Information Sciences 473 (2019) 166–177

Contents lists available at ScienceDirect

Information Sciences

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

An Evolutionary Context-aware Sequential Model for topic evolution of text stream

Ziyu Lu[a,*], Haihui Tan[b], Wenjie Li[b]

[a] Department of Computer Science, Central University of Finance and Economics, China
[b] Department of Computing, The Hong Kong Polytechnic University, China

ARTICLE INFO

Article history:
Received 4 April 2018
Revised 23 August 2018
Accepted 16 September 2018
Available online 17 September 2018

Keywords:
Evolutionary clustering
Evolutionary clustering
Recurrent Chinese Restaurant Process
Long Short Term Memory

ABSTRACT

Social media acts as the platform for users to acquire information and spreads out breaking news. The overwhelming amount of fast-growing information makes it a challenge to track the subsequences of the breaking news or events and find the corresponding user opinions towards special aspects. Tracking the evolution of an event and predicting its subsequent trends play an important role in social media. In this paper, we propose an Evolutionary Context-aware Sequential model (ECSM) to track the evolutionary trends of the streaming text and investigate their focused context-aware topics. We integrate two novel layers into the Recurrent Chinese Restaurant Process (RCRP), respectively one context-aware topic layer and one Long Short Term Memory (LSTM) based sequential layer. The context-aware topic layer can help capture the global context-aware semantic coherences and the sequential layer is exploited to learn the local dynamics and semantic dependencies during the dynamic evolutionary process. Experimental results on real datasets show that our method significantly outperforms the state-of-the-art approaches.

© 2018 Elsevier Inc. All rights reserved.

1. Introduction

With the development of social media such as Twitter,[1] Facebook[2] and Sina Weibo,[3] those platforms have been the new ways for people to acquire information like events or breaking news. Also, users in the social platforms can express their opinions and share their comments via publishing short and instant posts. However, the increasing volume of overwhelming information makes it difficult to track the subsequences of the breaking news or the evolution of events. It is challenging to find the corresponding user opinions towards special aspects or subevents. Therefore, tracking the evolution of an event or predict its subsequent trends in social media are very important in social media.

In recent decades, event tracking and topic evolution have attracted a lot of attentions [10,15,24]. Topic tracking and evolution aims to discover the evolution trends in the streaming text on social media and capture the dynamic change of focused contexts. Several models have been proposed for topic tracking and evolution in social media [13,22,25]. For example, Wang et al. [25] proposed a generative topic model for research theme evolution. Wang et al. proposed a Topic Over Time (TOT) model [24] based on Latent Dirichlet Allocation (LDA) [5]. TOT explicitly assumes time is generated from

* Corresponding author.
E-mail address: luziyuhku@gmail.com (Z. Lu).

[1] https://twitter.com.
[2] https://www.facebook.com.
[3] https://www.weibo.com/.

https://doi.org/10.1016/j.ins.2018.09.027
0020-0255/© 2018 Elsevier Inc. All rights reserved.

topics, which jointly models time and word, thus enabling itself to discover time-aware topics as well as topic temporal strength. Also, Dynamic Topic Model (DTM) [4] has been designed to address the problem of topic evolution by modeling topics changing over time. The above mentioned models required the fixed number of clusters. In addition, some evolutionary models have been designed. Ahmed and Xing proposed a non-parametric evolutionary clustering model called the Recurrent Chinese Restaurant Process (RCRP) [2] to mine the dynamic clusters (trends) over time. And it allowed for the occurrence of a new cluster at any time. Later, Ahmed et al.[1] presented a unified framework to group incoming news articles into temporary but focused storylines. They designed a hybrid model by combining the evolutionary clustering approach Recurrent Chinese Restaurant Process (RCRP) and the topic model Latent Dirichlet Allocation (LDA). The topics from LDA can address various aspects of the story and describe the content of each cluster. However, existing evolutionary models only considered the global semantic coherences based on the co-occurrences of words. They ignored the local dynamics and semantic dependencies when discovering the evolutionary (dynamic) trends.

In this paper, we propose an Evolutionary Context-aware Sequential model (ECSM) to discover the potential evolutionary trends and the temporally focused sub-special aspects of the streaming text in social media. Our proposed model is based on the dynamic evolutionary model RCRP and integrates two novel layers into RCRP, respectively one LDA-like topic layer and one Long Short Term Memory (LSTM) based sequential layer . The LDA-like topic layer is capable to capture the context-aware semantic coherences from the co-occurrences of words. The LSTM based sequential layer is exploited to learn the temporal dynamics and local semantic dependencies of the evolving trends. To our best knowledge, this is the first study to integrate local dynamics and semantic dependencies into the dynamic clustering model for topic tracking and evolution. The integrated layers can discover fine-grained topic dependencies in the evolving topics and temporally focused sub-special aspects in dynamic clustering.

The main contributions of our paper are listed as follows:

- An Evolutionary Context-aware Sequential model (ECSM) for evolutionary clustering is proposed. Both the temporal dynamics and local semantic sequential dependencies can be captured through the integrated two novel context-aware layers.
- The LDA-like topic layer is designed to capture the context-aware aspects of the discovered evolving trends. The LSTM sequential layer considers both the temporal dynamics and local semantic dependencies among the evolving trends, which enables our model to further discovers more focused aspects.
- We perform extensive experiments on one Twitter dataset and two academic paper datasets. The experiment results show that our method outperforms the state-of-the-art approaches significantly.

The rest of the paper is organized as follows: related work is presented in Section 2. In Section 3, we formulate our problem and introduce some background knowledge about RCRP and LSTM. In Section 4, we demonstrate our proposed Evolutionary Context-aware Sequential model (ECSM). In Section 6, we describe our experiment design and report the experiment evaluation results. Finally, we conclude our paper and discuss future work in Section 7.

2. Related work

Many models have been proposed for evolutionary clustering (dynamic clustering) [2,4,13,24]. Dynamic topic model (DTM) [4] is a traditional dynamic clustering method and has been extensively used in topic and event evolution [13,14]. For example, Kawamae [13] proposed a trend analysis model (TAM) for capturing the evolution of trends. TAM focuses on the differences between temporal words and other words in each document to detect topic evolution over time. TAM introduces a latent trend class variable into each document and a latent switch variable into each token for handling these differences. The trend class has a probability distribution over temporal words, topics, and a continuous distribution over time, where each topic is responsible for generating words. However, dynamic topic model is a parametric model which requires to fix the number of clusters at each epoch. In order to relax the constraints of the fixed number of topics, a non-parametric model, Recurrent Chinese Restaurant Process (RCRP), is proposed by Ahmed and Xing [2]. It considers the cluster parameter dynamics and cluster popularity over time. In RCRP, the number of clusters at each epoch is unbounded. The clusters can remain, die out or emerge over time. Based on RCRP, several models have been designed. For example, Ahmed et al. [1] developed a storyline model by combining the evolutionary clustering approach Recurrent Chinese Restaurant Process (RCRP) and the topic model Latent Dirichlet Allocation (LDA) [5], which can group incoming news articles into temporary but focused storylines. The intuitions behind RCRP are that clusters (e.g. events or stories) are composed of documents and the popularity of clusters varies over time. And it allows for the occurrence of a new cluster at any time. The topics from LDA describe the content of each cluster, and documents are drawn from the associated story. After an event occurs (the story), several articles are addressing various aspects of the story. To analyze a stream of the incoming news, the new cluster might be generated and topic mixtures describes the cluster best. Diao and Jiang [6] introduced a duration-based discount into RCRP. And the time decaying factor is used to discount the probability of choosing a cluster (story). Older clusters were chosen with smaller probabilities. It distinguished the topical-related or busty events. Therefore, a tweet's content is either topical or event-related. Tang et al. [22] combined the RCRP model with a trend analysis model (TAM) [13] with RCRP, and integrated user interests in the generative process for personalized recommendation. However, the proposed evolutionary clustering methods only considered the dynamic change in trends over time but ignored the semantic sequential depen-

dency when tacking the evolving trends of streaming texts. The semantic sequential dependencies have important impacts on discovering the local semantic dependencies and temporal dynamics of the evolving trends.

Recently, some Recurrent Neural Networks (RNN) like Long-Short Term Memory [11] and Gated Recurrent Unit (GRU) have been used to characterize semantic dependencies and extensively applied in sequence modeling, such as short text conversation [20], event prediction [12,17,18], etc. Event prediction aims to predict future events after some event happened. And some sequential event prediction models have been proposed. For example, Radinsky and Horvitz [18] described and evaluated models for sequential event prediction. Its main contributions lied on the automated extraction and generation of sequences of events from corpora. Also, some RNN models have been combined with some context-aware topic models. For example, TopicRNN [7] integrated the RNN model into the topic model to capture both the global semantic coherences and local semantic dependencies of clusters. However, those context-aware sequential models have never considered the dynamics and cannot be used for topic evolution or dynamic clustering.

3. Preliminaries

3.1. Problem definition

We have the text stream D, which is divided into T epochs, $D = \{D_1, D_2, \ldots D_T\}$. At each epoch t, D_t consist of a sequence of streaming texts, $D_t = \{d_{t1}, d_{t2}, \ldots d_{tM}\}$, M is the number of texts (documents) at the t epoch. There are flexible clusters S over T time and a set of K latent topic set $Z = \{z_1, z_2, \ldots z_k\}$. The K topics represent the focused aspects of the evolving clusters, e.g. reasons or effects. And each topic z is defined as a distribution over the vocabulary V. Each dynamic cluster s has a distribution over the topics Z.

Our problem is: Given the text stream $D = \{D_1, D_2, \ldots D_T\}$ in T epochs, to discover the evolving trends (clusters) S_t at each epoch and each temporal cluster's focused topics Z.

3.2. Background

In this section, we present some background knowledge about the Recurrent Chinese Restaurant Process (RCRP) [2] and the Long Short Term Memory (LSTM) [11].

3.2.1. Recurrent Chinese Restaurant Process (RCRP)

The Recurrent Chinese Restaurant Process (RCRP) [2] is a non-parametric model for the evolutionary clustering. It is based on the Chinese Restaurant Process (CRP). In the CRP process, parameters $\theta_{1:n}$ are drawn from G a distribution of Dirichlet process $DP(G_0, \gamma)$, in which G_0 is the base measure and γ is the concentration parameter. In the CRP metaphore, there is a Chinese Restaurant with an infinite numbers of tables. Customer x_i enters the restaurant and sits on the table s that has n_s customers with probability $\frac{n_s}{i-1+\gamma}$, and shares the dish ϕ_s, or picks a new table with probability $\frac{\gamma}{i-1+\gamma}$, and orders a new dish sampled from the base measure G_0. Therefore, in CRP, the parameters θ is defined as follows:

$$\theta_i|\theta_{1:i-1}, G_0, \gamma, \sim \sum_s \frac{n_s}{i-1+\gamma}\delta(\phi_s) + \frac{\gamma}{i-1+\gamma}G_0 \tag{1}$$

in which ϕ_s is the dish distribution for the cluster (table) s.

The RCRP operates in epochs. It assumes that the customers enter the restaurant in a given day and they are not allowed to stay behind the end of this day. The owner of the restaurant records on each table the dish served on this table and the number of customers who shared it. Dishes correspond to chains, and the variation corresponds to the dynamic evolution of the chain. At day t, customer i can pick up an *empty* table s, which was used to serve dish $\phi_{s,t-1}$ with probability $\frac{n_{s,t-1}}{N_{t-1}+i-1+\gamma}$; then he chooses the current flavor of the dish $\phi_{s,t}$, according to $\phi_{s,t} \sim P(\cdot|\phi_{s,t-1})$. If the table s has already $n^i_{s,t}$, then he joins them with probability $\frac{n_{s,t-1}+n^i_{s,t}}{N_{t-1}+i-1+\gamma}$ and shares the current flavor of the dish there. Or he can pick a *new empty* table that was not used in the previous day $t-1$ with probability $\frac{\gamma}{N_{t-1}+i-1+\gamma}$, a new cluster (table) s^+ and orders a dish $\phi_{s^+,t} \sim G_0$. N_{t-1} is the number of customers at day $t-1$. The RCRP is defined as $RCRP(\gamma, G_0)$ in Eq. (2).

$$\theta_{t,i}|\theta_{t-1}, \theta_{t,1:i-1}, G_0, \gamma, \sim \sum_s \frac{1}{N_{t-1}+i-1+\gamma} \times \left[\sum_{s\in I_{t-1}\cup I^i_t} (n_{s,t-1}+n^i_{s,t})\delta(\phi_{s,t}) + \gamma G_0\right] \tag{2}$$

When RCRP is used for evolutionary clustering of texts [1,2], it can find the evolving clusters s. $\phi_{s,t}$ represents the word distribution for the cluster s_t at epoch t. The graphical representation of the Recurrent Chinese Restaurant Process (RCRP) can be shown in Fig. 1. The generative process is as follows:

For each time period t from 1 to T:

– For each document d_{ti} in epoch t

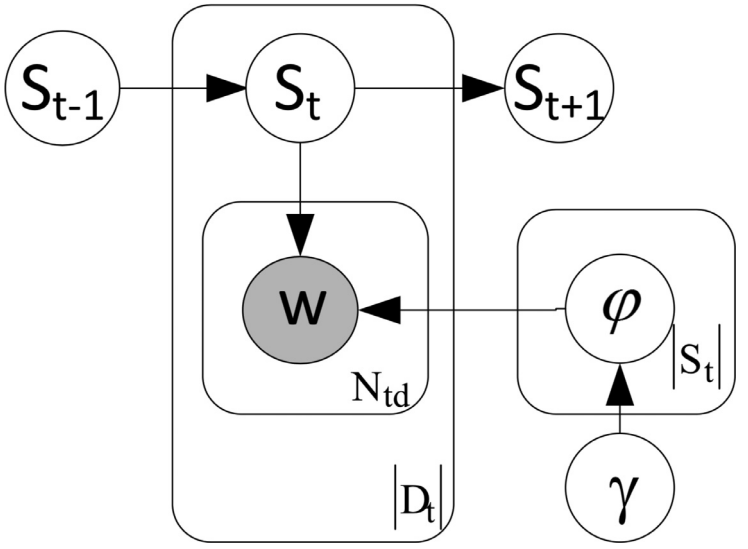


Fig. 1. RCRP.

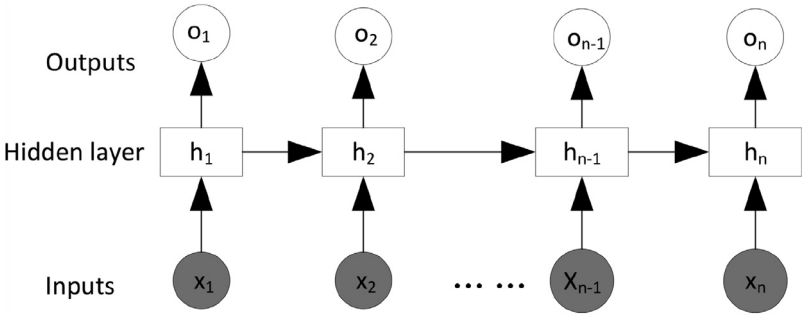


Fig. 2. Structure of RNN.

– draw the cluster indicator $s_{td}|s_{1:t-1}, s_{t,d_1:d_{i-1}}$ from $RCRP(\gamma, G_0)$
– if s_{td} is a new cluster, draw a distribution over words $\phi_{s,t}|G_0 \sim Dir(\gamma)$
– For each word w_{di} in document d, draw $w_{di} \sim \phi_{s_{td},t}$

3.2.2. Long Short Term Memory (LSTM)

We add one LSTM sequential layer in RCRP. LSTM is a special type of Recurrent Neural Networks and has been extensively used [11]. Fig. 2 shows the classic structure of RNN. It accepts a sequence of inputs $[x_1, x_2, \cdots, x_n]$ and outputs $[o_1, o_2, \cdots, o_n]$ in a sequence. Also there is a hidden layer to represent the hidden states. LSTM is an advanced model of RNN with additional structures, e.g. forget gates, cell states, etc. LSTM is capable to learn the long term semantic dependencies. The input for LSTM can be a sequence of the word *representation* (e.g. word embedding) in a document d, $\hat{d} = [x_1, x_2, \cdots, x_n]$. x_i is the word embedding of the word w_i and n is the length of d. LSTM computes the hidden vector $H = [h_1, h_2, \cdots, h_n]$ by iterating the following equations:

$$
\begin{aligned}
i_n &= \sigma(W_{xi}x_n + W_{hi}h_{n-1} + W_{ci}c_{n-1} + b_i) \\
f_n &= \sigma(W_{xf}x_n + W_{hf}h_{n-1} + W_{cf}c_{n-1} + b_f) \\
c_n &= f_n c_{n-1} + i_n \tanh(W_{xc}x_n + W_{hc}h_{n-1} + b_c) \\
o_n &= \sigma(W_{xo}x_n + W_{ho}h_{n-1} + W_{co} + b_o) \\
h_n &= o_n \tanh(c_n)
\end{aligned}
$$

in which σ is the logistic sigmoid function, and i, f, o and c are respectively the input gate, forget gate, output gate and cell activation vectors, all of which are the same size as the hidden vector h. W terms denote weight matrices, e.g. W_{vi} is the input-input gate weight matrix and W_{hi} is the hidden-input gate matrix. The b terms denote bias vectors, e.g. b_h is the hidden bias vector.

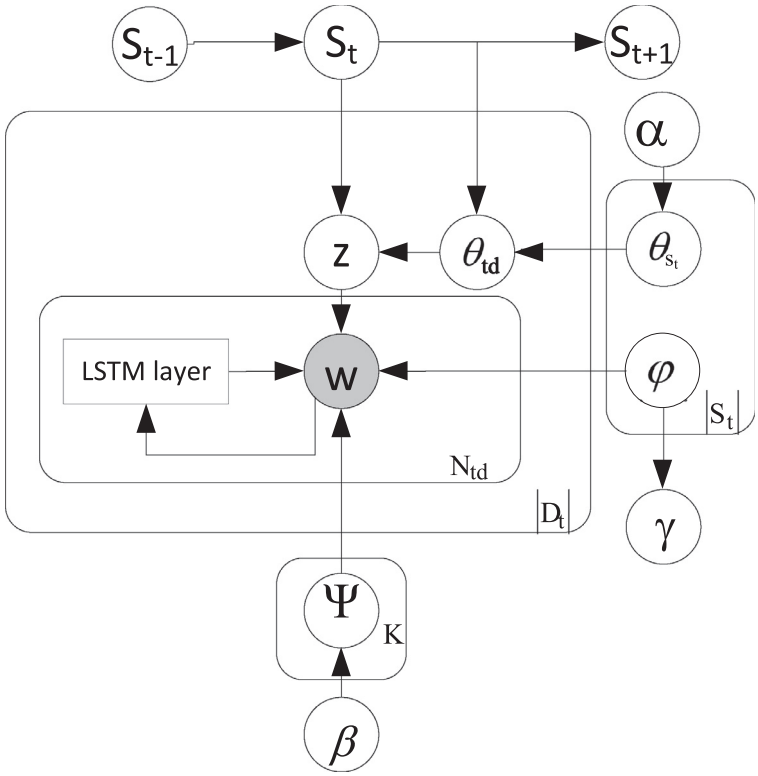


Fig. 3. Evolutionary Context-aware Sequential model (ECSM).

Table 1
Notations in Evolutionary Context-aware Sequential model.

Symbols	Description
T, K	Number of Epochs, Number of Topics
s, S	the trend variable, the trend set
M_t	Number of documents at each epoch
z	the latent topic variable
θ_s	topic distribution for cluster s
θ_{td}	topic distribution for document d at epoch t
ϕ_s	word distribution for cluster s
ψ_z	word distribution for topic z
α, β	prior parameters for distribution θ, ψ
γ	prior parameters for distribution ϕ

4. Model

In this section, we introduce our proposed Evolutionary Context-aware Sequential model (ECSM), which can predict the future trends or evolving aspects for an event. The intuition behind our model is that we assume an event might evolve and different aspects will be mentioned as time goes on. Our model is based on the Recurrent Chinese Restaurant Process (RCRP). RCRP can model the dynamic evolution process of an event and detect the event stories. Also, we integrates two new layers into RCRP, respectively one context-aware topic layer and a LSTM sequential layer. The context-aware topic layer is used to learn the global semantic coherences and capture the context-aware aspects of the evolving trends. The context-aware aspects are potential focused aspects of the evolving trends, e.g. reasons, effects, user opinions and actions. The LSTM sequential layer is exploited to learn the temporal dynamics and semantic dependencies during the dynamic evolutionary clustering process.

The graphical representation of our model ECSM is presented in Fig. 3 and the notations are listed in Table 1. Assume we have T epochs, s represents the trend (cluster) variable in the RCRP framework and there are S_t trends at the t epoch. Each trend s has its own word distribution ψ_s. In the context-aware topic layer, the latent topic variable z is to find the context-aware aspects of the evolving trends S. The number of topics is K. The topic selection depends on the trend-topic distribution θ_{s_t} of the evolving trends at the current epoch S_t. And each topic z has a word distribution ϕ_z. At one epoch, the temporal stories s_t might cover different aspects, e.g. an earthquake's effects or defensive measures when digesting incoming tweets.

Therefore, each trend s_t has a distribution θ_{s_t} to represent the topic strength and has its own word distribution $\phi_{s,t}$. Also, we exploit the LSTM layer to learn the long range semantic dependencies and capture the local dynamics. LSTM accepts the word (embedding) representation x of word w in a sequence and computes the local semantic dependencies. Therefore, both the global semantic coherences (the context-aware topics) and the local semantic dependencies are integrated to discover the evolving trends. The generative process of our proposed Evolutionary Context-aware Sequential model (ECSM) is as follows:

For each epoch t from 1 to T:

- for each document d_{ti} in epoch t, $d_{ti} \in \{d_{t1}, d_{t2}, \cdots, d_{tM_t}\}$
  - draw the cluster indicator s_{td} from $RCRP(\gamma, \phi_0)$
  - if s_{td} is a new cluster,
    - draw word distribution $\phi_{s_{new},t}|\phi_0 \sim Dir(\gamma)$
    - draw a cluster distribution over topic proportions θ_{s_t} from $Dir(\alpha)$
  - draw topic proportions θ_{td} from $Dir(\theta_{s_{td}})$
  - draw topic z_{td} from $Multi(\theta_{s_{td}})$
  - for each word w_{tdi} in the document, $i = 1, 2, \cdots, N_{td}$
    - compute LSTM hidden state h_i as Eq. (3)
    - draw w_i based on Eq. (4)

$$h_i = f(h_{i-1}, w_{i-1}) \tag{3}$$

$$p(w_i|w_{i-1}, \ldots, w_1; z_{td}; s_{td}, h_n) \propto \exp(v_{w_i}^T h_i + \psi_{z_{td}} \phi_{s_{td},t}) \tag{4}$$

The LSTM layer in Fig. 3 uses the basic LSTM model as introduced in Section 3.2.2. The same functions in Eq. (3) are used to compute the hidden state h. At each step, the LSTM Layer accepts one word w_{i-1} and computes the hidden state h_i. The next word w_i is generated based on both the context-aware semantics from the topic layer and the local sequential semantics from the LSTM sequential layer.

5. Inference

In this section, we describe the inference algorithm for our proposed model. We want to learn the latent variables $\{s, z\}$ and parameters $\Theta = \{\theta, \phi, \psi, \mathbf{\Omega}\}$. $\mathbf{\Omega}$ indicates the parameters used in LSTM, e.g. weights. Our observation is the sequence of documents at each epoch, $D^t = \{d_{t1}, d_{t2}, \ldots d_{tM}\}$. Each document d consist of a sequence of words $W_d = [w_1, w_2, \cdots, w_{N_{td}}]$. N_{td} is the number of words in the document d_{td}. The likelihood of the streaming text is defined as follows:

$$\mathcal{L}(\Theta, D) = \sum_{d=1}^{M^t} \sum_{s=1}^{|S^t|} p(s_{td}) \sum_{z} p(z|s_{td}) \prod_{i=1}^{N_{td}} p(w_i|w_{1:i-1}, h_i, z_{td}, \theta, \phi, \psi, \mathbf{\Omega}) \tag{5}$$

We adopt a hybrid inference approach for inference by combining sampling and variance optimization. We use a Gibbs-EM [23] to alternate between collapsed Gibbs sampling [2,9] and stochastic gradient decent [3,8] for estimating parameters in our model. In the E-step, Gibbs sampling is used to learn the hidden variables $\{s, z\}$ by fixing the other parameters $\mathbf{\Omega}$ in LSTM. In the M-step, we perform the gradient decent to learn LSTM model parameters $\mathbf{\Omega}$ by fixing the topic and story assignments. The inference algorithm is shown in Algorithm 1. T is the number of epochs and L is the number of iterations.

5.1. E-step

In the E-step, we apply the Gibbs Sampling learning process to infer $\{\theta, \phi, \psi\}$ while we fix the LSTM model parameters $\mathbf{\Omega}$ to be updated in the gradient descent step. We calculate the posterior probability as follows:

$$\begin{aligned} &p(s_{td}, z_{td}|\mathbf{s_{t-1}}, \mathbf{s_{t,1:d-1}}^{\neg td}, \mathbf{z}^{\neg td}, W_{td}) \\ &= P(s_{td}|\mathbf{s_{t-1}}, \mathbf{s_{t,1:d-1}}^{\neg td}) p(z_{td}|s_{td}, \mathbf{z}^{\neg td}, W_{td}) \\ &\prod_{i=1}^{N_{td}} P(w_{tdi}|z_{td}, s_{td}, \mathbf{\Omega}) \end{aligned} \tag{6}$$

In Eq. (6), the first term $P(s_{td}|\mathbf{s_{t-1}}, \mathbf{s_{t,1:d-1}}^{\neg td})$ is calculated as Eq. (2) in Section 3.2.1. The rest terms are computed as in Eq. (7).

Algorithm 1 Gibbs-EM inference algorithm for ECSM.

Input: The text stream at T epochs, $D = \{D_1, D_2, \ldots D_T\}$, number of iterations L
Output: Estimated parameters $\hat{\theta}, \hat{\phi}, \hat{\psi}, \hat{\mathbf{\Omega}}$

```
Initialize the estimated parameters θ̂, φ̂, ψ̂, Ω̂
for Iteration=1 to L do
    for t=1 to T do
        for each document d in the text stream at t epoch D^t do

            E-step: Update the trend assignment s_td and topic assignment z_td based on Equation 6.

            M-step: Compute gradients δ^t for LSTM parameters Ω and obtain Ω^t by stochastic gradient method [8].
        end for
    end for
end for
Obtain the estimated parameters θ̂, φ̂, ψ̂, Ω̂ based on Equation 11, Equation 10 and Equation 9.
```

$$p(z_{td} = z | s_{td}, \mathbf{z}^{\neg td}, W_{td}) \prod_{i=1}^{N_{td}} P(w_{tdi} | z_{td}, s_{td}, \mathbf{\Omega})$$
$$= \frac{n_{tdz}^{\neg td} + \frac{n_{s_{td}z}^{\neg td} + \alpha}{\sum_z n_{s_{td}z}^{\neg td} + \alpha}}{n_{td}^{\neg td} + 1} \prod_{i=1}^{N_{td}} \left[\frac{n_{zw_{tdi}}^{\neg tdi} + \beta}{\sum_w n_{zw_{tdi}}^{\neg tdi} + \beta} \phi_{s_{td},t}(w_{tdi}) + \nu_{w_{tdi}}^T h_i \right] \tag{7}$$

n_{tdz} is the number of times that a topic z is sampled from the document d at epoch t and $n_{s_{td}z}$ is the number of times that a topic z is sampled from the distribution $\theta_{s_{td}}$ specific to cluster s_{td} at epoch t. n_{zw} is the number of times that a word w is sampled from the word distribution ψ_z specific to the topic z. $\phi_{s_{td},t}(w)$ is the probability of the word w sampled from the cluster s_{td} word distribution at epoch t. x_w is the word embedding representation for the word w and h_i is the computed hidden state as in Eq. (3). Superscript $\neg td$ denoted a quantity excluding the current instance $\neg td$.

After sufficient iterations, the updates rules as follows:

$$\hat{\theta}_{td} = \frac{n_{tdz}^{\neg td} + \frac{n_{s_{td}z}^{\neg td} + \alpha}{\sum_z n_{s_{td}z}^{\neg td} + \alpha}}{n_{td}^{\neg td} + 1} \tag{8}$$

$$\hat{\psi}_{zw} = \frac{n_{zw_{tdi}}^{\neg tdi} + \beta}{\sum_w n_{zw_{tdi}}^{\neg tdi} + \beta} \tag{9}$$

$$\hat{\phi}_{s_{td},t} = \frac{n_{s_t w_{tdi}}^{\neg tdi} + \gamma}{\sum_w n_{s_t w_{tdi}}^{\neg tdi} + \gamma} \tag{10}$$

$$\theta_{t,s} | \theta_{t-1}, \theta_{t,1:i-1}, \phi_0, \gamma, \sim \sum_s \frac{1}{n_{s,t-1} + i - 1 + \gamma} \times \left[\sum_{s \in I_{t-1} \cup I_t^i} (n_{s,t-1} + n_{s,t}^i) \delta(\phi_{s,t}) + \gamma \phi_0 \right] \tag{11}$$

The parameters $\hat{\theta}_{t,s}$ and $\hat{\psi}_{s,t}$ follow the sample update rules as RCRP [2] (Eq. (2) in Section 3.2.1). i is the document index.

5.2. M-step

We learn the LSTM model parameters $\mathbf{\Omega}$ given the current estimates in the previous E-step, by using the stochastic gradient decent [8]. In each iteration, the input is a sequence of word representation $V_{td} = \{x_{w_{td2}}, x_{w_{td3}}, x_{w_{tdN_{td}}}, \cdots, x_{w_{tdn}}\}$ for the document d at epoch t and the current assigned topic index z. The gradients δ of the parameters $\mathbf{\Omega}$ (some weights) in LSTM are computed by LSTM backforward pass via stochastic gradient decent [8].

Table 2
Statistics (after preprocessing).

Dataset	Twitter	DBLP_IS	DBLP_DBDM
Number of documents	1730	2635	2073
Number of Words	1775	2020	1982
Timespan	10 days	13 years	9 years

6. Experiments

6.1. Dataset

We use three datasets to perform experimental evaluation. One is the TREC 2016 Microblog Track.[4] Another two are DBLP academic paper datasets [21] in the Information Security field, and Database and Data Mining field . We denote them respectively as DBLP_IS and DBLP_DBDM. We only use the title of each paper as the texts. The TREC 2016 Microblog Track dataset consists of the Twitter's sample tweet streams (approximately 1% of public tweets) during the official evaluation period from August 2, 2016 to August 11, 2016. 56 judged interest profiles (topics) of tweets have been given. For each interest profile, we use the relevant tweets. In DBLP_IS dataset, there are about 14 years' academic papers, from 2001 to 2013. In DBLP_DBDM dataset, it has 9 years' academic papers, from 2005 to 2013. We preprocess the three datasets by removing the non-English texts, replacing special symbols in them, tokenizing the sentences, stemming words. We also cleaned words which have less than two occurrences in the dataset. After preprocessing and cleaning steps, the statistics of the three datasets are shown in Table 2.

In the Twitter dataset, there is a total of 1730 tweets over 10 days and the vocabulary size is 1775. DBLP_IS consists of 2635 papers over 13 years and the number of words is 2020. In DBLP_DBDM, there are 2073 papers over 9 years and the vocabulary size is 1982. We use the first 80% of data as the training set and the remaining data as the test set. In Twitter, the first 8 days (from August 2, 2016 to August 9, 2016) are used as training data and tweets from the last two days (August 10, 2016 to August 11, 2016) are used as test set. In DBLP_IS, we treat the first 10 years' data as training set and the remaining three years' papers as test set. In DBLP_DBDM, the first 7 years' data are used as training set and the remaining two years' papers are used as test set.

6.2. Compared methods

We will compare our model with the following methods.

- Dynamic topic model (**DTM**) [4]: DTM is a dynamic topic model. It requires a fixed topic number and captures the dynamic topic evolution.
- Recurrent Chinese Restaurant Process (**RCRP**) [2]: RCRP is a non-parametric evolutionary clustering method introduced in Section 3.2.1.
- Storyline model (**Storyline**)[1]: The storyline model [1] is based on RCRP by introducing one topic layer to represent the cluster popularity. As we do not have entities in our dataset, we implement a simple version of the storyline model by removing the entity layer.
- Evolutionary Context-aware Sequential Model (**ECSM**): ECSM is our proposed model in Section 4. It considers the global semantic contexts (the topic layer) and local semantic dependencies (the sequential layer) in the evolutionary clustering.

For DTM, we used the typical setting $\alpha = 50/K, \quad \beta = 0.01$ as in previous work [5,19]. K is the number of topics. For Storyline and our method ECSM, we also use $\alpha = 50/K, \quad \beta = 0.01$ for the topic model part.[5] The prior parameters γ for RCRP, Storyline and ECSM are set as $\gamma = 0.1$ as in previous work [1,2]. The initial number of clusters in RCRP, Storyline and our method ECSM is 3.[6] In our proposed model, LSTM acts as the sequential layer. The setting for LSTM adopts the optimal values through parameter tunning. The number of hidden units is 50 and the learning rate in LSTM is 0.1.

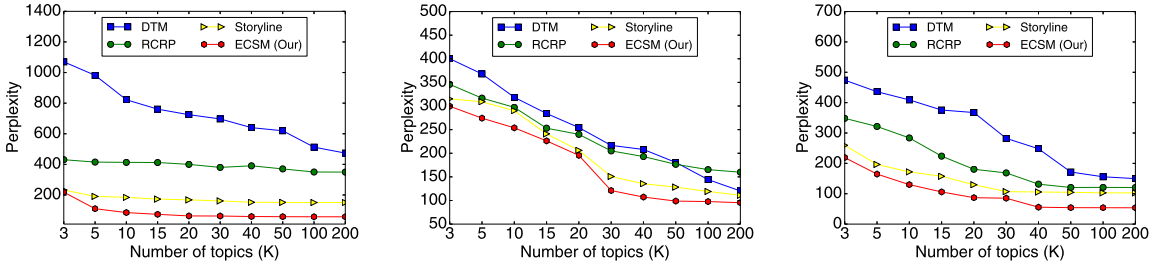
6.3. Evaluation metrics

Two metrics are used for evaluation. The first one is perplexity. The second one is the point-wise mutual information (PMI).

Perplexity measures how well the model fits the test data and evaluates the predictive power of a model. A lower perplexity score indicates the stronger predictive power. For a test set X_{test}, the perplexity is as follows:

$$perplexity(X_{test}) = \exp\left\{-\frac{\sum_{d\in X_{test}} \log P(\mathbf{w}_d)}{\sum_{d\in X_{test}} M_d}\right\} \tag{12}$$

[4] http://trecrts.github.io/.
[5] Different hyperparameters have very few impacts on experiment performances.
[6] The number of clusters are flexible and dynamic. The initial number of clusters has on impacts on final performances.



(a) Twitter (b) DBLP_DBDM (c) DBLP_IS

Fig. 4. Perplexity results.

Table 3
Temporal perplexity results for all datasets.

Dataset	Epoch	DTM	RCRP	Storyline	ECSM
Twitter	Aug. 8	231.24	265.39	209.03	**155.56**
	Aug. 9	2183.73	154.90	127.16	**61.42**
DBLP_IS	2011	204.75	394.56	300.43	**100.27**
	2012	1028.02	86.87	154.83	**77.28**
	2013	1217.59	129.29	78.25	90.28
DBLP_DBDM	2012	203.09	189.31	196.13	**181.59**
	2013	1169.64	414.77	323.46	**200.63**

where M_d is the number of words in a document d.

Topic Coherence: PMI (point-wise mutual information) measures the semantic coherences of learned topics [16]; the PMI score for the given topic z is calculated as the average relatedness of each pair of words in the set of top M representative words of a given topic:

$$PMI(z) = \frac{2}{M \times (M-1)} \sum_{1 \le i < j \le M} \log \frac{p(w_i, w_j)}{p(w_i)p(w_j)} \tag{13}$$

where $p(w)$ is the probability of word w to appear in the corpus and $p(w_i, w_j)$ is the joint probability of w_i and w_j. For each method, we average over PMI scores for all the learned topics. Higher PMI values are better since they indicate a coherent topic.

6.4. Experiment results

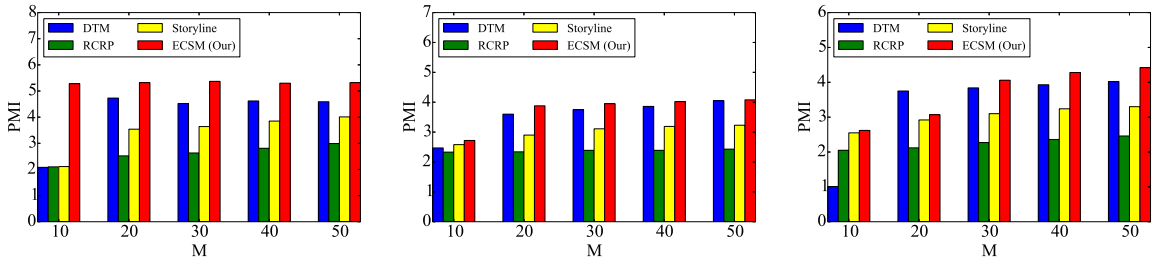
6.4.1. Perplexity results

Fig. 4 shows the perplexity results of all methods over different number of topics for the three datasets. We can observe that our proposed model ECSM has significantly lower perplexity results compared to all other methods for all datasets. It indicates that our method has the higher predictive power. As the number of topics increases, the perplexity scores decrease and gradually have small changes. From the results, we can find that RCRP based methods are better than DTM which has the fixed number of topics. It might show that the flexible number of clusters are essential and some clusters might emerge or disappear. And our method and the storyline model have better results than the raw RCRP method. It indicates that the introduced topic layer has positive impacts on learning topics and tracking its evolution. As our method ECSM has superior results than the storyline model, it demonstrates that the integrated sequential LSTM layer can capture local dynamics which is very essential in the evolutionary clustering process.

In addition, we calculate the temporal perplexity results over the test epoch to evaluate the temporal predictive powers. For each dataset, we calculate the temporal perplexity results at each test epoch. Namely, we treat documents at one test epoch as the temporal test set. For example, there are two test days in Twitter dataset, respectively, August 8, and August 9, 2016. We compute the perplexity results for each test day and report the separate results (according to the perplexity formula in Eq. (12), the overall results are not the average of the temporal results). In DBLP_IS, we report the temporal perplexity results from year 2011 to year 2013. For DBLP_DBDM, the test epochs are year 2012 and year 2013. We use the number of topics 20 for Twitter and DBLP_IS, and the number of topics 30 for DBLP_DMDB as examples to show the temporal results.[7] The temporal perplexity results are shown in Table 3.

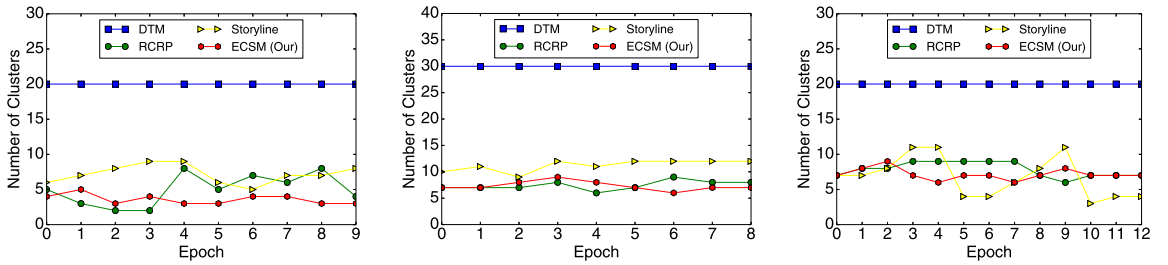
From Table 3, we can see that our method is also superior to the compared methods in separate test epochs. Especially, our method ECSM has small differences among the temporal perplexity results over different test epochs. It indicates that

[7] we select results from those numbers of topics as examples because there is small change in the perplexity scores when the numbers are larger.



(a) Twitter (b) DBLP_DBDM (c) DBLP_IS

Fig. 5. PMI results.



(a) Twitter (b) DBLP_DBDM (c) DBLP_IS

Fig. 6. Number of clusters at each epoch.

our model are more robust and flexible in discovering the topic evolutionary. DTM has good results in the first test epochs and has very different results in later epochs. DTM has the poor predictive power in the future epochs. Therefore, it indicates the effectiveness of the flexible dynamic models.

6.4.2. PMI results

Fig. 5 shows the PMI scores of all models over different M for the three datasets when the number of topics is 20 for Twitter and DBLP_IS, and the number of topics is 30 for DBLP_DMDB. M is the number of top probable words as used in Eq. (13). We investigate the results when $M = 10, 20, 30, 40, 50$. PMI measures topic coherences of the learned topics. Higher PMI values are better. PMI scores with larger M will be larger. The results show that our methods are significantly superior to the other methods in all cases for all datasets. Among the compared methods, the results are different from perplexity results. DTM have better PMI results than RCRP and the storyline model. It indicates probabilistic topic models can have more coherent topics as they model the word occurrences. But it might sacrifice the dynamic characteristics. The PMI results at different test epochs are similar as the results in Fig. 5. Therefore we omit them.

6.5. Topic evolution

6.5.1. Number of clusters

We investigate the change of the number of Clusters during the topic evolutionary process. Fig. 6 shows the results about the number of topics over all epochs for the three datasets. Similarly, we use example results when the number of topics is 20 for Twitter and DBLP_IS and 30 for DBLP_IS. In DTM, the number of clusters are equal to the number of topics and it is fixed over all epochs. For other methods, the number of clusters will change as clusters can emerge and disappear (the initial number of clusters is 3). The results show that the number of clusters might change during the evolutionary process. It might indicate models which allows clusters emerge or die out are more flexible. Form the results, we can see that the evolution patterns are very different for different datasets. Compared with other methods, our model has the relatively low frequencies of change and the evolving clusters have better dynamic coherences because our model integrates the sequential layer to consider the local dynamics and semantic dependencies.

6.5.2. Evolution analysis

In order to show the evolutionary process, we randomly select one cluster covering full epochs in the DBLP_IS dataset (the number of topics is 20) and list the example cluster's top-10 words at each epoch. The exemplary results for the selected cluster from year 2002 to 2013 are shown in Table 4. We can see that there might exist some differences in the top-10 words when time goes on. At the first two years, the studied topics are relatively similar but have some different words, e.g. Code, key. In the following year (2004), more new words come out, such as hash and compress, or attack and

Table 4
An exemplary evolution results for Information Security from 2002 to 2013.

Year	Top-10 words
2002	security, trust, software, dynamic, service forens, vulner, commun, curv, cloud
2003	security, code, software, key, check curve,dynamic,vulner,service,extend
2004	attack, security, extend, check, feature software, detection, binary, compress, hash
2005	detect, security, attack, service, trust new, code, vulner, IP, software, enfore
2006	security, network, attack, trust, software IP, malwar, detect, protocol, public
2007	security, network, access, protocol, attack internet, mobile, service, anonym, authent
2008	security, network, privacy, time, attack wireless, mobile, service, anonym,authent
2009	security, attack, detect, design, network system, code, scheme, function, model
2010	secuirty, attack, detect, new, framework, network, system, model, analysis,authent
2011	security, attack, detect, analysis, function, network, system, privacy, model,scheme
2012	security, attack, detect, analysis, privacy network, system, learning, recognition, function
2013	security, attack, detect, control, protocol network, system, recognition, model, scheme

detection. It might be because that more researchers focus on the coding problems and new attack detection in the security field. Then in the following years, topics might change to IP wireless issues or privacy in mobile platforms. At the last years, some words like model, function or learning have emerged. The research topics in the security community might have more attentions on model design and the network system.

7. Conclusions

In this paper, we propose an Evolutionary Context-aware Sequential model (ECSM) to discover the potential evolutionary trends and temporally focused sub-special aspects of the streaming text. Experimental evaluation on two metrics show that our proposed model can have higher predictive powers and topic coherences for evolutionary clustering. Especially, our method is more robust. In addition, the examplary results about topic evolution indicate that our method can capture the topic evolutionary over different time epochs. The experimental results show that the integration of the context-aware topic layer and the sequential layer can help track the dynamic change and semantic dependencies in topic evolutionary.

In the future, we plan to explore other complex sequential models (e.g. GRU) to further investigate temporal effects and sequential dependencies. We will consider the effects from the context-aware layer and sequential layer on more complex dynamic clustering model, e.g. hierarchical clustering models. In addition, we will explore the influences from other entity information, e.g. social relationship in Twitter or co-author relationship in academic datasets.

Acknowledgment

We would like to thank the reviewers for their invaluable comments. This work was supported by Research Grants Council of Hong Kong (No. PolyU 152094/14E), National Nature Science Foundation of China (61672445), The Hong Kong Polytechnic University (G-YBJP), and the Youth Development Fund of Central University of Finance and Economics (No. QJJ1730).

References

[1] A. Ahmed, Q. Ho, J. Eisenstein, E. Xing, A.J. Smola, C.H. Teo, Unified analysis of streaming news, in: Proceedings of the 20th International Conference on World Wide Web. WWW '11, 2011, pp. 267–276.
[2] A. Ahmed, E.P. Xing, Dynamic non-parametric mixture models and the recurrent chinese restaurant process: with applications to evolutionary clustering, in: Proceedings of the SIAM International Conference on Data Mining, SDM 2008, April 24–26, 2008, Atlanta, Georgia, USA, 2008, pp. 219–230.
[3] L. Bai, E.R. Hancock, Graph kernels from the Jensen–Shannon divergence, J. Math. Imaging Vision 47 (1–2) (2013) 60–69.
[4] D.M. Blei, J.D. Lafferty, Dynamic topic models, in: Proceedings of the 23rd International Conference on Machine Learning. ICML '06, ACM, New York, NY, USA, 2006, pp. 113–120, doi:10.1145/1143844.1143859.
[5] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent dirichlet allocation, J. Mach. Learn. Res. 3 (2003) 993–1022.
[6] Q. Diao, J. Jiang, Recurrent chinese restaurant process with a duration-based discount for event identification from twitter, in: Proceedings of the 2014 SIAM International Conference on Data Mining, Philadelphia, Pennsylvania, USA, April 24–26, 2014, 2014, pp. 388–397.
[7] A.B. Dieng, C. Wang, J. Gao, J.W. Paisley, Topicrnn: a recurrent neural network with long-range semantic dependency, CoRR abs/1611.01702 (2016).
[8] J. Duchi, E. Hazan, Y. Singer, Adaptive subgradient methods for online learning and stochastic optimization, J. Mach. Learn. Res. 12 (2011) 2121–2159.
[9] T.L. Griffiths, M. Steyvers, Finding Scientific Topics, National Academy of Sciences, 2005, pp. 5228–5235.
[10] Q. He, B. Chen, J. Pei, B. Qiu, P. Mitra, L. Giles, Detecting topic evolution in scientific literature: how can citations help? in: Proceedings of the 18th ACM Conference on Information and Knowledge Management. CIKM '09, 2009.
[11] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780.
[12] L. Hu, J. Li, L. Nie, X.-L. Li, C. Shao, What happens next? Future subevent prediction using contextual hierarchical lstm, in: Proceedings of the Thirty–First AAAI Conference on Artificial Intelligence., AAAI '17, 2017.
[13] N. 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. WSDM '11, 2011, pp. 317–326.
[14] N. Kawamae, R. Higashinaka, Trend detection model, in: Proceedings of the 19th International Conference on World Wide Web. WWW '10, 2010.
[15] Q. Mei, C. Zhai, Discovering evolutionary theme patterns from text: an exploration of temporal text mining, in: Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining. KDD '05, 2005.
[16] D. Newman, J.H. Lau, K. Grieser, T. Baldwin, Automatic evaluation of topic coherence, in: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics., HLT '10, 2010, pp. 100–108.

[17] K. Radinsky, S. Davidovich, S. Markovitch, Learning causality for news events prediction, in: Proceedings of the 21st International Conference on World Wide Web. WWW '12, 2012.
[18] K. Radinsky, E. Horvitz, Mining the web to predict future events, in: Proceedings of the Sixth ACM International Conference on Web Search and Data Mining, WSDM '13, 2013.
[19] M. Rosen-Zvi, T. Griffiths, M. Steyvers, P. Smyth, The author-topic model for authors and documents, in: Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence. UAI '04, 2004, pp. 487–494.
[20] L. Shang, Z. Lu, H. Li, Neural responding machine for short-text conversation, in: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics. ACL '15, 2015.
[21] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, Z. Su, Arnetminer: extraction and mining of academic social networks, in: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '08, 2008.
[22] X. Tang, C.C. Yang, Tut: a statistical model for detecting trends, topics and user interests in social media, in: Proceedings of the 21st ACM International Conference on Information and Knowledge Management. CIKM '12, ACM, 2012.
[23] H.M. Wallach, Topic modeling: beyond bag-of-words, in: Machine Learning, Proceedings of the Twenty-Third International Conference (ICML 2006), Pittsburgh, Pennsylvania, USA, June 25–29, 2006, 2006, pp. 977–984.
[24] X. Wang, A. McCallum, Topics over time: a non-markov continuous-time model of topical trends, in: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '06, 2006.
[25] X.W. Wang, C. Zhai, D. Roth, Understanding evolution of research themes: a probabilistic generative model for citations, in: KDD'13, ACM, 2013, pp. 1115–1123, doi:10.1145/2487575.2487698.