t \times (J - T_j)) \). We also use Normalized Detection Cost (NDC) defined as \( P^M(\eta) + 4.9 \times P^A(\eta) \) based on an intuition that false alarms are more unfavorable [2][12].

4. RESULTS

Figure 1 presents NDCs for the similarity thresholds from 0.040 to 0.055 with 0.001 step. The top panel gives NDCs averaged over 40 TDT-topics of TDT 2002, and further averaged over 30 results of CGS except for TFIDF. The bottom panel gives the results for TDT 2003. Based on Figure 1, we choose \( \eta = 0.04 \), because smaller values are advantageous to TDT 2002, and larger values are to TDT 2003. Figure 1 shows that TFIDF is efficient for TDT 2003. A detailed inspection reveals that TFIDF is likely to work for the TDT-topics whose topic explication includes characteristic words, e.g. for TDT-topic 41016 having an explication where the words “Basque” and “ETA” appear, and for 41021 whose explication includes the words “Ghana” and “Kufuor.”

Based on Figure 1, we can conclude that hyperparameter estimation improves LDA, because both LYNDA and HypLDA work better than LDA. However, there seem no significant differences between LYNDA and HypLDA. The differences will later be revealed by inspecting several TDT-topics separately. Further, Figure 1 shows that TOT results in a weak performance. However, if we inspect TDT-topics separately, we will know that TOT is comparable with LYNDA under a specific condition, and nevertheless that LYNDA is more robust with respect to overfitting to timestamp data.

In each of Figures 2 and 3, we select four TDT-topics and clarify detailed differences. Figure 2 presents the results for the TDT-topics where LYNDA succeeds, and Figure 3 for the TDT-topics where LYNDA fails. Each scatter graph is tagged with a TDT-topic ID and includes markers each of which corresponds to a pair of a false alarms probability (horizontal axis) and a miss probability (vertical axis). One cross marker is given for TFIDF, and 30 markers are for other methods, because we have 30 results of CGS except for TFIDF. The line graph on the left side of each scatter graph shows the number of on-topic documents at each date ranging from December 1 in 2000 to January 31 in 2001, where horizontal grids are put at unit interval.

First, we compare LYNDA with HypLDA and LDA. Fig-
ures 2 and 3 provide the following observations. In these figures, we give the TDT-topics whose on-topic documents exhibit a distribution with one distinguished peak along time axis. However, smaller peaks are also observed. When the height of smaller peaks is far less than that of the largest peak, LYNDA succeeds (e.g. 40007 and 41026 in Figure 2). However, even when the height of smaller peaks is considerably less than that of the largest peak, LYNDA fails as long as some smaller peaks are placed far from the largest peak (e.g. 40026 and 41025 in Figure 3). When the height of smaller peaks is comparable with that of the largest peak, LYNDA fails (e.g. 40003 and 41014 in Figure 3). However, even when the height of smaller peaks is comparable with that of the largest peak, LYNDA succeeds as long as smaller peaks are located near the largest peak (e.g. 40004 and 40038 in Figure 2). Based on these observations, we can draw a claim: For a pair of documents having similar timestamps, LYNDA works.

Second, to compare LYNDA with TOT2 and TOT5, we observe from Figure 2 that TOT achieves the results better than or comparable with LYNDA for many of the presented TDT-topics, e.g. 40007 and 41026. That is, TOT also works for a pair of documents having similar timestamps. However, recall that Figure 1 shows weaker results for TOT, where NDCs are averaged over all TDT-topics. This means that TOT gives poor results for the TDT-topics which are not included in Figure 2, i.e., the TDT-topics whose on-topic documents do not show a concentrated timestamp distribution. Therefore, we can draw another claim: For a pair of documents located far from each other along time axis, TOT is likely to give the results worse than other methods, though LYNDA can give the results at least comparable with others.

We can combine these two claims as follows: While TOT excessively favors the TDT-topics whose on-topic documents show a concentrated timestamp distribution, LYNDA receives such TDT-topics with moderate favor. In this sense, LYNDA helps us to prevent from overfitting to timestamp data.

5. CONCLUSIONS

In this paper, we give a new approach for using document timestamps in LDA-based document modeling. We clarify when our approach succeeds by discussing the correlation between timestamp similarities and document similarities by comparing the results of various competing methods. However, we reveal differences between the methods only with respect to link detection task. Further investigation will be required to confirm the efficiency of our approach.

6. REFERENCES