Hierarchical Dirichlet Gaussian Marked Hawkes Process for Narrative Reconstruction in Continuous Time Domain

Yeon Seonwoo, **Sungjoon Park** and **Alice Oh**
Department of Computing, KAIST, Republic of Korea
{yeon.seonwoo, sungjoon.park}@kaist.ac.kr, alice.oh@kaist.edu

Abstract

In news and discussions, many articles and posts are provided without their related previous articles or posts. Hence, it is difficult to understand the context from which the articles and posts have occurred. In this paper, we propose the Hierarchical Dirichlet Gaussian Marked Hawkes process (HD-GMHP) for reconstructing the narratives and thread structures of news articles and discussion posts. HD-GMHP unifies three modeling strategies in previous research: temporal characteristics, triggering event relations, and meta information of text in news articles and discussion threads. To show the effectiveness of the model, we perform experiments in narrative reconstruction and thread reconstruction with real world datasets: articles from the New York Times and a corpus of Wikipedia conversations. The experimental results show that HD-GMHP outperforms the baselines of LDA, HDP, and HDHP for both tasks.

1 Introduction

Online news sites and discussion forums generate large volumes of articles and discussions, which we can call “events”. To fully understand the discussions and the news stories, one often needs a larger context for that text, such as what related posts and relevant articles have been posted before. For instance, to understand a news article about the *presidential elections*, we would need to know the history of the candidates’ political actions through relevant previous articles. While there are some news articles with a curated set of related articles and discussion threads with a well-organized structure, there are many more articles and discussion threads for which the structure is absent or incomplete. In this context, automatically reconstructing the narrative of articles and thread structure is an important problem.

Generally, textual information and various meta information such as location and keywords are used as features to solve this problem of narrative reconstruction. With these features, previous research mainly focus on three modeling strategies. First, they model the triggering relationship of events to identify which preceding events led to the occurrence of the current event. Second, they use meta information such as location and keywords. Third, they consider the temporal characteristics in the event stream, such that events in close temporal proximity are more likely to be related. However, there is no method that effectively considers all three of these. In narrative reconstruction, there are several approaches that focus on using meta information and temporal characteristics with clustering methods (Zhou et al., 2016; Tang et al., 2015; Ahmed et al., 2011), and there are several approaches using the Hawkes process to model the temporal characteristics (Du et al., 2015; Mavroforakis et al., 2017; Jankowiak and Gomez-Rodriguez, 2017). In thread reconstruction, there are approaches that focus on modeling triggering relationships of events and using meta information (Kim et al., 2010; Louis and Cohen, 2015; Wang et al., 2011b).

In this paper, we propose a novel Gaussian Marked Hawkes Process (GMHP) that effectively reconstructs the narrative structure of articles and the thread structure of discussions considering all three modeling strategies. GMHP uses the Hawkes process to model events in continuous time, a Gaussian distribution for modeling the meta information of text, and the mixture of Gaussian for modeling the triggering relationships of events. The detailed modeling strategies are described as follows. We use the Hawkes process to model time in the continuous domain, as the Hawkes process is a stochastic process used to understand a sequence of events in continuous time

(Iwata et al., 2013; Rong et al., 2015). To use meta information, we represent text and meta information in a general vector form and use the Hawkes process to handle the vector of event information with a Gaussian distribution. To model the triggering relationships, we assume a model structure parameterized by each preceding event so that an event can be directly generated from a probability distribution parameterized by preceding events.

The GMHP models a single narrative or thread in event streams. To find the narratives or threads from a mixture of event streams, we combine our GMHP model with the Hierarchical Dirichlet Process to build HD-GMHP.

We evaluate the effectiveness of our model with two real world datasets: articles from the New York Times, and discussion threads from Wikipedia. In the New York Times dataset, we perform a narrative reconstruction experiment and compare the results with the human annotated narrative labels. In the Wikipedia discussion corpus, we perform two kinds of thread reconstruction experiment. One is grouping posts in the same thread. The other is reconstructing the post-reply structure of the posts. From these experiments, we see that our model outperforms the state-of-the-art model, the hierarchical Dirichlet Hawkes process (HDHP) (Mavroforakis et al., 2017).

The contributions of our research are threefold. First, we propose the Gaussian Marked Hawkes Process that effectively models a single narrative (event stream) with all three modeling strategies used in previous research. Second, we propose HD-GMHP, a combination of the GMHP model with the HDP to reconstruct the narratives of articles and the thread structure of discussions from a mixture of event streams. Finally, we propose a novel inference algorithm of the HD-GMHP with the Sequential Monte Carlo method (Doucet et al., 2001).

2 Related Work

Narrative Reconstruction: One major approach to reconstructing narratives from news articles is clustering articles by using a variant of the Chinese Restaurant Process (CRP). Related work such as (Zhou et al., 2016; Tang et al., 2015; Ahmed et al., 2011) models chronologically ordered news articles with text and various meta information including author, organization, keywords, and location. They use the CRP, distant-dependent CRP (Blei and Frazier, 2011). There is research that uses recurrent CRP (Ahmed and Xing, 2008) and exponential time decaying kernel to model probability of time difference between two relevant events. But they use discrete time information instead of continuous form and handcrafted parameters of the kernel (Ahmed et al., 2011).

There is another approach that reconstructs narratives by directly extracting important sentences from articles. (Xu et al., 2013) proposes a model that considers the sentence and image level narrative reconstruction as an optimization problem and solves it by maximizing the divergence of narratives with some constraints. (Wang et al., 2016) solves the narrative reconstruction problem as a sentence recommendation problem and uses matrix factorization. But these existing models focus on how to handle text and meta information of articles, while our model uses the Hawkes process to effectively model continuous time information of events.

Discussion Thread Reconstruction: There are several approaches to reconstruct threads from a corpus of unstructured discussions. (Wang et al., 2011a) uses Conditional Random Field to reconstruct reply structure in discussion corpus. (Balali et al., 2014) uses content, time and author information as features of a single post with rank SVM to reconstruct thread structure. (Dehghani et al., 2013; Aumayr et al., 2011) uses SVM and a decision tree with meta information of posts.

However, a major limitation in these previous research is that they are assuming that for each post, the main thread where it belongs is given. That is, the problem they solve is finding the post for which a post is immediately replying, rather than treating the corpus as a single set of posts with no known information about the threads, the initial post of each thread, and the posts that belong to each thread. This limitation of the previous research means those approaches are not applicable in more general online conversation data, such as IRC or a Facebook group chat which is a massive unstructured online discussion for which the initial post of a thread is not labeled. Unlike this strong assumption in previous research, we use a more general assumption that the initial posts are unknown, so our approach would be applicable to a wider, more general discussion data. Also, as in the narrative reconstruction research area, previous research focuses on how to handle text and meta information in posts. Again, unlike previous research, our research uses the Hawkes process to model continuous time information.

Continuous Time Modeling: The Hawkes process, a stochastic process that models continuous time information of events with event occurrence history, is an effective solution to model events in continuous time. One of the main research themes in the Hawkes process literature is finding which events trigger which other events. (He et al., 2015) models the topic diffusion patterns in a social network by inferring the triggering node with the Hawkes process. The Hawkes process is also used to model social event streams (Rong et al., 2015) and to classify rumors (Lukasik et al., 2016), and a combination of the Hawkes process and the Dirichlet mixture model is used to cluster event streams (Xu and Zha, 2017).

Recent research clusters text streams with the Hawkes process and the Chinese Restaurant Process or the Chinese Restaurant Franchise (Mavroforakis et al., 2017; Du et al., 2015). They use the bag-of-words representation of text in their model, while (Jankowiak and Gomez-Rodriguez, 2017) proposes a Hawkes process model that can handle a more general vector representation of events. The main difference of our model compared to this research is that we add the triggering relationship of two events. With this addition, our model can reconstruct narratives with an explicit relation of two documents.

3 Hawkes Processes

Before we describe our proposed model, we briefly explain the Hawkes process, one of two main stochastic processes used in our model. We leave out the explanation of the HDP due to space.

The Hawkes process (Hawkes, 1971) is a subclass of temporal point processes, whose functional form for intensity with exponential decaying kernel is represented as

$$\lambda^*(t) = \lambda_0(t) + \int_0^t \alpha\beta e^{-\beta(t-s)} dN(s),$$

where the intensity, $\lambda^*(t)$ represents the conditional probability of an event occurrence within time window $[t, t + dt)$. The Hawkes process is used to model the number of occurrences of events where one event can trigger other events. In the equation above, the base intensity $\lambda_0(t)$ models the intensity of events that occur on their own initiative whereas $\alpha\beta e^{-\beta(t-s)}$ models the intensity of events that are triggered by the previous event that occurred at time s. Here, multiplication of α and β represents influence of the previous event and β represents decaying rate of the influence. Thus, the effect of the previous event exponentially decays with respect to the time difference. From the definition of intensity $\lambda^*(t)$, the derived likelihood form of the Hawkes process is as follows,

$$f(D|\Theta) = e^{-\Lambda(T)} \prod_{i=1}^{n} \lambda^*(t_i), \qquad (1)$$

where $\Lambda(T) = \int_0^T \lambda^*(t)dt$.

4 Problem Setting

In this section, we define the event stream and the narrative and the thread reconstruction problems.

Definition of Event Stream: If a text appears at time t_i, we define the event s_i as $(t_i, \vec{e}_i, z_i, x_i)$. Here, $\vec{e}_i$ is the feature vector of the text, x_i is the latent global cluster indicator of event s_i which represents the cluster for events with similar text information, and z_i is the latent local cluster indicator for events that are temporally related in the same cluster. We define event stream S as $[s_1, .., s_n]$.

Assumptions: 1) We assume that two events in same local cluster occur in near time and have similar feature vectors $\vec{e}$. These properties are called temporal and spatial locality. 2) We assume hierarchy structure of a global cluster and a local cluster. That is, one global cluster can consist of multiple local clusters.

Problem Formulation: We formulate the spatial locality of two events in the same local cluster with a Gaussian distribution. If two events s_i and s_j are in the same local cluster and $t_i > t_j$, then we assume the later event $\vec{e}_i$ is generated from one of two relations,

$$\vec{e_i} \sim \mathcal{N}(\vec{e_j}, \Sigma_v), \quad \vec{e}_i, \sim \mathcal{N}(\vec{e}_0, \Sigma_0).$$

Here, $\vec{e}_0$ is the base event vector and Σ_0 is the covariance matrix of the cluster. Σ_v is the covariance matrix of the Gaussian distribution generated by a past event in the cluster.

We use the Hawkes process to formulate the temporal locality of two events in the same local cluster. If event s_i and s_j are in the same local cluster and $t_i > t_j$, then t_i is generated from in either following relations,

$$t_i \sim \text{Poisson Process}(\mu),$$
$$t_i - t_j \sim \text{Hawkes}(\alpha, \beta).$$

Here, if t_i is generated from Hawkes process of parameter α and β with time t_j and $\vec{e}_i$ is generated from $\vec{e}_j$, then we say that event s_j is the *parent event* of event s_i.

We formulate the hierarchy structure of the global cluster and the local cluster with Hierarchical Dirichlet Process (Teh et al., 2006). If the parameters θ_{z_i} of local clusters $z_1, z_2, .., z_n$ are equal to the parameters of the global cluster Θ_x, then we say that there is a hierarchy between all the local clusters and the global cluster. And this hierarchy structure can be written as follows,

$$\Theta_x = \theta_{z_1} = \theta_{z_2} = ... = \theta_{z_n}.$$

Now, we define the narrative reconstruction and the thread reconstruction problem as a problem of inferring the latent variables in S.

5 Model

We now describe clustering a mixture of event stream S with the Gaussian Marked Hawkes Process and the hierarchical Dirichlet process. We first propose Gaussian Marked Hawkes Process (GMHP) that models temporal and spatial locality assumptions that described in section 4. And after defining the GMHP, we propose Hierarchical Dirichlet Gaussian Marked Hawkes Process (HD-GMHP), a combination of the GMHP with the Hierarchical Dirichlet Process. The GMHP models event streams with the same local cluster z and HDP groups the local clusters to one global cluster x.

5.1 Gaussian Marked Hawkes Processes

5.1.1 Model Description

In GMHP, we assume events are generated by a past event or by their own initiative. If event s_i is generated by event s_j, then we say that event s_j is the parent event of event s_i. If event s_i occurs on their own initiative, the index of the parent event is 0. We define the intensity function with the given parent event c_i as follows,

$$\lambda(t_i|c_i) = \begin{cases} \mu & \text{if } c_i = 0 \\ \alpha\beta e^{-\beta(t_i - t_{c_i})} & \text{otherwise} \end{cases} \quad (2)$$

To model the spatial locality of two D-dimensional event vectors $\vec{e}_i$, $\vec{e}_{c_i}$, we define probability distribution for $\vec{e}_i$ as follows,

$$p_{c_i}(\vec{e}_i) = \begin{cases} \mathcal{N}(\vec{e}_i|\vec{e}_0, \Sigma_0) & \text{if } c_i = 0 \\ \mathcal{N}(\vec{e}_i|\vec{e_{c_i}}, \Sigma_v) & \text{otherwise} \end{cases} \quad (3)$$

Here, $\vec{e}_0$ is the base event vector for when $c_i = 0$. Σ_0 and Σ_v are covariance matrix for when an event occurs by their own initiative or occurs by past event. From the above definitions, we can calculate the intensity of the event vector $\vec{e}$ at time t as follows,

$$\lambda_{\vec{e}}(t) = \mu\mathcal{N}(\vec{e}|\vec{e}_0, \Sigma_0) + \sum_{t_j<t} \lambda(t|j)\mathcal{N}(\vec{e}|\vec{e}_j, \Sigma_v). \quad (4)$$

The total intensity of GMHP can be obtained by integrating the above intensity with the event vector $\vec{e}$.

$$\lambda(t) = \int_{\mathbb{R}^D} \lambda_{\vec{e}}(t)\, d\vec{e} = \mu + \sum_{t_j<t} \lambda(t|j). \quad (5)$$

5.1.2 Parameter estimation

From equation 1, the likelihood of the observed event stream can be computed as follows,

$$f(D|\theta) = e^{-\Lambda(T)} \prod_{i=1}^{n} \sum_{0\le j<i} p_j(\vec{e}_i)\lambda(t_i|j), \quad (6)$$

where $\Lambda(T) = \mu T + \sum_{i=1}^{n} \alpha(1 - e^{-\beta(T-t_i)})$.

Since the likelihood of GMHP is hard to maximize, instead of using the likelihood, we define a likelihood with the given parent events as follows,

$$\begin{aligned} f(D|C,\theta) =& e^{-\Lambda(T)} \times \prod_{i=1}^{n} \{(\mu\mathcal{N}(\vec{e}_i|\vec{e}_0, \Sigma_0))^{C_{i0}} \times \\ & \prod_{j=1}^{i-1} (\alpha\beta e^{\beta(t_i-t_j)}\mathcal{N}(\vec{e}_i|\vec{e}_j, \Sigma_v))^{C_{ij}}\}, \end{aligned} \quad (7)$$

where C_{ij} becomes 1 when $c_i = j$ and 0 otherwise. By maximizing equation 7, we can estimate the parameter $\theta = \{\mu, \alpha, \vec{e}_0, \Sigma_0, \Sigma_v\}$. The inference step of the parent events is described in section 6.

5.2 Modeling a Mixture of GMHP with the HDP

When clustering a mixture of streams using the Hawkes process, the exponential triggering function prevents two events with a large time difference from being assigned to the same global cluster. To solve this problem, (Mavroforakis et al., 2017) uses the HDP instead of using the Dirichlet process used in (Du et al., 2015). The hierarchy structure of the HDP assigns a cluster label with a probability proportional to the size of the cluster. This allows assignment of two events with a large time difference to the same cluster. For the same reason, we use the HDP to model mixture of the GMHP. We consider each GMHP in mixture as a table in the Chinese Restaurant Franchise metaphor. Since the intensity of k'th GMHP, $\lambda_k(t)$ represents how likely an event occurs in table k at time t, we use the intensity as the number of customers in the CRF metaphor. The whole generative process of HD-GMHP is as follows.

1. Initialize the number of local clusters $K = 0$, the number of global clusters $M = 0$.
2. For $n \in 1, 2, ..., N$
   (a) Draw t_n from Hawkes$(\lambda_0 + \sum_{k=0}^{K} \lambda_k)$
   (b) Draw z_n as follows.
   $$z_n \sim \lambda_0 \delta(K+1) + \sum_{k=1}^{K} \lambda_k \delta(k) \quad (8)$$
   (c) If $z_n = K + 1$, assign global cluster x_n, which is interpreted as parameter(θ_{x_n}) for local cluster z_n, and Increment K. Here, N_m is number of local cluster in global cluster m.
   $$x_n \sim \gamma \delta(M+1) + \sum_{m=1}^{M} N_m \delta(m) \quad (9)$$
   (d) If $x_n = M + 1$, increment M and draw new parameter as follows.
   $\alpha_{x_n} \sim \Gamma(\alpha_a, \beta_a)$, $\mu_{x_n} \sim \Gamma(\alpha_\mu, \beta_\mu)$
   $\frac{1}{\Sigma_0^{x_n}} \sim \Gamma(\alpha_0, \beta_0)$, $\frac{1}{\Sigma_v^{x_n}} \sim \Gamma(\alpha_v, \beta_v)$
   $\vec{e}_{0,x_n} \sim \mathcal{N}(\vec{e}_0, \Sigma_0^{x_n} / \vec{\lambda}_{e0})$
   Note that we assume that the covariance matrix $\Sigma_v^{x_n}$, and $\Sigma_0^{x_n}$ are diagonal.
   (e) Draw c_n and $\vec{e}_n$. Here, $g_{x_n}(t) = \alpha_{x_n} \beta e^{-\beta(t_n - t)}$.
   $$c_n \sim \mu_{x_n} \delta(N_{z_n} + 1) + \sum_{j=1}^{N_{z_n}} g_{x_n}(t_j) \delta(j) \quad (10)$$
   if $c_n = N_{z_n} + 1$, then replace c_n with 0 and sample event vector.
   $$\vec{e}_n \sim \begin{cases} \mathcal{N}(\vec{e}_{0,x_n}, \Sigma_0^{x_n}) & \text{if } c_n = 0 \\ \mathcal{N}(\vec{e}_{c_n}, \Sigma_v^{x_n}) & \text{otherwise} \end{cases} \quad (11)$$

λ_0, γ, $\vec{e}_0$, $\vec{\lambda}_{e0}$, (α_a, β_a), (α_μ, β_μ), (α_0, β_0), and (α_v, β_v) are the hyperparameters used in HD-GMHP.

6 Inference

To infer the latent variables z and x for each event from an observed event stream $s_o^{1:n}$ where $s_o^i = (t_i, \vec{e}_i)$ with observation time T, we propose an online inference algorithm with Sequential Monte Carlo (SMC) (Doucet et al., 2001). To calculate the posterior of the latent variables z and x for each timestamp t_i in the inference, we need the estimated parameter to calculate the intensity at each time t_i, $\lambda(t_i)$. As described in section 5.1.2, the parameter estimation step needs the parent event information. In our proposed inference, the parent events are inferred from SMC. The inference algorithm is summarized in algorithm 1.

Algorithm 1 Inference

```
Input: Stream data $S_o$
Initialize $w_1^i = \frac{1}{P}, i \in \{1, 2, ..., P\}$.
for $n = 1$ to $N$ do
    for $i = 1$ to $P$ do
        Update $\Theta$ as described in section 6.2.
        Sample $(x, z, c)_n^i$ with equation 16, 10
        Update $w_n^i$ with equation 17
    end for
    Normalize $w_n^{1:P}$
    if $||w_n||_2^{-2} < thresh$ then
        Resample particles
    end if
end for
```

6.1 Sequential Monte Carlo with parent event inference

To approximate the posterior of the latent variables, SMC samples the latent variables from the proposal distribution and calculates the weight of each sampled variables which is called the particle weight. To infer the parent event in SMC, we define the particle weight of our modified SMC as follows:

$$w_n^i = \frac{p(\psi_{1:n}^i | s_o^{1:n})}{q(\psi_{1:n}^i | s_o^{1:n})} \frac{p(c_{1:n}^i | \psi_{1:n}^i, s_o^{1:n})}{q(c_{1:n}^i | \psi_{1:n}^i, s_o^{1:n})} \quad (12)$$

Here, ψ_n^i is (x_n^i, z_n^i). Let the left part on the right hand term and right part on the right hand term of the equation 12 are $w_{\psi}{}_n^i$ and $w_c{}_n^i$. Then the terms can be calculated by $w_{\psi}{}_n^i = \eta_\psi w_{\psi}{}_{n-1}^i$ and $w_c{}_n^i = \eta_c w_c{}_{n-1}^i$, where the η_ψ is

$$\frac{p(\vec{e}_n, t_n, \psi_n | s_o^{1:n-1}, \psi_{1:n-1}^i)}{q(\psi_n^i | \psi_{1:n-1}^i, s_o^{1:n})}. \quad (13)$$

and η_c is

$$p(\vec{e}_n | t_n, s_o^{1:n-1}, \delta_{1:n}^i) \frac{p(c_n^i | t_n, \delta_{1:n-1}^i, s_o^{1:n-1}, \psi_n^i)}{q(c_n^i | \delta_{1:n-1}^i, s_o^{1:n}, \psi_n^i)}. \quad (14)$$

Here, δ_n^i is (x_n^i, z_n^i, c_n^i).

We use $p(\psi_n|\psi_{1:n-1}, s_o^{1:n})$ as the proposal distribution of ψ_n^i in the equation 13 to minimize the variance of w_n^i (Doucet et al., 2000) and $p(c_n|\delta_{1:n-1}, \psi_n, t_n, s_o^{1:n-1})$ as the proposal distribution of c_n^i in the equation 14. From the above proposal distribution, $\eta_{c_n^i}$ can be calculated as $\eta_{c_n^i} = p(\vec{e}_n|t_n, s_o^{1:n-1}, c_{1:n}^i, \psi_{1:n})$ and $\eta_{\psi_n^i}$ can be calculated by the following form .

$$\begin{aligned}\eta_{\psi_n^i} =& p(t_n|\psi_{1:n-1}, z_n, s_o^{1:n-1}) \\ & \times \sum_{z_n}(p(\vec{e}_n|\psi_{1:n-1}, z_n, t_n, s_o^{1:n-1}) \\ & \times p(z_n|t_n, \psi_{1:n-1}, s_o^{1:n-1}))\end{aligned} \tag{15}$$

From the proposal distribution of ψ_n^i, we can sample ψ_n^i as follows:

$$\begin{aligned}p(\psi_n|rest) \propto & p(\vec{e}_n|\psi_{1:n}, t_n, s_o^{1:n-1}) \\ & \times p(\psi_n|\psi_{1:n-1}, t_n, s_o^{1:n-1}) \\ & \times p(t_n|\psi_{1:n-1}, s_o^{1:n-1})\end{aligned} \tag{16}$$

Here, the term $p(\vec{e}_n|\psi_{1:n}, t_n, s_o^{1:n-1}) \times p(\psi_n|\psi_{1:n-1}, t_n, s_o^{1:n})$ can be simply calculated by the student's t-distribution derived from the conjugate relation between the parameter $\{\vec{e}_{k0}, \Sigma_{0,k}, \Sigma_{v,k}\}$ and the normal-inverse-gamma and inverse-gamma prior in the generative process of HD-GMHP.

From $\eta_{c_n^i} = p(\vec{e}_n|t_n, s_o^{1:n-1}, c_{1:n}^i, \psi_{1:n})$ and 15, the particle weight can be updated by the following,

$$\begin{aligned}w_n^i \propto & w_{n-1}^i \\ & \times p(t_n|s_o^{1:n-1}, \psi_{1:n}^i) p(\vec{e}_n|c_n, t_n, s_o^{1:n-1}, \psi_{1:n}) \\ & \times \sum_{z_n}(p(\vec{e}_n|z_n, \psi_{1:n-1}, t_n, s_o^{1:n-1}) \\ & \times p(z_n|t_n, \psi_{1:n-1}, s_o^{1:n-1})).\end{aligned} \tag{17}$$

When calculating the probability of t_n in 17, we assume the parameters $\mu_{1:K}, \alpha_{1:K}$ are given (Carvalho et al., 2010). From the likelihood of GMHP, the probability term $p(t_n|\psi_{1:n}, s_o^{1:n-1})$ in equation 17 can be calculated by $\lambda_{z_n}(t_n)e^{-\Lambda(t_n, t_{n-1})}$. Where $\Lambda(t_n, t_n - 1)$ is

$$\begin{aligned}&\lambda_0(t_n - t_{n-1}) + (t_n - t_{n-1})\sum_{k=1}^{K}\mu_k \\ &+ \frac{1}{\beta}(1 - e^{-\beta(t_n - t_{n-1})})\sum_{k=1}^{K}\lambda_k(t_{n-1}).\end{aligned} \tag{18}$$

In the case of the probability term $p(\vec{e}_n|c_n, rest)$ and $p(\vec{e}_n|z_n, rest)$ in 17, as explained in the sampling process of ψ_n, we can calculate the terms by student's t-distribution. With the particle weight update rule 17 and the parameter update rule described in section 6.2, we infer latent variables with algorithm 1.

6.2 Updating Parameter

From the equation 7 and the prior of the parameters used in GMHP, we can estimate the parameters by following form.

$$\alpha_m = \frac{\alpha_a - 1 + \sum_{x_i=m}\sum_{0<j<i} C_{ij}}{\beta_a + \sum_{x_i=m}(1 - e^{-\beta(T - t_i)})} \tag{19}$$

$$\mu_m = \frac{\alpha_\mu - 1 + \sum_{x_i=m} C_{i0}}{\beta_\mu + \sum_{\theta_k=\Theta_m}(T - t_{0,k})} \tag{20}$$

$$\vec{e}_{0,m} = \frac{\vec{e}_0 \circ \vec{\lambda}_{e0} + \sum_{x_i=m} C_{i0}\vec{e}_i}{\vec{\lambda}_{e0} + \sum_{x_i=m} C_{i0}} \tag{21}$$

$$\begin{aligned}diag(\Sigma_0^m) =& \{\vec{\lambda}_{e0} \circ (\vec{e}_{0,m} - \vec{e}_0)^2 + 2\vec{\beta}_0 \\ & + \sum_{x_i=m} C_{i0}(\vec{e}_i - \vec{e}_{0,m})^2\} \\ & \times \{2\vec{\alpha}_0 + 3 + \sum_{x_i=m} C_{i0}\}^{-1}\end{aligned} \tag{22}$$

$$diag(\Sigma_v^m) = \frac{2\vec{\beta}_v + \sum_{x_i=m}\sum_{0<j<i} C_{ij}(\vec{e}_i - \vec{e}_j)^2}{2 + 2\vec{\alpha}_v + \sum_{x_i=m}\sum_{0<j<i} C_{ij}} \tag{23}$$

6.3 Approximation

To reduce the computation time in the inference algorithm, we use several approximation strategies.

6.3.1 Marginal distribution Approximation

To calculate $p(\vec{e}_n|z_n, \psi_{1:n-1}, t_n, s_o{}^{1:n-1})$ in the equation 17, we need marginalization of $p(\vec{e}_n, c_n|z_n, \psi_{1:n-1}, t_n, s_o^{1:n-1})$ which takes time complexity of $O($n of events in $z_n)$ and cause the time complexity of the equation 17 to be $O(n)$. To reduce the time complexity, we note that event vector $\vec{e}_n$ is sampled from a Gaussian mixture that the influence of each Gaussian distribution is exponentially decreases. We assume the marginal distribution $p(\vec{e}_n|z_n, \psi_{1:n-1}, t_n, s_o^{1:n-1})$ can be approximated to $p(\vec{e}_n|c_{1:n} = 0, z_n, \psi_{1:n-1}, t_n, s_o^{1:n-1})$. From the approximation, we can calculate the posterior predictive with student's t-distribution. The result of approximation is as follows,

$$p(\vec{e}_n|z_n, \psi_{1:n-1}, t_n, s_o^{1:n-1}) = t_{\nu_n}(\vec{e}_n|\vec{m}_n, \frac{\vec{\kappa}_n + 1}{\vec{\kappa}_n \nu_n} \vec{S}_n), \quad (24)$$

where

$$\nu_n = 2\alpha_0 + N_{z_n}, \quad \kappa_n = \vec{\lambda}_{e_0} + N_{z_n},$$

$$\vec{m}_n = \frac{\vec{\lambda}_{e_0} \circ \vec{e}_0 + \sum_{z_i = z_n} \vec{e}_i}{\vec{\kappa}_n},$$

$$\vec{S}_n = 2\beta_0 + \sum_{z_i = z_n} \vec{e}_i^2 + \vec{\lambda}_{e_0} \circ \vec{e}_0^2 - \kappa_n \vec{m}_n^2.$$

To calculate $p(\vec{e}_n|c_n, t_n, s_o^{1:n-1}, \psi_{1:n})$ in the equation 17, we need to calculate posterior predictive for each past event. To reduce the computation time in the process of calculation, we approximate the probability distribution $p(\vec{e}_n|c_n, t_n, s_o^{1:n-1}, \psi_{1:n})$ as follows.

$$p(\vec{e}_n|c_n, t_n, s_o^{1:n-1}, \psi_{1:n}) \approx \begin{cases} \mathcal{N}(\vec{e}_{0,x_n}, \Sigma_0^{x_n}) & \text{if } c_n = 0 \\ \mathcal{N}(\vec{e}_{c_n}, \Sigma_v^{x_n}) & \text{otherwise} \end{cases} \quad (25)$$

6.3.2 Sampling c_n from recent W events

Sampling c_n has time complexity of $O(N_{z_n})$. To reduce the time complexity to $O(1)$, we sample c_n from recent W events in the local cluster z_n.

7 Experiment

In this section, we demonstrate the narrative reconstruction and thread reconstruction performance of our model on a corpus of the New York Times articles and the Wikipedia conversation dataset.

7.1 Dataset

New York Times Dataset: We collected 112,538 New York Times news articles from January 2016 to July 2017. The dataset contains the text, timestamp, the news section, and the keywords. These keywords are semantic tags specified by the newsroom to indicate the main topics of the articles. We select news articles in sections "U.S.", "World", "Opinion", and "Sports" that contain at least one of the top ten most frequently used keywords. The statistics of these keywords are described in table 1. Further, we select articles with more than ten words in its body. The final number of articles used in our experiment is 16,858. The dataset is publicly available [1].

Table 1: Statistics of keywords. "N" column lists the number of articles with the corresponding keyword.

Keyword	N
Trump, Donald J	7940
Presidential Election of 2016	5737
United States Politics and Government	4986
Republican Party	2371
Clinton, Hillary Rodham	2330
Baseball	2058
United States International Relations	1817
Terrorism	1618
Obama, Barack	1551
Russia	1400

Wikipedia Conversation Dataset is released by (Danescu-Niculescu-Mizil et al., 2012). The dataset contains the timestamp, the initial post of the conversation, "reply to" link information, and the text information of each post in conversation threads in Wikipedia talk pages. We select threads that have ten or more posts from September 2010 to December 2010. The final number of posts used in our experiment is 2,004 and the final number of threads is 154.

7.2 Preprocessing

To apply our model to the real world datasets, we represent each event with time information and an event vector. For the time information, we take the first article or post and set the time as zero, the last article or post as one, and scale the timestamps of all other articles and posts accordingly. To extract the event vectors, we use different vectorization methods for the two datasets. For the NYT dataset, we use the document topic vector from LDA (Blei et al., 2003). For the Wikipedia dataset, because there are only a few words in each post, we cannot use the LDA topic vector, so we use the averaged word embedding vector (Mikolov et al., 2013) of the words used in each post.

[1] https://github.com/yeonsw/NYT-dataset

Table 2: Narrative reconstruction results in NYT dataset and post grouping results in Wikipedia conversation dataset.

		AMI	ARI
NYT	LDA + DBSCAN	0.0627	0.0117
	HDP + DBSCAN	0.0260	0.0203
	HDHP	0.1768	0.0746
	HD-GMHP (100D)	**0.2479**	**0.1416**
Wiki	W2V + DBSCAN	0.0055	0.0001
	HDHP	0.4240	0.3512
	HD-GMHP (100D)	**0.5848**	**0.3834**

7.3 Task

Narrative reconstruction: To demonstrate the narrative reconstruction performance of our model, we apply the inference method to our corpus of NYT articles. We use a set of multiple keywords of each article as the ground truth label. Then we run our model and consider the set of articles with the same global cluster information as one narrative. We compare the results with the ground truth labels using the common clustering metrics AMI and ARI (Hubert and Arabie, 1985; Vinh et al., 2010) to evaluate the narrative reconstruction performance of our model. We compare HD-GMHP with the following baselines: LDA and HDP with DBSCAN, and the Hierarchical Dirichlet Hawkes Process (HDHP) (Mavroforakis et al., 2017) which is a state-of-the-art model for text and continuous timestamps of an event. Also, to measure the similarity of each recovered narrative and the ground truth narrative, we use the F1 score of the top ten narratives.

Thread reconstruction: In this experiment, we use two evaluation criteria. One is post grouping and the other is reply structure recovery, which is simply the recovery of the child nodes. Here, we use a different child node recovery task compared to the child node recovery used in previous research. In our task, we do not give the initial post of each thread, while previous research does. This makes thread reconstruction problem more general and more difficult.

In post grouping, we use the initial post of each of the posts as the ground truth label and measure the clustering metrics used in the NYT dataset. In the child node recovery experiment, we use the parent event information inferred from our method as the recovered tree structure of the threads. We measure the performance with node precision and node recall metrics (Wang et al., 2011a; Dehghani et al., 2013). We compare our model with the following baselines: HDHP, and a naive baseline that reconstructs threads in the form of a single linked list of posts in chronological order.

Table 3: F1-score of each label

Label	N	HD-GMHP	HDHP
Baseball	2011	0.8114	0.8899
Trump, Donald &Politics and Government	1664	0.1833	0.2157
Terrorism	1260	0.6052	0.4059
Trump, Donald &Election	1110	0.2537	0.1939
Trump, Donald	994	0.0975	0.1227
Politics and Government	822	0.1677	0.1215
Clinton, Hillary &Election &Trump, Donald	755	0.1754	0.1402
Election	714	0.1378	0.1280
Clinton, Hillary &Election	665	0.1669	0.1157
Russia	637	0.3177	0.2223
Micro F-score	N/A	**0.2874**	0.2189
Macro F-score	N/A	**0.3637**	0.3165

7.4 Metrics

AMI, ARI are commonly used to measure clustering performance (Hubert and Arabie, 1985; Vinh et al., 2010). $\mathbf{P_{node}}$, $\mathbf{R_{node}}$ measure local similarity between two thread structures (Wang et al., 2011a).

$$\mathrm{P_{node}} = \frac{1}{N}\sum_{i=1:N}\frac{|\mathrm{child_{GT}}(i) \cap \mathrm{child_{E}}(i)|}{|\mathrm{child_{E}}(i)|}$$
$$\mathrm{R_{node}} = \frac{1}{N}\sum_{i=1:N}\frac{|\mathrm{child_{GT}}(i) \cap \mathrm{child_{E}}(i)|}{|\mathrm{child_{GT}}(i)|}$$

where, $\mathrm{child_{GT}}(i)$ and $\mathrm{child_{E}}(i)$ are the sets of children of node i in the ground truth thread structure and the recovered thread structure, respectively. The author (Wang et al., 2011a) also proposed $\mathrm{P_{path}}$, $\mathrm{R_{path}}$ to measure the similarity of the global structure of two threads. The path metrics are sensitive to the recovered initial post of each thread, but since we do not give the initial post of each thread in our experiment, the path metrics are

Table 4: Reply structure recovery results in Wikipedia conversation dataset.

	$\mathrm{P}_{\mathrm{node}}$	$\mathrm{R}_{\mathrm{node}}$	$\mathrm{F1}_{\mathrm{node}}$
Naive Baseline	0.3223	**0.6501**	0.4310
HDHP	0.5598	0.5834	0.5714
HD-GMHP	**0.6433**	0.5468	**0.5911**

no longer proper in our experiment. So we measure the node metrics only.

7.5 Results

Table 2 shows the clustering accuracy of our method and the baseline methods in real world datasets. We average the results with five runs for each model. The highest value for each metric is indicated with boldface. From the results, we establish that our model outperforms the baseline methods in both the NYT narrative reconstruction task and the Wikipedia thread reconstruction task.

For the NYT, to see the accuracy of our model in more detail, we compute and show the F-scores for the top ten most frequent labels and the micro and macro averages in table 3. To compute the F-score between the true labels and the recovered cluster labels, we select the cluster with the highest F-score as the corresponding cluster. From the results, we establish that our model performs better than the baseline model, HDHP.

Table 4 shows the thread reconstruction results of our model and the baseline models in the Wikipedia conversation dataset. Since the HDHP model does not infer the parent event, we reconstruct threads in the form of chronologically ordered linked list of posts in each local cluster that inferred from HDHP. From the $\mathrm{F1}_{\mathrm{node}}$ score of the results, we establish our model performs better than other baseline models.

To demonstrate the robustness of HD-GMHP on dimensional change of the input vector, we measure the performance of each task in using 50, 100, and 150 dimensional vectors. The results are described in table 5 and 6. From the results, we verify there are no drastic changes in performance in both the NYT dataset and the Wikipedia dataset.

8 Conclusion

In this paper, we defined the narrative and thread reconstruction problems as clustering problems. To cluster the event streams with continuous time information and triggering event information, we proposed the Gaussian Marked Hawkes process that models event streams with additional event information represented in a vector form. Furthermore, we combined our model GMHP with the HDP to cluster event streams (HD-GMHP). We showed that our model performs better than several baseline methods in both narrative reconstruction in a dataset of NYT articles and thread reconstruction in a dataset of Wikipedia conversations.

Table 5: Model Robustness on dimensional change of input vectors in NYT dataset.

	AMI	ARI
HD-GMHP (50D)	0.2310	0.1518
HD-GMHP (100D)	0.2479	0.1416
HD-GMHP (150D)	0.2421	0.1191

Table 6: HD-GMHP model robustness on dimensional change of input vector in Wikipedia conversation dataset.

	AMI	ARI	$\mathrm{P}_{\mathrm{node}}$	$\mathrm{R}_{\mathrm{node}}$
50D	0.5836	0.3782	0.6466	0.5554
100D	0.5848	0.3834	0.6433	0.5468
150D	0.5948	0.3670	0.6450	0.5473

Acknowledgments

Yeon Seonwoo is supported by NCSoft Corporation. Sungjoon Park and Alice Oh are supported by the Engineering Research Center Program through the National Research Foundation of Korea (NRF) funded by the Korean Government MSIT (NRF-2018R1A5A1059921).

References

Amr Ahmed, Qirong Ho, Jacob Eisenstein, Eric Xing, Alexander J Smola, and Choon Hui Teo. 2011. Unified analysis of streaming news. In *WWW*.

Amr Ahmed and Eric Xing. 2008. Dynamic nonparametric mixture models and the recurrent chinese restaurant process: with applications to evolutionary clustering. In *SIAM International Conference on Data Mining*.

Erik Aumayr, Jeffrey Chan, and Conor Hayes. 2011. Reconstruction of threaded conversations in online discussion forums. In *ICWSM*.

A Balali, H Faili, and M Asadpour. 2014. A supervised approach to predict the hierarchical structure of conversation threads for comments. *The Scientific World Journal*, 2014.

David M Blei and Peter I Frazier. 2011. Distance dependent chinese restaurant processes. *Journal of Machine Learning Research*, 12(Aug):2461–2488.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent Dirichlet Allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Carlos M Carvalho, Michael S Johannes, Hedibert F Lopes, Nicholas G Polson, et al. 2010. Particle learning and smoothing. *Statistical Science*, 25(1):88–106.

Cristian Danescu-Niculescu-Mizil, Lillian Lee, Bo Pang, and Jon Kleinberg. 2012. Echoes of power: Language effects and power differences in social interaction. In *WWW*.

Mostafa Dehghani, Azadeh Shakery, Masoud Asadpour, and Arash Koushkestani. 2013. A learning approach for email conversation thread reconstruction. *Journal of Information Science*, 39(6):846–863.

Arnaud Doucet, Nando De Freitas, and Neil Gordon. 2001. An introduction to sequential monte carlo methods. In *Sequential Monte Carlo methods in practice*, pages 3–14. Springer.

Arnaud Doucet, Nando De Freitas, Kevin Murphy, and Stuart Russell. 2000. Rao-blackwellised particle filtering for dynamic bayesian networks. In *UAI*.

Nan Du, Mehrdad Farajtabar, Amr Ahmed, Alexander J Smola, and Le Song. 2015. Dirichlet-hawkes processes with applications to clustering continuous-time document streams. In *SIGKDD*.

Alan Hawkes. 1971. Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, pages 83–90.

Xinran He, Theodoros Rekatsinas, James Foulds, Lise Getoor, and Yan Liu. 2015. Hawkestopic: A joint model for network inference and topic modeling from text-based cascades. In *ICML*.

Lawrence Hubert and Phipps Arabie. 1985. Comparing partitions. *Journal of classification*, 2(1):193–218.

Tomoharu Iwata, Amar Shah, and Zoubin Ghahramani. 2013. Discovering latent influence in online social activities via shared cascade poisson processes. In *SIGKDD*.

Martin Jankowiak and Manuel Gomez-Rodriguez. 2017. Uncovering the spatiotemporal patterns of collective social activity. In *SIAM International Conference on Data Mining*.

Su Nam Kim, Li Wang, and Timothy Baldwin. 2010. Tagging and linking web forum posts. In *CoNLL*.

Annie P Louis and Shay B Cohen. 2015. Conversation trees: A grammar model for topic structure in forums. In *EMNLP*.

Michal Lukasik, PK Srijith, Duy Vu, Kalina Bontcheva, Arkaitz Zubiaga, and Trevor Cohn. 2016. Hawkes processes for continuous time sequence classification: an application to rumour stance classification in twitter. In *ACL*.

Charalampos Mavroforakis, Isabel Valera, and Manuel Gomez-Rodriguez. 2017. Modeling the dynamics of learning activity on the web. In *WWW*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *NIPS*.

Yu Rong, Hong Cheng, and Zhiyu Mo. 2015. Why it happened: Identifying and modeling the reasons of the happening of social events. In *SIGKDD*.

Siliang Tang, Fei Wu, Si Li, Weiming Lu, Zhongfei Zhang, and Yueting Zhuang. 2015. Sketch the storyline with charcoal: A non-parametric approach. In *IJCAI*.

Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei. 2006. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581.

Nguyen Xuan Vinh, Julien Epps, and James Bailey. 2010. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *Journal of Machine Learning Research*, 11(Oct):2837–2854.

Hongning Wang, Chi Wang, ChengXiang Zhai, and Jiawei Han. 2011a. Learning online discussion structures by conditional random fields. In *SIGIR*.

Li Wang, Marco Lui, Su Nam Kim, Joakim Nivre, and Timothy Baldwin. 2011b. Predicting thread discourse structure over technical web forums. In *EMNLP*.

William Yang Wang, Yashar Mehdad, Dragomir R Radev, and Amanda Stent. 2016. A low-rank approximation approach to learning joint embeddings of news stories and images for timeline summarization. In *NAACL*.

Hongteng Xu and Hongyuan Zha. 2017. A dirichlet mixture model of hawkes processes for event sequence clustering. In *NIPS*.

Shize Xu, Shanshan Wang, and Yan Zhang. 2013. Summarizing complex events: a cross-modal solution of storylines extraction and reconstruction. In *EMNLP*.

Deyu Zhou, Haiyang Xu, Xin-Yu Dai, and Yulan He. 2016. Unsupervised storyline extraction from news articles. In *IJCAI*.