Time-Varying Dynamic Topic Model: A Better Tool for Mining Microblogs at a Global Level

Jun Han, Beihang University, Beijing, China
Yu Huang, Beihang University, Beijing, China
Kuldeep Kumar, Bond University, Gold Coast, Australia
Sukanto Bhattacharya, Deakin Business School, Deakin University, Geelong, Australia

ABSTRACT

In this paper the authors build on prior literature to develop an adaptive and time-varying metadata-enabled dynamic topic model (mDTM) and apply it to a large Weibo dataset using an online Gibbs sampler for parameter estimation. Their approach simultaneously captures the maximum number of inherent dynamic features of microblogs thereby setting it apart from other online document mining methods in the extant literature. In summary, the authors’ results show a better performance of mDTM in terms of the quality of the mined information compared to prior research and showcases mDTM as a promising tool for the effective mining of microblogs in a rapidly changing global information space.

KEYWORDS

Dynamic Topic Models, Gibbs Sampler, LDA, Metadata, Microblogs, Microposts, Perplexity Measure, Text Mining

1. INTRODUCTION

The global information space is a rapidly morphing one where newer, faster and more individualized modes of person-to-person electronic information transmittance are emerging on a fairly regular basis. Microblogs is one such mode of global information transmittance that is becoming very popular in spreading and sharing breaking news, personal updates and spontaneous ideas. Microblogs such as Weibo and Twitter allow users to exchange small packets of information content and reflect the general public’s reactions to major events occurring around the globe. Due to an explosion of online data generated via social networking, a vast amount of user generated content has accumulated on the popular social networking websites. This has spawned a big demand for automated text mining models to delve into this large online collection of short textual elements. Blei, Ng & Jordan (2003) showed for the first time that statistical admixture topic models were a promising text mining tool in this regard.

Topic models are a generative probabilistic model, which posits that each topic (theme) can generate words according to certain probability and each document or collection of data is assumed to be a fixed-dimensional mixture model of topics. Dirichlet distribution is usually used to model the variability in the topic mixing vector of the documents and the word mixing vector of the topics, although other alternatives have been explored in the literature (Li & McCallum, 2006; Blei & Lafferty, 2007). Standard topic models such as latent Dirichlet allocation (LDA), a generative probabilistic

DOI: 10.4018/JGIM.2018010106

Copyright © 2018, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.
model for collections of discrete data such as text corpora (Blei, Ng & Jordan, 2003), are usually applied to model long textual documents and not suitable for microblog posts as these are short, noisy and highly correlated with the authors. Intuitively, posts published by the same user have a higher probability to belong to the same topic. Rosen-Zvi, Griffiths, Steyvers & Smyth (2004) expanded topic distributions from document-level to user-level to include authorship information. Based on it, Zhao et al. (2011) proposed twitter-LDA (twitter latent Dirichlet allocation), assuming that each post is assigned a single topic and some words can be background words. Diao, Jiang, Zhu & Lim (2012) further improved the twitter-LDA presenting the TimeUserLDA (time and user latent Dirichlet allocation) to detect topics that arrived in bursts. However, all these studies took the number of topics to be fixed – thus they failed to capture the inherently dynamic characteristics of microblog posts.

To accommodate the temporal information of document collections, a number of temporal topic models have been proposed. Wang & McCallum (2006) developed a Topics over time (TOT) model and found trends in time-sensitive topics using a continuous distribution over time-stamps. However, the number of topics were fixed over time in the TOT model. Blei & Lafferty (2006) proposed a dynamic topic model (DTM), which focuses on the change of topic composition i.e. word distributions. But in this case too, the number of topics were fixed. Moreover, the authors assumed that the parameters are conditionally distributed normally with a mean equal to the corresponding parameter at the previous time instance. However, since the normal distribution is not a conjugate to the multinomial distribution, their model does not yield a simple solution to the problems of inference and estimation. Ahmed & Xing (2010) proposed an infinite dynamic topic model (iDTM) that can accommodate the evolution of all aspects of the latent structure such as the number of topics, distribution and popularity. The iDTM also addressed the birth and death of topics over a timeline following a Markovian assumption and was therefore a significant improvement over the preceding models. There are just a handful of other published works on temporal evolution of topics (Masada et al., 2009; Hong, Dom, Gurumurthy & Tsioutsiosliklis, 2011).

None of the prior works have been able to adequately capture all the key inherent features of microblog posts. Our work is primarily motivated by this prominent gap in the extant literature, which our paper purports to fill. Here we consider the problem of mining short textual elements; in particular microblog posts (or microposts). Microposts are short, diverse, real-time, highly correlated with the authors and particularly come as temporal streams. Given a very vast number of posts due to the online data explosion owing to ever-increasing popularity of social networking, we develop and test an online, adaptive, metadata-enabled dynamic topic model (mDTM) along the lines initially posited by Li, Kveton, Wu & Kashyap (2012) which considers all the following factors:

1. Posts come as temporal streams, so the number of topics, the topics’ distribution and popularity should be time-evolving
2. The context is short, so we assume each post is just about one topic
3. Posts are diverse, users talk about everything on microblogs, so the number of topics is large and changeable
4. Users publish posts both about personal interests and global events, thus necessitating a separation of “personal” and “global” posts
5. Posts are real time and the amount is vast, thus batch processing of all posts at once is inefficient if not infeasible

We derive an online Gibbs sampler to estimate the key parameters and apply mDTM to a large Weibo dataset which has been a popular data source in recent empirical researches on ascertaining enterprise value impacts of micro-blogging in the global business arena (Huang, Zhang, Li, & Lv, 2014). Our results show a better performance in terms of the quality of the mined information as compared to popular topic models and thus helps to better demonstrate the practical utility of microblogs in the global information space.
Information Technology and the Internationalization of the Firm
[www.igi-global.com/article/information-technology-internationalization-firm/3579?camid=4v1a](www.igi-global.com/article/information-technology-internationalization-firm/3579?camid=4v1a)

Impact of Vendor Selection on Firms' IT Outsourcing: The Korea Experience
Jae-Seung Han and Sang-Yong Tom Lee (2012). *Journal of Global Information Management* (pp. 25-43).
[www.igi-global.com/article/impact-vendor-selection-firms-outsourcing/65097?camid=4v1a](www.igi-global.com/article/impact-vendor-selection-firms-outsourcing/65097?camid=4v1a)