A Study on Time Series Topic Popularity Extraction Methods with Topic Modeling

Student No.: 201826098
Name: MUHAMMAD HASEEB UR REHMAN KHAN

Understanding large document datasets is a fundamental natural language processing (NLP) problem. A large document collection is often generated as an accumulation for a long time span, which naturally has a time series structure. One of the important aspects of understanding such document collection is an estimation of time series topic popularity, which means the amount of mention of topics in each time slice. Topic modeling is an unsupervised NLP technique that constructs a set of topics pervaded in a given document dataset by a grouping process like a clustering. Particularly for modeling time series documents, Dynamic Topic Model (DTM) has been proposed to capture dynamic changes of topics over time. DTM is considered to be suitable to model time series documents since the basic topic model called Latent Dirichlet Allocation (LDA) assumes a static set of topics.

However, DTM has a drawback that is a high computation cost, whereas LDA is far faster thanks to its simplicity in the model architecture. For this reason, people have a motivation to employ LDA rather than DTM even for time series document collections. The collections of topics extracted by DTM and LDA are different, but little insight has been known about how they are different in practice.

In this paper, we extensively compare the topics extracted by LDA for time series document collections with the topics induced by DTM through a new objective analysis. Topic drifting and popularity are two fundamental aspects of time series topic analysis. We conducted experiments with multiple 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, also 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 topic popularity analysis. On the other hand, LDA topics with no time transition information provided concrete results of topic popularity. Thus, for time series topic popularity analysis, LDA is the accurate choice from both models.

Academic Advisors: Principal: Kei Wakabayashi
Secondary: Atsuyuki Morishima