Master's Thesis in Graduate School of Library, Information and Media Studies

A study on time series topic popularity extraction methods with topic modeling

July 2020

201826098

Khan Muhammad Haseeb UR Rehman

A study on time series topic popularity extraction methods with topic modeling

Khan Muhammad Haseeb UR Rehman

Graduate School of Library,
Information and Media Studies

University of Tsukuba

July 2020

A study on time series topic popularity extraction methods with topic modeling

Student No.: 201826098
Name: Khan Muhammad Haseeb UR Rehman

Topic modeling is extensively used for the natural language processing (NLP) problems of summarizing, organizing, and understanding large document datasets. LDA is widely used for the collection of topics, whereas DTM is famous for the time-series topic analysis. However, by estimating the number of occurrences of topics in each time slice, we can obtain time-series topic popularity using standard LDA. Therefore, if this can be extracted with LDA, then why do we need DTM which has a very high computation cost? The purpose of this research is to determine, either time-series topic information can be extracted from LDA or we need DTM. Topic drifting and popularity are two fundamental aspects of time-series topic analysis. we conducted experiments with different datasets to check the reliability of the information extracted from both models. We used Jensen-Shannon (JS) similarity-based analysis to check for information overlap, and overall and time-series correlation analysis as an inverse approach to extract DTM information from LDA topics. Lastly, we constructed time-series topic popularity graphs for both models from the document-topic distributions and compared the results. Our results show that there is notable DTM topic drifting information in some cases and sometimes no or vague topic drifting. Topic drifting embedded in DTM topics makes this model less favorable for time-series topic popularity analysis. On the other hand, LDA topics with no time transition information provided concrete results of time-series topic popularity. Thus, our results favor the usage of LDA.

Academic Advisors: Principal: Atsuyuki Morishima
Secondary: Kei Wakabayashi

Contents

List of Figures

Chapter 1

Introduction

Latent Dirichlet Allocation (LDA) [1] and Dynamic Topic Model(DTM) [2] are widely used topic models that revolutionized the solving of unsupervised topic modeling-based NLP problems. Situations that need the assistance of topic models often involve time-series document collections, including Twitter posts, news articles, and academic paper archives, because a continuous accumulation of documents typically yields a massive amount of text data. By focusing on the nature of time-series, many useful applications can be developed, such as bursty topic detection [3], trend analysis [4, 5, 6], topic evolution analysis [2, 7, 8, 9, 10], and topic transition pattern mining [11], etc.

To capture the time-series features of topics, DTM and its related-models [9, 10] assume dynamic drift of distributions. Although the DTM-based models appropriately find topics over time, they require expensive computational cost, which can be a critical drawback in some applications. On the other hand, there is a large body of work developing efficient inference algorithms for LDA because of its simpler architecture compared to DTM. While both models learn and work differently and even give different results, some practitioners and researchers employ LDA instead to analyze the time-series nature to take advantage of its efficiency, and these attempts seem to be successful according to the literature [6].

The question that arises in this background is; if time series topics information can be extracted by using LDA, which is faster than DTM, then why do we need to use DTM? To answer the above-mentioned question this research is conducted with a problem statement that "*Can time-series topic information of DTM be extracted from LDA?*" To the best of our knowledge, there have been no studies that extensively compared the information extracted using LDA with that of DTM.

In this paper, we examine the differences between LDA topics and DTM topics by using multiple datasets and model configurations. For this, we must compare two sets of topics from both models. Topic drifting and topic popularity are fundamental time-series information that can be extracted from DTM. Topic drifting is the topic transition over time and popularity is the measure of topic proportion at each time slice. The challenging part in topic transition analysis is that, DTM topic set has a sequential structure whereas LDA topic set has no type of sequential information. To map the unstructured topic set with DTM topics, we used a probability distribution similarity method.

Based on this matching, we analyzed both topic sets and in this process, we encountered with fragmentation issue, which we will describe later (**Figure ??**). DTM provides the time evaluation of topics, which means one single DTM topic can shift to a new subject if compared with the initial time's topic subject, where as an LDA topic's theme remains the same because LDA has no time aspect. This shifting in DTM topics is called fragmentation. In this experiment, we found that some DTM topics contain the information of two or more LDA topics; in other words, they have two or more fragmented topics.

To extract topic drifting from LDA topics, we compared both models, trying different approaches including correlation analysis and time-series topic correlation. For topic popularity, we built time-series population graphs for the topics of both models. Because both models have different types of information, there are pros and cons for each model. LDA extracts the focus on the collection of topics, whereas DTM can find connections between different themes and how subjects interchange within the same domain or topic.

Even though DTM has the edge of finding topic transitions over time for time-series data, in most cases, constructing only population graphs for LDA topics is enough for time-series analysis (e.g., events insights extraction from social media documents [6]). Some specific problems in which topic transition extraction is mandatory requires DTM despite its high computation cost (e.g., determining the focuses and trends of protected technological innovations across the entire disease landscape [12]).

Chapter 2

Related Work

2.1 Study of Aa

2.1.1 Study of Bb

Chapter 3

Proposal

Chapter 4

User Study

Chapter 5

Disucussion

Chapter 6

Conclusion

Acknowledgement

References

[1] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[2] David M Blei and John D Lafferty. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, pages 113–120. ACM, 2006.

[3] Daichi Koike, Yusuke Takahashi, Takehito Utsuro, Masaharu Yoshioka, and Noriko Kando. Time series topic modeling and bursty topic detection of correlated news and twitter. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 917–921, 2013.

[4] Noriaki Kawamae. Trend analysis model: trend consists of temporal words, topics, and timestamps. In *Proceedings of the fourth ACM international conference on Web search and data mining*, page 317–326, 2011.

[5] Hao Zhang, Gunhee Kim, and Eric P Xing. Dynamic topic modeling for monitoring market competition from online text and image data. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, page 1425–1434, 2015.

[6] Khan Muhammad Haseeb UR Rehman, Kei Wakabayashi, and Satoshi Fukuyama. Events insights extraction from twitter using lda and day-hashtag pooling. In *Proceedings of the 21th International Conference on Information Integration and Web-based Applications & Services*, pages 240–244. ACM, 2019.

[7] Janani Kalyanam, Amin Mantrach, Diego Saez-Trumper, Hossein Vahabi, and Gert Lanckriet. Leveraging social context for modeling topic evolution. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, page 517–526, 2015.

[8] Wei Xie, Feida Zhu, Jing Jiang, Ee-Peng Lim, and Ke Wang. Topicsketch: Real-time bursty topic detection from twitter. *IEEE Transactions on Knowledge and Data Engineering*, 28(8):2216–2229, 2016.

[9] Hesam Amoualian, Marianne Clausel, Eric Gaussier, and Massih-Reza Amini. Streaming-lda: A copula-based approach to modeling topic dependencies in document streams. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, page 695–704, 2016.

[10] Ayan Acharya, Joydeep Ghosh, and Mingyuan Zhou. A dual markov chain topic model for dynamic environments. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, page 1099–1108, 2018.

[11] Younghoon Kim, J. Jiawei Han, and Cangzhou Yuan. Toptrac: Topical trajectory pattern mining. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, page 587–596, 2015.

[12] Ming Huang, Maryam Zolnoori, Joyce E Balls-Berry, Tabetha A Brockman, Christi A Patten, and Lixia Yao. Technological innovations in disease management: text mining us patent data from 1995 to 2017. *Journal of medical Internet research*, 21(4):e13316, 2019.