Bag of Timestamps: A Simple and Efficient Bayesian Chronological Mining

Tomonari Masada[1], Atsuhiro Takasu[2], Tsuyoshi Hamada[1], Yuichiro Shibata[1], and Kiyoshi Oguri[1]

[1] Nagasaki University, 1-14 Bunkyo-machi, Nagasaki, Japan
{masada,hamada,shibata,oguri}@cis.nagasaki-u.ac.jp
[2] National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo, Japan
takasu@nii.ac.jp

Abstract. In this paper, we propose a new probabilistic model, *Bag of Timestamps (BoT)*, for chronological text mining. BoT is an extension of latent Dirichlet allocation (LDA), and has two remarkable features when compared with a previously proposed *Topics over Time (ToT)*, which is also an extension of LDA. First, we can avoid overfitting to temporal data, because temporal data are modeled in a Bayesian manner similar to word frequencies. Second, BoT has a conditional probability where no functions requiring time-consuming computations appear. The experiments using newswire documents show that BoT achieves more moderate fitting to temporal data in shorter execution time than ToT.

1 Introduction

Topic extraction is an outstanding agenda item in practical information management. Many researches provide efficient probabilistic methods where topics are modeled as values of latent variables. Further, researchers try to avoid *overfitting* via Bayesian approaches. Latent Dirichlet allocation (LDA) [5] is epoch-making in this research direction. In this paper, we focus on documents with *timestamps*, e.g. weblog entries, newswire articles, patents etc. *Topics over Time (ToT)* [9], one of the efficient chronological document models, extends LDA with per-topic beta distributions defined over the time interval normalized to $[0, 1]$. However, since beta distribution parameters are directly estimated with no corresponding priors, ToT may suffer from overfitting to temporal data. In general, we can observe overfitting to temporal data in two ways. First, same topics are assigned to word tokens from documents having similar (i.e., close) timestamps even when those word tokens relate to dissimilar semantical contents. Second, different topics are assigned to word tokens from documents having dissimilar timestamps even when those word tokens relate to semantically similar contents. We will later show that ToT suffers from the latter case in our experiments.

In this paper, we propose *Bags of Timestamps (BoT)* as a new probabilistic model for Bayesian topic analysis using temporal data. Our approach is similar to mixed-membership models [6], where reference data of scientific publications are treated in a Bayesian manner along with word frequency data. In BoT, we attach an array, called *timestamp array*, to each document and fill this array with tokens of document timestamps. Further, a Dirichlet prior is also introduced for timestamp multinomials, not only for word multinomials. We will show that BoT can realize more moderate fitting to temporal data than ToT.

2 Previous Works

Recent researches provide efficient and elaborated document models for utilizing temporal data. In Dynamic Topic Models [4], a vector is drawn, at each position on time axis, from a normal distribution conditioned on the previously drawn vector. This vector is used to obtain topic and word multinomials. However, normal distribution is not conjugate to multinomial, and thus inference is too complicated. On the other hand, Multiscale Topic Tomography Models [8] segment the given time interval into two pieces recursively and construct a binary tree whose root represents the entire time span and each internal node represents a shorter time interval. Leaf nodes are associated with a Poisson distribution for word counts. These two researches do not explicitly discuss overfitting to temporal data. In this paper, we focus on this problem and propose a new probabilistic model, *Bag of Timestamps (BoT)*. We can compare BoT with Topics over Time (ToT) [9], because both are an extension of LDA. ToT has a special feature that it can model continuous time with beta distributions. However, beta distribution parameters are estimated in a non-Bayesian manner. This may lead to overfitting to temporal data. BoT discards sophisticated continuous time modeling and takes a simpler approach. We attach an array, called *timestamp array*, to each document and fill timestamp arrays with tokens of document timestamps. Each timestamp token is drawn from a per-topic multinomial in the same manner as word tokens. Therefore, our appraoch is similar to mixed-membership models [6] where references of scientific articles are generated along with word tokens within the same LDA framework. We can introduce a Dirichlet prior not only for per-topic word multinomials but also for per-topic timestamp multinomials. Therefore, we can expect that overfitting to temporal data will be avoided. Further, BoT requires no time-consuming computations of gamma functions and of power functions with arbitrary exponents, which are required by ToT.

However, ToT may avoid overfitting with likelihood rescaling [10]. Here we introduce notations for model description. z_{ji} is the latent variable for the topic assigned to ith word token in document j. n_{jk} is the number of word tokens to which topic k is assigned in document j, and n_{kw} is the number of tokens of word w to which topic k is assigned. A symmetric Dirichlet prior for per-document topic multinomials is parameterized by α, and a symmetric Dirichlet prior for per-topic word multinomials is parameterized by β. t_j is the observed variable for the timestamp of document j. In case of ToT, t_j takes a real value from $[0,1]$. ψ_{k1} and ψ_{k2} are two parameters of a beta distribution for topic k. By applying likelihood rescaling for beta distribution, the conditional probability that z_{ji} is updated from k' to k in Gibbs sampling for ToT is $\frac{n_{jk}+\alpha-\Delta_{k=k'}}{\sum_k n_{jk}+K\alpha-1} \cdot \frac{n_{kx_{ji}}+\beta-\Delta_{k=k'}}{\sum_w n_{kw}+W\beta-\Delta_{k=k'}} \cdot \left\{ \frac{(1-t_j)^{\psi_{k1}-1} t_j^{\psi_{k2}-1}}{B(\psi_{k1},\psi_{k2})} \right\}^\tau$ where $\Delta_{k=k'}$ is 1 if $k=k'$, and 0 otherwise. τ is a parameter for likelihood rescaling and satisfies $0 \leq \tau \leq 1$. When $\tau = 1$, we obtain the original ToT. When $\tau = 0$, ToT is reduced to LDA. Therefore, we can control the degree of fitting to temporal data by adjusting τ. BoT will be also compared with ToT after likelihood rescaling in our experiments.

3 Bag of Timestamps

We modify LDA to realize an efficient chronological document modeling and obtain BoT as described in the following. First, a topic multinomial is drawn for each document from a corpus-wide symmetric Dirichlet prior parameterized by α. Second, for each document, topics are drawn from its topic multinomial and assigned to the elements of both word array and timestamp array. y_{js} is the latent variable for the topic assigned to sth timestamp token in document j. Let n_{jk} be the number of both word and timestamp tokens to which topic k is assigned in document j. Third, we draw words from word multinomials in the same manner with LDA, where per-topic word multinomials are drawn from a corpus-wide symmetric Dirichlet prior parameterized by β. Fourth, we draw timestamps from timestamp multinomials, where per-topic timestamp multinomials are also drawn from another corpus-wide symmetric Dirichlet prior paremeterized by γ. Then the conditional probability that z_{ji} is updated from k' to k is $\frac{n_{jk}+\alpha-\Delta_{k=k'}}{\sum_k n_{jk}+K\alpha-1} \cdot \frac{n_{kx_{ji}}+\beta-\Delta_{k=k'}}{\sum_w n_{kw}+W\beta-\Delta_{k=k'}}$ as in case of LDA [7]. Further, let n_{ko} be the number of tokens of timestamp o to which topic k is assigned. Then the conditional that y_{js} is updated from k' to k is $\frac{n_{jk}+\alpha-\Delta_{k=k'}}{\sum_k n_{jk}+K\alpha-1} \cdot \frac{n_{ky_{js}}+\gamma-\Delta_{k=k'}}{\sum_o n_{ko}+O\gamma-\Delta_{k=k'}}$. BoT is more than just introducing timestamps as new vocabularies because two distinct Dirichlet priors are prepared for word multinomials and timestamp multinomials.

By using document timestamps, we determine observed configuration of timestamp arrays as follows. We assume that all documents have a timestamp array of the same length L. The timestamp arrays of documents having timestamp o are filled with $L/2$ tokens of timestamp o, $L/4$ tokens of $o-1$, and $L/4$ tokens of $o+1$, where $o-1$ and $o+1$ are two timestamps adjacent to o along time axis. When a document is placed at either end of the given time interval, we leave $L/4$ elements of its timestamp array empty. Obviously, this is not the only way to determine configuration of timestamp arrays. However, we use this configuration in this paper for simplicity.

4 Experiments

We conduct experiments to reveal differences between BoT and ToT. The number of topics is fixed to 64. α is set to $50/K$, and β is 0.1 for both ToT and BoT [7]. γ, playing a similar role with β, is set to 0.1. We test three timestamp array lengths: 32, 64, and 128. In this order, dependency on temoral data gets stronger. We abbreviate these three settings as BoT32, BoT64, and BoT128. In ToT, we test 0.5 and 1.0 for τ. These two settings are referred to by ToTsqrt and ToTorig. Preliminary experiments have revealed that 300 iterations in Gibbs sampling are enough for all settings. We use the following three data sets.

"MA" includes 56,755 Japanese newswire documents from Mainichi and Asahi newspaper web sites (`www.mainichi.co.jp`, `www.asahi.com`). Document dates range from Nov. 16, 2007 to May 15, 2008. While we collapse the dates into 32 timestamps for BoT, the dates are mapped as is to $[0,1]$ for ToT. We use MeCab [1] for morphological analysis. The number of word tokens and that of unique words are 7,811,336 and 40,355, respectively. It takes nearly 90 minutes (resp. 105 minutes) for 300 iterations in case of BoT64 (resp. BoT128) on a single core of Intel Q9450. The same number of iterations require 135 minutes for both ToTorig and ToTsqrt. "S" consists of 30,818 Korean newswire documents from Seoul newspaper web site (`www.seoul.co.kr`). Document dates range from Jan. 1 to Dec. 31 in 2006. We map the dates to real values from $[0,1]$ for ToT, and to 32 timestamps for BoT. KLT version 2.10b [2] is used for Korean morphological analysis, and we obtain 5,916,236 word tokens, where 40,050 unique words are observed. The execution time of BoT64 (resp. BoT128) is 60 minutes (resp. 70 minutes) for 300 iterations. Both ToTorig and ToTsqrt require 100 minutes. "P" includes 66,050 documents from People's Daily (`people.com.cn`). The dates range from May 1 to 31 in 2008. We use a simple segmenter prepared for Chinese word segmentation bakeoff [3] with its prepared dictionary. So far as comparison between BoT and ToT is concerned, we think that this segmenter is enough. "P" includes 41,552,115 word tokens and 40,523 unique words. We map the dates to $[0,1]$ for ToT and discretize them into 24 timestamps for BoT. The execution time of 300 iterations is about 360 minutes (resp. 400 minutes) for BoT64 (resp. BoT128). Both ToTorig and ToTsqrt require 690 minutes.

Fig. 1 includes six graphs showing the results for "MA" under various settings. The horizontal axis represents time, and the vertical axis shows the percentage of the assignments of each topic to word tokens from the documents at each position on time axis. Every graph has a plot area divided into 64 regions. Each region corresponds to a distinct topic. When a region occupies a larger area, the corresponding topic is assigned to more word tokens. Top left graph is the result for LDA. While the graph shows poor dynamism in time axis direction, this is not a weakness. LDA is efficient in distinguishing topics enduring over time (e.g. stock news, weather forecasts, news of a person's death). However, LDA cannot control the degree of fitting to temporal data. Top right, middle left, and middle right graphs are obtained by BoT32, BoT64, and BoT128, respectively. More intensive temporal dynamism appears in middle right than in top right. We can say that the degree of fitting to temporal data is controlled by adjusting timestamp array length. Bottom left and bottom right graphs show the results for ToTorig and ToTsqrt, respectively. Each region has a smooth contour, because ToT can model continuous time. However, many topics are localized on narrow segments of time axis. Namely, same topics are rarely assigned to word tokens from distant positions on time axis. With respect to ToTorig (bottom right), only 7˜20 topics among 64 are observed at every position of time axis. In contrast, by using ToTsqrt (bottom left), we can find 44˜51 topics at each

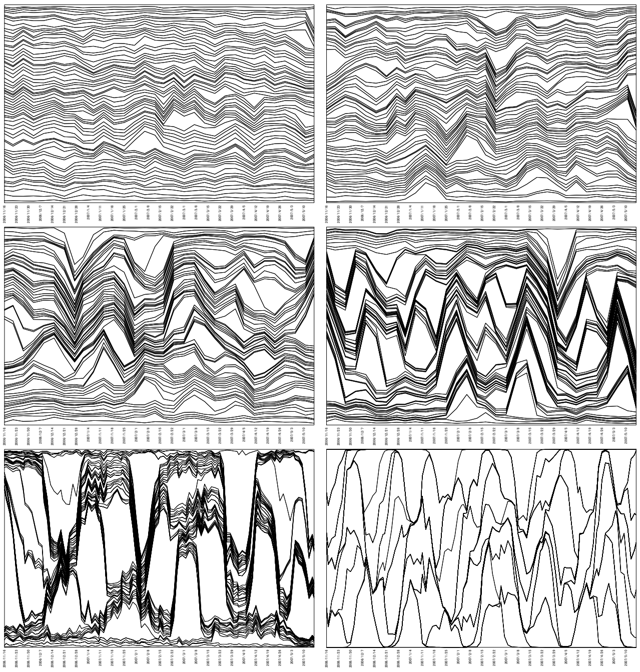

Fig. 1. Graphs of topic distribution changes for MA dataset.

position of time axis. However, many of these 44˜51 topics are assigned to very few word tokens. Likelihood rescaling seems to result in partial success. For both "S" and "P", we have obtained similar results.

We can also estimate the degree of overfitting to temporal data as follows. For each topic k, we sort words by n_{kw}. Many of top-ranked words may represent semantical content of the corresponding topic. We select 100 top-ranked words for each k. As the number of topics is 64, we have 64 lists of 100 top-ranked words. Further, for each word, we count the number of topics whose top-ranked word lists include the word. For example, when we apply LDA to "MA", the word "sakunen (last year)" appears in 25 among 64 lists. This kind of words may not help in focusing on a specific semantical content. In contrast, "bouei (military defence)" appears only in two among 64 lists. When many different topics are assigned to a word focusing on a specific content, poor topic extraction will result. For "P", "aoyun (Olympic games)" appears only in four among 64 lists for both LDA and BoT64. However, when we use ToTorig, "aoyun" appears in 17 among 64. This suggests that ToT assigns different topics to word tokens only because they appear in documents far apart in time. This corresponds to the second overfitting case described in Section 1. Table 1 includes other examples.

We can give another evidence. For "MA", 64 lists of 100 top-ranked words consists of 3,189 and 2,934 unique words when we use LDA and BoT64, respectively. However, when we use ToTorig, only 1,177 unique words are observed.

Table 1. Examples of words presenting the difference between BoT and ToT.

dataset	word	**LDA**	**BoT64**	**ToTsqrt**	**ToTorig**
MA	Minshu-tou (a political party)	5	5	8	16
(Japanese)	Ilaku (Iraq)	1	1	1	6
S	Lee Seung-Yeob (a baseball player)	1	1	1	7
(Korean)	Hannara (a political party)	2	3	3	14
P	Sichuan (a province)	16	20	28	32
(Chinese)	Shenghuo (Olympic Flame)	10	10	17	21

Namely, the same word appears in many different lists of top-ranked words. For both "S" and "P", we have obtained similar results. ToT tends to assign different topics to word tokens from documents having dissimilar timestamps even when they relate to a similar semantical content. ToT may be efficient when the given document set rarely includes outstanding contents ranging over a long period of time. However, when we are interested in both the semantical contents ranging over a long period of time and those localized on a small portion of time axis, BoT is a better choice, because BoT can respect both semantical similarity and temporal similarity, at least for Chinese, Japanese, and Korean documents.

5 Conclusion

In this paper, we propose a new probabilistic model to realize an efficient and intuitive control of the degree of fitting to temporal data of documents. As future work, we plan to model timestamp array lengths also in a Bayesian manner to realize more flexible control of the degree of fitting to temporal data.

References

1. http://mecab.sourceforge.net/
2. http://nlp.kookmin.ac.kr/HAM/kor/
3. http://www.sighan.org/bakeoff2005/
4. D. Blei and J. Lafferty. Dynamic Topic Models. in *Proc. of ICML'06*, pp.113-120, 2006.
5. D. Blei, A. Ng and M. Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, Vol. 3, pp. 993-1022, 2003.
6. E. Erosheva, S. Fienberg and J. Lafferty. Mixed-membership models of scientific publications. in *Proc. Natl. Acad. Sci.*, Vol. 101 (suppl. 1), pp.5220-5527, 2004.
7. T. Griffiths and M. Steyvers. Finding Scientific Topics. in *Proc. Natl. Acad. Sci.*, Vol. 101 (suppl. 1), pp.5228-5235, 2004.
8. R. Nallapati, W. Cohen, S. Ditmore, J. Lafferty and K. Ung. Multiscale Topic Tomography. in *Proc. of KDD'07*, pp.520-529, 2007.
9. X. Wang and A. McCallum. Topics over Time: A Non-Markov Continuous-Time Model of Topical Trends. in *Proc. of KDD'06*, pp. 424-433, 2006.
10. X. Wang, N. Mohanty and A. McCallum. Group and Topic Discovery from Relations and Text. in *Proc. of LinkKDD'05*, pp.28-35, 2005.