# dhCM: Dynamic and Hierarchical Event Categorization and Discovery for Social Media Stream

JINJIN GUO and ZHIGUO GONG, State Key Lab of IoTSC and University of Macau, China
LONGBING CAO, University of Technology Sydney, Australia

The online event discovery in social media based documents is useful, such as for disaster recognition and intervention. However, the diverse events incrementally identified from social media streams remain accumulated, ad hoc, and unstructured. They cannot assist users in digesting the tremendous amount of information and finding their interested events. Further, most of the existing work is challenged by jointly identifying incremental events and dynamically organizing them in an adaptive hierarchy. To address these problems, this article proposes *d*ynamic and *h*ierarchical *C*ategorization *M*odeling (dhCM) for social media stream. Instead of manually dividing the timeframe, a multimodal event miner exploits a density estimation technique to continuously capture the temporal influence between documents and incrementally identify online events in textual, temporal, and spatial spaces. At the same time, an adaptive categorization hierarchy is formed to automatically organize the documents into proper categories at multiple levels of granularities. In a nonparametric manner, dhCM accommodates the increasing complexity of data streams with automatically growing the categorization hierarchy over adaptive growth. A sequential Monte Carlo algorithm is used for the online inference of the dhCM parameters. Extensive experiments show that dhCM outperforms the state-of-the-art models in terms of term coherence, category abstraction and specialization, hierarchical affinity, and event categorization and discovery accuracy.

CCS Concepts: • **Computing methodologies** → **Bayesian network models**; *Knowledge representation and reasoning*; • **Information systems** → **Document topic models;**

Additional Key Words and Phrases: Hierarchical categorization, document stream, online inference, Bayesian nonparametrics, kernel estimation, event categorization, event discovery

**ACM Reference format:**
Jinjin Guo, Zhiguo Gong, and Longbing Cao. 2021. dhCM: Dynamic and Hierarchical Event Categorization and Discovery for Social Media Stream. *ACM Trans. Intell. Syst. Technol.* 12, 5, Article 57 (September 2021), 25 pages.
https://doi.org/10.1145/3470888

This work was supported by National Key D&R Program of China (2019YFB1600704), FDCT (FDCT/0068/2020/AGJ, FDCT/0045/2019/A1, FDCT/0007/2018/A1, FDCT/0018/2019/AKP), GSTIC (201907010013, EF005/FST-GZG/ 2019/GSTIC), University of Macau (MYRG2018-00129-FST), and GDST (2019B111106001).

Authors' addresses: J. Guo and Z. Gong (corresponding author), University of Macau, Taipa, Macau S.A.R, China; emails: {yb57414, fstzgg}@um.edu.mo; L. Cao, University of Technology Sydney, Ultimo, New South Wales 2007, Sydney, Australia; email: LongBing.Cao@uts.edu.au.

2157-6904/2021/09-ART57 $15.00
https://doi.org/10.1145/3470888

## 1 INTRODUCTION

An event (e.g., a sports game or an art show) is an unusual activity engaging a large crowd of participants in certain areas within a short time duration. Online event discovery from social media data has been recognized as a significant task in real life. For example, by detecting an emergent disaster online, an alarm can be quickly sent to the emergency management at the very first moment for timely actions. Today, social media services play a major role in creating first-hand reports on influential events, and the geo-temporal-tagged data acquired from social media platforms (e.g., Flickr and Twitter) can well match the 3W (what, when, where) perspectives of events in the real world. Hence, extensive studies have been conducted to discover online events from social media streams [7, 16, 53, 55, 56].

However, the flat set of miscellaneous events remains randomly accumulated, ad hoc, and unstructured when exposed to social media stream, which cannot be quickly digested by end users. Hence, a dynamic structure to incrementally organize events to an unbounded hierarchy is highly desired by users, which offers a real-time bird's view of the data stream and quickly navigates users to their interests [36]. Such a dynamic organization process highly resembles the scenario of human cognition [33], where a human being continuously learns knowledge from the surroundings across time. The learned knowledge is not randomly accumulated in the human brain but rather is automatically organized and stored in an adaptive structure in one's private knowledge base, which supports informed judgment making.

Motivated by the preceding discussion, we bridge the gap between online event discovery and hierarchical organization, and propose a dynamic and hierarchical categorization model (dhCM) for social media stream. In the solution, both categories at different levels of abstraction and event assignments are jointly identified to organize the document stream automatically.

### 1.1 Limitations of Existing Techniques

In previous work, online event detection and hierarchical event organization have been studied separately. In the case of online event detection, to deal with continuous document streams, a common practice [2, 24] is to manually discretize the timeframe into a sequence of slices and extract topics from each slice sequentially. The main deficiency is that it could not offer real-time feedback along the document stream, since the algorithms start the detection task only when a slice is fully loaded with documents. Thus, some work [10, 16] explores the Hawkes process and density estimation to capture the temporal influence between continuous documents. However, the flat set of events detected by these online approaches still remains ad hoc and unstructured without exploring their hierarchical relationships. Therefore, a *hierarchical* organization of online events is highly desired to deal with online document stream.

In the case of topic hierarchy modeling, some works [4, 14, 29, 58] represent the typical unsupervised methods, which identify the topics from a document collection and project them to a hierarchical structure in a uniformed fashion. The recent advances by some authors [20, 26, 27, 43] represent the supervised methods, where a user's prior or external knowledge is required to guide the construction of term taxonomy. Despite their success in hierarchy construction, most of them are significantly challenged by the online setting, failing to accommodate the incremental events from document stream. Therefore, not only a hierarchical structure to effectively organize the online events but also a *dynamic* and *adaptive* structure to accommodate the increasing document stream are in demand.

### 1.2 Contributions

To satisfy the requirements, we propose a dynamic and hierarchical categorization modeling (dhCM) in social media data stream, which consists of two integrated components to jointly identify online events from stream and dynamically categorize them into the hierarchy with adaptive growth.

To deal with the streaming data, we design a component of multimodal event miner to continuously identify diverse events in a joint space. Instead of manually dividing the timeframe, we apply the density estimation technique to flexibly capture the temporal influence. In this framework, we assume the time (temporal tag) of generating each document as a sample point, and the temporal influence of events is smoothly estimated by aggregating these samples via Gaussian kernels [17]. To relieve the sparsity of short texts from social media, we assign one event to each document and leverage both word tokens and word embeddings to capture the event semantics. In the spatial dimension, we associate each event with multiple regions to detect its geographical influence by the Gaussian mixture. By modeling the textual, temporal, and spatial features in a joint space, the online events are incrementally identified by multi-dimensional features.

In addition, the component of hierarchical categorization modeling with adaptive growth is designed to automatically organize document streams into categories of different granularities. Inspired by the merit of *multi-level* clustering and nonparametrics from a **nested Dirichlet Process (nDP)** [35], we develop a hierarchical structure with multiple levels of abstraction, where the root node represents the most coarse category, nodes at lower levels turn to be more fine-grained categories, and the bottom leaf nodes represent the identified events. When documents arrive in a streaming manner, the category trajectories from coarse to specific are sampled from the hierarchy and automatically assign the documents to the proper categories.

By unifying two components together, we formulate a dynamic and hierarchical categorization model for social media data stream. In the solution, when documents continuously arrive, dhCM identifies a sequence of categories with different abstraction and event assignments to organize them automatically, resulting in a general-to-specific category hierarchy with adaptive growth. To address the biases caused by the streaming-arrivals of social media data, the **Sequential Monte Carlo (SMC)** [8, 9] algorithm is exploited to conduct an *online* inference to the dynamic categorical structure with multiple levels of granularities. In SMC, a set of weighted particles are exploited to provide real-time estimates to the document stream so as to well match the requirement of the online setting.

To summarize, we make the following contributions:

- A dynamic and hierarchical categorization modeling for social media data stream is proposed, where both the categories at different levels of abstraction and the event assignments are identified.
- Without a time-division setting, a density estimation technique is developed to deal with continuous document stream by smoothly capturing the temporal dependency. By jointly modeling contents, temporal dependency, and spatial features, events are incrementally identified in the joint space.
- The adaptive hierarchical structure at multiple levels of abstraction is designed, where the documents are automatically organized to the categories with different levels of abstraction. The category hierarchy is capable of accommodating new events and dynamically updating its structure across time.
- A comprehensive set of experiments is conducted on three real-world social media datasets. The results show that the proposed model outperforms the state-of-the-art ones in terms of term coherence, category abstraction/specialization, hierarchical affinity, event categorization, and category accuracy.

## 2 PRELIMINARY

In this section, we introduce two fundamental prior techniques for building a dynamic and hierarchical categorization model: density estimation and Bayesian nonparametrics techniques.

### 2.1 Density Estimation

Suppose a sequence of time points, $t_1, t_2, \ldots, t_N$ within the interval $[t_a, t_b]$, is randomly generated under a density distribution $\lambda(t)$, then $\lambda(t)$ can be estimated by using samples as $\hat{\lambda}_N(t)$ [12]

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

where $\kappa(t_i, t, \sigma) = e^{-\frac{(t-t_i)^2}{2\sigma^2}}$ is a Gaussian kernel fitted over the time point $t_i$ with bandwidth $\sigma$, which indicates the influence range of the kernel. The formulation shows the density estimation via the kernels is nonparametric since no parametric distributions are imposed on the estimation [41]. As noticed, the only variable, which should be defined, is the bandwidth $\sigma$ of the kernel. Theoretically, the setting of $\sigma$ depends on the number of data points $N$ within the interval. It should satisfy the condition $\lim_{N\to\infty} \sigma_N = 0$ to ensure the unbiased estimation. At the same time, the other condition $N \cdot \sigma_N \to \infty$ is also required to ensure $\lim_{N\to\infty} \hat{\lambda}_N(t) = \lambda(t)$ consistently. Therefore, the value of $\sigma_N$ should be set smaller if the number of samples $(N)$ is large.

### 2.2 Bayesian Nonparametrics

In this section, we give a brief introduction to two members of the Bayesian nonparametric family. One is the **Dirichlet Process (DP),** and the other is the nDP, which serve as the prior for a rich family of models and allow the model complexity (e.g., number of latent clusters, number of latent variables) to grow as more data are observed [18].

*2.2.1 Dirichlet Process.* The DP [39] is one member of the nonparametric stochastic processes, which is used to cluster the data samples without specifying the parameters. Let $H$ be a base distribution on the measurable space, and $\alpha$ is a positive concentration parameter. We write $G \sim DP(\alpha, H)$ to denote that $G$ is a draw from DP parameterized by $\alpha$ and $H$.

The *stick-breaking* representation provides a constructive definition for samples drawn from a DP. A draw $G$ from $DP(\alpha, H)$ can be written as [39]

$$G = \sum_{k}^{\infty} \pi_k \delta_{\phi_k}, \qquad \text{with} \quad \{\phi_k\}_{k=1}^{\infty} \overset{iid}{\sim} H, \quad \{\pi_k\} \sim GEM(\alpha),$$

where the unique atoms $\phi_k$ are drawn independently from $H$ and the corresponding weights $\{\pi_k\}$ follow the stick-breaking construction. The $GEM$ distribution is defined as $\pi_i = \omega_i \prod_{j=1}^{i-1}(1 - \omega_j)$ and $\omega_i \sim Beta(1, \alpha)$.

*2.2.2 Nested Dirichlet Process.* The nDP [35] introduces an automatic *multi-level* clustering over groups and observations, where observations are divided into different groups, and they are further grouped into clusters within each group.

We first define a set of distributions $\{G_1^0, \ldots, G_k^0\}$ and each represents a group, which is expressed in the *stick-breaking* representation [35],

$$G_k^0 = \sum_{l=1}^{\infty} \pi_{kl} \delta_{\phi_{kl}}, \qquad \text{with} \quad \phi_{kl} \overset{iid}{\sim} H, \quad \{\pi_{kl}\}_{l=1}^{\infty} \sim GEM(\alpha_0),$$

and then draw group=specific distributions from the preceding mixture $\{G_k^0\}$,

$$G_j^1 \sim G_B^1 \equiv \sum_{k=1}^{\infty} \beta_k \delta_{G_k^0}, \qquad \{\beta_k\} \sim GEM(\alpha_1).$$

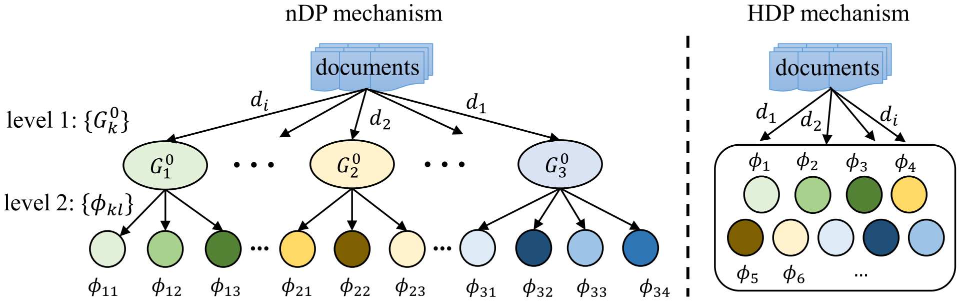


Fig. 1. The comparison between the nDP and HDP in the task of documents clustering.

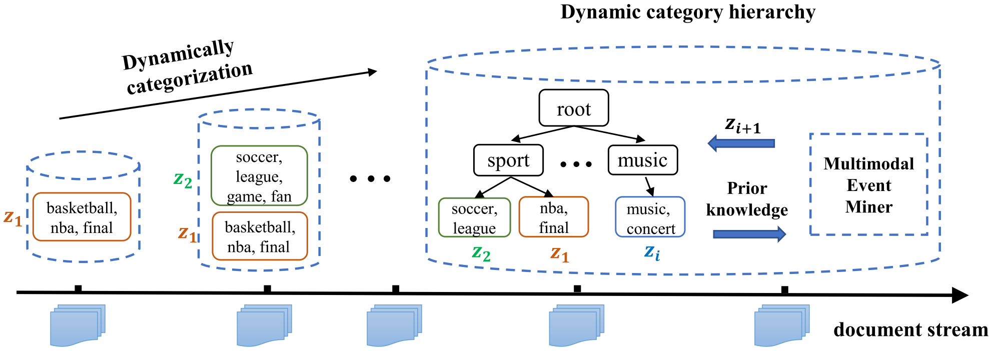


Fig. 2. An illustration of the proposed dynamic and hierarchical categorization modeling framework over a stream of document collections.

We denote the generation process as $\{G_j^1\} \sim nDP(\alpha_1, \alpha_0, H)$. It ensures no-zero probability of $\{G_j^1\}$ when the same selection of $G_k^0$ is clustered into the group $G_k^0$, and observations are grouped into $l$ clusters within each group $G_k^0$ according to the selections $\{\phi_{kl}\}$.

To obtain an intuitive understanding of *multi-level* clustering of nDP, we use the comparison between nDP and **Hierarchical Dirichlet Process (HDP)** in the task of clustering the collection of documents. Indicated by Figure 1, under the mechanism of nDP, documents are grouped into different latent categories $\{G_k^0\}$ (document-level clustering) according to their contents, and words from documents further form different topics $\{\phi_{kl}\}$ (word-level clustering). In contrast, such *multi-level* clustering is not available from HDP, where only word-level clusterings $\{\phi\}$ are identified that are globally shared by all documents.

## 3 THE DHCM MODEL

In this section, we first state our problem and give the definitions for the key concepts. Then we present the whole framework of the dynamic and hierarchical categorization modeling.

### 3.1 Problem Formulation

Figure 2 presents an overview of the proposed dynamic and hierarchical categorization modeling, which consists of two integrated components: (1) multimodal event miner and (2) hierarchical categorization modeling. Specifically, as the documents arrive continuously, the component of hierarchical categorization modeling offers the proper trajectory of category nodes with different granularities, and the component of multimodal event miner determines the leaf node assignments for the documents based on the joint features of semantics, time, and location. The two components work together to automatically organize the document stream into a category hierarchy with adaptive growth.

Let $D = \{d_1, d_2, \ldots, d_n\}$ be a continuous stream of social media posts (e.g., Flickr image, tweet) that arrive in a chronological order. Each document is represented as a triplet $d_i = (\mathbf{w}, t_i, \mathbf{l})$, where $\mathbf{w}$ is a set of words denoted by users, $t_i$ is the posting time, and $\mathbf{l}$ captures the geo-coordinates.

We give the definitions for the key concepts in the following.

*Definition 3.1 (Category Hierarchy $\mathcal{H}$ ).* This is a tree-structured category hierarchy, where only the bottom leaf nodes represent the identified events from the document stream, and the rest of the nodes at different levels denote the categories with different granularities. Moreover, the parent-child nodes in $\mathcal{H}$ follow the category-subcategory relation such that nodes near the root are general and abstract, whereas nodes near the bottom are locally specific.

*Definition 3.2 (Event $z$).* This is a group of semantically relevant documents with close temporal and spatial coherence, which is defined in the joint space. Formally, an event $z$ is semantically described by a word distribution $\phi_z = \{P(w|z)\}_{w \in V}$ , temporally specified by the density of kernels $\lambda(t) \approx \sum_i \kappa(t, t_i)$, and spatially depicted by a mixture of regions $\sum_{r=1}^{N} P(r|z) = 1$. It is located as the leaf node in the category hierarchy.

*Definition 3.3 (Category $c$).* This is a group of semantically similar documents and locates at different levels in the hierarchy except the leaf nodes. When traversing down the hierarchy, the documents within the parent category node are further grouped into child categories, leading to the general-to-specific category hierarchy from top to bottom. Naturally, a category is represented by the documents within this category, which is defined as multinomial distribution over words $\phi_c = \{P(w|c)\}_{w \in V}$.

### 3.2 Multimodal Event Miner

The component of multimodal event miner is designed to identify the online events on the joint space of semantics, time, and geography.

*Content modeling.* The short textual content generated by social media contains limited word occurrences. For example, a Flickr message consists of no more than 10 textual tags, and similarly, a tweet message contains less than 140 characters. The traditional practice of associating all topics with a document results in the sparsity problem [24, 51]. Following the previous study [15, 16, 50], we assume that the textual content of a Flickr or microblog message concentrates on one topic. This assumption has been shown effective to improve the clustering performance by mitigating the sparsity problem [23].

Word embeddings [28] have demonstrated the effectiveness to capture semantic regularities in language. In addition to the raw text input, the proposed model is flexible to accommodate the word embeddings and explore the category hierarchy in the embedding space. Therefore, if the input is in the form of word tokens, a multinomial distribution is opted to generate the word occurrence; otherwise, a Gaussian distribution [6, 49] is activated. Under the one event assumption, the words in document $d$ are generated as follows:

$$P(\mathbf{d}|z, \cdot) = \prod_{i=1}^{n_d} P(w_i|z) = \begin{cases} \prod_{i=1}^{n_d} Mult(w_i|\phi_z) & w_i \text{ is a word token} \\ \prod_{i=1}^{n_d} \mathcal{N}(\mathbf{w_i}|\mu_\mathbf{z}, \Sigma_\mathbf{z}), & \mathbf{w_i} \text{ is the word embedding} \end{cases} \quad (1)$$

where $n_d$ records the number of words, $Mult()$ denotes the multinomial distribution, and $\mathcal{N}()$ is the Gaussian distribution with mean vector $\mu$ and covariance matrix $\Sigma$ .

*Online temporal density.* With the social media data streaming in, not only the textual contents but also the *density* of their timestamps serve as a good clue to inferring online events. For example, when an event appears in the spotlight, more posts may be triggered to report it. Such a *self-excitation* phenomenon [10] often leads to many closely related textual contents generated cohesively within a short time duration. Therefore, based on the nature of events in the time dimension, we take advantage of their mutual influence to infer the temporal density of online events.

In light of the temporal cohesiveness of social media data for the same event, we treat each arriving document as a sample. Under the mechanism of temporal density, we suppose the first document $d_1$ arrives at time $t_1$, the second document $d_2$ arrives at $t_2$, and repeatedly the current document $d_j$ arrives at $t_j$. When all these received documents are taken as samples in $[t_1, t_j]$, then the density at $t_j$ can be roughly estimated as [12]

$$\hat{\lambda}_N(t_j) \propto \sum_{i=1}^{j} \kappa(t_i, t_j, \sigma_j), \tag{2}$$

where $\sigma_j$ indicates the bandwidth, which is related to the average density of sampling points in the temporal space $[t_1, t_j]$.

However, this estimation dynamically brings the influences of all prior documents to $d_j$ in a natural manner, which results in expensive computation. Hence, a sliding window $[t_j - L, t_j]$ is naturally introduced to only account those 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 $L$). Within the sliding window $[t_j - L, t_j]$, the optimal value of $\sigma_j$ is dynamically tuned according to the technique [16].

Let $\mathbf{W} = \{d_{j-L_j}, d_{j-L_j+1}, \ldots, d_j\}$ denote the sequence of documents in the sliding window $[t_j - L, t_j]$, and $L_j$ is the number of documents in it. The density of event $z$ at time $t_j$ is estimated by

$$P(t_j|z, \mathbf{W}, \alpha) \propto \begin{cases} \dfrac{\lambda_{z_k}(t_j)}{\sum_{i=j-L_j}^{j} \lambda_{z_i}(t_j) + \alpha} & z_k \text{ is one of the existing} \\ \dfrac{\alpha}{\sum_{i=j-L_j}^{j} \lambda_{z_i}(t_j) + \alpha}, & z \text{ is new} \end{cases} \tag{3}$$

where $\lambda_{z_k}(t_j) := \sum_{i=j-L_j}^{j} \kappa(t_i, t_j, \sigma_j)\delta_{z_k}$ denotes the aggregated influence on the current document $d_j$ from those documents with event indicator $z(d_i) = z_k$ in the sliding widow.

*Spatial patterns.* In practice, a real event may occur in one spatial region or multiple regions [15, 16]. To capture the footprints of diverse events from social media data, we assume each event is geographically distributed over a set of regions, and each region is modeled by a bi-variant Gaussian parameterized as $(\mu, \Sigma)$. Thus, the spatial distribution of an event is represented by a mixture of Gaussians. Such spatial modeling allows complex and diverse spatial patterns of an event. Formally, given the event indicator $z$, the probability of generating geo-coordinates $\mathbf{l}$ for the document $d$ is computed as

$$P(\mathbf{l}|z) = \sum_{r=1}^{N} P(r|\pi_z) \cdot \mathcal{N}(\mathbf{l}|r), \tag{4}$$

where $r$ denotes the geospatial region drawn from the multinomial with $\pi_z$, $\mathcal{N}(\mathbf{l}|r)$ indicates the probability density of $\mathbf{l}$ drawn from a region represented by a bi-variant Gaussian distribution, and $N$ is the number of regions.

*Multimodal event modeling.* In summary, given the document $d_j$ and its associated sliding window $\mathbf{W}$, the event assignment of document $d_j$ is identified based on the preceding multi-dimensional features in the joint space, which is denoted as

$$P(d_j|z, \mathbf{W}, \cdot) \propto \prod_{i=1}^{n_d} P(w_i|z) \cdot P(t_j|z, \mathbf{W}, \cdot) \cdot P(\mathbf{l}|z). \tag{5}$$

Through the component of multimodal event detection, each document arrival is annotated by the corresponding event indicator.

### 3.3 Hierarchical Categorization Modeling

The nDP mechanism automatically induces a two-level clustering (i.e., category and topics) over the social media stream. However, such a confined two-level clustering might be inadequate to summarize the continuous data stream. Therefore, to solve this limitation, we design a deep categorization model with adaptive growth to flexibly accommodate the increasing complexity in the online setting.

First, a category hierarchy with maximum depth $L$ is defined by deepening the nDP to the $L$-level clustering, which is represented as

$$\{G_k\} \sim hnDP(\underbrace{\alpha_{L-1}, \alpha_{L-2}, \ldots, DP(\alpha_0, H_0)}_{\text{depth of L}}), \tag{6}$$

where **hnDP**() is the defined hierarchical nDP with deep multi-level clustering, and $\{\alpha_i\}_{i=0}^{L-1}$ is the specified hyperparameters. From top to bottom, the clustering at each level is expressed in the *stick-breaking* representation,

$$\begin{aligned}
G_k^0 &\sim \sum_k^{\infty} \beta_k^{L-1} \delta_{G_k^1}, \quad \text{with} \quad \{\beta_k^{L-1}\} \sim GEM(\alpha_{L-1}), \quad \text{at 0 level} \\
G_k^1 &\sim \sum_k^{\infty} \beta_k^{L-2} \delta_{G_k^2}, \quad \text{with} \quad \{\beta_k^{L-2}\} \sim GEM(\alpha_{L-2}), \quad \text{at 1 level} \\
&\cdots \\
G_k^{L-1} &\sim \sum_l^{\infty} \beta_l^0 \delta_{\phi_l}, \quad \text{with} \quad \{\beta_l^0\} \sim GEM(\alpha_0), \quad \{\phi\} \sim H_0, \quad \text{at } L-1 \text{ level},
\end{aligned} \tag{7}$$

where the category node $G_k^i$ at the $i$-th level ($0 \leq i < L-1$) is the summarization over the $i+1$ level $\{G_k^{i+1}\}$, and the leaf nodes $\{\phi\}$ at the bottom level represent the identified fine-grained events. Hence, it is naturally deemed that the nodes at higher level are general and abstract while nodes near the bottom level turn to be specific.

### 3.4 The Integration of Two Components

Imagine a category hierarchy with the depth of $L$ at time $t$ as presented in Figure 3(a). When a new document $d_j$ arrives at $t_j$ ($t_j > t$), a set of categories with different levels of abstraction as well as the event indicator are sampled sequentially. The updating process starts with the root category in the hierarchy and ends up with leaf nodes. We summarize its potential sampling trajectories in three ways:

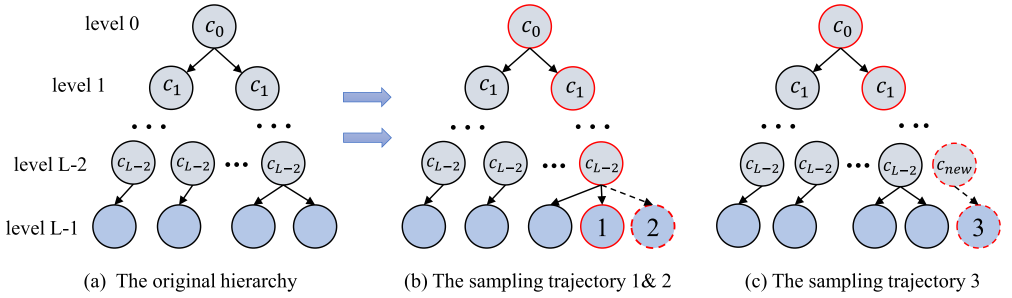

level 0
level 1
level L-2
level L-1
$c_0$
$c_1$
$c_{L-2}$
$c_{new}$
1
2
3
(a) The original hierarchy
(b) The sampling trajectory 1& 2
(c) The sampling trajectory 3

Fig. 3. The dynamic updating process for accommodating new arriving documents, in which the bottom nodes are distinguished from other nodes in the blue color. (a) The original category hierarchy. (b, c) Three possible sampling trajectories denoted with red color.

(1) Starting from the root node $c_0$, one of its existing children $c_1$ at level 1 is first sampled, and sampled repeatedly in this way until one of the existing leaf nodes with the most similarity is chosen, which is denoted as trajectory 1 in Figure 3(b). Only in this trajectory does the structure of the category hierarchy remains unchanged.
(2) Continuing with trajectory 1 at level $c_{L-2}$, one new event is triggered to be created and cover the content from document $d_j$ if the existing leaf nodes are not semantically similar to $d_j$. In trajectory 2, the category hierarchy is reused and a new event is created at the bottom level.
(3) As the existing category nodes do not match the content of document $d_j$, trajectory 3 is triggered to create a sequence of new category from level $M$ ($0 < M \leq L-2$) to level $L-2$, and a new event at the bottom level is accordingly generated under the new category, which is indicated in Figure 3(c). In this trajectory, the category hierarchy updates with new categories created at different levels to explain the content of document $d_j$.

Therefore, given the document $d_j$, the probability of obtaining these three trajectories $\tau$ could be denoted in the following:

$$P(\tau|d_j,\mathcal{H}) \propto P(c|\mathcal{H}) \cdot P(d_j|c) \propto \begin{cases} P(c_{L-2}|\mathcal{H})P(z|c_{L-2})P(d_j|z) & \text{trajectory 1} \\ P(c_{L-2}|\mathcal{H})P(z_{new}|c_{L-2})P(d_j|z_{new}) & \text{trajectory 2} \\ \prod_{i}^{L-2} P(c_{new}^{i}|\mathcal{H})P(z_{new}|c_{new}^{L-2})P(d_j|z_{new}), & \text{trajectory 3} \end{cases} \quad (8)$$

where $c_{L-2}$ denotes one of the existing category nodes at level $L-2$, and $c_{new}^{i}$, $z_{new}$ indicate the new category node at level $i$ and new leaf node respectively.

The updating process repeatedly handles the later arriving documents after time $t_j$, and the hierarchical category structure with unbound mechanism correspondingly updates its structure to accommodate new categories or new events. Finally, a hierarchical categorization with multiple levels of abstraction is formulated to summarize the document stream, where the leaves of the hierarchy represent the identified events, and the rest of the category nodes denote the corresponding categories with different granularities.

*Generative process for dhCM.* The graphical representation of this dynamic and hierarchical categorization modeling is illustrated in Figure 4. In this framework, with new documents streaming in, the proposed dhCM offers the proper category-to-event trajectories to the documents based on integration of hierarchical sampling and their multi-dimensional features. The former determines category trajectories of different granularities accomplished by the component of categorization

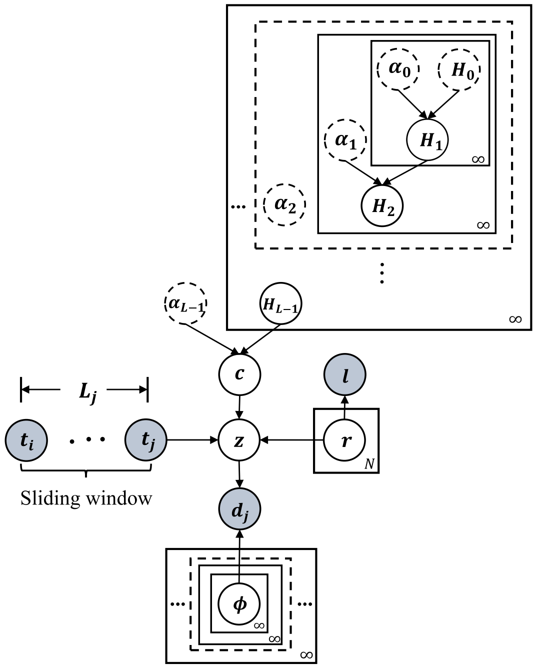


Fig. 4. Graphical representation of a dynamic and hierarchical categorization model, where the circles with dotted lines denote the specified hyperparameters, circles with shaded color denote the observed variables (timestamps within the sliding window, words, and geo-location), and the rest indicate the latent variables.

modeling, whereas the latter exploits the density estimation technique to deal with data stream and determines the event nodes via the multimodal event miner. Therefore, the two integrated components are unified to form a dynamic and hierarchical categorization modeling for the social media data stream. The generative process of dhCM is summarized in the following:

For each document $d_j$ arriving at time $t_j$ with geo-coordinates $\mathbf{l}$,

(1) Sample the trajectories $\tau$ from the category hierarchy,
(2) If the category node $c_{L-2}$ at level $L-2$ sampled exists,
  i. Sample one of existing events $z$,
  ii. Otherwise create a new event $z_{new}$ under $c_{L-2}$,
(3) If the category node $c^{L-2}_{new}$ sampled is new,
  i. Create a new event $z_{new}$,
(4) Given the event $z$,
  i. Draw a region indicator $r \sim P(r|z)$,
   Draw the geo-coordinates $\mathbf{l} \sim \mathcal{N}(\mathbf{l}|r)$,
  ii. Draw the time $t_j \sim P(t_j|z,\cdot)$,
  iii. For each word $w_i$ in the document,
   $w_i \sim P(w_i|z)$.

## 4 ONLINE INFERENCE

Our goal is to conduct an *online* inference for the posterior distribution $P(\tau_{1:n}, z_{1:n}|d_{1:n})$, where $d_{1:n}$ indicates the document sequence received from time $t = t_1$ to $t = t_n$, and $\tau_{1:n}, z_{1:n}$ represent their corresponding category trajectories and event indicators from time $t_1$ to $t_n$, respectively. To adapt to continuous social media stream, an *online* inference approach is highly desired to estimate the latent variables.

Let $P(\tau_{1:n}, z_{1:n-1}|d_{1:n-1})$ denote the posterior distribution of document samples from time $t = t_1$ to $t = t_{n-1}$. When the new document $d_n = (\mathbf{w}, t_n, \mathbf{l})$ arrives at time $t_n$, this posterior would update it to yield the most recent value $P(\tau_{1:n}, z_{1:n}|d_{1:n})$ by reusing the past samples from $P(\tau_{1:n}, z_{1:n-1}|d_{1:n-1})$. Hence, we are motivated to leverage the SMC method [8, 9] to infer the sampling process.

In SMC, there is a set of particles (samples) to maintain the posterior distribution, and each particle is associated with a weight to indicate how well it explains the data. If the weight of the particle is high, it means the particle has better explanatory power of the data. The weight $\omega_n^f$ of each particle $f \in \{1, \ldots, F\}$ is defined as the ratio between the true posterior and a proposal distribution $Q(\tau_{1:n}, z_{1:n}|d_{1:n})$ at time $t_n$. Thus, the particle weight at time $t_n$ is denoted as $\omega_n^f = P(\tau_{1:n}, z_{1:n}|d_{1:n})/Q(\tau_{1:n}, z_{1:n}|d_{1:n})$. The true posterior is recursively expressed as

$$\begin{aligned} P(\tau_{1:n}, z_{1:n}|d_{1:n}) &= \frac{P(d_n|\tau_{1:n}, z_{1:n}, d_{1:n-1})P(\tau_{1:n}, z_{1:n}|d_{1:n-1})}{P(d_n|d_{1:n-1})}, \\ &\propto P(d_n|\tau_n, z_n)P(\tau_n, z_n|\tau_{n-1}, z_{n-1})P(\tau_{1:n}, z_{1:n-1}|d_{1:n-1}). \end{aligned} \tag{9}$$

The proposal distribution is defined in the following form:

$$Q(\tau_{1:n}, z_{1:n}|d_{1:n}) = Q(\tau_n, z_n|\tau_{1:n}, z_{1:n-1}d_{1:n})Q(\tau_{1:n}, z_{1:n-1}|d_{1:n-1}). \tag{10}$$

To minimize the variance of resulting particle weight, we take $Q(c_n, z_n|\tau_{1:n}, z_{1:n-1}, d_{1:n})$ to be the posterior distribution $P(\tau_n, z_n, r_n|\tau_{n-1}, z_{n-1})$ [10], and the resulting particle weight is updated as

$$\omega_n^f \propto \omega_{n-1}^f \cdot P\left(d_n^f|z_n^f, \cdot\right), \tag{11}$$

where $P(d_n^f|z_n^f, \cdot)$ determines the probability of generating document $d_j$ given the event $z$, which is referred to Equation (5), and $\tau_n^f$, $z_n^f$ denote the category trajectory and event indicator, respectively, in the particle $f$ at time $t_n$.

Within each particle, the trajectory from the category hierarchy and the event node at the bottom level are jointly sampled to avoid redundant computational cost [15]. Then the region sampling within an event for the arriving document is followed. The overall structure of SMC is described in Algorithm 1.

The sampling trajectory and event node are denoted in the following [59]:

$$P(\tau, z|\mathcal{H}, d_n, \cdot) = \begin{cases} \dfrac{\prod_{i=1}^{L-2} m_{c_i} P(d_n|c_i) \cdot m_z}{\prod_{i=0}^{L-2}(n_i + \alpha_i)} P(d_n|z, \cdot) & \text{both } c_i, z \text{ are existing} \\ \dfrac{\prod_{i=1}^{L-2} m_{c_i} P(d_n|c_i) \cdot \alpha_{L-2}}{\prod_{i=0}^{L-2}(n_i + \alpha_i)} P(d_n|z_{new}, \cdot) & c_i \text{ is existing} \\ \dfrac{\prod_{i=1}^{j} m_{c_i} P(d_n|c_i) \prod_{j}^{L-1} \alpha_j}{\prod_{i=0}^{L-2}(n_i + \alpha_i)} P(d_n|z_{new}, \cdot), & c_j (j \leq L-1) \text{ is new} \end{cases} \tag{12}$$

where $P(d_n|c_i)$ indicates the Dirichlet probability of producing $d_n$ from category $c_i$ in the case of word token input or multivariant t-student probability for word embedding input, $P(d_n|z, \cdot)$ is the probability of generating $d_n$ from event $z$ referred to Equation (5), $c_i$ is the category node at level $i$ from the trajectory $\tau$, $m_{c_i}$ records the number of documents within the category node $c_i$, $m_z$ counts the documents assigned to the event node $z$, and $n_i$ records the total number of documents at level $i$.

**ALGORITHM 1:** SMC inference over social media document stream

```
Input: Document streams $d_1, d_2, \ldots, d_n$.
Output: A dynamic category hierarchy $\mathcal{H}$ with adaptive growth.
Initialize $\omega_1^f$ to $\frac{1}{F}$ for all $f \in \{1, \ldots, F\}$.
for each document $d_i$, $i \in \{1, 2, \ldots, n\}$ do
    for $f \in \{1, \ldots, F\}$ do
        Sample trajectory and event indicator $\tau_i^f, z_i^f$ by Equation (12).
        Sample region indicator $r_i^f$ by Equation (4).
        Update particle weight $\omega_i^f$ by Equation (11).
    end for
    Normalize particle weights.
    if $||\omega_i||_2^{-2} = 1/\sum_{f=1}^{F}(\omega_i^f)^2 < threshold$ then
        Resample the particles.
    end if
end for
```

*Complexity analysis.* In the online setting, an important question concerns the *efficiency* of dealing with continuous documents together with sampling from category hierarchy. For each document, a category trajectory from coarse to fine is identified, and the time complexity is $O(|N_{max}|)$, where $N_{max}$ records the total nodes of the category hierarchy. In the component of multimodal event miner, the complexity of learning the temporal influences between documents is $O(|L_{max}|)$, where $L_{max}$ denotes the maximum document number in the sliding window. The complexity of learning spatial pattern for an event is $O(N)$, where $N$ is the specified region number. To sum up, the overall time complexity to deal with one document is summarized as $O(|N_{max}|(|L_{max}| + N))$.

Since the region number $N$ is quite small compared with $|L_{max}|$, the time complexity of processing one document is estimated as $O(|N_{max}||L_{max}|)$. Thus, we conclude the time complexity of each document in dhCM is mainly affected by two aspects. One is the historical document number in the sliding window, and the other is the size of the category hierarchy. When the prior documents uniformly arrive, the number of past documents $|L_{max}|$ could be treated as a constant. Therefore, the time complexity of processing each document is approximately linear to the size of the hierarchical structure. With more events detected from the stream, the hierarchy grows slowly because of the convergence and stops growing after the long-term accumulation. Thus, the entire node number $O(|N_{max}|)$ from the hierarchy remains almost unchanged. Such a time complexity is accepted for a large volume of data stream in the online setting.

## 5 EXPERIMENTS

In this section, the real-world datasets and comparison models are first introduced. We then conduct empirical evaluations to demonstrate the effectiveness of the proposed model from different perspectives, and the case studies are followed to show the dynamic and hierarchical categorization modeling in both text and embedding space.

### 5.1 Experimental Setup

*Dataset.* Our experiments are based on three real-world social media datasets, which are crawled from two different platforms: Flickr and Twitter. Each Flickr message consists of a set of textual tags (words), location coordinates, and the taken time. In our preparation, some noisy words (e.g., camera names, extremely high frequency words) are removed from Flickr datasets. The Twitter

Table 1. Statistics of Three Real Datasets

| Dataset | Time Span | #document | #vocabulary |
|---|---|---|---|
| Paris | 06/01/2010–08/22/2010 | 21,435 | 1,887 |
| NY | 01/01/2010–09/26/2010 | 155,943 | 18,208 |
| LA | 08/01/2014–11/30/2014 | 1.1M | 0.4M |

dataset is contributed from the work of Zhang et al. [55], and each tweet is associated with words, timestamps, and location coordinates.

In the city of Paris, we collect 21,435 documents after the preprocessing step, which is from June 1, 2010, to August 22, 2010. We refer to it as Paris. The other corpus, referred as NY for short, consists of 155,943 documents from New York City, which ranges from January 1, 2010, to the end of September in 2010. The Twitter dataset, referred to as LA, consists of around 1.1 million geotagged messages published in Los Angeles during the period from August 1, 2014, to November 30, 2014. After the preprocessing, each post from three datasets is annotated by words, timestamps, and location coordinates. The details of the datasets are presented in Table 1.

*Baselines.* We classify the state-of-the-art baselines into two types. One is based on document stream, which includes *online* competitors with adaptive structures to detect the incremental events or category hierarchy (e.g., the proposed dhCM). The other is the *static* hierarchical competitors with given structures, which are ascendant over the online algorithms by taking the whole dataset as the input instead of the document stream. The static competitors are included as important comparisons to measure the effectiveness of the proposed model in the context of document stream.

The following are *online* baselines:

- *TrioVec:* This is a flat approach for online local event detection based on multimodal embeddings [53], where the geo-topic clusters are first detected based on a Bayesian mixture model and then high-quality events are extracted from them.
- *MS:* This refers to the model-based online clustering of short streams without the hierarchy [50], where a Dirichlet multinomial mixture model is leveraged to deal with continuous document streams and identify the increasing clusters.
- *hLDA-O:* This is the **hierarchical Latent Dirichlet Allocation (hLDA)** [4, 14], a static benchmark for constructing a flexible topic hierarchy with infinite depth and branch width. Since hLDA itself does not involve the temporal dependency and online inference over the streaming data, we extend it to the *online* version by integrating the density-based estimation and online inference techniques, which is exactly the same as the proposed model. We refer to the extended model as hLDA-O.
- *hPAM-O:* This is the **hierarchical Pachinko Allocation Model (hPAM)** [29], a static benchmark for building a topic hierarchy with the specified structure in both depth and branch width. Similar to hLDA-O, we extend it to an *online* approach, referred to as hPAM-O.
- *hDCT:* Built on a sequence of predefined time slices, the original **Dynamic Clustering Topic (DCT) model** [24] continuously detects online topics from the short text stream without the hierarchy. To explore the hierarchical relationship between topics, we manually merge the similar topics if their cosine similarity exceeds a predefined threshold after each time slice, the new merged topics are taken as category nodes at a high level, and category nodes are further merged at a higher level repeatedly. Such a dynamic topic hierarchy, formed from bottom to top based on DCT, is referred to as hDCT.
- *hrCRP:* The **recurrent Chinese Restaurant Process (rCRP)** [2] is a nonparametric and dynamic approach for flat topic discovery based on a sequence of time slices. Similar to

hDCT, we extend rCRP to the hierarchical structure by manually merging similar topics. Such a hierarchical and dynamic structure is referred to as hrCRP.

The following are *static* hierarchical baselines:

- ***DirBN:*** This refers to the **Dirichlet Belief Network (DirBN)** [58], which is a *static* multi-layer generative process on word distributions of topics, where each layer consists of a set of topics and they are fully connected between consecutive layers.
- ***CluHTM:*** This is the Hierarchical Topic Modeling based on the CluWords [42, 43], where the non-negative Matrix Factorization is iteratively exploited at each level to generate a topic hierarchy.
- ***LSTM:*** The **Long Short-Term Memory (LSTM)** network is a widely used recurrent neural network [19]. We iteratively apply it at each level to classify the documents from the coarse to fine-grained categories given the document labels [15].

Besides the preceding competitive baselines, the following are our proposed models:

- *dhCM:* This is the proposed dynamic and hierarchical Categorization Model for document streams. Its static version is referred to as dhCM-S, which is taken as the baseline to measure the effectiveness of the density estimation technique.
- *dhCM-E:* This is a variant of the proposed dhCM, which takes the word embedding representation as the input and outputs an incremental category hierarchy in the embedding space. Its static version is referred to as dhCM-E-S.

The nonparametric methods (e.g., hLDA-O, dhCM, and dhCM-E) are capable of growing adaptively to accommodate the increasing complexity, whereas the rest depend on the specified structure in both depth and branch width. For a fair comparison, we set the maximum depth for all baselines as $L_{max} = 4$, one single root node at level 0, and the number of nodes from parametric approaches at level 1 and level 2 and level 3 is set the same as the proposed dhCM. The hyperparameters from competitor models are suggested from the empirical study. The hierarchical approaches (e.g., hLDA-O, hPAM-O, dhCM, and dhCM-E) are approximated by the SMC algorithm, where the particle number is set as 100 in the Paris dataset, 30 in the NY dataset, and 10 in the LA dataset. The other two baselines, hDCT and hRCRP, are built on a sequence of time slices, which are estimated by Gibbs sampling over each time slice. According to our empirical study, the iteration of Gibbs sampling for hDCT, hrCRP, dhCM-S, and dhCM-E-S is set as 1,000 in the Paris dataset, 500 in the NY dataset, and 100 in the LA dataset, respectively. The parameter setting from DirBN, CluHTM, TrioVec, and MS refers to the original papers and open source code. To make a fair comparison, the approaches TrioVec, CluHTM, dhCM-E, and dhCM-E-S leverage the pretrained word embeddings [31].

## 5.2 Term Coherence

*Term coherence* aims at quantifying how semantically coherent the top terms within a category or an event node are. A significant metric for measuring term coherence is the PMI-score [30, 34], which highly matches with human judgment. Formally, the average *PMI* (Pointwise Mutual Information) between the top-$k$ terms within the topic is defined as

$$C = \frac{2}{k(k-1)\sum_{i=1}^{k-1}\sum_{j=i+1}^{k} PMI(w_i, w_j)}, \qquad PMI(w_i, w_j) = \frac{log\frac{P(w_i, w_j)+\epsilon}{P(w_i)P(w_j)}}{-log(P(w_i, w_j)+\epsilon)}, \tag{13}$$

where $P(w_i, w_j)$ denotes the probability of co-occurrence of $w_i$ and $w_j$ in one document and $P(w_i)$ is the probability of $w_i$ appearing in the document. A higher $PMI$ value indicates the terms

Table 2. Term Coherence Results on the Paris Dataset

| | Online & Flat | | Online & Hierarchical | | | | | | Static & Hierarchical | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | TrioVec | MS | hLDA-O | hPAM-O | hDCT | hrCRP | dhCM | dhCM-E | DirBN | CluHTM | LSTM |
| **Level 1** | — | — | 0.43 | **0.44** | 0.42 | 0.41 | 0.43 | **0.44** | **0.44** | 0.40 | 0.42 |
| **Level 2** | — | — | 0.48 | 0.47 | 0.47 | 0.47 | **0.50** | **0.50** | 0.47 | 0.44 | 0.48 |
| **Level 3** | 0.47 | 0.47 | 0.47 | 0.48 | 0.47 | 0.46 | 0.48 | **0.49** | 0.48 | 0.46 | 0.48 |
| **Avg.** | 0.47 | 0.47 | 0.46 | 0.463 | 0.453 | 0.447 | 0.47 | **0.477** | 0.463 | 0.433 | 0.46 |

The best performance is denoted in boldface.

Table 3. Term Coherence Results on the NY Dataset

| | Online & Flat | | Online & Hierarchical | | | | | | Static & Hierarchical | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | TrioVec | MS | hLDA-O | hPAM-O | hDCT | hrCRP | dhCM | dhCM-E | DirBN | CluHTM | LSTM |
| **Level 1** | — | — | 0.35 | 0.35 | 0.34 | 0.34 | 0.36 | 0.36 | 0.35 | **0.37** | **0.37** |
| **Level 2** | — | — | 0.39 | 0.41 | 0.41 | 0.41 | 0.42 | **0.45** | 0.43 | 0.44 | 0.39 |
| **Level 3** | 0.42 | 0.41 | 0.41 | 0.37 | 0.43 | 0.42 | 0.45 | **0.47** | 0.44 | 0.45 | 0.41 |
| **Avg.** | 0.42 | 0.41 | 0.383 | 0.377 | 0.393 | 0.39 | 0.41 | **0.427** | 0.407 | 0.42 | 0.39 |

The best performance is emphasized in boldface.

Table 4. Term Coherence Results on the LA Dataset

| | Online & Flat | | Online & Hierarchical | | | | | | Static & Hierarchical | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | TrioVec | MS | hLDA-O | hPAM-O | hDCT | hrCRP | dhCM | dhCM-E | DirBN | CluHTM | LSTM |
| **Level 1** | — | — | 0.34 | 0.33 | 0.33 | 0.33 | 0.35 | **0.36** | 0.35 | 0.32 | 0.34 |
| **Level 2** | — | — | 0.36 | 0.36 | 0.39 | 0.37 | **0.40** | **0.40** | 0.36 | **0.40** | 0.38 |
| **Level 3** | 0.38 | 0.38 | 0.36 | 0.35 | 0.38 | 0.36 | 0.39 | **0.40** | 0.38 | 0.39 | 0.38 |
| **Avg.** | 0.38 | 0.38 | 0.353 | 0.347 | 0.367 | 0.353 | 0.38 | **0.387** | 0.363 | 0.37 | 0.367 |

The best performance is denoted in boldface.

within the topics are more consistent and interpretable. To obtain an unbiased result, we resort to the large-scale external Wikipedia data [34] to measure the top-10 coherence values for all competitor models. Since there is only one root node at level 0, we omit it and focus on the topics at level 1, level 2, and level 3. We first calculate the coherence score of each topic and obtain the average coherence results for each level. The coherence results from all competitor models on the three datasets are presented in Tables 2, 3, and 4 respectively.

Indicated by Tables 2, 3, and 4, the coherence scores from hierarchical models turn to be larger from level 1 with terms drifting from general to specific. The two flat event detectors (TrioVec and MS) only contain the coherence results of the identified events, corresponding to events from the hierarchical structures at level 3. By examining the performance comparisons, we have the following remarks:

- Compared with the online and hierarchical competitors (hLDA-O, hPAM-O, hDCT, and hrCRP), it is noted that both dhCM and dhCM-E achieve the highest coherence score at each level of the hierarchy on the three datasets. Although hLDA-O and hPAM-O are equipped with the same density estimation as well as inference techniques, dhCM and dhCM-E beat them with an evident increase in the coherence results. Such comparison implies that the categorization modeling in dhCM and dhCM-E is effective to obtain a more coherent and interpretable category hierarchy for the document stream.
- Although the ascendant static baselines (DirBN, CluHTM, and LSTM) are averagely advantageous over other online hierarchical competitors including hLDA-O, hPAM-O, hDCT, and hrCRP on the NY and LA datasets, dhCM still achieves a comparable result with the best

static competitor, and particularly dhCM-E outperforms them across three levels on the three datasets. Such performance further supports that the density estimation in dhCM and dhCM-E is valid to deal with document stream without compromising their performance in the task of term coherence.
- In terms of event coherence at level 3, the proposed dhCM and dhCM-E outperform the two flat topic detectors (TrioVec and MS) on the three datasets. The comparison implies not only the terms within the categories at level 1 and level 2 but also the identified events in dhCM and dhCM-E are meaningful and interpretable.

Through the comparisons with both online and static competitors, the coherence results on the three datasets show that the proposed integration of density estimation and categorization modeling in dhCM and dhCM-E are effective to produce a coherent and interpretable category hierarchy for the document stream.

### 5.3 Node Abstraction/Specialization

In general, the nodes located at the top of a hierarchical structure are regarded as the most general categories, whereas the nodes placed near the leaf are deemed as more specific categories. To quantitatively measure such a *general-to-specific* property when traversing down the hierarchical structure, we borrow an evaluation metric *node abstraction/specialization* from the work of Kim et al. [22].

Let $\phi_0$ represent the word distribution of root node at level 0 from the hierarchy, which is considered as the most general topic of the corpus. The topic $\phi_k$ denotes a category node anywhere in the hierarchy. We measure the semantic distance between $\phi_0$ and $\phi_k$ to see how far away the topic $\phi_k$ has drifted from the most general topic $\phi_0$ in terms of word distribution. The semantic distance between two word distributions $\phi_0$ and $\phi_k$ is defined as follows:

$$D(\phi_0||\phi_k) = 0.5 \cdot D_{KL}(\phi_0||\phi_k) + 0.5 \cdot D_{KL}(\phi_k||\phi_0), \tag{14}$$

where $D_{KL}(\phi_k||\phi_0)$ is the Kullback-Leibler divergence from $\phi_0$ to $\phi_k$. A higher value indicates the word distribution of $\phi_k$ has drifted farther away from $\phi_0$, which implies the category is more specialized. To make a fair comparison, we use a global root node $\phi_G$ from the entire dataset to compare with the nodes at different levels from all competitors. The probability of each term $v$ in $\phi_G$ is the ratio between its occurrence number and sum of occurrence numbers of all terms $P(v|\phi_G) = n_v / \sum_i n_i$. The category specialization value for the $N_i$ nodes at level $i$ (i=1, 2, . . .) is defined as

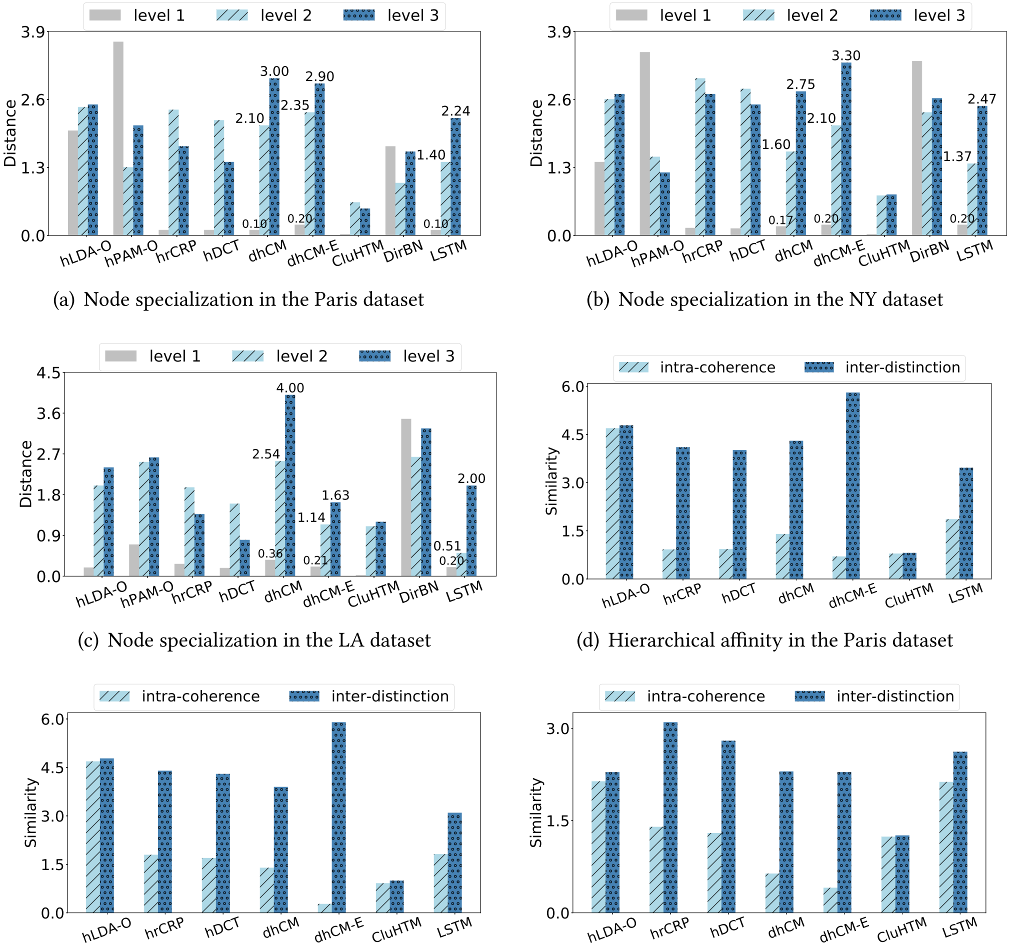
$$Dis(level\, i) = \frac{1}{|N_i|} \sum_{k=1}^{|N_i|} D(\phi_G||\phi_k), \tag{15}$$

which is the averaging distance from the global root $\phi_G$.

We compute the category specialization scores of all competitors and the results on three datasets are presented in Figure 5(a), (b), and (c), respectively.

The discussion on *node abstraction/specialization* results over three datasets is in the following:

- Clearly, the proposed dhCM, dhCM-E, and LSTM accomplish the continuously increasing node specialization performance from level 1 to level 3, which are annotated in Figure 5(a), (b), and (c). Such results imply that category nodes in dhCM and LSTM gradually turn to be more specialized as the level goes deeper, which confirms their advantage in the construction of hierarchical structures.
- In hLDA and CluHTM, the node specialization values at deeper levels are higher than the level 1, but their difference between level 2 and level 3 is not as evident as dhCM and LSTM.



(a) Node specialization in the Paris dataset

(b) Node specialization in the NY dataset

(c) Node specialization in the LA dataset

(d) Hierarchical affinity in the Paris dataset

(e) Hierarchical affinity in the NY dataset

(f) Hierarchical affinity in the LA dataset

Fig. 5. Node abstraction/specialization and hierarchical affinity results on three datasets.

After inspecting the results of hLDA-O, we find that the words from a short document in hLDA-O are mostly focused on a single topic rather than evenly distributed along the multi-topic path, and thus the property of general-to-specific topic hierarchy in hLDA-O is not distinctive.

- Distinct from the preceding observations, hDCT and hrCRP gain an evident increase from level 0 to level 1, whereas the node specialization value declines at level 2. We attribute it to the topic merging mechanism via the cosine similarity in these two models.
- Unfortunately, the results from hPAM-O and DirBN are not satisfactory in terms of the general-to-specific feature. The poor performance of hPAM-O and DirBN stems from their fully connected topic layers in the hierarchical design.

### 5.4 Hierarchical Affinity

*Hierarchical affinity* [22] is another significant principle to measure the affinity of a hierarchical structure. It consists of two aspects: (1) *intra-coherence*, the sibling categories descending from

Table 5. Event Accuracy Results over the Three Datasets

| | TrioVec | MS | hDCT | hrCRP | DirBN | dhCM | dhCM-S | dhCM-E | dhCM-E-S |
|---|---|---|---|---|---|---|---|---|---|
| **Paris** | 0.62 | 0.52 | 0.49 | 0.51 | 0.46 | 0.65 | **0.67** | 0.71 | **0.72** |
| **NY** | 0.73 | 0.60 | 0.63 | 0.65 | 0.54 | 0.75 | **0.76** | 0.79 | **0.79** |
| **LA** | 0.68 | 0.46 | 0.48 | 0.49 | 0.41 | 0.63 | **0.64** | 0.69 | **0.71** |

The best performance is denoted in boldface.

the common parent node should be more semantically coherent, and (2) *inter-distinction*, child categories are distinct from those non-parent categories. This widely used metric depicts a clear affinity of a category hierarchy [27].

Since the root node from all hierarchical structures fully connects with its children at level 1, we start with nodes at level 1. Let $\phi_k$ be a parent node, $\phi_{ki}$ be the set of its direct child nodes, and $\phi_{ji} (j \neq k)$ be the set of its child nodes descending from other parent node $j$ at the same level. To evaluate the *hierarchical affinity*, we compare the closeness between $\phi_k$ and its children $\phi_{ki}$ against $\phi_k$ and non-children $\phi_{ji} (j \neq k)$. The *hierarchical affinity* is quantified by Equation (14). Hence, a small value could be interpreted that two nodes are more semantically close.

Because DirBN and hPAM-O are designed with fully connected topic layers in depth, they are excluded from this metric. The *hierarchical affinity* results from the rest of the baselines are presented in Figure 5(d), (e), and (f), and the corresponding analysis is in the following:

- The performance of hDCT and hrCRP is satisfactory and stable over three datasets, and it gives credit to the merging mechanism. In these two approaches, only semantically relevant child topics are formed to produce their parent node via the cosine similarity, which naturally results in small values in intra-coherence and large numbers in inter-distinction. We take them as reference to other competitors.
- Both dhCM and dhCM-E achieve good results in the hierarchical affinity where the distance between child nodes and their parent nodes are much smaller against those non-parent nodes. Such performance confirms the superiority of categorization modeling in the task of organizing documents hierarchically. Compared to dhCM, the improvement of dhCM-E is benefited from the incorporation of word embedding, which enhances the clustering process of documents and words at different levels.
- The performance of LSTM over three datasets is acceptable, where the intra-coherence measuring parent-to-child affinity is smaller than that inter-distinction between parent and non-child relationship.
- However, in hLDA-O and CluHTM, the difference between intra-coherence and inter-distinction is not notable. One reason for hLDA-O is that the multi-topic assignment is not appropriate for the clustering of short texts. In CluHTM, the iterative non-negative matrix factorization exploited in CluHTM fails to distinguish the intra-coherence and inter-distinction between topics at two consecutive levels.

### 5.5 Event Detection Accuracy

Event detection is the significant base for building a high-quality category hierarchy. To evaluate the *event detection accuracy* from the baselines, three experts are invited to measure the accuracy of events by examining their semantics, timestamps, and geographical areas. Only when all of them agree the detected event is true is it considered as true. The hierarchical models hDCT, hrCRP, DirBN, dhCM, and dhCM-E and their static versions are included because their leaf nodes exactly correspond to the events. Together with another two flat online competitors, TriVec and MS, the event detection results on the three datasets are summarized in Table 5.

Table 6. Category Accuracy Results over the Three Datasets

| | hLDA-O | hDCT | hrCRP | CluHTM | LSTM | dhCM | dhCM-S | dhCM-E | dhCM-E-S |
|---|---|---|---|---|---|---|---|---|---|
| **Paris** | 0.23 | 0.66 | 0.76 | 0.47 | 0.54 | 0.75 | 0.75 | **0.77** | **0.77** |
| **NY** | 0.30 | 0.65 | 0.71 | 0.56 | 0.73 | 0.77 | 0.77 | **0.80** | **0.80** |
| **LA** | 0.25 | 0.58 | 0.65 | 0.36 | 0.36 | 0.65 | 0.66 | **0.78** | **0.78** |

The best performance is denoted in boldface.

*Discussion.* First, except for the flat competitors (TrioVec and MS), the proposed dhCM, dhCM-E, and their static versions once again achieve better event accuracy on the three datasets than those hierarchical competitors (hDCT, hrCRP, and DirBN). The evident distinction between them confirms the superiority of the proposed dynamic categorization modeling and density estimation technique. Second, compared with TrioVec and MS, the flat version of online event detectors, dhCM and dhCM-E still retain their advantage, implying our proposed dhCM and dhCM-E are effective to detect the incremental events. Third, regarding the static dhCM-S and dhCM-E-S, which take the whole dataset as the input, dhCM and dhCM-E gain comparable results. For example, dhCM-E achieves the accuracy of 0.71 while achieving 0.72 from dhCM-E-S on the Paris dataset, and both achieve the accuracy of 0.79 on the NY dataset. The slight difference between them further demonstrates that the density estimation technique in the proposed model is effective to deal with the document stream and identify the true events.

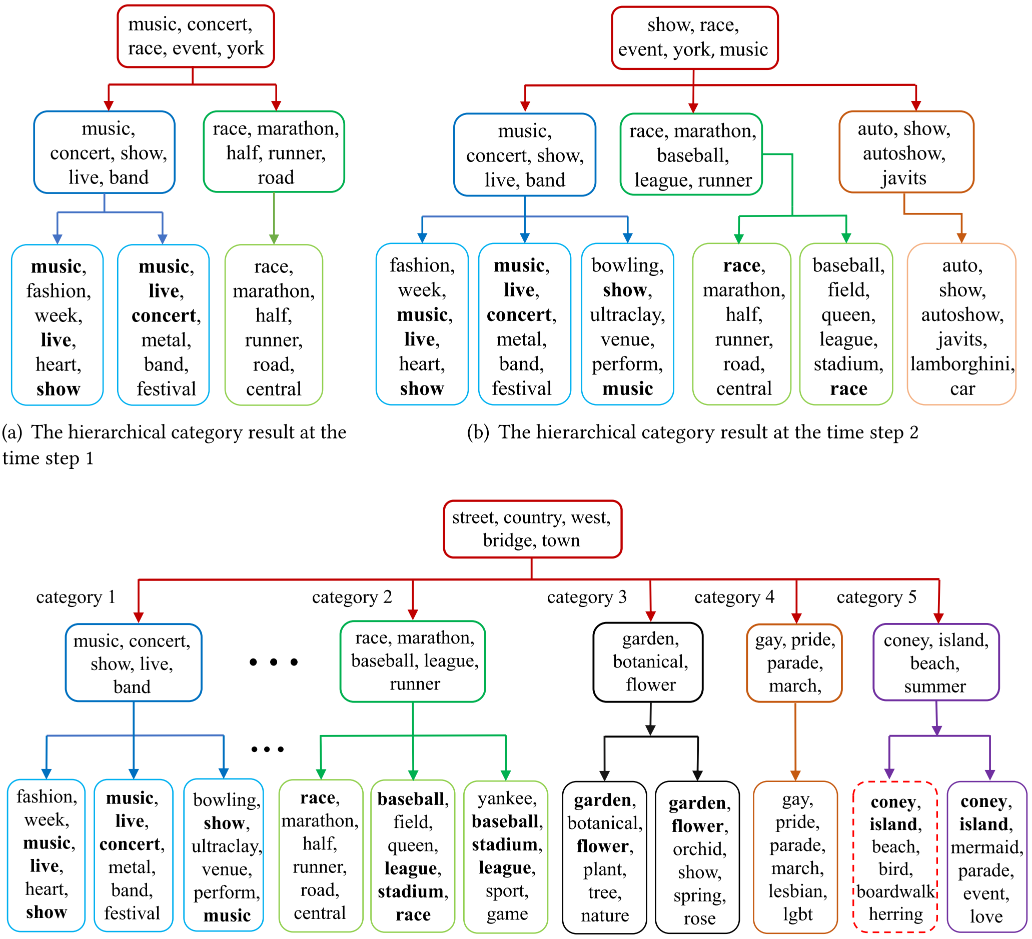
## 5.6 Category Accuracy

*Category accuracy* is an important metric to identify whether a child category belongs to its parent category when traversing the overall category hierarchy [26, 27]. Following the empirical study [15, 26, 27], we invite five graduate students to determine the accuracy of hierarchical relations. By presenting top-10 terms of both child category and parent category, each annotator independently decides whether the child category belongs to its parent category. The final results are the average scores of the five annotators. The competitors hPAM-O and DirBN are omitted due to the fully connected layers in the hierarchy, and the static dhCM-S and dhCM-E-S are included to further measure the effectiveness of the density estimation. The results of category accuracy on the three datasets are presented in Table 6.

*Discussion.* First, among all competitors, dhCM, dhCM-E, and their static versions achieve the highest accuracy in the task of hierarchical relation on the three datasets, implying the merits of the proposed hierarchical categorization modeling in the task of parent-to-child relationship identification. The enhanced performance of dhCM-E and dhCM-E-S is benefited from the pretrained word embeddings. Second, taking their static versions as references, dhCM obtains nearly the same accuracy results as the dhCM-S (e.g., 0.75 on the Paris dataset and 0.77 on the NY dataset), and dhCM-E earns the same results with dhCM-E-S on the three datasets. Such comparisons further validate the effectiveness of the density estimation technique in dealing with the social media data stream. Third, the performance of hrCRP and hDCT is followed, which benefits greatly from the merging mechanism between relevant topics. The performance of hLDA-O on the category accuracy is poor. The deteriorative performance is caused by the sparsity problem of short texts.

## 5.7 Qualitative Analysis

To have a better understanding of the dynamic and hierarchical category, we present two intuitive examples from dhCM and hLDA-O on the NY dataset, which are denoted in Figure 6 and Figure 7, respectively.


music, concert, race, event, york
music, concert, show, live, band
race, marathon, half, runner, road
music, fashion, week, live, heart, show
music, live, concert, metal, band, festival
race, marathon, half, runner, road, central
show, race, event, york, music
music, concert, show, live, band
race, marathon, baseball, league, runner
auto, show, autoshow, javits
fashion, week, music, live, heart, show
music, live, concert, metal, band, festival
bowling, show, ultraclay, venue, perform, music
race, marathon, half, runner, road, central
baseball, field, queen, league, stadium, race
auto, show, autoshow, javits, lamborghini, car
street, country, west, bridge, town
category 1
category 2
category 3
category 4
category 5
music, concert, show, live, band
race, marathon, baseball, league, runner
garden, botanical, flower
gay, pride, parade, march,
coney, island, beach, summer
fashion, week, music, live, heart, show
music, live, concert, metal, band, festival
bowling, show, ultraclay, venue, perform, music
race, marathon, half, runner, road, central
baseball, field, queen, league, stadium, race
yankee, baseball, stadium, league, sport, game
garden, botanical, flower, plant, tree, nature
garden, flower, orchid, show, spring, rose
gay, pride, parade, march, lesbian, lgbt
coney, island, beach, bird, boardwalk herring
coney, island, mermaid, parade, event, love

(a) The hierarchical category result at the time step 1

(b) The hierarchical category result at the time step 2

(c) The hierarchical category result at the time step 3

Fig. 6. The example of dynamic and hierarchical category from dhCM.

Figure 6 illustrates the incremental category hierarchy from the dhCM along the timeline, where each category or event is described by a set of most probable words. At timestep 1, three different events at the bottom level are detected, and two on music are naturally grouped into one category and the remaining one forms its own category. With the new arrivals of documents, the category hierarchy naturally grows with more events and categories. At step 3, more diverse categories (e.g., category 1, category 2, category 3, and category 4) are detected and well separated from each other. More importantly, their descendant events, corresponding to the real activities occurring in cohesive temporal and spatial range, are closely connected by the common words that are highlighted in boldface. Category 5 and its descendant nodes serve as the failure case, where the former shown with a dotted line depicts a static topic while the latter talks about the event of a mermaid parade. Since a large number of documents clustered into the former node, which share the strong semantics and spatial pattern with the latter, are mistakenly recognized as the sister of an event. In comparison, Figure 7 shows the example of category hierarchy from hLDA-O, where the categories are not good summaries of their children. Specifically, the dotted

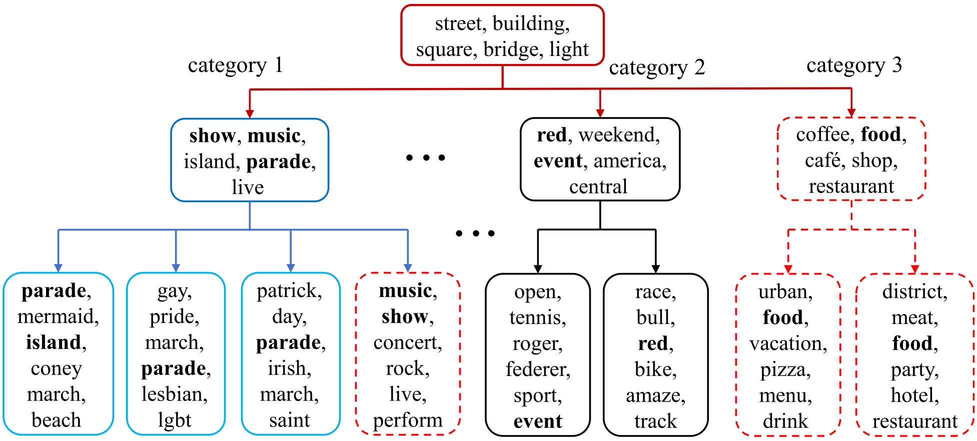
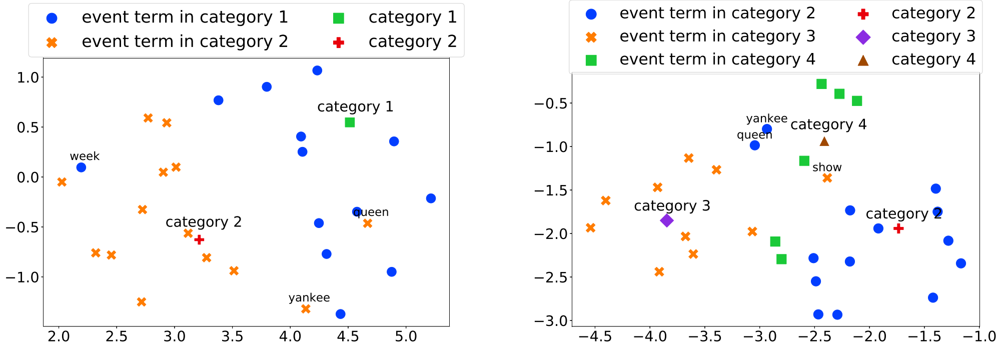

street, building, square, bridge, light
category 1
category 2
category 3
**show**, **music**, island, **parade**, live
• • •
**red**, weekend, **event**, america, central
coffee, **food**, café, shop, restaurant
• • •
**parade**, mermaid, **island**, coney march, beach
gay, pride, march, **parade**, lesbian, lgbt
patrick, day, **parade**, irish, march, saint
**music**, **show**, concert, rock, live, perform
open, tennis, roger, federer, sport, **event**
race, bull, **red**, bike, amaze, track
urban, **food**, vacation, pizza, menu, drink
district, meat, **food**, party, hotel, restaurant

Fig. 7. Example of the hierarchical category from the baseline hLDA-O.



(a) Visualization of categories 1 and 2

(b) Visualization of categories 2, 3, and 4

Fig. 8. Joint embedding space visualization of categories and their events.

node under category 1 is about a music show, which is different from its siblings. Additionally, category 3 and its descendant nodes record the static topics instead of real events.

Such a comparison confirms that the proposed hierarchical categorization modeling is capable of identifying events via the multimodal miner from the document stream, and dynamically categorizing them into an interpretable hierarchy.

*Embedding space visualization.* To further validate whether the categories align with their corresponding events in the embedding space, we use the first four categories in Figure 6(c) and apply t-SNE [25] to visualize the embedding space in Figure 8, where each category is denoted by its inferred embedding representation, and each event is denoted by the most probable words.

It is observed that categories are well distinguished in the embedding space, and each category is surrounded by the representative terms from its events. In Figure 8(a), category 1 on music is well distinguished from category 2 on race in the embedding space, and both are close to their representative terms. In Figure 8(b), three categories and their representative terms are well identified. This shows that the proposed dynamic and hierarchical categorization modeling not only incrementally identifies events and its categories from the social media stream but also captures a coherent and interpretable category hierarchy.

## 6 RELATED WORK

There are mainly two research lines that are highly related to our work. One is the hierarchical topic modeling including unsupervised and supervised methods, and the other is online topic modeling.

*Hierarchical topic modeling.* hLDA [4, 14] is an unsupervised method, which only requires the document collection as the input and produces a hierarchical topic structure with adaptive growth. Each document in hLDA is generated by multiple topics along the path from the hierarchy. Based on hLDA, Kim et al. [22] relieve the restriction of a single path and generates the document by the mixture of any topic in the hierarchy. [40] further propose a nested HDP, consisting of multiple levels of HDP, to address the nonparametric multi-level admixture problem. Distinct from the preceding adaptive hierarchical structures, other works [29, 47, 58] require the specified depth and width at each level to define the hierarchical structure. Among them, the hPAM [29] is defined on a three-level hierarchy with a specified number of topics at each level. Zhao et al. [58] propose a DirBN, where the topics at the bottom level are propagated to upper levels. In addition, Wang et al. [44] construct a phrase-based hierarchy through recursively clustering term co-occurrence network. Wang et al. [47] build a heterogeneous text graph including word, document, and topic nodes, and leverage the graph convolutional networks (GCNs) to learn the weighted edges between topics. Although these approaches are successful in the topic hierarchy, their predefined structures are significantly challenged in the dynamic context. In contrast, our unsupervised model dhCM is able to accommodate the document streams and dynamically update the hierarchical structure. However, Dubey et al. [11] combine both hierarchical structure [13] and time evolving together, to dynamically cluster the documents into a hierarchy. It is a dynamic clustering of documents and fails in extracting the latent topics from document streams.

In addition, other works [20, 27, 37, 38, 43, 54] require users to provide extra keywords or external knowledge as supervision to guide the construction of hierarchy. Based on the specified key terms, the method in the work of Zhang et al. [54] recursively applies the adaptive clustering process to push down the specific words into the child nodes and builds a term taxonomy. The work of Shen et al. [37] takes the given keyword taxonomy as supervision and recursively expands the entity set into a term taxonomy. The recent advance in the work of Meng et al. [27] also takes the specified category tree as guidance and extracts the representative term for each specified category, whereas the work of Huang et al. [20] expands the given keyword taxonomy in both depth and width by learning and transferring the given relations. Effective as these term taxonomies, they mainly focus on exploring the hierarchical relationship between terms, which majorly differs from our work. In addition, these methods require that users first have good prior knowledge of the whole corpus. However, in the context of document streams, such prior knowledge of the whole corpus is often unavailable.

In addition, the supervised approaches in other works [32, 45] focus on the labeled document collection and develop the supervised topic hierarchy. Shin and Moon [38] incorporate the domain knowledge as the prior and improve the hierarchical clustering. The method of Viegas et al. [43] takes the pretrained word embedding as the input and explores the matrix factorization iteratively to build a hierarchical structure. Different from these text-based hierarchies, the work of Shang et al. [36] takes both texts and networks as input to jointly learn the term embedding, then clusters terms into child nodes and forms the term taxonomy. The work of [57] takes the metadata and label hierarchy as supervision to categorize the document collection. Still, these models target at static document collection and ignore the dynamics of document stream in the online setting.

*Online topic modeling.* We summarize the related work into two categories according to the inference method. One is based on the SMC inference process, which sequentially processes the continuous documents. Canini et al. [5] introduce online inference of SMC based on LDA with a given topic number. Ahmed et al. [1] relieve the fixed topic number and apply SMC to detect a hybrid of topics and storylines. Other works [10, 16] induce a kernel-based algorithm to capture the temporal influence. Among them, a Hawkes process algorithm [10] is used to capture the dynamics of events by using a specific number of Gaussian kernels.

The other is Gibbs sampling based approaches, which are built on a predefined sequence of time slides. The main deficiency of this technique is that nobody knows the optimal setting for the temporal slices. Furthermore, only when the time slice is fully saturated with document arrivals can it start the detection task. Given the sequence of time slices, other methods [2, 24] learn the topic generation in the current slice depending on the results from the previous. Yin et al. [50] group the short texts into clusters in each batch based on the DP.

Although these methods are able to extract the events across time, they ignore hierarchical relations among them, and consequently the flat set of events from these models remain ad hoc and structureless without further processing. In comparison, dhCM not only detects the events from the document streams but also dynamically categorizes them into proper levels of abstraction.

In addition, there are other online studies. Other methods [46, 48] study the dependency of topic distribution, whereas still others [3, 21] learn the dependency of both topic distribution and topic transition between consecutive documents. Given the query window, some works [7, 52, 53, 56] detect a set of online local events. Still, these studies ignore the hierarchical relationship among the unstructured online events, and thus they could not present a dynamic hierarchical structure as dhCM does across time.

## 7 CONCLUSION

By bridging the gap between online event detection and hierarchical organization, we propose a dynamic and hierarchical categorization model for social media stream. As the document stream arrives, both the categories with different levels of abstraction and event assignments are automatically identified. This work includes two key contributions. First, the novel hierarchical structure by dhCM is able to dynamically organize documents into a category hierarchy with flexible growth, to accommodate the increasing streams in the online setting. Such a dynamic and adaptive merit well fits the continuous streams in the online setting. Second, the density estimation technique is proposed to flexibly learn the temporal dependency from the past document without manually dividing the timeframe. Our extensive experiments have confirmed that our proposed model is capable of identifying a high-quality categorization hierarchy.

Since the incorporation of word embeddings is effective to enhance the performance in various NLP tasks. Without the aid of the pretrained word embeddings, a future attempt of this work is to integrate the learning of word representations with the training of adaptive category hierarchy.

## REFERENCES

[1] Amr Ahmed, Qirong Ho, Choon Hui Teo, Jacob Eisenstein, Alex Smola, and Eric Xing. 2011. Online inference for the infinite topic-cluster model: Storylines from streaming text. In *Proceedings of AISTATS*. 101–109.

[2] Amr Ahmed and Eric Xing. 2008. Dynamic non-parametric mixture models and the recurrent Chinese restaurant process: With applications to evolutionary clustering. In *Proceedings of SDM*. 219–230.

[3] Hesam Amoualian, Marianne Clausel, Eric Gaussier, and Massih-Reza Amini. 2016. Streaming-LDA: A Copula-based approach to modeling topic dependencies in document streams. In *Proceedings of SIGKDD*. ACM, New York, NY, 695–704.

[4] David M. Blei, Thomas L. Griffiths, and Michael I. Jordan. 2010. The nested Chinese restaurant process and Bayesian nonparametric inference of topic hierarchies. *Journal of the ACM* 57, 2 (2010), 7.

[5] Kevin Canini, Lei Shi, and Thomas Griffiths. 2009. Online inference of topics with latent Dirichlet allocation. In *Artificial Intelligence and Statistics*. 65–72.

[6] Rajarshi Das, Manzil Zaheer, and Chris Dyer. 2015. Gaussian LDA for topic models with word embeddings. In *Proceedings of ACL*. 795–804.

[7] Debanjan Datta. 2020. Small Survey Event Detection. arXiv:2011.05801 (2020).

[8] Arnaud Doucet, Nando De Freitas, Kevin Murphy, and Stuart Russell. 2000. Rao-Blackwellised particle filtering for dynamic Bayesian networks. In *Proceedings of UAI*. 176–183.

[9] Arnaud Doucet, Nando de Freitas, and Gordon Neil. 2001. *Sequential Monte Carlo Methods in Practice*. Springer, New York, NY.

[10] 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 *Proceedings of SIGKDD*. ACM, New York, NY, 219–228.

[11] Kumar Dubey, Qirong Ho, Sinead A. Williamson, and Eric P. Xing. 2014. Dependent nonparametric trees for dynamic hierarchical clustering. In *Proceedings of NIPS*. 1152–1160.

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

[13] Zoubin Ghahramani, Michael I. Jordan, and Ryan P. Adams. 2010. Tree-structured stick breaking for hierarchical data. In *Proceedings of NIPS*. 19–27.

[14] Thomas L. Griffiths, Michael I. Jordan, Joshua B. Tenenbaum, and David M. Blei. 2004. Hierarchical topic models and the nested Chinese restaurant process. In *Proceedings of NIPS*. 17–24.

[15] Jinjin Guo and Zhiguo Gong. 2016. A nonparametric model for event discovery in the geospatial-temporal space. In *Proceedings of CIKM*. ACM, New York, NY, 499–508.

[16] Jinjin Guo and Zhiguo Gong. 2017. A density-based nonparametric model for online event discovery from the social media data. In *Proceedings of IJCAI*. 1732–1738.

[17] Alexander Hinneburg and Daniel A. Keim. 1998. An efficient approach to clustering in large multimedia databases with noise. In *Proceedings of SIGKDD*, Vol. 98. ACM, New York, NY, 58–65.

[18] Nils Lid Hjort, Chris Holmes, Peter Müller, and Stephen G. Walker. 2010. *Bayesian Nonparametrics*. Vol. 28. Cambridge University Press.

[19] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation* 9, 8 (1997), 1735–1780.

[20] Jiaxin Huang, Yiqing Xie, Yu Meng, Yunyi Zhang, and Jiawei Han. 2020. CoRel: Seed-guided topical taxonomy construction by concept learning and relation transferring. In *Proceedings of SIGKDD*. 1928–1936.

[21] Tomoharu Iwata, Shinji Watanabe, Takeshi Yamada, and Naonori Ueda. 2009. Topic tracking model for analyzing consumer purchase behavior. In *Proceedings of IJCAI*, Vol. 9. 1427–1432.

[22] Joon Hee Kim, Dongwoo Kim, Suin Kim, and Alice Oh. 2012. Modeling topic hierarchies with the recursive Chinese restaurant process. In *Proceedings of CIKM*. ACM, New York, NY, 783–792.

[23] Chenliang Li, Yu Duan, Haoran Wang, Zhiqian Zhang, Aixin Sun, and Zongyang Ma. 2017. Enhancing topic modeling for short texts with auxiliary word embeddings. *TOIS* 36, 2 (2017), 1–30.

[24] Shangsong Liang, Emine Yilmaz, and Evangelos Kanoulas. 2016. Dynamic clustering of streaming short documents. In *Proceedings of SIGKDD*. ACM, New York, NY, 995–1004.

[25] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research* 9 (Nov. 2008), 2579–2605.

[26] Yu Meng, Jiaxin Huang, Guangyuan Wang, Zihan Wang, Chao Zhang, Yu Zhang, and Jiawei Han. 2020. Discriminative topic mining via category-name guided text embedding. In *Proceedings of WWW*. 2121–2132.

[27] Yu Meng, Yunyi Zhang, Jiaxin Huang, Yu Zhang, Chao Zhang, and Jiawei Han. 2020. Hierarchical topic mining via joint spherical tree and text embedding. In *Proceedings of ACM SIGKDD*.

[28] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of NIPS*. 3111–3119.

[29] David Mimno, Wei Li, and Andrew McCallum. 2007. Mixtures of hierarchical topics with Pachinko allocation. In *Proceedings of ICML*. ACM, New York, NY, 633–640.

[30] David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. 2010. Automatic evaluation of topic coherence. In *Proceedings of NAACL*. 100–108.

[31] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of EMNLP*. 1532–1543.

[32] Adler J. Perotte, Frank Wood, Noemie Elhadad, and Nicholas Bartlett. 2011. Hierarchically supervised latent Dirichlet allocation. In *Proceedings of NIPS*. 2609–2617.

[33] Russell Revlin. 2012. *Cognition: Theory and Practice*. Macmillan.

[34] Michael Röder, Andreas Both, and Alexander Hinneburg. 2015. Exploring the space of topic coherence measures. In *Proceedings of WSDM*. ACM, New York, NY, 399–408.

[35] Abel Rodriguez, David B. Dunson, and Alan E. Gelfand. 2008. The nested Dirichlet process. *Journal of the American Statistical Association* 103, 483 (2008), 1131–1154.

[36] Jingbo Shang, Xinyang Zhang, Liyuan Liu, Sha Li, and Jiawei Han. 2020. Nettaxo: Automated topic taxonomy construction from text-rich network. In *Proceedings of WWW*. 1908–1919.

[37] Jiaming Shen, Zeqiu Wu, Dongming Lei, Chao Zhang, Xiang Ren, Michelle T. Vanni, Brian M. Sadler, and Jiawei Han. 2018. HiExpan: Task-guided taxonomy construction by hierarchical tree expansion. In *Proceedings of SIGKDD*. ACM, New York, NY, 2180–2189.

[38] Su-Jin Shin and Il-Chul Moon. 2017. Guided HTM: Hierarchical topic model with Dirichlet forest priors. *IEEE Transactions on Knowledge and Data Engineering* 29, 2 (2017), 330–343.

[39] Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei. 2004. Hierarchical Dirichlet processes. *Journal of the American Statistical Association* 101 (2004), 1–30.

[40] Lavanya Sita Tekumalla, Priyanka Agrawal, and Indrajit Bhattacharya. 2013. Nested Hierarchical Dirichlet Processes for Multi-Level Non-Parametric Admixture Modeling. *ECML* (2013).

[41] Christiaan M. Van der Walt and Etienne Barnard. 2017. Variable kernel density estimation in high-dimensional feature spaces. In *Proceedings of AAAI*. 2674–2680.

[42] Felipe Viegas, Sérgio Canuto, Christian Gomes, Washington Luiz, Thierson Rosa, Sabir Ribas, Leonardo Rocha, and Marcos André Gonçalves. 2019. CluWords: Exploiting semantic word clustering representation for enhanced topic modeling. In *Proceedings of WSDM*. 753–761.

[43] Felipe Viegas, Washington Cunha, Christian Gomes, Antônio Pereira, Leonardo Rocha, and Marcos Goncalves. 2020. CluHTM-semantic hierarchical topic modeling based on CluWords. In *Proceedings of ACL*. 8138–8150.

[44] Chi Wang, Marina Danilevsky, Nihit Desai, Yinan Zhang, Phuong Nguyen, Thrivikrama Taula, and Jiawei Han. 2013. A phrase mining framework for recursive construction of a topical hierarchy. In *Proceedings of SIGKDD*. ACM, New York, NY, 437–445.

[45] Chi Wang, Jialu Liu, Nihit Desai, Marina Danilevsky, and Jiawei Han. 2015. Constructing topical hierarchies in heterogeneous information networks. *Knowledge and Information Systems* 44, 3 (2015), 529–558.

[46] Yu Wang, Eugene Agichtein, and Michele Benzi. 2012. TM-LDA: Efficient online modeling of latent topic transitions in social media. In *Proceedings of SIGKDD*. ACM, New York, NY, 123–131.

[47] Zhengjue Wang, Chaojie Wang, Hao Zhang, Zhibin Duan, Mingyuan Zhou, and Bo Chen. 2020. Learning dynamic hierarchical topic graph with graph convolutional network for document classification. In *Proceedings of AISTATS*.

[48] Xing Wei, Jimeng Sun, and Xuerui Wang. 2007. Dynamic mixture models for multiple time-series. In *Proceedings of IJCAI*, Vol. 7. 2909–2914.

[49] Guangxu Xun, Yaliang Li, Wayne Xin Zhao, Jing Gao, and Aidong Zhang. 2017. A correlated topic model using word embeddings. In *Proceedings of IJCAI*. 4207–4213.

[50] Jianhua Yin, Daren Chao, Zhongkun Liu, Wei Zhang, Xiaohui Yu, and Jianyong Wang. 2018. Model-based clustering of short text streams. In *Proceedings of SIGKDD*. ACM, New York, NY, 2634–2642.

[51] Jianhua Yin and Jianyong Wang. 2014. A Dirichlet multinomial mixture model-based approach for short text clustering. In *Proceedings of SIGKDD*. ACM, New York, NY, 233–242.

[52] Chao Zhang, Dongming Lei, Quan Yuan, Honglei Zhuang, Lance Kaplan, Shaowen Wang, and Jiawei Han. 2018. GeoBurst+: Effective and real-time local event detection in geo-tagged tweet streams. *ACM Transactions on Intelligent Systems* 9, 3 (2018), 34.

[53] Chao Zhang, Liyuan Liu, Dongming Lei, Quan Yuan, Honglei Zhuang, Tim Hanratty, and Jiawei Han. 2017. TrioVecEvent: Embedding-based online local event detection in geo-tagged tweet streams. In *Proceedings of SIGKDD*. ACM, New York, NY, 595–604.

[54] Chao Zhang, Fangbo Tao, Xiusi Chen, Jiaming Shen, Meng Jiang, Brian Sadler, Michelle Vanni, and Jiawei Han. 2018. TaxoGen: Unsupervised topic taxonomy construction by adaptive term embedding and clustering. In *Proceedings of SIGKDD*. ACM, New York, NY, 2701–2709.

[55] Chao Zhang, Keyang Zhang, Quan Yuan, Haoruo Peng, Yu Zheng, Tim Hanratty, Shaowen Wang, and Jiawei Han. 2017. Regions, periods, activities: Uncovering urban dynamics via cross-modal representation learning. In *Proceedings of WWW*. 361–370.

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

[57] Yu Zhang, Xiusi Chen, Yu Meng, and Jiawei Han. 2021. Hierarchical Metadata-Aware Document Categorization under Weak Supervision. In *WSDM*. ACM, 770–778.

[58] He Zhao, Lan Du, Wray Buntine, and Mingyuan Zhou. 2018. Dirichlet belief networks for topic structure learning. In *Proceedings of NIPS*. 7955–7966.

[59] Daiane Aparecida Zuanetti, Peter Müller, Yitan Zhu, Shengjie Yang, and Yuan Ji. 2018. Clustering distributions with the marginalized nested Dirichlet process. *Biometrics* 74, 2 (2018), 584–594.

Received November 2020; revised May 2021; accepted June 2021