A Density-Based Nonparametric Model for Online Event Discovery from the Social Media Data

Jinjin Guo ^{1}, **Zhiguo Gong** $^{2\,*}$
Department of Computer and Information Science, University of Macau, Macau SAR
{^{1}yb57414,^{2}fstzgg}@umac.mo

Abstract

In this paper, we propose a novel online event discovery model DP-density to capture various events from the social media data. The proposed model can flexibly accommodate the incremental arriving of the social documents in an online manner by leveraging Dirichlet Process, and a density based technique is exploited to deduce the temporal dynamics of events. The spatial patterns of events are also incorporated in the model by a mixture of Gaussians. To remove the bias caused by the streaming process of the documents, Sequential Monte Carlo is used for the parameter inference. Our extensive experiments over two different real datasets show that the proposed model is capable to extract interpretable events effectively in terms of perplexity and coherence.

1 Introduction

The geo-temporal-tagged data in social media services (i.e. Flickr, Twitter, Blog) are well matching the 3W perspective (what, when, where) of events in the real world, thus, are popularly used for the task of event discovery. In general, most of the existing discovery techniques only work on a fixed static collection of the dataset [Guo and Gong, 2016; Zhao *et al.*, 2016; Pan and Mitra, 2011]. This seriously restrains their application scope in the context of online streaming data from the social media, which are incrementally generated by users along the time.

Practically, users desire the mining work (i.e. document summarization, item recommendation, emergency detection) could be achieved in an online manner. The requirement motivated researchers towards the online techniques, and generated several works for the online event discovery. However, the existing algorithms reveal several limitations when applying them to social media. In this paper, we are going to tackle the problem and propose a robust parameter-free online algorithm for the event discovery from the social media data.

1.1 Limitations of the Existing Online Techniques

In order to deal with incremental streaming texts, both Recurrent Chinese Restaurant Process(RCRP)[Ahmed and Xing, 2008] and Dynamic Clustering Topic model(DCT)[Liang *et al.*, 2016] are built on a sequence of continuous epochs along the time frame. The generation of topics in the current epoch are jointly influenced by the data distribution from prior and current epochs. One of the main deficiencies of RCRP and DCT is that no one knows what should be the optimal setting for the length of the epoch.

To solve the problem, [Du *et al.*, 2015] modeled the temporal dynamics of events with an approach named Dirichlet-Hawkes Process. Specifically, the Hawkes Process utilized a fixed number of Gaussian kernels with different predefined bandwidths (variances), and set the means of those kernels to some points (called reference time points in the paper) along the timeframe. The current document is temporally influenced by the history of the stream by summing all those weighted kernels. However, it is not clear how to set the reference time points, the different bandwidths of the kernels, and the number of those kernels.

Besides the limitations addressed above, none of them take into account the geospatial influences of the documents. It is intuitive that geographical features of social messages are capable to better define and distinguish events. In this paper, we are going to tackle those problems.

1.2 Contributions

In this paper, we propose a nonparametric model for the online event discovery from social media data.

We assume the document stream is generated under a density function over the temporal space. By regarding the generating time (temporal tag) of each document as a sampling point in the temporal space, we can estimate the density function $\lambda(t)$ using the summation of Gaussian kernels [Hinneburg *et al.*, 1998]. When a document d arrives at time t, we regard the temporal influence on d from the stream as $\lambda(t)$.

For the online event discovery, the number of the events (topics) may dynamically increase with more documents arriving. To accommodate this complexity, Dirichlet Process is exploited to infer the parameters in the model. When the new arrival of a document tells a different story from the previous, a new event is likely to be generated according to Dirichlet Process.

The social media data are not only tagged with temporal features, but also geospatial features. It is intuitive that a real world event may occur in one region or across several re-

[*] Corresponding author.

gions and one region may encompass multiple events along the time. We assume that events detected from the social media are distributed over a set of regions, and a mixture of Gaussians is thus employed to model the geographical distributions of events.

We integrate all those components together to formulate the Density Based Dirichlet Process for the online event discovery. Sequential Monte Carlo [Arnaud Doucet, 2001] is applied to perform the inference.

To summarize, we make the following contributions

- We incorporate the textual contents, timestamps and geo-locations of the social media data in a uniformed model for the online event discovery.
- We combine Dirichlet Process and density-based temporal dynamic technique together, where Dirichlet Process is used to capture the diversity of events while density estimation technique is exploited to learn the temporal influence of the document stream. To our best knowledge, this is the first model with parameter-free for the online event discovery in the joint space crossing temporal and geographical dimensions.
- We perform extensive experiments on two different real datasets to evaluate the proposed model, which show that our model is capable to discover meaningful events in terms of low perplexity and high coherence.

1.3 Organization of the Paper

The rest of this paper is organized as follows. In Section 2, we give a brief introduction of Dirichlet Process and density estimation method. Section 3 presents the proposed model and inference process. We discuss the experimental results in Section 4 and introduce the related work in Section 5. Section 6 concludes our work.

2 Preliminaries

In this section, we give a brief introduction to two fundamental techniques used in the paper, Dirichlet Process and the sample-based density estimation algorithm.

2.1 Dirichlet Process

Dirichlet Process [Teh *et al.*, 2006] is one member of the nonparametric stochastic processes, which is used to model the data in a "rich get richer" fashion without specifying the parameter (e.g. topic number). We write $G \sim DP(\alpha, G_0)$ to denote the draw from a DP which is parameterized by a concentration parameter α and base distribution G_0. We can draw samples $\theta_{1:n}$ from it since G itself is a distribution. Specifically, we resort to the scenario of Chinese Restaurant Process to illustrate this process, where each data point is regarded as a customer of entering a Chinese Restaurant with infinite tables and dishes (topics). When the first customer arrives, she can randomly select one empty table (cluster), sit and order one dish. Then, the second customer can either join with the first customer and share the dish, or she can start a new table and order a new dish. In this way, when the n^{th} customer arrives, she can select one table from k occupied tables with probability proportional to the number of guests already seated there, or start a new table with probability proportional to α. Formally, the conditional probability can be written as

$$CRP(\theta_n|\theta_1, \theta_2, ...\theta_{n-1}) \propto \begin{cases} \dfrac{m_k}{n-1+\alpha} & \theta_k \text{ exists} \\ \dfrac{\alpha}{n-1+\alpha} & \theta \text{ is new,} \end{cases}$$

where m_k represents the number of guests selecting θ_k and n is the total number of customers in the restaurant by now.

2.2 Density Estimation

Suppose points $t_1, t_2, ..., t_N$ in an interval $[a, b]$ are randomly generated under a density distribution $\lambda(t)$, then $\lambda(t)$ can be estimated using those samples as $\hat{\lambda}_N(t)$ [Fukunaga and Hostetler, 1975]

$$\hat{\lambda}_N(t) \propto \sum_{i=1}^{N} \kappa(t_i, t, \sigma),$$

where $\kappa(t_i, t, \sigma) = e^{-\frac{(t-t_i)^2}{2\sigma^2}}$ is Gaussian kernel, and σ is called bandwidth of the Gaussian kernel, which indicates the influence range of the kernel. Theoretically, the value of σ varies depending on N, that is $\sigma = \sigma_N$, such that $\lim_{N\to\infty} \sigma_N = 0$ in order to guarantee the unbiasedness of the estimate. A condition $N \cdot \sigma_N \to \infty$ is required to ensure $\lim_{N\to\infty} \hat{\lambda}_N(t) = \lambda(t)$ consistently [Fukunaga and Hostetler, 1975]. Therefore, the value of σ_N should be set smaller if the number of samples (N) is larger.

3 Proposed Model

If we take each new arriving document $d_j(\mathbf{w}, \mathbf{l}, t)$ as a sample, it is intuitive to utilize Dirichlet Process for deriving topics (events) incrementally. Since the documents for the same event are temporally cohesive, we use the estimated density over time dimension to bring the temporal influence from the prior documents to the current one. Further, the documents for the same event may also be spatially clustered, a mixture of Gaussians is used for modeling the spatial distribution of events.

The overall model is presented in Figure 1, where notations in shade are observed variables, notations in dotted circles are hyper parameters, and the remains are latent variables. In detail, $(t_i, ..., t_j)$ are the arriving time points of documents in the L-length sliding window (which will be discussed in the following subsection), $\mathbf{w}$ and $\mathbf{l}$ are the bag of words and location of the current document, respectively; z is the event (topic) which is generated jointly from the influence of prior documents in the sliding window and Dirichlet Process; r is the region variable with mean μ and variance $\mathbf{\Sigma}$; $\mathbf{w}$ and $\mathbf{l}$ are generated from respective distribution φ and π given event z.

3.1 Sliding Window and the Dynamic Bandwidth

In the situation of online modeling, documents are received continuously in a stream manner. We suppose the first document d_1 in the stream arrives at time t_1 and the current document d_j arrives at t_j. As introduced before, if we take all those received documents until t_j as the samples in $[t_1, t_j]$ then, the density at t_j can be roughly estimated as:

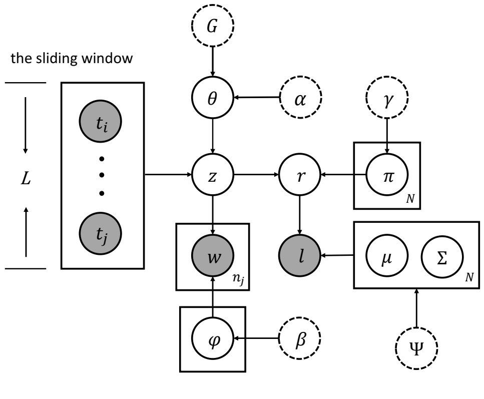


Figure 1: Graphical representation of the proposed model

$$\hat{\lambda}_N(t_j) \propto \sum_{i=1}^{j} \kappa(t_i, t_j, \sigma_j),$$

where σ_j indicates the bandwidth, which is associated with both the number of points in $[t_1, t_j]$ and length of the temporal space ($|t_j - t_1|$).

This estimation dynamically brings the influences of all those prior documents on the current document d_j in a natural manner. However, there are two problems when using it directly in the online modeling: (1) the dynamic increasing of the temporal space $[t_1, t_j]$ which requires all previous documents to be in the computation, and (2) the previously determined optimal value of σ_j may become improper along the document streaming.

To deal with the first problem, we introduce a sliding window $[t_j - L, t_j]$ to account only documents arriving after time $t_j - L$ for the density estimation at t_j. The intuition of this measure is based on the significant decay of the Gaussian kernel influence to $\hat{\lambda}_N(t_j)$ when $|t_i - t_j|$ is large enough (i.e. larger than L).

To solve the second problem, we suppose the arriving of documents is generated uniformly under the density function $\lambda(t)$ in a relatively short time duration with fixed length Γ (i.e. a duration of one or two months). However, the arriving of documents may become nonuniform in a larger timeframe because of the the significant increase of users in the social media system along the time. Let σ_j denote the bandwidth in the timeframe $[t_j - \Gamma, t_j]$, and N_j denote the number of documents received in this period. Then, σ_j will be dynamically tuned to maintain $\sigma_0 N_0 \approx \sigma_j N_j$, where σ_0 is the optimal setting for the bandwidth in the initial period $[t_1, t_1 + \Gamma]$.

3.2 Generating Topics

The joint process of Dirichlet Process and temporal density is parameterized by an intensity parameter α and a base distribution G over a given space θ. Each event is associated with an unique θ .

Let $D = \{d_{j-L_j}, d_{j-L_j+1}, ..., d_j\}$ denote the set of documents in the sliding window of the current document d_j, and L_j is the number of documents in it. The probability of sampling θ for document d_j is computed as:

$$P(\theta_j|\theta_{j-L_j}, ...\theta_{j-1}) \propto \begin{cases} \dfrac{\lambda_{\theta_k}(t_j)}{\sum_{i=j-L_j}^{j} \lambda_{\theta_i}(t_j) + \alpha} & \text{reuse } \theta_k \\ \dfrac{\alpha}{\sum_{i=j-L_j}^{j} \lambda_{\theta_i}(t_j) + \alpha} & \theta \text{ is new,} \end{cases} \tag{1}$$

where $\lambda_{\theta_k}(t_j) := \sum_{i=j-L_j}^{j} \kappa(t_i, t_j, \sigma_j)\mathbb{I}[\theta(t_i) = \theta_k]$ denotes the aggregated influence to the current document d_j from the documents in the corresponding sliding widow, whose event assignments are θ_k.

Therefore, the combination of Dirichlet Process and temporal density can not only infer the event number automatically, but also take into account the temporal influence of the documents in the stream.

3.3 Generating Location and Word

It is intuitive that a real world event may occur in one region or across several regions, and one region may encompass multiple events along the time. To capture the footprints of events from the social media, we assume that each event geographically is distributed over a set of regions, and each region is modeled by a Bi-Variant Gaussian parameterized as $(\mu, \mathbf{\Sigma})$. Thus the spatial distribution of an event is represented by a mixture of Gaussians, and geographical popularity of a region to a specific event is determined by its weight. Such spatial modeling allows a complex and diverse spatial pattern of an event. In detail, given an event indicator z of document d, the probability of generating location $\mathbf{l}$ is computed as:

$$P(\mathbf{l}|z, rest) = \sum_{r=1}^{N} P(r|\pi_z) \cdot \mathcal{N}(\mathbf{l}|r),$$

where r denotes the geospatial region, π_z denotes the weights of the Gaussian mixture for topic z, and N is the number of regions.

In general, documents generated from social media (e.g. Twitter, Weibo and Flickr photos) share a common property that each of them is short and always tells one story. Hence in our problem setting, each document is only assigned to one event. Under such an assumption, given an event z the probability of generating the content $\mathbf{w}$ of document d is computed as:

$$P(\mathbf{w}|z, rest) = \prod_{i=1}^{n_d} P(w_i|\varphi_z).$$

3.4 SMC Inference Process

Given the documents arriving in a stream manner, our goal is to conduct an online computation of the posterior distribution $P(z_{1:n}, r_{1:n}|d_{1:n})$, where $d_{1:n}$, $z_{1:n}$ and $r_{1:n}$ represent all the past documents, their event indicators and region assignments.

At time $t = n - 1$, let $P(z_{i:n-1}, r_{i:n-1}|d_{i:n-1})$ denote the posterior distribution in the corresponding sliding window. With the new document d_n arriving, the posterior would yield the most recent value $P(z_{i:n}, r_{i:n}|d_{i:n})$ by reusing $P(z_{i:n-1}, r_{i:n-1}|d_{i:n-1})$, which motivates us to apply Sequential Monte Carlo (SMC) method [Doucet *et al.*, 2000; Arnaud Doucet, 2001] to infer the sampling process.

The overall structure of Sequential Monte Carlo (SMC) is described in Algorithm1. Briefly, the posterior approximation is maintained as a set of *particles*, each of which represents a hypothesis about the latent variables. There is an importance weight associated with each particle to indicate how well the hypothesis explains the data. The main ingredient for designing an SMC is the proposal distribution $Q(z_{i:n}, r_{i:n}|d_{i:n})$ which can be regarded as the biased posterior estimation given the streaming documents in our online model. The weight ω_n^f of each particle $f \in \{1, ..., F\}$ is therefore defined as the ratio of the true posterior distribution over the proposal distribution $\omega_n^f = \frac{P(z_{i:n}, r_{i:n}|d_{i:n})}{Q(z_{i:n}, r_{i:n}|d_{i:n})}$. In order to minimize the variance of resulting particle weights, a commonly used technique in SMC [Arnaud Doucet, 2001; Ahmed *et al.*, 2011b], is to take $Q(z_n, r_n|z_{i:n-1}, r_{i:n-1}, d_{i:n})$ as the posterior $P(z_n, r_n|z_{i:n-1}, r_{i:n-1}, d_{i:n})$. Therefore the unnormalized importance weight for particle ω_n^f can be updated as

$$\omega_n^f \propto \omega_{n-1}^f \cdot P(d_n|z_n^f, r_n^f, d_{i:n-1}).$$

Since each document is denoted by a triplet $d_i = \{\mathbf{w}_i, t_i, \mathbf{l}_i\}$, the above equation can be detailed as

$$\omega_n^f \propto \omega_{n-1}^f \cdot P(\mathbf{w}_n|z_n^f, rest) \cdot \mathcal{N}(\mathbf{l}_n|r_n^f, rest). \qquad (2)$$

The event and region indicator can be alternatively sampled for each newly arriving document in the following way.

Sampling event indicator z. When the new document $d_n = \{\mathbf{w}_n, t_n, \mathbf{l}_n\}$ arrives, the probability of its topic assignment is computed as

$$P(z_n|z_{i:n-1}, r_{i:n-1}, d_{i:n}, rest) \propto P(z_n|t_n, rest) \cdot$$
$$P(\mathbf{w}_n|z_n, rest) \cdot \sum_r^N P(r|\pi_{z_n}) \mathcal{N}(\mathbf{l}_n|r), \qquad (3)$$

where $P(z_n|t_n, rest)$ is computed by Equation 1. Because of the one-event restriction for each document in our context, the probability of generating textual content $\mathbf{w}$ for document d can be computed as

$$P(\mathbf{w}|z, rest) \propto \frac{\prod_{v=1}^{V} \prod_{q=0}^{n_{dv}} (n_{zv} + q + \beta)}{\prod_{q=0}^{n_d} (n_z^v + q + \beta|V|)}, \qquad (4)$$

where n_{dv} records the occurrence number of word v in document d, n_{zv} describes the co-occurrence number of word v and event z, n_z^v denotes the word number assigned to event z, and n_d records word number in document d.

Sampling region indicator r. Given the event indicator and new document, the probability of sampling region is

$$P(r|z_n, d_n, rest) \propto P(r|\pi_{z_n}) \cdot \mathcal{N}(\mathbf{l}_n|r). \qquad (5)$$

4 Experimental Evaluation

All the experiments are conducted on a computer with Intel Core i7 2.93GHz CPU and 8GB RAM, and all the algorithms are implemented using Visual C#.

4.1 Experimental Setup

Dataset. Our experiments are based on two real datasets, which are crawled from Flickr through its API[1]. The first dataset is collected for Paris, which consists of more than 26 thousands photos from June to August 22th in 2010. The second dataset is for New York (NY), which includes more than 40 thousands photos from January to March in 2010. For each dataset, the photos without any textual description are discarded, and we split the Flickr tag (phrase) into single word and remove the irrelevant words (e.g. camera names). After the preprocessing, the statistics of the two datasets are as shown in Table 1.

Algorithm 1 SMC algorithm over document stream

```
Initialize $\omega_1^f$ to $\frac{1}{F}$ for all $f \in \{1, ..., F\}$
for each document $d_n$, $n = 1, 2, ...$ do
    for $f \in \{1, ..., F\}$ do
        Sample event indicator $z_n^f$ by Equation 3
        Sample region indicator $r_n^f$ by Equation 5
        Update each particle weight $\omega_n^f$ by Equation 2
    end for
    Normalize particle weights
    if $||\omega_n||_2^{-2} = 1/\sum_{f=1}^{F} (\omega_n^f)^2$ < threshold then
        Resample particles
    end if
end for
```

Table 1: The statistics of the datasets

Dataset	Time span	images	vocabulary
Paris	06/01/10-08/22/10	21436	2042
NY	01/01/10-03/31/10	33780	2366

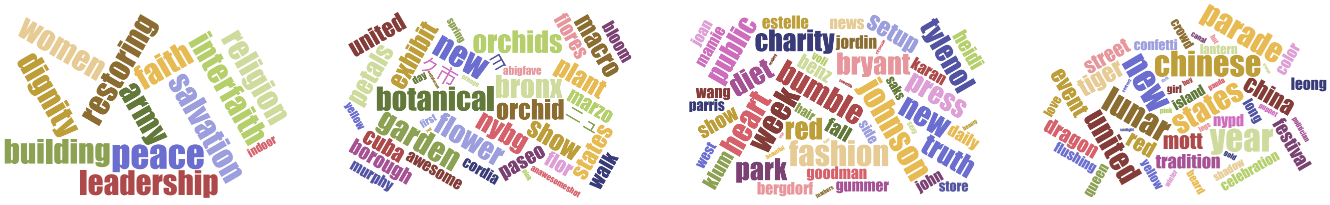
Compared methods.

- RCRP-geo. RCRP is regarded as one of the benchmark algorithms in modeling online event detection [Ahmed and Xing, 2008]. Since RCRP does not involve spatial distributions of events, we extend it to the spatio-temporal space by associating each topic with regions using a mixture of Gaussians(the same as the proposed model). The extended model is referred to as RCRP-geo.
- DCT-geo. DCT is one of the-state-of-art algorithms for modeling online topics [Liang *et al.*, 2016]. Similar to RCRP-geo, we extend it to spatial-temporal space. The topic number is predefined as the same value as DP-density.
- DP-density-fixed. It is a variation of the proposed model by fixing the event number as predefined. This model is used to measure the effectiveness of the proposed approach in inferring the right number of events.

4.2 Event Discovery from the Datasets

Content Analysis. First of all, to have an intuitive idea of events discovered from streaming datasets, Figure 2 shows some interesting events extracted from New York city, including Women Summit, Orchid Show, New York Fashion Week and Chinese New Year Parade. Even though tag vocabulary attached to Women Summit photos are very concise, still the semantic feature of this event is quite clear. From the perspective of interpretability and word coherence, our proposed model can effectively discover the meaningful events.

[1] http://www.flickr.com/services/api



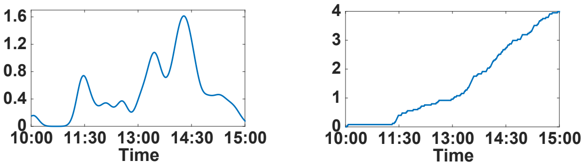
(a) Women Summit

(b) Orchid Show

(c) Fashion Week

(d) Chinese New Year Parade

Figure 2: Events discovered from New York



(a) Density distribution

(b) Cumulative density distribution

Figure 3: Event of Chinese New Year Parade in the temporal space

Temporal Dynamics. Figure 3 (a) and (b) plot the density and cumulative density distribution of Chinese New Year Parade on February 21th, respectively. As we notice, during the first one and half hour, indicated from the density value of Figure 3 (a), document arrivals are sparse, which is also confirmed by the gradient of cumulative density distribution in Figure 3 (b). Starting from 11:30 to 13:00, this event starts to be popular since the density value approximately keeps stable around 0.4. In the duration from 13:00 to 14:30, more documents are densely received and density value ranges from 0.4 to the peak of 1.6. During the last interval, the arrivals of documents are decreased as shown in the figure.

4.3 Perplexity Evaluation

In this section, we quantitatively evaluate the performance of the proposed method.

Metric *perplexity* [Blei *et al.*, 2003] is widely used to measure how well a probabilistic model predicts a sample. A lower perplexity indicates a better generalization of the model. Distinct from the conventional random selection of the testing dataset in the offline models, a set of documents are collected at each regular interval along the time, the aggregated testing collection accounts for 10% of the overall dataset. Specifically, the text perplexity is defined as

$$perplexity = exp\{-\frac{\sum_{d=1}^{M} \log p(\mathbf{w}_d)}{\sum_{d=1}^{M} N_d}\},$$

where M represents the testing collection, N_d records the number of words in document d, and $\mathbf{w}_d$ denote the words in document d.

Perplexity over varying temporal epochs. Figure 4 plots the perplexity value via varying numbers of epochs over both New York and Paris dataset. As we notice, our proposed model DP-density achieves a significant decrease in perplexity compared with competitor models over the two datasets. The performances of RCRP-geo and DCT-geo are greatly affected by the manual setting of epoch numbers, both of which gain their optimal value at the epoch number of 50 in two datasets. Specifically, DCT-geo outperforms RCRP-geo at any setting of epoch numbers in New York dataset.

Perplexity over varying event numbers. Figure 5 shows the comparison between the proposed DP-density and its variation DP-density-fixed. In the case of Paris dataset, the automatically inferred event number by the proposed model ranges from 55 to 62 through many trials, which is consistent with the performance of DP-density-fixed in terms of perplexity metric. In the New York dataset, perplexity of DP-density-fixed tells the optimal event number around 75, which is coincided with the automatical derived result (ranging from 70 to 75) by the proposed approach.

4.4 Term Coherence

We proceed to validate the interpretability of each discovered event using the coherence measure. Similar to [Guo and Gong, 2016], we define the coherence of the top-10 terms of each event as the average of PMI(Pointwise Mutual Information), which is calculated based on the coherence computation from [Röder *et al.*, 2015]. A higher PMI-Score indicates the terms within an event are more coherent and consistent to describe the event.

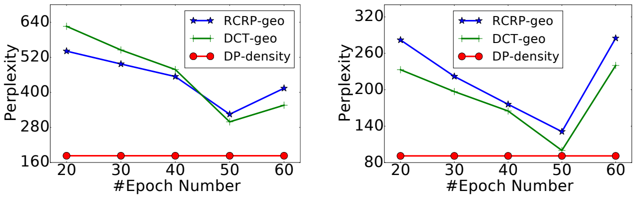
The coherence results are presented in Figure 6. It is observed that the coherence score of DP-density is always higher with any setting of epoch numbers, while the performances of RCRP-geo and DCT-geo decrease with the growing epoch number. In Paris dataset, RCRP-geo performs better than DCT-geo before the setting of 50 epoch number, and gains close result with DCT-geo as the epoch number increases. In New York dataset, DCT-geo outperforms RCRP-geo at any setting of epoch number, which is consistent with perplexity performance.

4.5 Efficiency

Figure 7 presents the event number with time evolution and document efficiency in Paris dataset. Since the event number is automatically learned in our proposed model, it is observed the event number grows gradually over time and keeps constant at 63 in Figure 7 (a) . Correspondingly, we expect average time cost of processing each document keeps roughly stable after running a time period, which is confirmed by the Figure 7 (b) from 18000th document after the long build-up period.

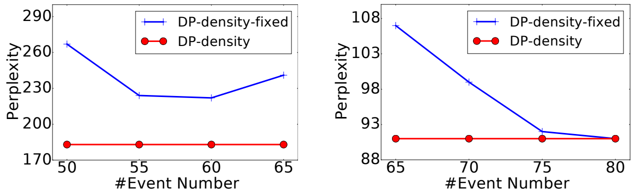
5 Related Work

Most of online event detection models are extensions of Latent Dirichlet Allocation(LDA) [Blei *et al.*, 2003]. We summarize the related work into two categories, Sequential



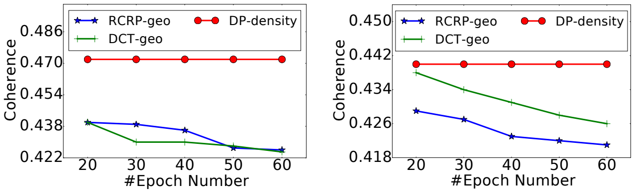
(a) Paris dataset (b) NY dataset

Figure 4: Perplexity performance over varying epoch numbers



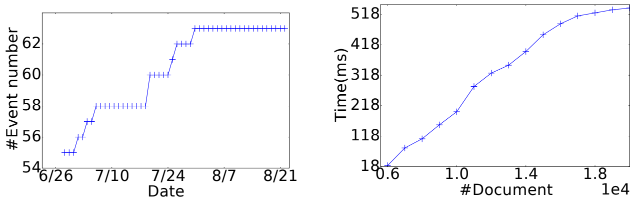
(a) Paris datast (b) NY dataset

Figure 5: Perplexity performance over varying event numbers



(a) Paris dataset (b) NY dataset

Figure 6: Coherence measure over two datasets



(a) Event number with time evolution (b) Document efficiency

Figure 7: Event number with time evolution and document efficiency in Paris dataset

Monte Carlo (SMC) based and Gibbs Sampling based inference approach according to the way of estimation.

Sequential Monte Carlo based inference process [Doucet *et al.*, 2000; Arnaud Doucet, 2001], requires one time pass over the streaming data since it sequentially processes the documents according to their arriving time. [Canini *et al.*, 2009] introduced online inference of SMC based on LDA while [May *et al.*, 2014] improved rejuvenation step of SMC. Whereas these two methods suffer from the manual setting of topic number in the case of online modeling. [Ahmed *et al.*, 2011b; 2011a] relieved such a setting and applied SMC inference to detect a hybrid of topics and storylines. However, all these methods fail to depict dynamic density of topics along the time. [Du *et al.*, 2015] exploited Dirichlet-Hawkes Process model to detect online topics from streaming news articles. But how to give the specific the number of kernels, and their respective locations and widths is an challenge, since documents are received continuously along the time. Another drawback is that it fails to investigate the function of spatial patterns of topics. In comparison, we proposed a density based online model in the joint spatio-temporal space, and temporal dynamics of events are flexibly learned from estimated density distribution in a sliding window.

Gibbs Sampling based inference process, requires multiple passes over the streaming collection to obtain a stable sampling for each document. Thus it is unsuitable for streaming data. Given the set of predefined time epochs, both RCRP[Ahmed and Xing, 2008] and DCT[Liang *et al.*, 2016] detected the most recent topic depending on learned results from previous epochs. One of main deficiencies of RCRP, DCT and their related work [Blei and Lafferty, 2006; Cheng *et al.*, 2014] is that it requires an explicit division of streaming documents into unit episode. Rather than divide time, [Wei *et al.*, 2007; Iwata *et al.*, 2009; Wang *et al.*, 2012; Amoualian *et al.*, 2016] built online topic modelings on the assumption of dependency between temporal consecutive documents. Among them, [Wei *et al.*, 2007; Wang *et al.*, 2012] modeled the dependency of topic distribution while [Iwata *et al.*, 2009; Amoualian *et al.*, 2016] modeled dependency of both topic distribution and topic transition between consecutive documents. However, the dependency assumption is rigid, since two consecutive documents from a dense document stream is possibly contributed by two simultaneous events and there is no temporal dependency between them. [Yin and Wang, 2016] presented an online clustering scheme in which the number of clusters (topics) is flexible to grow with data but limited by a predefined threshold. In contrast, the proposed DP-density aims to detect online events from social media data where both event number and event density can be flexibly learned without restrictions.

Besides, there are other online studies focused on specific events. [Zhang *et al.*, 2016] aimed to detect real-time local events from geo-tagged tweet streams. [Sakaki *et al.*, 2010] explored real-time earthquake detection while [Li *et al.*, 2012] detected crime and disaster events(CDE) with self-adaptive crawler.

6 Conclusion

We present a novel online event discovery model for social media data. The key aspects of our model are (1) flexibly inferring event number using Dirichlet Process to accommodate the complexity of continuous document arrivals, (2) dynamically learning the temporal dynamics of events using density estimation, (3) integrating combination of DP and density estimation into topic modeling in the joint spatio-temporal space. Our extensive experiments have demonstrated that our proposed model is able to discover interpretable events in terms of low perplexity and high coherence.

Acknowledgments

This work was supported in part by Fund of Science and Technology Development of Macau Government under FDCT/116/2013/A3 and FDCT/007/2016/AFJ and in part by University Macau Research Committee under MYRG2015-00070-FST and MYRG2017-00212-FST.

References

[Ahmed and Xing, 2008] Amr Ahmed and Eric P. Xing. Dynamic non-parametric mixture models and the recurrent chinese restaurant process: with applications to evolutionary clustering. In *SDM*, pages 219–230, 2008.

[Ahmed *et al.*, 2011a] Amr Ahmed, Qirong Ho, Jacob Eisenstein, Eric Xing, Alexander J Smola, and Choon Hui Teo. Unified analysis of streaming news. In *WWW*, pages 267–276, 2011.

[Ahmed *et al.*, 2011b] Amr Ahmed, Qirong Ho, Choon Hui Teo, Jacob Eisenstein, Alexander J Smola, and Eric P Xing. Online inference for the infinite topic-cluster model: Storylines from streaming text. In *AISTATS*, pages 101–109, 2011.

[Amoualian *et al.*, 2016] Hesam Amoualian, Marianne Clausel, Éric Gaussier, and Massih-Reza Amini. Streaming-lda: A copula-based approach to modeling topic dependencies in document streams. In *KDD*, pages 695–704, 2016.

[Arnaud Doucet, 2001] Nando de Freitas & Neil Gordon Arnaud Doucet, editor. *Sequential Monte Carlo Methods in Practice*. Springer, 2001.

[Blei and Lafferty, 2006] David M. Blei and John D. Lafferty. Dynamic topic models. In *ICML*, pages 113–120, 2006.

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

[Canini *et al.*, 2009] Kevin Robert Canini, Lei Shi, and Thomas L. Griffiths. Online inference of topics with latent dirichlet allocation. In *AISTATS*, pages 65–72, 2009.

[Cheng *et al.*, 2014] Xueqi Cheng, Xiaohui Yan, Yanyan Lan, and Jiafeng Guo. Btm: Topic modeling over short texts. *TKDE*, 26(12):2928–2941, 2014.

[Doucet *et al.*, 2000] Arnaud Doucet, Nando De Freitas, Kevin Murphy, and Stuart Russell. Rao-blackwellised particle filtering for dynamic bayesian networks. In *UAI*, pages 176–183, 2000.

[Du *et al.*, 2015] Nan Du, Mehrdad Farajtabar, Amr Ahmed, Alexander J Smola, and Le Song. Dirichlet-hawkes processes with applications to clustering continuous-time document streams. In *KDD*, pages 219–228, 2015.

[Fukunaga and Hostetler, 1975] Keinosuke Fukunaga and Larry Hostetler. The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Transactions on information theory*, 21(1):32–40, 1975.

[Guo and Gong, 2016] Jinjin Guo and Zhiguo Gong. A nonparametric model for event discovery in the geospatial-temporal space. In *CIKM*, pages 499–508, 2016.

[Hinneburg *et al.*, 1998] Alexander Hinneburg, Daniel A Keim, et al. An efficient approach to clustering in large multimedia databases with noise. In *KDD*, volume 98, pages 58–65, 1998.

[Iwata *et al.*, 2009] Tomoharu Iwata, Shinji Watanabe, Takeshi Yamada, and Naonori Ueda. Topic tracking model for analyzing consumer purchase behavior. In *IJCAI*, volume 9, pages 1427–1432, 2009.

[Li *et al.*, 2012] Rui Li, Kin Hou Lei, Ravi Khadiwala, and Kevin Chen-Chuan Chang. Tedas: A twitter-based event detection and analysis system. In *ICDE*, pages 1273–1276, 2012.

[Liang *et al.*, 2016] Shangsong Liang, Emine Yilmaz, and Evangelos Kanoulas. Dynamic clustering of streaming short documents. In *KDD*, pages 995–1004, 2016.

[May *et al.*, 2014] Chandler May, Alex Clemmer, and Benjamin Van Durme. Particle filter rejuvenation and latent dirichlet allocation. In *ACL (2)*, pages 446–451, 2014.

[Pan and Mitra, 2011] Chi-Chun Pan and Prasenjit Mitra. Event detection with spatial latent dirichlet allocation. In *JCDL*, pages 349–358, 2011.

[Röder *et al.*, 2015] Michael Röder, Andreas Both, and Alexander Hinneburg. Exploring the space of topic coherence measures. In *WSDM*, pages 399–408, 2015.

[Sakaki *et al.*, 2010] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In *WWW*, pages 851–860, 2010.

[Teh *et al.*, 2006] Yee Whye Teh, Michael I Jordan, Matthew J Beal, and David M Blei. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101(476), 2006.

[Wang *et al.*, 2012] Yu Wang, Eugene Agichtein, and Michele Benzi. TM-LDA: efficient online modeling of latent topic transitions in social media. In *KDD*, pages 123–131, 2012.

[Wei *et al.*, 2007] Xing Wei, Jimeng Sun, and Xuerui Wang. Dynamic mixture models for multiple time-series. In *IJCAI*, volume 7, pages 2909–2914, 2007.

[Yin and Wang, 2016] Jianhua Yin and Jianyong Wang. A text clustering algorithm using an online clustering scheme for initialization. In *KDD*, pages 1995–2004, 2016.

[Zhang *et al.*, 2016] Chao Zhang, Guangyu Zhou, Quan Yuan, Honglei Zhuang, Yu Zheng, Lance Kaplan, Shaowen Wang, and Jiawei Han. Geoburst: Real-time local event detection in geo-tagged tweet streams. In *SIGIR*, pages 513–522, 2016.

[Zhao *et al.*, 2016] Kaiqi Zhao, Lisi Chen, and Gao Cong. Topic exploration in spatio-temporal document collections. In *SIGMOD*, pages 985–998, 2016.