A Graphical Model for Topical Impact over Time

Zhiya Zuo
University of Iowa
Iowa City, IA
zhiya-zuo@uiowa.edu

Kang Zhao
University of Iowa
Iowa City, IA
kang-zhao@uiowa.edu

ABSTRACT

After being published, a document, whether it is a research paper or an online post, can make an impact when readers cite, share, or endorse it. A document may not make its greatest impact right after its publication, and some documents' impact can last a long period of time. This study develops a graphical model to capture the temporal dynamics in the impact of latent topics from a corpus of documents. Specifically, we modeled citation counts using Poisson distributions with Gamma priors. We conducted experiments on papers published in (i) D-Lib Magazine and (ii) The Library Quarterly from 2007 to 2017. Comparing with ToT, we found that our model produced more robust results on topical trends over time. The results also showed that prevalence and impact of the same topic are not correlated. Enabling better understanding and modeling of topical impact over time, this model can be used for the design of digital libraries and social media platforms, as well as evaluation of scientific contributions and policies.

CCS CONCEPTS

• **Information systems → Document topic models**; • **Applied computing → Document analysis**;

KEYWORDS

Topic models, temporal impact, text mining

ACM Reference Format:
Zhiya Zuo and Kang Zhao. 2018. A Graphical Model for Topical Impact over Time. In *JCDL '18: The 18th ACM/IEEE Joint Conference on Digital Libraries, June 3–7, 2018, Fort Worth, TX, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3197026.3203891

1 INTRODUCTION

Growing numbers of documents have imposed challenges on information retrieval and text mining. In the context of digital libraries, huge amounts of data have brought challenges to content searching and navigation. With a lower dimensional representation containing essential statistical characteristics for a collection of documents, information processing can be more scalable and efficient, under minimal human supervision. Topic modeling algorithms, therefore, are applied to aid the process of document/metadata annotation and labeling (e.g., [4]). In fact, there has been evidence on the effectiveness of topic modeling on such tasks [3].

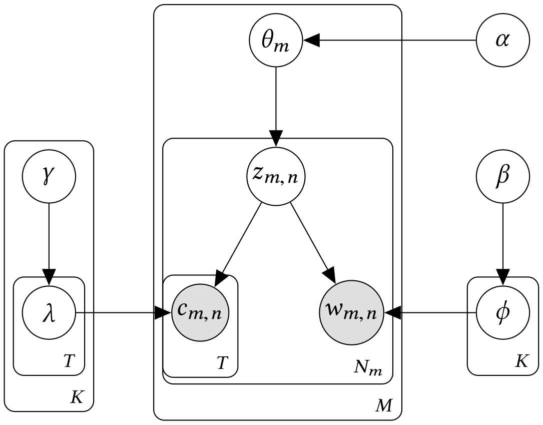

Figure 1: Graphical representation of TIoT model

While detection of latent topics among large corpora has been a remarkable success, the rise and fall of various topics are equivalently important. Blei and Lafferty [1] proposed a Markovian dynamic topic model where topics across time are dependent on their precedents. However, topics may vary in different time slices. Meanwhile, Topics over Time (ToT) model by Wang and McCallum [5] treated time as continuous variables with Beta distributions. Topics are invariant, while the shape of per topic Beta distributions portrait temporal topical prevalence. Leveraging citation data, Inheritance Topic Model (ITM) [2] aims at capturing topic evolutions.

However, none managed to explicitly model temporal dynamics of impact at topic level. While topical prevalence and evolution are helpful in identifying trend, impact can provide additional insights. Breakthrough research, for example, may not be the mainstream area, but can attract a significant amount of citations. In this study, we propose Topical Impact over Time (TIoT), to fill this gap. For academic publications, impact can be quantified by citation counts. The generative process corresponding to the graphical model in Figure 1 is as follows:

- Sample topic distributions $\boldsymbol{\theta} \sim Dirichlet(\alpha)$ for documents
- Sample word distribution $\boldsymbol{\phi} \sim Dirichlet(\beta)$ for topics
- For each document m:
  - For the n^{th} word w_{mn}:
    - Sample a topic $z_{mn} \sim Multinomial(\boldsymbol{\theta}_m)$
    - Sample a word $w_{mn} \sim Multinomial(\boldsymbol{\phi}_{z_{mn}})$
    - Sample citations $c_t \sim Poisson(\lambda_{kt})$ for $t \in \{1..T\}$

Exact inference cannot be done on this model. We applied collapsed Gibbs sampling for approximate inference, where $\boldsymbol{\theta}$, $\boldsymbol{\phi}$, and

JCDL '18, June 3–7, 2018, Fort Worth, TX, USA
© 2018 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-5178-2/18/06.
https://doi.org/10.1145/3197026.3203891

$\boldsymbol{\lambda}$ are integrated out. The full conditional probability for TIoT is:

$$P(z_{mn}|\mathbf{w},\mathbf{c},z_{-(\boldsymbol{mn})},\alpha,\beta,\gamma) \propto (n_{mz_{mn}}+\alpha-1)\times\frac{(n_{z_{mn}w_{mn}}+\beta-1)}{(n_{z_{mn}}+V*\beta)}$$
$$\times\prod_{t=1}^{T}\left[\frac{\Gamma(\sum_d c_{dt}n_{dz_{mn}}+\gamma_1)}{\Gamma(\sum_d c_{dt}n_{dz_{mn}}+\gamma_1-c_{mt})}\frac{(n_{z_{mn}}+\gamma_2-1)^{\sum_d c_{dt}n_{dz_{mn}}\gamma_1-c_{mt}}}{(n_{z_{mn}}+\gamma_2)^{\sum_d c_{dt}n_{dz_{mn}}\gamma_1}}\right] \tag{1}$$

where n_{mz} is the number of words in m assigned to topic z; n_{zw} is the number of times word w assigned to topic z; n_z is the number of times topic z sampled; c_{dt} is the document d's citations at time t; T is the number of timestamps; K is the number of topics; M is the number of documents; N_m is the size of m. For simplicity, we assume α, β, and γ are fixed values. It is noteworthy that citations for each word are the same as their belonging documents. Since we are over-using citation information for N_m times, we will raise the last term in Equation 1 to the power of $1/N_m$ to make text and citation modality comparable, as in Wang and McCallum [5].

2 RESULTS AND DISCUSSIONS

We conducted experiments with TIoT model on two datasets: (i) D-Lib Magazine (DLM) and (ii) The Library Quarterly (TLQ). We harvested abstracts and annual citations of papers from year 2007 to 2017 from Scopus. The numbers of topics K are set to 20 for both datasets. For simplicity, we followed the convention in Wang and McCallum [5] on fixed symmetric Dirichlet ($\alpha = 50/K = 2.5; \beta = 0.1$) and weak gamma priors are used ($\gamma_1 = \gamma_2 = 0.005$ for all $\boldsymbol{\lambda}$). Topics presented below are extracted from one sample at the 1000$^{\text{th}}$ iteration of a single Gibbs sampler. We ran ToT as a baseline model with the same setting. In the following, we use ToT's Beta densities as topical prevalence and TIoT's $\boldsymbol{\lambda}$ as topical impact over time. Topics from ToT and TIoT are paired by Jensen-Shannon divergence for the purpose of comparison.

First, we show how TIoT successfully extracts the impact trend by jointly modeling topics and citations. Figure 2 shows prevalence and impact for two topics in TLQ: (i) *policy & library* (top keywords include *political*, *factor*, and *election*) and (ii) library service to immigrants (top keywords include *immigrants*, *ala*, and *library*). For *policy & library*, specifically, a steep increase in citations happened right after year 2013 - in fact, two 2013 papers in TLQ on this topic were cited by a total of over 40 times. It is also interesting to note that the 2016 U.S. presidency election may result in the small increase from 2016 to 2017. TIoT also shows the burst of impact after year 2013, where citations start to accumulate gradually for papers on *library service to immigrants*. ToT, on the other hand, failed to provide useful information with the U-shaped Beta distributions.

Finally, we emphasize that topical prevalence and impact are different aspects. To illustrate this, we show in Figure 3 that there is hardly any correlation at the two arbitrarily selected timestamps in DLIB data. In both subplots, moreover, there is at least one outlier topic with high impact but relatively low prevalence. Results are consistent for both DLIB and TLQ data.

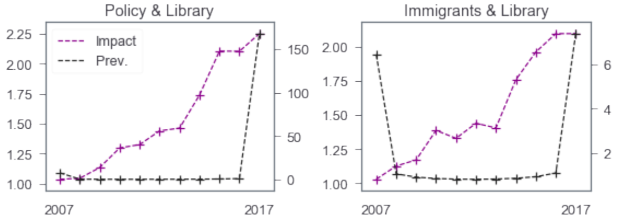

Figure 2: Two topics discovered by ToT (black) and TIoT (magenta) in TLQ dataset. Left y axes are Poisson parameters λ that measures average citation counts (i.e., impact) for that topic in each year; right y axes are Beta distribution probability densities that exhibits temporal prevalence

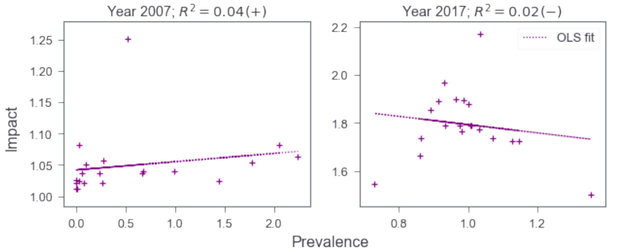

Figure 3: Correlation between topical prevalence and impact for all 20 topics in DLIB dataset in year 2007 and 2017. The dotted lines are fitted on the "+" data points by ordinal least squares. The sign of correlation is shown after R^2 value.

3 CONCLUSION

This paper presented the preliminary results of TIoT model and showed its promising capability in the area of digital libraries. For example, TIoT can be used for detecting trending topics and suggesting impactful papers in a bibliographical database. Annual citations, however, may increase over time, simply because of the increasing numbers of academic articles and researchers. In the future, it is important to normalize citations to make them comparable across time. Finally, it is interesting to apply TIoT on online social media and evaluate the extracted topical impact over time.

ACKNOWLEDGMENTS

The authors thank Dr. Kate Cowles and anonymous reviewers for providing valuable comments and Elsevier developer team for providing access to citation data.

REFERENCES

[1] David M. Blei and John D. Lafferty. 2006. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning - ICML '06*. ACM Press, 113–120. https://doi.org/10.1145/1143844.1143859

[2] Qi He, Bi Chen, Jian Pei, Baojun Qiu, Prasenjit Mitra, and Lee Giles. 2009. Detecting topic evolution in scientific literature. *Proceeding of the 18th ACM conference on Information and knowledge management - CIKM '09* (2009), 957. https://doi.org/10.1145/1645953.1646076

[3] David Newman, Youn Noh, Edmund Talley, Sarvnaz Karimi, and Timothy Baldwin. 2010. Evaluating topic models for digital libraries. *Proceedings of the 10th ACM/IEEE-CS Joint Conference on Digital Libraries* (2010), 215–224. https://doi.org/10.1145/1816123.1816156

[4] Suppawong Tuarob, Line C. Pouchard, and C. Lee Giles. 2013. Automatic tag recommendation for metadata annotation using probabilistic topic modeling. *Proceedings of the 13th ACM/IEEE-CS joint conference on Digital libraries - JCDL '13* (2013), 239. https://doi.org/10.1145/2467696.2467706

[5] Xuerui Wang and Andrew McCallum. 2006. Topics over time: A non-Markov continuous-time model of topical trends. *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2006), 424–433. https://doi.org/10.1145/1150402.1150450