Decision Support Systems 70 (2015) 1–14

Contents lists available at ScienceDirect

Decision Support Systems

journal homepage: www.elsevier.com/locate/dss

Detecting short-term cyclical topic dynamics in the user-generated content and news

Hsin-Min Lu *

Department of Information Management, National Taiwan University, Taipei 106, Taiwan

ARTICLE INFO

Article history:
Received 6 June 2014
Received in revised form 22 October 2014
Accepted 30 November 2014
Available online 6 December 2014

Keywords:
Topic models
Gibbs sampling
Temporal dynamics
Context dependent
Cyclical dynamics

ABSTRACT

With the maturation of the Internet and the mobile technology, Internet users are now able to produce and consume text data in different contexts. Linking the context to the text data can provide valuable information regarding users' activities and preferences, which are useful for decision support tasks such as market segmentation and product recommendation. To this end, previous studies have proposed to incorporate into topic models contextual information such as authors' identities and timestamps. Despite recent efforts to incorporate contextual information, few studies have focused on the short-term cyclical topic dynamics that connect the changes in topic occurrences to the time of day, the day of the week, and the day of the month. Short-term cyclical topic dynamics can both characterize the typical contexts to which a user is exposed at different occasions and identify user habits in specific contexts. Both abilities are essential for decision support tasks that are context dependent. To address this challenge, we present the Probit-Dirichlet hybrid allocation (PDHA) topic model, which incorporates a document's temporal features to capture a topic's short-term cyclical dynamics. A document's temporal features enter the topic model through the regression covariates of a multinomial-Probit-like structure that influences the prior topic distribution of individual tokens. By incorporating temporal features for monthly, weekly, and daily cyclical dynamics, PDHA is able to capture interesting short-term cyclical patterns that characterize topic dynamics. We developed an augmented Gibbs sampling algorithm for the non-Dirichlet-conjugate setting in PDHA. We then demonstrated the utility of PDHA using text collections from user generated content, newswires, and newspapers. Our experiments show that PDHA achieves higher hold-out likelihood values compared to baseline models, including latent Dirichlet allocation (LDA) and Dirichlet-multinomial regression (DMR). The temporal features for short-term cyclical dynamics and the novel model structure of PDHA both contribute to this performance advantage. The results suggest that PDHA is an attractive approach for decision support tasks involving text mining.

1. Introduction

Making informed decisions in our current fast-changing environment often demands the timely and comprehensive analysis of large amounts of text data. Researchers have attempted to address this challenge by developing text mining approaches such as topic models [1]. Topic models, including latent Dirichlet allocation (LDA) and its variations [2,3], have been applied to discover coherent topics, analyze trends in academic publications [4], and examine user-generated content from different sources [5].

Additional contextual information that signals the unobservable structures beneath the observed textual data can help a model extract better latent topics. Thus, an emerging research direction in topic models is to incorporate contextual information such as authors' identities and timestamps to extract latent topics that reveal the influence of changing contexts. For example, incorporating the authors' identities into a topic model [6] can improve performance because an author's specialties can reveal additional information regarding the latent topics of the document.

Incorporating contextual information into topic models also provides a direct route for topic models to support decision making. As a generative probabilistic model, the learned topic model is essentially the joint distribution of contextual information and textual data. By computing the conditional distributions of contextual information given observed textual data, a topic model can provide crucial information that supports decision-making employing contextual information. For example, by incorporating additional information of microblog users who shared news articles online, a model is able to recommend news articles to other microblog users who may have similar interests [7].

Despite recent progressions, few studies have focused on the short-term cyclical topic dynamics that connect changing topic occurrences to the time of day, the day of the week, and the day of the month. Content generated by social media and mobile platforms often reveal strong short-term cyclical dynamics because users' day-to-day routines heavily influence the contexts of use, which contribute to the variations of topic

* Tel.:+886 2 33661184.
E-mail address: luim@ntu.edu.tw.

http://dx.doi.org/10.1016/j.dss.2014.11.006
0167-9236/

occurrence. By including short-term cyclical dynamics in a topic model, we are able to better characterize the cyclical dynamics that reflect users' activities, habits, and preferences [8], three factors that can improve decision support tasks such as market segmentation [9] and product recommendation [10].

To fill this gap, we introduce a new family of topic models that can discover short-term cyclical patterns from a document collection. The proposed Probit-Dirichlet hybrid allocation (PDHA) provides a general framework with which to link discrete and continuous document-specific exogenous temporal features to topic distributions. PDHA includes features for daily, weekly, and monthly cyclical patterns as a way of capturing short-term dynamics. In addition to the topic–token distributions and document–topic mixes provided by a typical topic model, PDHA learns the coefficients of these temporal features through a multinomial Probit-like structure; these coefficients can reveal the occurrences of topic changes within a day, week, or month. Unlike the topic over time (TOT) [11] and dynamic topic model (DTM) [12], which focus more on the long-term evolution of topics, PDHA can model short-term cyclical variations that may be harder to capture using other flavors of topic models. Moreover, our PDHA model includes random variables for document-specific topic tendencies. These random variables allow each document to deviate from the mean tendency specified by the temporal features while preserving a common theme for each document.

In the subsequent sections, we first review previously proposed time-dependent topic models. We then present the PDHA model and discuss in detail the Gibbs sampling algorithm. Afterward, we present experimental results that incorporate daily, weekly, and monthly cyclical patterns. We conclude with a short discussion of future research directions.

2. Literature review

Topic models [1,4] are a family of algorithms aimed at discovering latent structures in large document collections. Based on the assumption that observed tokens are governed by latent topics, topic models define a generative process that first generates the mixture of topics in a document and then select observed tokens conditioned on latent topics. The data generating process provides a rich structure that is capable of capturing meaningful latent topical structures in documents. Compared to their predecessors, such as the probabilistic latent semantic indexing (pLSI), topic models do not have the over-fitting problem and outperform pLSI in terms of perplexity [1].

The original topic models are often referred to as the latent Dirichlet allocation (LDA) because they adopt the conjugate prior for multinomial distribution, the Dirichlet distribution, to simplify computation. The idea of capturing short-term cyclical dynamics is related to the research stream that incorporates additional time-dependent information to improve LDA models. We review selected time-dependent topic models in this section. We refer readers to Blei [13] for a general introduction of topic models.

2.1. Time-dependent topic models

We start with an overview of the LDA model and then extend it to time-dependent topic models. For a document collection that contains D documents indexed by integers 1, 2, …, D, LDA assumes that the N_d tokens in the document d, $w_d = (w_{d1}, w_{d2}, \ldots, w_{dN_d})$, were generated by first drawing the topic mix $\theta_d \sim Dir(\alpha)$, where $Dir(\alpha)$ is a Dirichlet distribution with symmetric concentration parameter α. The topic mix θ_d is a vector of length J, where J is the total number of topics in a document collection. Each element of θ_d is the probability of selecting the corresponding topic for a position in document d. All elements of θ_d sum to one.

The second step is to determine the topic for a token at position i, $1 \le i \le N_d$, by drawing $z_{di} \sim Multinomial(\theta_d)$. This process assumes that given the topic mix θ_d, the latent topic for each token in document d is independent of one another. Finally, a token at position i is determined by drawing from the corresponding topic–token distribution $w_{di} \sim Multinomial\left(\phi_{z_{di}}\right)$, where $\phi_{z_{di}}$ is a vector determining the probability that a token may appear given z_{di}, the topic at position i of document d. The length of each ϕ_j is the vocabulary size W for $j = 0, 1, \ldots, J-1$. The model assumes that each ϕ_j is generated from a Dirichlet distribution with a symmetric concentration parameter β.

The generative process can be represented using the plate notation shown in Fig. 1. The shaded circle indicates observed variables, and the open circles indicate latent variables and parameters. Starting from the upper left, Panel (A) in Fig. 1 provides a summary for the data-generating process described above.

In the subsequent discussion, variables such as z_{di} and θ_d should be regarded as latent variables because the number of these variables grow with the size of the dataset [14]. Other variables, including α, β, and $\phi_{z_{di}}$ are regarded as parameters. The joint distribution of observed tokens, latent topic variables and other parameters conditional on α and β is given by:

$$p(\theta, \phi, Z, w|\alpha, \beta) = \prod_{d=1}^{D} p(\theta_d|\alpha) \prod_{j=0}^{J-1} p\left(\phi_j|\beta\right) \prod_{i=1}^{N_d} p(z_{di}|\theta_d) p\left(w_{di}|\phi_{z_{di}}\right), \quad (1)$$

where $\theta = (\theta_1, \theta_2, \ldots, \theta_D)$, $\phi = (\phi_1, \phi_2, \ldots, \phi_J)$, $w = (w_1, w_2, \ldots, w_D)$, $Z = (z_1, z_2, \ldots, z_D)$, and $z_d = \left(z_{d1}, z_{d2}, \ldots, z_{dN_d}\right)$. One challenge presented by topic models is to design efficient and effective algorithms for estimating θ, ϕ, and Z given w, α, and β. We will review model estimation methods later.

The LDA model does not explicitly include temporal features. However, simple post-processing can be used to determine time trends. As demonstrated by Griffiths and Steyvers [4], the estimated θ_d for individual documents can be averaged by year to identify the trending topics across the sample period. This post-processing approach, however, is unable to take advantage of the potential time-dependent clusters naturally occurring in datasets.

Two types of time-dependence structures, upstream and downstream, can incorporate temporal features (see Fig. 2) [3]. The upstream structure allows the temporal features (e.g., timestamps) to influence the topic–mix distribution of a document, thereby determining the latent topics and tokens in a document. The downstream structure, in contrast, generates both tokens and timestamps conditioned on a latent topic. We first introduce TOT, a downstream model, followed by two upstream models, temporal collection (TC) and DTM.

The TOT model (see Panel (B) of Fig. 1) associates the document timestamp to every token in the document. It assumes that the topic mix of a document determines the latent topic at each position, which subsequently determines the observed tokens and the timestamp [11]. This model has a downstream structure because the topic mix (θ_d) determines the distribution of observed tokens (w_{di}) and timestamps (t_{di}) [3]. The additional timestamp variables in TOT allow the discovery of time-sensitive topics. One example is discovering topics over 21 decades of U.S. Presidential State-of-the-Union Addresses. The LDA model combines statements about the Mexican-American War (1846–1848) with those about World War I. The result is in contrast with topics discovered by TOT. TOT is able to localize statements about the Mexican-American War [11] and considers statements about World War I as belonging to a different topic because of the time gap between the two wars.

The temporal collection (TC) model [5] is based on similar ideas but instead adopts an upstream structure. The timestamp variable t enters the topic model under the assumption that the parameters of the prior distribution of topic mix, α, are a function of t. As a result, t influences the topic mix of document d, θ_d, the latent topics, and the observed tokens. TC adopts the gamma distribution to model time-dependent topic occurrence.