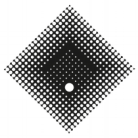


Mining Event-Oriented Topics in Microblog Stream with Unsupervised Multi-View Hierarchical Embedding

Mining Event-oriented Topics in Microblog Stream with Unsupervised Multi-view Hierarchical Embedding

MIN PENG, Wuhan University
JIAHUI ZHU, Wuhan University
HUA WANG, Victoria University
XUHUI LI, Wuhan University
YANCHUN ZHANG, Victoria University
XIUZHEN ZHANG, RMIT University
GANG TIAN, Wuhan University

This paper presents an *unsupervised multi-view hierarchical embedding* (UMHE) framework to sufficiently reveal the intrinsic topical knowledge in social events. Event-oriented topics are highly related to such events as it can provide explicit descriptions of what have happened in social community. In many real-world cases, however, it is difficult to include all attributes of microblogs, more often, textual aspects only are available. Traditional topic modeling methods are failed to generate event-oriented topics with the textual aspects, since the inherent relations between topics are often overlooked in these methods. Meanwhile, the metrics in original word vocabulary space might not effectively capture semantic distances. Our UMHE framework overcomes the severe information deficiency and poor feature representation. The UMHE first develops a multi-view Bayesian rose tree to preliminarily generate prior knowledge for latent topics and their relations. With such prior knowledge, we design an unsupervised translation-based hierarchical embedding method to make a better representation of these latent topics. By applying self-adaptive spectral clustering on the embedding space and the original space concomitantly, we eventually extract event-oriented topics in word distributions to express social events. Our framework is purely data-driven and unsupervised, without any external knowledge. Experimental results on TREC Tweets2011 dataset and Sina Weibo dataset demonstrate that the UMHE framework can construct hierarchical structure with high fitness, but also yield topic embeddings with salient semantics, therefore, it can derive event-oriented topics with meaningful descriptions.

CCS Concepts: • **Information systems** → **Data stream mining**;

General Terms: Algorithms, Experimentation

This work is supported by the Natural Science Foundation of China, under grant No.61472291, grant No.61272110, and Natural Science Foundation of Hubei Province, China, under grant No.ZRY2014000901.

Author's addresses: M. Peng (pengm@whu.edu.cn), J. Zhu (zhujiahui@whu.edu.cn), G. Tian (tiang2008@whu.edu.cn, corresponding author), School of Computer, Wuhan University, Wuhan, China; X. Li (lixuhui@whu.edu.cn), School of Information Management, Wuhan University, Wuhan, China; H. Wang (hua.wang@vu.edu.au) and Y. Zhang (yanchun.zhang@vu.edu.au), Centre for Applied Informatics, Victoria University, Melbourne, Australia; X. Zhang (xiuzhen.zhang@rmit.edu.au), School of CS&IT, RMIT University, Melbourne, Australia.

Additional Key Words and Phrases: Event-oriented topic, multi-view hierarchical embedding, unsupervised learning, Bayesian rose tree

ACM Reference format:
Min Peng, Jiahui Zhu, Hua Wang, Xuhui Li, Yanchun Zhang, Xiuzhen Zhang, and Gang Tian. 2017. Mining Event-oriented Topics in Microblog Stream with Unsupervised Multi-view Hierarchical Embedding. *ACM Trans. Knowl. Discov. Data.* X, X, Article 39 (July 2017), 29 pages.
DOI: XXXXXXX.XXXXXXX

1 INTRODUCTION

Microblogging services, such as Twitter[1] and Sina Weibo[2], have become popular and influential social media platforms where people share public information and express their opinions. For example, more than 80% of Sina Weibo users participated in the discussion of the *2014 FIFA World Cup Brazil* and over 76% of users posted feeds (the messages released by users in social networks, also known as user generated contents) on Sina Weibo while watching the games[3]. In real-world applications, microblog feeds always arrive in the form of a stream in chronological order. In such a stream, several events may be discussed at the same time. Detecting topics that are highly oriented to these events is therefore essential for people and organizations. Discovering the event-oriented topics in microblog stream enables people to get the essence of the whole news and concerns that have already occurred in the previous time. Additionally, by analyzing the relations of these event-oriented topics, organizations can gradually build up a wide variety of applications, such as buzz lifecycle modeling (Chang et al. 2016), information propagation control (Yu et al. 2015), web user profiling (Tang et al. 2010), market competition monitoring (Zhang et al. 2015), and so on.

Apparently, the crux of such problems is the mining of topics. There are mainly two forms of topics: (1) word phrases and (2) multinomial distributions over word vocabulary. The phrase-based topics are usually produced by natural language processing (NLP) (Li et al. 2012), while the distribution-based topics are usually released by probabilistic topic models based on latent Dirichlet allocation (LDA) (Blei et al. 2003). The phrase-based topics can effectively conserve the semantic interpretation (coherence) of the event aspects. However, the precision of extracting such phrases is often limited by the NLP itself. For example, the precision of the "segment" in (Li et al. 2012) may be influenced by the effectiveness of its multi-gram extraction. Besides, external knowledge, such as Wikidata[4]and social network link formation (Lou et al. 2013), may be indispensable in these methods so as to achieve higher performance. Distribution-based topics, on the other hand, can be generated in a more unsupervised way because of their coarse-grained features. However, the coherence of such kind of topics is not guaranteed because of the poor understanding in presentation (Tang et al. 2014). Recently, topic embedding (Li et al. 2016) becomes a popular distribution-based representation method in which topics are projected into low dense vectors. In this embedding space, the semantic relations of such topics can be well captured.

[1]https://twitter.com/.
[2]http://weibo.com/.
[3]http://www.199it.com/archives/255612.html.
[4]https://www.wikidata.org/wiki/Wikidata:Main_Page.

When encountered with microblog streams, there are mainly three challenges in mining such distribution-based event-oriented topics. First, the coherence of topics is intractable to fulfill. A good way of solving this problem is to directly model the coherent topics, rather than the latent topics. Coherent topic models (Chen et al. 2013) are the state-of-the-art forms of topic models aiming at generating semantically meaningful topics. However, they often incur highly complex inference and are always computationally expensive. Second, how to reasonably refactor such topics into event-oriented ones is limited by the current knowledge. Although the microblog stream is voluminous and can provide heterogeneous information, it is still difficult to obtain all attributes of the data in real-world applications (McMinn et al. 2013). Normally, only the textual aspect of the microblog feeds is available for mining. This incompleteness of the information compels us to make full use of the current knowledge. Third, how to unsupervisedly associate topics into a joint model for precisely capturing their relations is also an open-ended question. *Bayesian rose tree* (BRT) (Blundell et al. 2010), a multi-branch tree, has received great attention because of its high fitness of text hierarchical modeling. However, this kind of tree-based or graph-based models simply interpret relations in symbols, while incapable of providing continuous numeric operation. *Translation-based embedding* method (Bordes et al. 2013) in knowledge graph representation is a promising way of jointly modeling topics and their relations, but there are naturally no prior topics or relations for training, and to organize training examples of topics together with their relations is a daunting work in such a stream occasion.

The motivation of our work is therefore to devise an unsupervised topic representation framework to automatically generate topic and relation embeddings that can deal with the challenges above. To overcome the information deficiency, we focus on the level of topics and propose a multi-view algorithm to create features in two views from the original textual aspect. One view is the obvious topic-word distribution, like "*(0.2, World Cup), (0.15, Argentina), (0.15, Shoot), (0.10, Promotion), ...*", the other is the related feed collection, like "d_1, d_2, d_5,...", where d_i denotes a feed in this collection. In this way, the knowledge learned in one view can guide the learning of the other. And to perform unsupervisedly, we firstly use Bayesian rose tree from two views to automatically construct the tree-style structure in a probabilistic manner, thus yielding the topics together with their symbolic relations for prior information. For example, although the words of topic T_a "*(0.2, World Cup), (0.15, Argentina), (0.15, Shoot), (0.10, Promotion), ...*" and Topic T_b "*(0.18, World Cup), (0.16, Germany), (0.15, Final), (0.12, Argentina), ...*" are not completely equivalent, they still share a large number of common words. In this case, these two topics should be regarded as the same event "*World Cup*" and they may share the join relation which would not be considered in LDA. However, in our multi-view BRT, we can figure out this join relation and derive a triplet *(T_a, join, T_b)*. Then this prior information is incorporated to align topics and relations in a joint model via translation-based embedding method. Topic embedding and relation embedding provide salient semantic measurement of topics and can help aggregate coherence to make them more event-oriented. For example, topic "*(0.2, World Cup), (0.15, Argentina), (0.15, Shoot), (0.10, Promotion), ...*" and topic "*(0.25, Germany), (0.20, Klose), (0.15, Prize), (0.10, Fans), ...*" both refer to the same event "*World Cup*", but the literal words are invalid to show any similarities. By projecting to embedding space, the intrinsic similarity of these two topics could be well captured.

To sum up, in this paper, we propose an *unsupervised multi-view hierarchical embedding* framework that can generate topics with a high accordance to the events in a microblog stream. First, we apply LDA to

extract the feed-topic distribution and the topic-word distribution of all feed batches. Therefore, for each latent topic, there are two different view features: (1) the *latent word distribution*, and (2) the *relevant feed collection*. Second, we design a novel *multi-view Bayesian rose tree* (Mv-BRT) to refactor the latent topics into a hierarchy to form the prior knowledge. Finally, a translation-based hierarchical embedding is formulated to encode the topics and relations in low dense vectors to better capture their semantic coherence. Besides, we introduce the accelerating strategy in Mv-BRT and the acquisition of event-oriented topic distributions from topic embeddings. Experimental results show that our algorithm outperforms some of the state-of-the-art algorithms in both detection accuracy and time efficiency. In addition, the event-oriented topics are rather motivated and encouraging for conveying the semantic coherence.

The contributions of our work are the following:

- We devise an event-oriented topic mining framework to uncover the topics that are oriented to the real-world events in the microblog stream. It considers the inherent relations between topics and focuses on generating coherent topics which are more compact and meaningful than LDA's latent topics.
- We construct a multi-view Bayesian rose tree to refactor the latent topics in a probabilistic and symbolic manner. It automatically generates prior knowledge for topics together with relations and seamlessly resolves the impediments for unsupervised learning.
- We regard the topic-relation patterns in mocroblog stream as the entity-relation patterns in knowledge base, and propose a hierarchical translation-based embedding method to semantically vectorize topics and relations with the prior knowledge fed by Mv-BRT. To the best of our knowledge, we are the first to use translation-based embedding method to model the distributional representation of topics and relations. What's more, our method works in an unsupervised way, no external knowledge is needed.
- We accommodate the proposed framework to the event detection task, and evaluate it with extensive experiments.

The rest of the paper is organized as follows: Section 2 reviews the related work. Section 3 presents the problem definition and preliminary work. Section 4 elucidates our unsupervised multi-view hierarchical embedding framework. Section 5 demonstrates the experimental evaluation. At last, Section 6 concludes our results.

2 RELATED WORK

2.1 Topic Detection

One line of research in topic detection is based on the NLP or data mining. The topics or events in these algorithms are all word phrases or representative documents. For example, (Shou et al. 2013) introduced a summarization-based framework for incrementally capturing the evolution of topics in the tweet stream. (Yang et al. 2012) elaborated frequent patterns mining to efficiently perform topic lifelong learning in the compressed tweet stream. Other strategies, like the segment-based method (Li et al. 2012), sequence-based method (Yang et al. 2014), sentence-based method (Wang et al. 2012), and distant-supervision-based method (Foley et al. 2015) are also capable of dealing with the TDT task.

Another line of research is the variants of LDA. By introducing the temporal information, this kind of topic models, such as TOT (Wang and McCallum 2006), are able to extract topics in real time from a massive amount of texts, and usually perform better than the simple LDA in terms of perplexity metrics. With further improvement, models such as npTOT (Dubey et al. 2013) and ADLTM (Chen et al. 2016) begin to solve the detection and tracking in a multi-aspect way. The nonparametric topic model npTOT, which is the development of TOT, allows an unbounded number of topics and flexible topic-word distributions. Although all these topics described above are widely used, the real meanings or interpretations of such topics are overlooked, leading to the dissatisfaction of the event detection. Hence all of these topic models are incoherent topic models, which fail to release event-oriented topics.

2.2 Coherent Topic Models

Due to the probabilistic nature, the LDA-based topic model is not always capable of generating coherent topics (Tang et al. 2014). To overcome this problem, coherent topic models (Chen et al. 2013) are developed to improve the coherence of the LDA's latent topics. Usually, the coherent topic models are linked to the knowledge bases (Chen and Liu 2014a,b; Lorenzetti et al. 2016), because the domain information is helpful to enhance their semantic interpretation. The coherent topics, together with their hierarchical constructions, have been extensively applied in social media user generated content analysis (Zhu et al. 2014), achieving reasonably sound performance.

Another line of research, known as focused topic models ("focused" is the same as "coherent" in the topic model context), are keen on exploiting the posterior sparsity (Graca et al. 2009) of the document-topic and topic-word distributions. Retaining the topic-word distribution in a sparse coding form, such as the STC (Zhu and Xing 2011), or both the document-topic and topic-word distribution sparsely, such as the DsparseTM (Lin et al. 2014), are effective to reduce the semantic confusion of topics.

2.3 Topic Hierarchy

The strategy of organizing documents or topics into a hierarchical structure is popularly incorporated in many applications. Intuitive examples of such works include storyline generation (Wang et al. 2015) and taxonomy construction (Yang 2015). In the probabilistic camp, HDP (Teh et al. 2004), which is the hierarchical version of LDA, aims at organizing the latent topics without the limitation of a given number of topics. To better fit the TDT task, the HDP was developed into EvoHDP (Zhang et al. 2010) to discover evolutionary patterns. However, the hierarchical construction of HDP or EvoHDP requires complicated parameter estimation, and moreover, the topic coherence is also overlooked.

In this paper, we adopt the idea of Bayesian rose tree (BRT) (Blundell et al. 2010) to construct the topic hierarchy. BRT has achieved sound performance in keywords taxonomy (Liu et al. 2012) and text mining (Wang et al. 2013) due to its high fitness and smoothness. Unlike previous studies, we need to build the BRT for latent topics in this paper, which is inherently difficult due to the deficiency of the current knowledge.

2.4 Translation-based Embedding

Translation-based embedding methods are originally used in knowledge base representation. These methods model the relations as translation operations from head entities to tail entities in factual triplets and

therefore project them into low-dimensional dense vectors. The basic translation-based method is TransE (Bordes et al. 2013), in which the entities and relations are all represented in the same vector space. TransR (Lin et al. 2015), on the other hand, addresses the diversity of entities and relations, thus represents them in two separate spaces. When considering the heterogeneity of knowledge graph, TranSparse (Ji et al. 2016) employs different degrees of sparse matrices to map entities into spaces that can better fit the relation type.

All the translation-based embedding methods mentioned above belong to supervised learning, namely, they enforce the prerequisite that a large number of training examples should be fed to the model. However, in real-world microblog stream, there is no such external ground-truth available for training. So in this paper, we design an unsupervised translation-based hierarchical embedding framework, and all of its training examples (the patterns of topics and relations) are automatically produced by our multi-view Bayesian rose tree in advance.

3 PROBLEM DEFINITION AND PRELIMINARIES

In this section, we present definitions and preliminaries necessary for introducing our framework. Section 3.1 gives definitions and problem formulation. Preliminaries of this framework are introduced in Section 3.2 and Section 3.3, including topic modeling in LDA and multi-view similarity matrices construction.

3.1 Problem Definition

This paper aims at mining the topics that are highly accordant to the events in real-world microblog stream. To solve this problem, we should first give some definitions in our algorithm.

Following the proposal of (McMinn et al. 2013), we confirm the definition of event as follows:

Definition 3.1 (Event). An event is a public and significant issue that happens at some specific time and place. Usually, event appears in the form of word phrases.

Here "public" means that this issue should arouse public concerns, not just for personal purpose, while "significant" means that this issue should be widely discussed in social media. Therefore, in our datasets, all the first level keywords, such as "*World Cup*" in Sina Weibo and "*NBA*" in Twitter, can be considered as events.

Compared with event, topic usually reflects issues in a more specific scope. As we model topics based on LDA, the topics in our cases are all multinomial distributions over word vocabularies. However, LDA-based topics are usually less coherent to express certain events. So here is given our definition of the event-oriented topic:

Definition 3.2 (Event-oriented Topic). An event-oriented topic ET_i is a multinomial distribution over word vocabulary, i.e., $ET_i = \{p(w_j)|j = 1, 2, ..., V\}$, where V is the size of the vocabulary, and this word distribution can be interpreted with an event word phrase EP_i.

Suppose there are a diversity of candidate event word phrases $EP = \{EP_1, EP_2, ..., EP_E\}$, here the declaration that the word distribution of event-oriented topic ET_i can be interpreted by an event word phrase $EP_i \in EP$ implies that the point-wise mutual information (PMI) between ET_i and its corresponding EP_i should be higher than any other pair meanwhile exceeds a minimum threshold χ. That is to say, a

topic is event-oriented if we can find an event word phrase EP_i with $EP_i = \arg\max_{EP_j \in EP} \{PMI(EP_j||ET)\}$, s.t. $PMI(EP_j||ET) \geq \chi$. Given Definition 3.2, we hence customize the definition of event-oriented topic mining in the spirit of the traditional TDT task.

Definition 3.3 (Event-oriented Topic Mining). Given a microblog feed batch D, event-oriented topic mining is to extract a series of event-oriented topics $\{ET_1, ET_2, ..., ET_k\}$, each of which is a multinational distribution over the word vocabulary.

3.2 Latent Topic Modeling by LDA

In order to construct our framework in a multi-view manner, we should first generate multi-view information about the microblog feeds. Since our framework is topic-centric, we need to uncover different views about the topics, not the texts or the words. This can be equivalently solved by utilizing the LDA (Blei et al. 2003), which delineates the relations between texts, topics, and words. In the LDA model, each feed is considered a document describing one or more certain aspects, and all the feeds of a feed batch are considered to have K aspects, namely K topics. Given a feed batch D, LDA can extract K topics from the feed batch, and these K topics are the so-called latent topics, denoted as ϕ, each of which is a multinomial distribution over the vocabulary V with Dirichlet prior β. Together with the topic-word distributions are the feed-topic distributions θ. Here θ is an $N \times K$ matrix, while ϕ is a $K \times V$ matrix, N is the number of the feeds in D and V is the size of the vocabulary.

3.3 Multi-view Similarity Matrices

In fact, the matrix of feed-topic θ and the matrix of topic-word ϕ are able to describe the latent topics from two different views: *feed relevance* and *word distribution*. Thus by implementing LDA, we automatically generate multi-view information about topics and sufficiently make the most of the current textual knowledge. In this paper, these two matrices are used for designing a multi-view hierarchical structure to refactor those latent topics.

One important preparation of multi-view hierarchical structure is to figure out two kinds of similarity matrices for both views in case of the pervasive similarity computing. For the word distribution view, topics are directly featured by words with probability distributions, then a symmetrical similarity function based on Kullback-Leibler divergence (KLD) (Peng et al. 2015) is applied to measure the similarity of two latent topics ϕ_i and ϕ_j:

$$\kappa_{ij} = \frac{1}{(KLD(\phi_i||\phi_j) + KLD(\phi_j||\phi_i))/2 + 1} \tag{1}$$

Thus, a similarity matrix $\mathbf{W}_1 \in \mathbb{R}^{K \times K}$ is constructed via Equation (1) to describe the latent topics from the view of word distributions.

For the feed relevance view, we consider using topic relevant feed collection to construct the similarity matrix. Each column $\theta_i (i = 1, 2, ..., K)$ in the feed-topic matrix reveals the related feeds collection of the latent topics ϕ_i to some extent. Given a latent topic ϕ_i and all the microblog feeds $D = \{d_1, d_2, ..., d_N\}$, if the membership of feed d_i and topic ϕ_i is larger than a given similarity threshold δ, then d_i is included in ϕ_i's related feed collection RFC_i. In this way, a feed can be allocated to different latent topics to form the topic relevant feed collection. At last, the Jaccard metric is utilized to measure the similarity between

latent topics ϕ_i and ϕ_j:

$$\eta_{ij} = Jaccard(\phi_i||\phi_j) = \frac{|RFC_i \cap RFC_j|}{|RFC_i \cup RFC_j|} \tag{2}$$

As a result, another similarity matrix $\mathbf{W}_2 \in \mathbb{R}^{K\times K}$ is constructed via Equation (2) to describe the latent topics from the view of the feed relevance. In terms of these two views, the word distribution of the latent topic is more essential because we care more about the word descriptions, thus it is treated as the dominant view.

4 UNSUPERVISED MULTI-VIEW HIERARCHICAL EMBEDDING

In this section, we present our framework to uncover event-oriented topics at each time point with multi-view information fusion and unsupervised translation-based hierarchical embedding (in Figure 1). In Section 4.1, we develop the Bayesian rose tree in a multi-view manner, where knowledge from two different views is employed to sufficiently guide its hierarchical construction, and we name our hierarchical structure as *multi-view Bayesian rose tree* (Mv-BRT). In Section 4.2, we further introduce the update and acceleration strategies in Mv-BRT to improve efficiency. With the topics and their relations modeled by Mv-BRT, we therefore propose an unsupervised translation-based embedding method in Section 4.3 to project topics and relations into low-dimensional dense vectors for better semantic measurement. By implementing spectral clustering on those topic embeddings, we progressively extract the event-oriented topics from the inscrutable latent topics in Section 4.4.

4.1 Topic Hierarchy Construction with Multi-view BRT

4.1.1 Bayesian Rose Tree. As we have derived the features of latent topics from two views, a traditional way of refactoring is to directly perform multi-view clustering or multi-view NMF, as in (Guan et al. 2015; Zheng et al. 2014). But actually, this strategy is inept in our cases. For one reason, we have no prior knowledge about the number of event-oriented topics, or the number of clusters in clustering. For another reason, the latent topics are often incoherent, thus the coherence of the clusters is always not warranted. Refactoring does not mean just clustering or partition, but the coherence improvement of the topics. To this end, we employ the Bayesian rose tree to organize the latent topics.

Bayesian rose tree is a kind of multi-branch tree that allows one-to-many relations between latent topics. Compared with hierarchical topic models, BRT builds up a topic hierarchy with higher structure fitness and lower time consumption. Moreover, it can well describe the meaning of relations between topics, and this is instrumental in the subsequent topic embedding for better representation. Usually, BRT can be constructed in a greedy manner. For a BRT at a specific time point, all the latent topics $\phi = \{\phi_1, \phi_2, ..., \phi_K\}$ are its leaves, and the objective is to organize them into a multi-branch tree to better fit the text data. At first, each topic ϕ_i is treated as an individual tree on its own, namely, $T_i = \{\phi_i\}$. Then in each iteration, two trees are selected to greedily combine a new tree T_m by one of the three following basic operations (in Figure 2):

(1) Join: $T_m = \{T_i, T_j\}$, so T_m has two branches.

(2) Absorb: $T_m = \{ch(T_i), T_j\}$, so T_m has $|T_i| + 1$ branches.

(3) Collapse: $T_m = \{ch(T_i), ch(T_j)\}$, so T_m has $|T_i| + |T_j|$ branches.

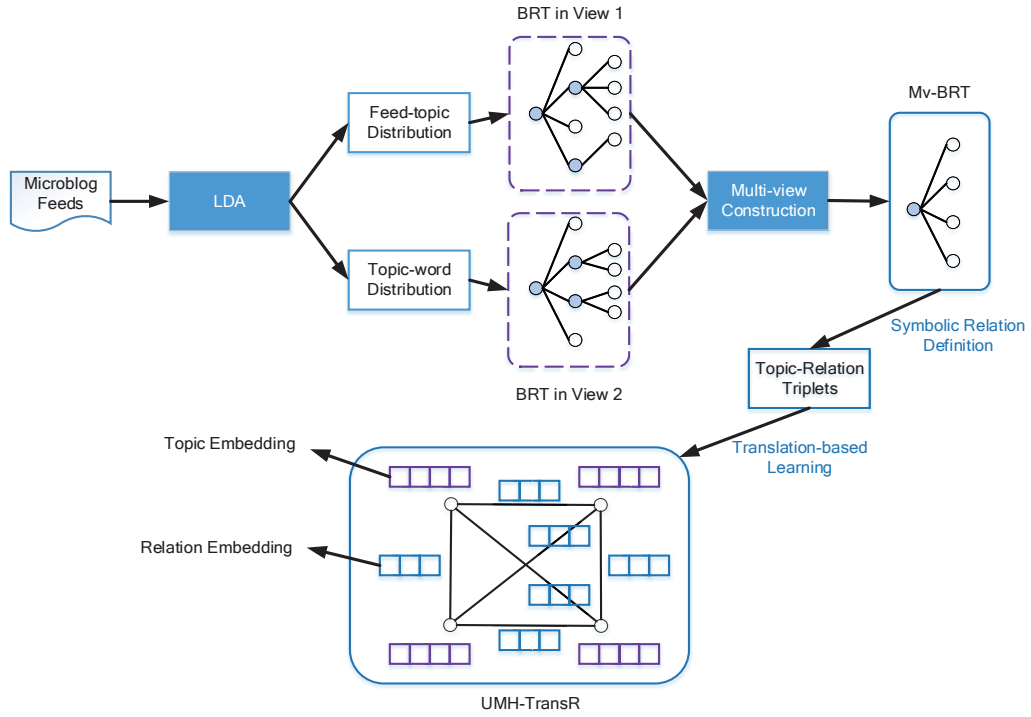


Fig. 1. Process of the unsupervised multi-view hierarchical embedding.

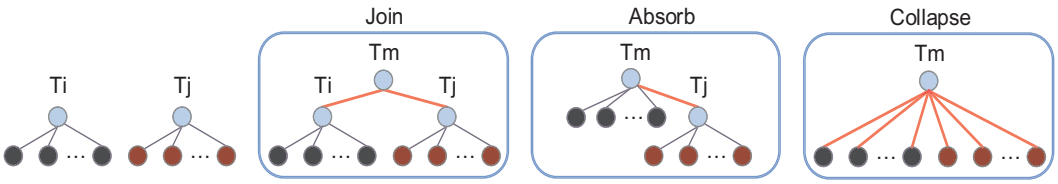


Fig. 2. Three operations of the BRT.

The combining objective is to maximize the following ratio of probability:

$$\frac{p(\phi_m|T_m)}{p(\phi_i|T_i)p(\phi_j|T_j)}, \tag{3}$$

where $\phi_m = [\phi_i, \phi_j]$ represents the latent topics under tree structure T_m, and $p(\phi_m|T_m)$ is the likelihood of ϕ_m under T_m. A Bayesian rose tree can be treated as a mixture of partitions, thus its likelihood will be the combination of all the partition likelihoods. However, in order to avoid exponential partitions, the tree likelihood is often defined in a recursive way. Previous works (Liu et al. 2015, 2012) have addressed that $p(\phi_m|T_m)$ can be calculated through dynamic programming:

$$p(\phi_m|T_m) = \pi_{T_m} f(\phi_m) + (1 - \pi_{T_m}) \prod\nolimits_{T_i \in ch(T_m)} p(\phi_i|T_i), \tag{4}$$

where $f(\phi_m)$ is the marginal probability of ϕ_m, $ch(T_m)$ is the children set of T_m, and π_{T_m} is the prior probability that all the topics in T_m are kept in the same partition.

4.1.2 Multi-view Bayesian Rose Tree. To accomplish the task of organizing topics and their relations, we design the multi-view Bayesian rose tree (Mv-BRT), a multi-view version of the basic Bayesian rose tree. Here comes the definition of our multi-view Bayesian rose tree:

Definition 4.1 (Multi-view Bayesian Rose Tree). For each view V_i, there exists a hierarchical relation set R_i for the data, then the multi-view Bayesian rose tree MT is the consensus of all the hierarchical relations under all the views, namely, $MT = C(\{R_1, R_2, ..., R_v\})$, where $C(\cdot)$ is the consensus operator and v is the number of views.

In our setting, the tree nodes are the topic distributions, rather than the document vectors in previous works (Liu et al. 2015, 2012). So when it comes to the specification of topic marginal probability $f(\phi_m)$ in Equation (4), the formulation may be quite different. In LDA, the latent topics are restricted by both the document-topic simplex and topic-word simplex. So the marginal probability $f(\phi_m)$ should also be defined in two views.

First, for the topic-word view, the topic distributions all satisfy Dirichlet-Multinomial distribution with multinomial parameter C and Dirichlet prior β, so the marginal distribution can be defined as:

$$f_{TW}(\phi_m) = \prod_{i=1}^{M} \frac{1}{\Delta(\beta + C)} \prod_{j=1}^{V} \phi_{ij}^{C_j + \beta_j - 1}, \tag{5}$$

where $\Delta(\beta + C) = \prod_{j=1}^{V} \Gamma(\beta_j + C_j) \Big/ \Gamma\left(\sum_{j=1}^{V} (\beta_j + C_j)\right)$ is the Dirichlet delta function, and M is the number of topics in ϕ_m under tree T_m.

Second, for the feed-topic view, the distribution of latent topics is not so easy to be deduced. Although it is obvious that the feed-topic distributions all satisfy Dirichlet-Multinomial distributions, it is still not clear what the topic-feed distributions are. In order to formulate the marginal probability $f(\phi_m)$ in the view of feed-topic, we must figure out the topic distributions over the feeds. From the generative process of LDA, it may be convincing that the topic-feed distributions satisfy multinomial distributions, as each word in feed is drawn by the topics with multinomial process.

To uncover the parameters of the topic-feed distributions, the feed-topic distribution should be rearranged. Each column $\theta_{\cdot,i} (i = 1, 2, ..., K)$ in the feed-topic distribution θ reveals the related feed collection of the latent topic ϕ_i. Then for a latent topic ϕ_i and all the microblog feeds $D = \{d_1, d_2, ..., d_N\}$, we can derive ϕ_i's related feeds collection RFC_i as in Section 3.3. In this way, a feed can be allocated to different latent topics to form the topic relevant feed collections. Thus the selection and allocation of feeds can be simulated by an N Bernoulli process.

For each topic ϕ_i, the probability of selecting a feed is denoted as q, then the probability of selecting all its related feeds collection RFC_i is:

$$p(X = |RFC_i|) = \begin{pmatrix} N \\ |RFC_i| \end{pmatrix} q^{|RFC_i|} (1 - q)^{N - |RFC_i|}, \tag{6}$$

where $|RFC_i|$ is the number of selected feeds. Additionally, the parameter q of this distribution can be inferred as:

$$\hat{q} = \frac{\sum_{i=1}^{M} |RFC_i|}{NM} \tag{7}$$

In real world applications, the total number of feeds N in a batch can be very large. Under this definition, the distribution of the selected cases can be approximated by the normal distribution according to the central limit theorem, that is:

$$p(X = |RFC_i|) \approx \frac{1}{\sqrt{2\pi Nq(1-q)}} \exp\left\{-\frac{(X-Nq)^2}{2Nq(1-q)}\right\} \tag{8}$$

Therefore, for the feed-topic view, the marginal distribution can be defined as:

$$f_{FT}(\phi_m) = \prod_{i=1}^{M} \frac{1}{\sqrt{2\pi Nq(1-q)}} \exp\left\{-\frac{(|RFC_i|-Nq)^2}{2Nq(1-q)}\right\} \tag{9}$$

Totally, the calculation of $p(\phi_m|T_m)$ in Equation (3) can be solved by combining both the feed-topic view and topic-word view:

$$p(\phi_m|T_m) = \pi_{T_m}\left(\sigma f_{FT}(\phi_m) + (1-\sigma) f_{TW}(\phi_m)\right) + (1-\pi_{T_m}) \prod\nolimits_{T_i \in ch(T_m)} p(\phi_i|T_i), \tag{10}$$

where σ is the weight of the feed-topic view.

In this way, we refactor the latent topics into a hierarchy in terms of both the feed-topic distributions and the topic word distributions. So we name our topic tree a *multi-view Bayesian rose tree*.

4.1.3 Multi-view Similarities Enhancement. To further improve the smoothness of the multi-view topic hierarchy, the similarity of the nodes (topics) should be considered. Obviously, there are also two kinds of similarities between the topics: the similarities from the feed-topic view and topic-word view.

For the feed-topic view, the similarity of each pair of topic can be measured by the Jaccard metric of their relevant feed collection:

$$topic\_sim1(\phi_i, \phi_j) = Jaccard(\phi_i||\phi_j) = \frac{|RFC_i \cap RFC_j|}{|RFC_i \cup RFC_j|} \tag{11}$$

For the topic-word view, the topics are in the form of word distributions, so a symmetrical similarity measure based on Kullback-Leibler divergence is directly introduced to measure the similarity:

$$topic\_sim2(\phi_i, \phi_j) = \frac{1}{(KLD(\phi_i||\phi_j) + KLD(\phi_j||\phi_i))/2 + 1} \tag{12}$$

Finally, the generation of the Mv-BRT is to maximize the following objective function with join, absorb and collapse operations:

$$\frac{p(\phi_m|T_m)}{p(\phi_i|T_i)p(\phi_j|T_j)} \cdot \left(\sigma \cdot topic\_sim1(\phi_i, \phi_j) + (1-\sigma) \cdot topic\_sim2(\phi_i, \phi_j)\right), \tag{13}$$

where the maximization of $p(\phi_m|T_m)$ is in Equation (10).

4.2 Update and Acceleration

After each operation, a new topic is formed to expound the composition of its descendant topics. In order to pertinently obtain the distribution of this combined topic, we formulate it in three cases respectively.

(1) For the join operation, where the new tree $T_m = \{T_i, T_j\}$ is directly combined by T_i and T_j, so the distribution of their ancestral topic is defined as:

$$CT_m = \frac{CT_i \cdot p(\phi_i|T_i) + CT_j \cdot p(\phi_j|T_j)}{p(\phi_i|T_i) + p(\phi_j|T_j)} \tag{14}$$

(2) For the absorb operation, where the new tree T_m is comprised of T_j and all the children of T_i, so the distribution of their ancestral topic is defined as:

$$CT_m = \frac{CT_j \cdot p(\phi_j|T_j) + \sum_{Ta \in ch(T_i)} CT_{Ta} \cdot p(\phi_{Ta}|Ta)}{p(\phi_j|T_j) + \sum_{Ta \in ch(T_i)} p(\phi_{Ta}|Ta)} \tag{15}$$

(3) For the collapse operation, where the new tree T_m is comprised of all the children of T_i and T_j, so the distribution of their ancestral topic is defined as:

$$CT_m = \frac{\sum_{Ta \in ch(T_i)} CT_{Ta} \cdot p(\phi_{Ta}|Ta) + \sum_{Tb \in ch(T_j)} CT_{Tb} \cdot p(\phi_{Tb}|Tb)}{\sum_{Ta \in ch(T_i)} p(\phi_{Ta}|Ta) + \sum_{Tb \in ch(T_j)} p(\phi_{Tb}|Tb)} \tag{16}$$

Here, CT_i is the topic-word distribution under tree T_i, and CT_j is the topic-word distribution under tree T_j respectively. Specifically, they should be distinguished from the topic set ϕ_i and ϕ_j, as the former two are the combined topics while the latter two are the whole topic sets.

Without any additional constraints, the combination of topics usually proceeds normally with the construction of the Mv-BRT. However, the generative process of Mv-BRT (in Equation (13)) is in a greedy manner, that is, before each combination, all pairs of trees should be considered as the candidates, as we would instinctively find the most probable one. In such a case, the generation of the Mv-BRT may suffer from the heavy cost. Thus the most urgent way to reduce the time consumption is to reduce the candidate pairs.

Here we incorporate the idea of K-nearest-neighbour to restrict the search space. In K-nearest-neighbour search, the K-dimensional tree (K-d tree) is usually employed to efficiently find the K nearest neighbours. However, when it comes to high dimensional data, such as the topic distributions in our case, the search of the K-d tree may degenerate to a linear strategy, leading to paltry improvement. Hence we employ the BBF algorithm (Beis and Lowe 1997) to overcome this problem, especially in the search process. BBF means *Best Bin First*, that is, all the branches are allocated with priorities, and those with higher priorities are more likely to be searched.

We adjust the BBF to our multi-view cases to perform efficient search, and we name our variant as Mv-KD-BBF. For each view, we first construct a K-d tree according to the similarity metrics described in previous sections. Then we compare the query node QN with the nodes from each K-d tree in a top-down manner. If the value of QN in current dimension is lower than the pivot node PN, then PN's left subtree is selected, otherwise, the right subtree is selected, and all of their similarity to QN will be appended to a priority queue PQ. When a leaf node LN is reached, the first round is finished and its similarity is also appended to PQ. Therefore, a back-tracking strategy from LN is started in parallel in all views, and if there exists a node that is more similar to the query node than LN, then LN is replaced by that node.

ALGORITHM 1: Nearest-Neighbour Search by Mv-KD-BFF

Input: Query node QN, K-d trees in all views, $\{KD_1, KD_2, ..., KD_v\}$.
Output: The nearest node NN for QN.
Select a dominant view and its K-d tree is KD_d;
Priority queue $PQ == NULL$;
// Priority queue construction;
repeat
  $A =$ *root of* KD_d;
  if *(*$QN.pivot \leq A.pivot$*)* **then**
    $A = A.left\_child$;
    Append A.right_child to PQ with similarity(A.right_child, QN);
  else
    $A = A.right\_child$;
    Append A.left_child to PQ with similarity(A.left_child, QN);
  end
  $NN = A$;
until *(*$ch(A) == NULL$*)*;
// Back-tracking;
for *(*each K-d tree KD_i*)* **do**
  repeat
    Select a node B from PQ with the highest priority;
    $NN = B$;
    if *(*$similarity(B.parent, QN) \geq similarity(B, QN)$*)* **then**
      $B = B.parent$;
      $NN = B$;
    end
    delete B.parent from QN;
  until $PQ == NULL$;
end

Note that back-tracking is performed in all views simultaneously, once a target node is reached in a view, then this back-tracking is completed. Together with the maximum back-tracking step, our Mv-KD-BBF can approximately find the nearest neighbors in a high dimensional space. The search of the Mv-KD-BBF is presented in Algorithm 1 with the time complexity $O((v+1) * logN)$, where v is the number of different views.

In addition to search, the reconstruction of the Mv-KD-BBF is also nontrivial. After each operation, two trees are combined to form a united tree, so the Mv-KD-BBF tree should also be reorganized accordingly. Spilltree (Liu et al. 2005) is sure to be an efficient structure for K-nearest-neighbour search, but it is not easy for update, as it needs to handle the overlapping regions. In Mv-KD-BFF, the update of the tree structure in each iteration is quite practicable, only with twice deletion and an insertion.

For the deletion of a Mv-KD-BFF node, a best fit node will be selected from all of its direct descendants. Here "best fit" means this selected node may be the rightmost node of its left subtree or the leftmost node of its right subtree. By replacement like this recursively, the deletion of the Mv-KD-BFF node is achieved

ALGORITHM 2: Node Deletion of Mv-KD-BFF

Input: Mv-KD-BFF tree T_M, node A.
Output: A new Mv-KD-BFF tree T_{MD}.
if *(A is a leaf node)* **then**
    delete A from T_M;
else
    repeat
        if *($RST(A)$!= $NULL$)* **then**
            $B = leftmost(RST(A))$; // $RST(A)$ is the right subtree of node A;
        else
            $B = rightmost(LST(A))$; // $LST(A)$ is the left subtree of node A;
        end
        $A = B$; // replace A with B;
    until *($ch(A)$ == $NULL$)*;
end

in a light-weight way. In consideration of the worst case in both left and right subtrees, the time complexity of the node deletion is $O(log2N)$. The deletion of the Mv-KD-BFF node is presented in Algorithm 2.

For the insertion of a Mv-KD-BFF node, it is quite similar to the insertion of a plain K-d tree. It iteratively compare each dimension of the inserted node to the existing nodes in a top-down search. When a leaf node is reached, the position of the insertion is also clear, thus the time complexity of the node insertion is $O(logN)$.

4.3 Topic and Relation Embedding

In this section, we will represent the topics and their relations in low-dimensional dense vectors to capture their semantic patterns. We employ the translation-based knowledge base embedding method to fulfill this task. Since topics and their relations are already produced by our Mv-BRT in Section 4.2, we can treat them as prior knowledge and take them to train our embedding model.

In order to perform the topic and relation embedding, we must first define the symbolic relations used in our method. As the relations are inherently generated by the join, absorb and collapse operations in Mv-BRT, we use those tree operations to express the symbolic meaning of the topic relations. According to the definitions of join and collapse operations, they are both symmetric relations, i.e., the left sub-tree and the right sub-tree are on equal terms. While for the absorb operation, it is asymmetric, and we therefore separate the absorb operation into left-absorb and right-absorb to posit its left sub-tree and right sub-tree with difference. Thus the symbolic relations used in our cases are *join*, *left-absorb*, *right-absorb*, and *collapse*. Once the symbolic relations are released, we can represent the Mv-BRT into plain triplets, each of which is a tuple like (*left-topic*, *relation*, *right-topic*).

Our embedding method is based on TransR (Lin et al. 2015), a translation-based model that represents topics and their relations into two different semantic spaces. It is quite reasonable to model topics and relations in distinct spaces on account of the fact that they are naturally disparate. For the topic space and relation space, there exists a space transformation, namely, a matrix that can bridge the gap between them.

The topics should be first projected into the relation space, then all the translation operations are done in this space.

In our method, we set the topic embeddings as $\mathbf{LTE} \in \mathbb{R}^g$ (left-topic), $\mathbf{RTE} \in \mathbb{R}^g$ (right-topic), and set the relation embeddings as $\mathbf{RE} \in \mathbb{R}^h$, where g is the dimension of topic embedding and h is the dimension of relation embedding. For each relation RE, we set its transformation (from topic space to relation space) matrix as $\mathbf{M}_{RE} \in \mathbb{R}^{g \times h}$. Hence, we derive the projected vectors of left-topic and right-topic:

$$\begin{aligned} \mathbf{LTE}_{RE} &= \mathbf{LTE} \cdot \mathbf{M}_{RE} \\ \mathbf{RTE}_{RE} &= \mathbf{RTE} \cdot \mathbf{M}_{RE} \end{aligned} \tag{17}$$

Then the score function is defined as:

$$f_{RE}\left(\mathbf{LTE}, \mathbf{RTE}\right) = \left\|\mathbf{LTE}_{RE} + \mathbf{RE} - \mathbf{RTE}_{RE}\right\|_2^2 \tag{18}$$

Negative sampling is helpful for training a robust model. In our cases, the negative sampling is quite different from (Lin et al. 2015), as the number of relations is limited and the number of topic pairs for each relation is quite large. During the negative sampling, We also replace the left-topic or right-topic of the correct triplet with another topic to form a wrong triplet, but two requirements should be satisfied: (1) this wrong triplet should not have appeared before, and (2) the new topic should be similar to the replaced one. Here the similarity between topics can be measured by the multi-view similarity metrics in Section 3.3. With negative sampling, we hence give our margin-based training objective:

$$L = \sum_{(LTE,RE,RTE) \in S} \sum_{(LTE*,RE,RTE*) \in S*} \max\left(0, \lambda + f_{RE}\left(LTE, RTE\right) - f_{RE}\left(LTE^*, RTE^*\right)\right), \tag{19}$$

where $(LTE, RE, RTE) \in S$ is the correct triplet set, $(LTE*, RE, RTE*) \in S*$ is the wrong triplet set, and λ is the margin.

We name our embedding method as unsupervised multi-view hierarchical TransR (UMH-TransR), because our embedding method does not need external training examples, and all the training examples are automatically generated by our Mv-BRT. The learning process of our UMH-TransR is the same as TransR, both employing stochastic gradient descent.

4.4 Event-oriented Topic Generation

With the above steps in Section 4, an unsupervised multi-view framework has been devised to represent the latent topics in salient semantic space. In this section, we utilize the latent topic embeddings to yield event-oriented topics via a self-adaptive spectral clustering. The "topics" before the generation of event-oriented topics are all referred to the latent topics unless with special explanation.

In order to cluster the topic embeddings, we first need to measure the similarity of each topic pair. Since all the topics are in the form of embeddings, we employ Euclidean distance to calculate the similarity, and derive a similarity matrix $\mathbf{ES} \in \mathbb{R}^{K \times K}$.

Another key point of the spectral clustering is how to automatically decide the number of the clusters. The decision of the cluster number is widely discussed in self-adaptive clustering. Here, we just take a simple but effective approach to estimate the cluster number. Generally, the top-N eigenvalues of the graph Laplacian of $\mathbf{ES}$ have strong discrimination ability for clustering (Kumar and DauméIII 2011). Based on

this, we design a clustering number estimation approach with following steps: 1) rank all eigenvalues of the graph Laplacian in descending order; 2) set a container with capacity $cont = \tau * evsum$, where τ is a ratio and $evsum$ is the sum of all the eigenvalues in the graph Laplacian; 3) put the eigenvalues into the container in turn until the sum of the eigenvalues in this container is larger than $cont$; 4) the number of the eigenvalues in the container is the estimated value of the clusters number.

Once the similarity metric and the cluster number are clear, the spectral clustering is in the way. Through clustering, we make a division of the topic embeddings and generate K^* clusters, each of which can represent an event-oriented topic. Since the space of topic embeddings is not the original word vocabulary, the event-oriented topics generated in this way can hardly expressed by words. In order to improve intuition and obtain topic-word like event-oriented topics, we need to copy the clustering of topic embeddings to the original latent topic-word distributions.

5 EXPERIMENTAL RESULTS

In this section, we evaluate the performance of the proposed unsupervised multi-view hierarchical embedding framework (UMHE). The experiments are conducted in four aspects: (1) Effectiveness of the event detection; (2) Semantic coherence of the event-oriented topics; (3) Fitness of the multi-view Bayesian rose tree; (4) Efficiency of the framework.

5.1 Datasets

Three microblog datasets are used in our experiments. The first is the TREC Tweets2011 dataset[5], which may be the most widely used one in microblog task. The second is the Sina Weibo dataset, which is crawled through Sina API[6]. The last is the real-time Twitter stream (Tweets2014 stream for short), which is crawled through Twitter API[7] and mainly used in our case study.

The TREC Tweet2011 dataset spans Jan.23, 2011 to Feb.8, 2011, including at least 50 events and approximately 16 million feeds (only two thirds are available). The dataset is divided into 17 feed batches, namely, one day per batch. The Sina Weibo dataset spans Jan.1, 2014 to Sep.15, 2014, including 65 events and approximately 6.6 million feeds. This dataset also generates feed batches with daily segmentation, totally 257 batches. The Tweets2014 stream spans March.1, 2014 to June.30, 2014, including an unlimited munber of events[8] and approximately 629.5 million feeds[9]. The detailed information of these three datasets is shown in Table I.

All the feed batches are pre-processed by text modeling techniques, such as word segmentation, word selection and vector space modeling. For the Chinese word segmentation, we resort to the Jieba[10]. For the word selection, only the top 500 most frequent words are reserved to model the feeds. All the algorithms are implemented in Python 3.5 with a server equipped with Intel Xeon E5-2620 v2 2.1GHz CPU, 32GB

[5]http://trec.nist.gov/data/tweets/.
[6]http://open.weibo.com/.
[7]https://developer.twitter.com/.
[8]It is difficult do discover all of the events included in such a large Twitter stream, so we manually label the detected events to identify whether it is a real-world event or just a noisy event.
[9]Here we omit the non-English tweets or short tweets, and finnaly get approximately 43.8 million feeds after pre-processing.
[10]https://github.com/fxsjy/jieba.

Table 1. Description of datasets.

	TREC Tweets2011	Sina Weibo	Tweets2014 Stream
Feeds	16 million	6.6 million	629.5 million
Batches	17	257	122
Time span	2011/01/23-2011/02/08	2014/01/01-2014/09/15	2014/03/01-2014/06/30
Events	50	65	/
Representative event	*2012 Olympics* *US Unemployment* *Barack Obama*	*World Cup* *Transformers IV* *MH 370*	*MH 370* *iPad* *Brian Adams*

RAM and a desktop with Intel Core i7-4790k 4.4GHz CPU, 24GB RAM, Nvidia GeForce GTX 960 GPU. When evaluating the time performance, all the implementations are under the same platform.

5.2 Event Detection Effectiveness

By performing our framework, a certain number of event-oriented topics can be constructed from the latent topics. Each event-oriented topic is semantic coherent and may express an event in real world.

Baselines. Eight baselines are incorporated for comparison to show the effectiveness of our framework UMHE:

(1) LDA (Blei et al. 2003), which is also the basis of our framework partially;

(2) LDA with its latent topics refactored in frequent patterns mining, denoted as LDA+FPM. Here we use the FP-growth (Han et al. 2000) to extract the frequent patterns with a simple filtering, so as to control the number of the candidate patterns;

(3) LDA with its latent topics refactored in similarity enhanced Bayesian rose tree (Blundell et al. 2010), denoted as LDA+BRT;

(4) HDP (Teh et al. 2004), which is a hierarchical version of LDA, also organize the topics in tree structure[11];

(5) KD-FF-TF (Peng et al. 2013), a feature fusion algorithm that is based on contents and user retweeting behaviors to extract high quality feeds to form social events. According to (Zhang et al. 2015), user retweeting behaviors are also important for social computing, so it is essential to include a baseline with this factor;

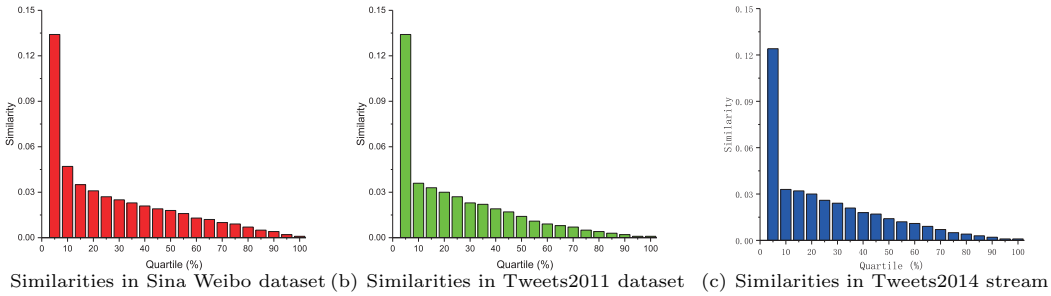
(6) Spectral Clustering (Kumar and DauméIII 2011) (denoted as SC), a simple but effective clustering algorithm, which is also designed to aggregate the texts into events;

(7) LTM (Chen and Liu 2014b), a coherent topic model that is recently proposed, also without any external knowledge;

(8) CenTM (Peng et al. 2015), a central topic model that also deals with event-oriented topics mining in microblog stream.

Parameter Setting. In LDA, HDP, CenTM, and our UMHE, the Dirichlet parameters are set as $\alpha = 0.1$ and $\beta = 0.01$, while in LTM, α is set to 1 as in (Chen and Liu 2014b).

[11] HDP is not parameterized by latent topic number K, as it will extract an unbounded number of topics according to the texts. So in our experiments, we set its K as the upper bound of its latent topic number.



(a) Similarities in Sina Weibo dataset (b) Similarities in Tweets2011 dataset (c) Similarities in Tweets2014 stream

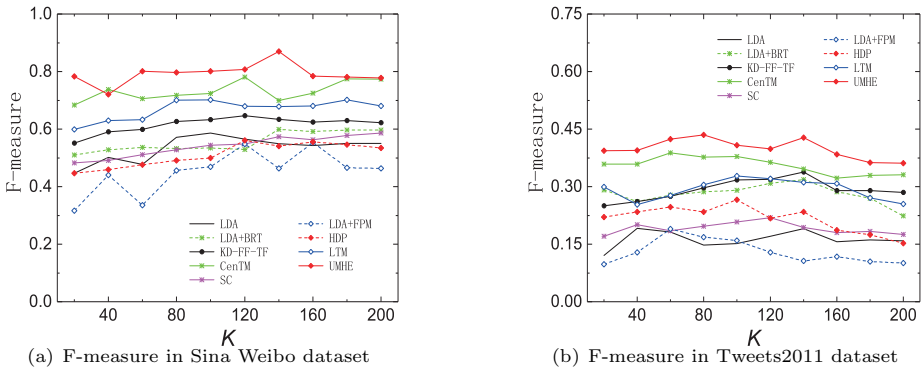
Fig. 3. Topic pair similarity distributions.

Specifically, in our UMHE, the similarity threshold δ to form topic relevant feed collection is set to 0.05 in Sina Weibo dataset, and 0.04 in Tweets2011 dataset and Tweets2014 stream. This parameter is set according to the distribution of topic pair similarities. With the statistics of ten feed batches (see in Figure 3), we can see that the top 10% similarities (of all topic pairs) are approximately above 0.047 (on average) in Sina Weibo dataset, 0.036 (on average) in Tweets2011 dataset, and 0.032 (on average) in Tweets2014 Stream.

The weight of feed-topic view σ in Mv-BRT is set to 0.2, this will be further discussed in Section 5.4. The percentage of high quality feeds selected in KD-FF-TF is set to 10%. The capacity ratio τ in spectral clustering is set to 0.6. Besides, we set the dimension of topic embedding and relation embedding to 100, 50 respectively, together with the setting of margin λ in UMH-TransR to 1.0.

Human Evaluation. It is desirable to employ human evaluation in such a subjective case. In order to make it possible for measuring the precision, recall and F-measure as used in information retrieval, we should generate a set of topical words for each event in advance. This can be solved by retaining the top-10 frequent words of all the feeds on that event. Note that in our setting, each feed is already labeled by an event, so it is feasible to do like this. We call these ten words the topical words of that event, and they are somehow regarded as the ground-truth which should be remembered by the evaluators when they are making the judgement. In our evaluation, 5 volunteers are invited to label the topics generated by every model. Hence we define the precision as the proportion of the correctly labeled topics in all the topics generated in that batch. Meanwhile, the recall can be defined as the proportion of the distinctive events identified from these topics.

Figure 4 displays the F-measure of our UMHE and eight baselines in a certain batch on both two datasets. From this figure, we can see that LDA, LDA+FPM, LDA+BRT, HDP, and spectral clustering give a rather poor performance. This is because that they consider little about the social context and are not suitable for short texts. Specifically, the LDA+FPM is the worst, as most of its topical information is missing with the frequent pattern mining. Despite of this weak point, the topical phrases generated by LDA+FPM are somehow succinct and coherent (This will be further discussed in Section 5.3). On the contrary, KD-FF-TF, LTM, CenTM and UMHE are relatively better than other four algorithms. However, the improvement of KD-FF-TF and LTM is limited, since only the characteristic of user behaviors or the co-occurrence of frequent words is taken into consideration. As for the CenTM, all of its topics are aggregated by the



(a) F-measure in Sina Weibo dataset (b) F-measure in Tweets2011 dataset

Fig. 4. F-measure of event detection with the ranging of latent topic number K.

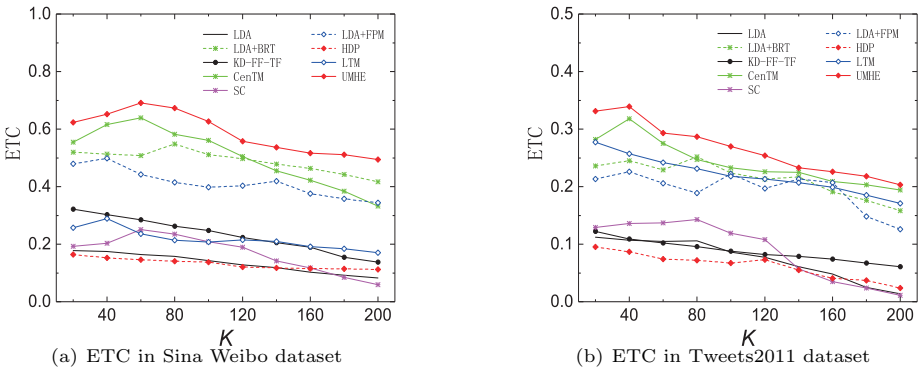
LDA's latent topics with multi-view clustering. Those central topics are often more event-oriented than the traditional topics as in LTM. What's more, our UMHE is the best among all these algorithms, both in Sina Weibo and Twitter datasets. In a global perspective, with the increasing of topic number K, the F-measure of all these nine algorithms stays nearly in a certain degree, only exhibiting a slightly declining trend on Twitter dataset. It means that in the preliminary process of our UMHE, the arbitrary setting of latent topic number K in LDA seems to have little side-effect on the final results.

Events-topics Consistency. With the increase of latent topic number K, quite a number of duplicate topics would be extracted, resulting in elusive interpretation. So it is better to generate topics that are equal to the events both in content and number. To measure the mapping result from topics to events, we adopt the events-topics consistency (ETC) as used in our previous work (Peng et al. 2015). The ETC is defined as:

$$ETC = \frac{|e \cap t|}{|e| + |t| - |e \cap t| + overlap(t) + 1}, \tag{20}$$

where e is the event set, t is the extracted topic set, and $overlap(t)$ is the number of topics that appears more than once. The average ETC results of all these methods are listed in Figure 5. Here note that KD-FF-TF and spectral clustering are non-probabilistic topic models, so there are no so-called latent topics in those algorithms, namely, without the setting of K. In order to do fair comparison with other algorithms, we enforce the K in KD-FF-TF as the selected number of high quality feeds, while in spectral clustering, K is set as the number of clusters.

Obviously, the ETCs of LDA, HDP, KD-FF-TF, LTM, and spectral clustering are not so satisfactory as the others, mainly due to the arbitrary setting of K. When the latent topics outnumbers greatly than the actual events, this kind of algorithms will suffer from heavy topic overlaps. And because of the hierarchical refinement or aggregation of latent topics, LDA+FPM, LDA+BRT, CenTM and UMHE achieve higher ETCs. The sound performance of LDA+BRT and our UMHE demonstrates that the Bayesian rose tree is considerably fit for refactoring latent topics into the event-oriented ones. Except for the LDA+FPM, the ETCs in nearly all of these algorithms decline slightly with the increase of latent topic number K. This means that the problem of overlaps would become more severe when given a larger number of latent topics.



(a) ETC in Sina Weibo dataset (b) ETC in Tweets2011 dataset

Fig. 5. Events-topics consistency with the ranging of latent topic number K.

Table 2. Map from events to topics.

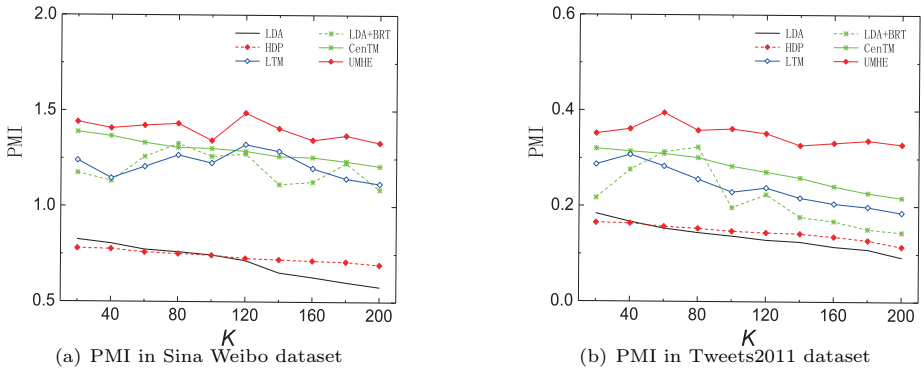
	LDA+BRT	**CenTM**	**UMHE**
Events	Topics	Topics	Topics
#1 World Cup	2	2	2
#2 Germany	2	1	1
#3 Argentina	1	1	0
#4 Netherlands	0	1	1
#5 Brazil	2	2	1
#6 CEE[a]	0	1	1
#7 Transformers IV	1	1	1
#8 Anti-corruption	0	0	1
#9 Swords of Legends[b]	1	1	1
Noise	0	0	0
Topic number	9	10	9
Total overlap	3	2	1
ETC	0.375	0.571	0.667

[a]CEE is short for college entrance examination.
[b]Swords of Legends is a popular television drama in China.

Since LDA+FPM employs frequent patterns filtering to reduce the duplicate topical phrases, their ETCs are quite stable.

Case Study of Events-topics Consistency. To intuitively understand the events-topics consistency, we present a map from events to topics. Here we choose the results of July 13, 2014 on Sina Weibo dataset, and due to the limitation of paper, we only compare our UMHE to the LDA+BRT and CenTM. In regard to the latent topic number K, we set it to 20 to achieve succinct results, as shown in Table II.

From Table II we can see that there are a total of nine real-world events on this day, and as the discussions of "*Germany*", "*Argentina*", "*Netherlands*", and "*Brazil*" are quite fervent, we separate them from the event "*World Cup*" to form other four events. When focusing on the total number of overlaps, we find that the topic overlaps of all these algorithms are controlled in a reasonable range. Apparently, our UMHE is the



(a) PMI in Sina Weibo dataset (b) PMI in Tweets2011 dataset

Fig. 6. PMI with the ranging of latent topic number K.

best, with only one duplicate topic. Compared with LDA+BRT, our UMHE is capable of discovering topics that are not so frequently occurred ("*CEE*" and "*Anti-corruption*") or easily submerged by other topics ("*Argentina*"). The result of CenTM is quite similar to our UMHE, since both of them incorporate the multi-view learning of latent topics. However, the embedding-based method (in UMHE) seems to be better at generating distinctive topics than the directly clustering-based method (in CenTM), since it considers the inherent topic relations and forms a more semantic feature representation.

5.3 Topic Coherence Evaluation

In this section, we compare the semantic coherence of the event-oriented topics generated by our UMHE with the ones generated by the baselines. Here the baselines and their parameter settings are the same as in Section 5.2 unless otherwise specified.

Automatic Evaluation with PMI. To quantitatively evaluate the topic coherence without human judgement is rather intractable. Perplexity used to measure the ability of topic model's generalization (Blei et al. 2003), that is, the fitness of predicting the unseen data. However, the perplexity prefers to measure the topic model in statistic behaviors, rather than in a semantic way (Chen and Liu 2014b). Since the coming of (Newman et al. 2010), point-wise mutual information (PMI) has become a major metric to measure the coherence of the topics. For example, (Chen and Liu 2014b) and (Röder et al. 2015) both adopt the PMI in their experiments to demonstrate the coherence of the topics. So in this paper, we also use the PMI to automatically evaluate the semantic coherence. Given a distribution-based topic ϕ, the PMI is defined as:

$$PMI\,(\phi) = \frac{2}{V(V-1)} \sum\nolimits_{1 \le i \le j \le V} \frac{p(w_i, w_j)}{p(w_i)p(w_j)}, \tag{21}$$

where V is the vocabulary size, $p(w_i, w_j)$ is the joint probability of word pair (w_i, w_j), and $p(w_i)$ is the marginal probability of w_i. In order to make it possible for computing the PMI, we regard the feed batch itself as the external test corpus. And since the topics in LDA+FPM and spectral clustering are not in the form of word distributions, we leave them out in this part. The average PMI of each algorithm is shown in Figure 6.

Table 3. Human evaluation of topic coherence.

	FIS	EDA	DWD	TWC	Final score
LDA	2.231	3.882	1.754	3.539	2.727
LDA+FPM	3.276	3.146	1.923	**4.415**	3.207
LDA+BRT	3.368	**4.112**	3.245	3.566	3.532
HDP	2.012	3.757	1.676	3.327	2.557
KD-FF-TF	3.023	3.844	2.268	3.972	3.226
SC	3.095	3.874	2.204	4.137	3.281
LTM	2.457	3.856	2.585	3.371	2.945
CenTM	3.334	3.893	3.394	3.507	3.492
UMHE	**3.784**	4.076	**3.860**	3.929	**3.887**

As shown in Figure 6, the PMI of our event-oriented topics (in UMHE) is higher than any other topics (in baselines) in most cases. Also with multi-view learning, the CenTM, however, behaves a bit poorly, and in Tweets2011 dataset, the situation is even worse. Although CenTM can provide discriminative topics, the semantic consistency of words inside the topic might not be ensured. This indicates that the semantic metric in original word vocabulary space might not work so well as in embedding space. Additionally, when comparing Figure 6(a) with Figure 6(b), we may find that the PMI in Sina Weibo dataset is higher than that in Tweets2011 dataset. This is partly because the co-occurrence of words in Twitter is not so abundant as in Sina Weibo.

Human Evaluation. Though the PMI is highly related to the coherence, it still works in a statistic fashion. Therefore, we also employ human evaluation to assess the coherence. The key of human evaluation in coherence is to settle down a series of indexes. In our experiments, we include 4 customized indexes: (1) first impression score (FIS); (2) event description ability (EDA); (2) discrimination within batch (DWD); (4) topic word consistency (TWC). The first impression is somehow very important to coherence evaluation. Presumably, when a topic, with all its words describing the same event, is displayed in front of the evaluator, it can be quickly identified as a coherent topic. While for an incoherent topic, it may take a bit longer time to go over all the words. To this end, we give the FIS a high weight, i.e., 0.4. While for the EDA, DWD, and SIF, the weights are all set to 0.2. The scores of these four indexes are all integers ranging from [0, 5]. Additionally, due to the diversity of understanding, different people may give the same topic with different scores. Thus, for each topic, the maximum score and the minimum score should be removed.

Following the human evaluation in section 5.2, we invite these five volunteers to rate the generated topics according to the four indexes above. Since all the volunteers are Chinese college students, we only evaluate the topics in Chinese Sina Weibo dataset to avoid language gap. Here we set K to 20, and the evaluation results of all the algorithms are listed in Table III. The FIS, EDA, DWD, and TWC are all displayed in average scores on the batches from July 1, 2014 to July 25, 2014.

After comparing the human evaluation results with PMI in Figure 6, we immediately find out that they are quite consistent with each other. This also confirms (Newman et al. 2010)'s assertion that PMI can be effective in measuring the coherence of topics.

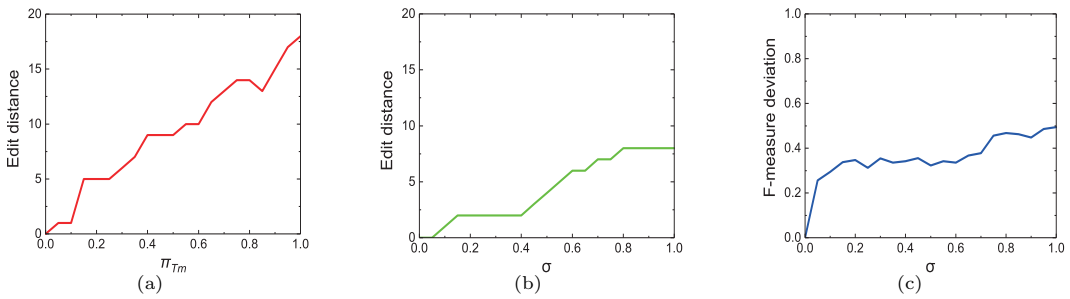


Fig. 7. Influence of different parameters.

5.4 Case Study of Hierarchical Structure

In this section, we try to evaluate the fitness of the Mv-BRT in our UMHE with other baselines in considering of the hierarchical structure. That is to say, we desire to understand the construction process of the Mv-BRT, and explain why it should organize the latent topics like this. Obviously, this is also an intractable work as there are few referential metrics. In view of this dilemma, we resort it to case study.

Parameter Analysis. Before the case study, some parameters in Mv-BRT should be well investigated. Here we present the influence of the partition granularity π_{Tm} and the weight of the feed-topic view σ. Specifically, we use the edit distance between the changed tree and the benchmark to measure the variation of trees. Meanwhile, for the σ, we also use the absolute deviation of event detection F-measure between the changed tree and the benchmark to figure out its best value.In experiments, the π_{Tm} and σ both range from 0 to 1 with step 0.05, and when focusing on one parameter, the other one is fixed. As for the benchmark, we set $\pi_{Tm} = 0$ and $\sigma = 0$.

Figure 7 presents the edit distance and F-measure deviation with the change of parameters in a batch when K is set to 100. The partition granularity π_{Tm} is important to the structure of Mv-BRT, when it changes, the topology of the tree also changes correspondingly. Different from the results in π_{Tm}, the topology of the tree is not so sensitive to σ. When σ is from 0.15 to 0.70, the results are quite consistent. So if we have no idea about which view is more important, a trial around the median may be a good choice. But here we set the weight of feed-topic view σ to 0.2, i.e., the weight of topic-word view is 0.8, as the topic-word view is usually more important. Why? Because the topic-word distributions are usually the so-called "topics".

Case Study of the Fitness in Mv-BRT. We perform case study on the Mv-BRT and the LDA+BRT generated in July 13, 2014 on Sina Weibo dataset as used in Section 5.3. Here we also set the topic number K to 20 to create a more readable tree. And for brevity, only the top-3 layers of the tree are displayed in the result.

In Figure 8, each leaf node in the tree represents a latent topic and each topic is presented with at most five significant words from its word distribution. In Mv-BRT, there are totally nine latent topics in presentation, five of them (labelled with "4", "5", "6", "7", "8") are all concerned with "*World Cup*", and these five topics can be correctly gathered together to form a more general one, resulting in the high fitness of tree structure. As we have stated that the event of "*Germany*", "*Argentina*", "*Netherlands*", and "*Brazil*"

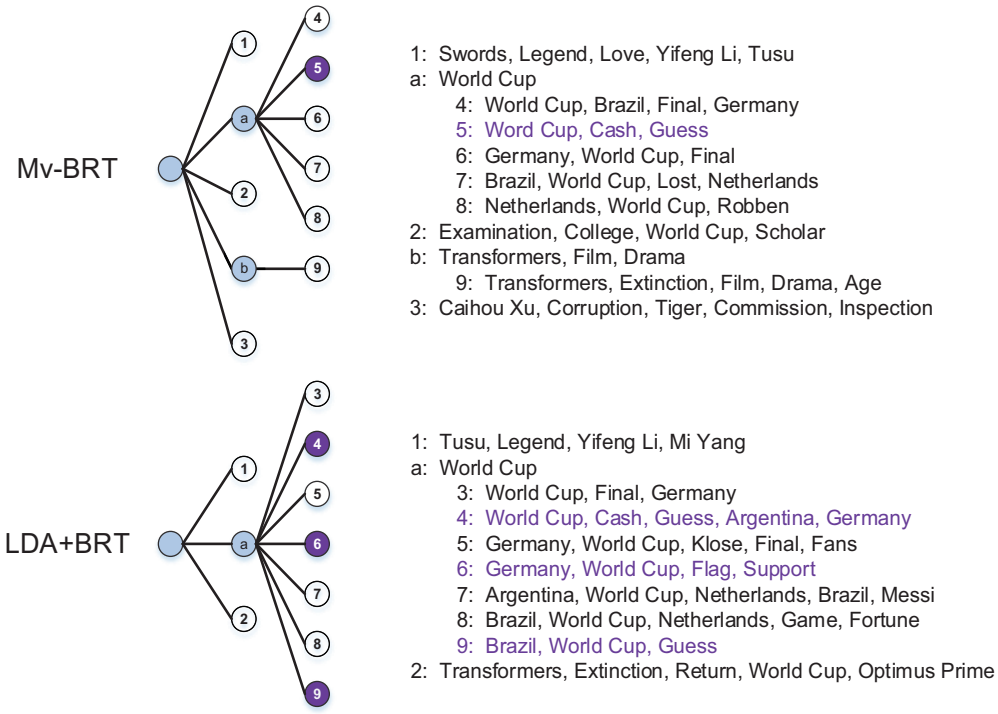


Fig. 8. Topical presentation of Mv-BRT and LDA+BRT. The nodes with numbers are the latent topics described by topical words. The nodes with letters are the inner nodes, which correspond to BRT operations. The nodes in purple represent the duplicate topics.

should be separated from the event "*World Cup*", so in Mv-BRT, the topic overlap is only 1 (the topic "5" in purple). While for the LDA+BRT, the problem of overlap is more severe, 3 topics (topic "4", "6", and "9") are identified as the duplicate ones. Besides, some events (like the "*College Entrance Examination*" labelled with topic "2" in Mv-BRT) could not be effectively detected by LDA+BRT. This is because that the multi-view learning in Mv-BRT could provide more semantic information about the events, while the traditional BRT is not so competent for this task. But anyway, the topics in both of these two algorithm are all organized correctly to compose a hierarchical structure according to their topical words. Thus, it is convincing to employ Bayesian rose tree or its variants in topic relation construction to form prior knowledge for learning embeddings.

5.5 Case Study of Tweets2014 Stream

In this section, we take a case study on the Tweets2014 stream to demonstrate the effectiveness of our UMHE when there is an unlimited number of events. Since we have no ground-truth events of this stream, we only conduct experiments on a certain batch. We select the batch of March.9, 2014, as the events in this day are overall more typical. It contains 1.471 million feeds and at least hundreds of events. When an event is released by UMHE, human evaluation will be adopted to identify whether it is a real-world social event

Table 4. Case Study on Tweets2014 Stream.

	KD-FF-TF	**SC**	**CenTM**	**UMHE**
Noisy Events @K=1000	822	745	691	673
Distinctive Events @K=1000	39	33	35	47
Top-5 Detected Events	MH 370 iPad Nadal Pray For MH370 Vote Jennette	iPad Hala Madrid The Walking Dead The 26th Taeyeon Day Austin Mahone	MH 370 iPad Pray For MH370 Fifth Harmony Heart Of A Champion	MH 370 Pray For MH370 iPad Ask Hanna For Cash Selenators
Top-5 Hashtags	*#Pray For MH370#* *#MH 370#* *#iPad#* *#Ask Hanna For Cash#* *#Heart Of A Champion#*			

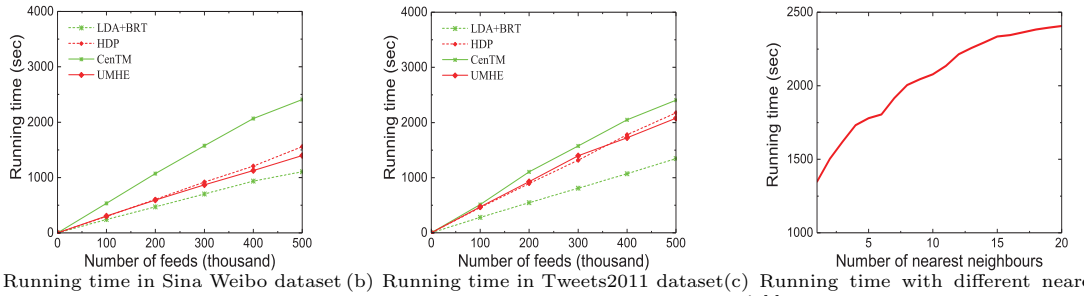
or just a noisy event. Table 4 presents the case study result of event detection on Tweets2014 stream. Here we set Topic number K as 1000 to create an opportunity for discovering abundant events.

From Table 4, we can see that the noisy events of the UMHE (673) is relatively lower than KD-FF-TF (822), SC (745) and CenTM (691). This somehow indicates that event-oriented topic models are more effective than the feature fusion methods (KD-FF-TF) and structure-based clustering methods (SC). However, due to the dynamic characteristics of real-world streams, the percentage of noisy events is extremely high, at least 60% of the extracted events are annotated as noise. Although the noisy events of these four methods are significantly different, their distinctive event numbers are quite similar. UMHE is a little better at avoiding duplicate events, but the progress is limited according to this case study. This also raises new challenges for our UMHE to detect more events from large-scale real-world microblog stream. Meanwhile, the top-5 detected events extracted by all methods cover all the most frequently discussed events in this batch. The hottest events, such as "*MH 370*", "*Pray For MH 370*", and "*iPad*", can be easily detected in all these four methods. But totally, the top-5 events detected by UMHE are most similar to the top-5 hashtags.

5.6 Running Time

Scalability is another important aspect for the event detection task, especially in real-world applications. To demonstrate the efficiency of the proposed UMHE, we compare its running time with some baselines. Here the baselines for comparison are the LDA+BRT, HDP, CenTM, as they all generate topics in a hierarchical way.

Figure 9 depicts the running time with the change of processed feed number in both datasets. Here K is set to 200 to generate an adequate number of latent topics. As is shown in Figure 9(a) and Figure 9(b), LDA+BRT is the most efficient in both datasets due to its simple implementation. Meanwhile UMHE also behaves well in terms of time efficiency, only slightly slower that the LDA+BRT, since it requires a lightweight iterative learning process in UMH-TransR. But together with the accuracy of event detection and the coherence of topic distribution, our UMHE framework is a good choice overall. Also based on LDA,



(a) Running time in Sina Weibo dataset (b) Running time in Tweets2011 dataset (c) Running time with different nearest neighbours

Fig. 9. Evaluation of running time.

HDP takes more time, since its topic inference is more complicated. As for the CenTM, its scalability is rather poor because of the two phase random walk.

To further evaluate the efficiency of Mv-BRT in our UMHE, we lay out the running time according to the number of the nearest neighbours in Mv-KD-BFF. Apparently, with the increase of the nearest heighbours in search space, the time consumption also grows slightly. And when it reaches a certain number, the running time seems to converge to a target level. This means that when the number of the nearest neighbours is large, Mv-BRT will degenerate to the plain multi-view BRT with no acceleration.

6 CONCLUSIONS

In this paper, we propose an unsupervised multi-view hierarchical embedding framework UMHE to address the problem of event-oriented topic mining in microblog stream with constraint of the only textual aspects. UMHE can precisely and efficiently aggregate the incoherent latent topics into ones with salient semantic interpretation under a translation-based hierarchical embedding method. Experimental results show that our UMHE can not only effectively identify the topics related to social events, but also drastically enhance their coherence. Meanwhile, UMHE can faithfully express topics together with their relations from hierarchical structure and automatically perform translation-based embedding, without any external knowledge or training examples. In the future, we are going to reconcile the UMHE to the applications with infinite number of views, and also study other efficient embedding strategies to improve its feature representation.

ACKNOWLEDGMENTS

This work is supported by the National Science Foundation of China, under grant No.61472291, grant No.61272110, and Natural Science Foundation of Hubei Province, China, under grant No.ZRY2014000901. Besides, the authors would like to thank Kai Zhang, Junjie Xi, Guanyin Zeng, Hongliang Shi, and Bo Li from Wuhan University for their careful evaluation.

REFERENCES

Jeffrey S. Beis and David G. Lowe. 1997. Shape indexing using approximate nearest-neighbour search in high-dimensional spaces. In *Proceedings of 13th IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'97)*. 1000–1006.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Machine Learn. Res.* 3 (2003), 993–1022.

Charles Blundell, Yee Whye Teh, and Katherine A. Heller. 2010. Bayesian rose trees. In *Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence (UAI'10)*. arXiv: 1203.3468.

Antoine Bordes, Nicolas Usunier, and Alberto Garcia-Duran. 2013. Translating embeddings for modeling multi-relational data. In *Proceedings of the 27th Annual Conference on Neural Information Processing Systems (NIPS'13)*. 2787–2795.

Yi Chang, Makoto Yamada, Antonio Ortega, and Yan Liu. 2016. Lifecycle modeling for buzz temporal pattern discovery. *ACM Trans. Knowl. Discov. Data* 11, 2 (2016), Article 20.

Yu Chen, Tom Diethe, and Peter Flach. 2016. ADLTM: A topic model for discovery of activities of daily living in a smart home. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI'16)*. 1404–1410.

Zhiyuan Chen and Bing Liu. 2014a. Mining topics in documents: standing on the shoulders of big data. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'14)*. 1116–1125.

Zhiyuan Chen and Bing Liu. 2014b. Topic modeling using topics from many domains, lifelong learning and big data. In *Proceedings of the 31st International Conference on Machine Learning (ICML'14)*. 703–711.

Zhiyuan Chen, Arjun Mukherjee, Bing Liu, Meichun Hsu, Malu Castellanos, and Riddhiman Ghosh. 2013. Discovering coherent topics using general knowledge. In *Proceedings of the 22nd ACM International Conference on Information and Knowledge Management (CIKM'13)*. 209–218.

Avinava Dubey, Ahmed Hefny, Sinead Williamson, and Eric P. Xing. 2013. A nonparametric mixture model for topic modeling over time. In *Proceedings of the 13th SIAM International Conference on Data Mining (SDM'13)*. 530–538.

John Foley, Michael Bendersky, and Vanja Josifovski. 2015. Learning to extract local events from the web. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'15)*. 423–432.

Joao V. Graca, Kuzman Ganchev, Ben Taskar, and Fernando Pereira. 2009. Posterior vs parameter sparsity in latent variable models. In *Proceedings of the 23rd Annual Conference on Neural Information Processing Systems (NIPS'09)*. 664–672.

Ziyu Guan, Lijun Zhang, Jinye Peng, and Jianping Fan. 2015. Multi-view concept learning for data representation. *IEEE Trans. Knowl. Data Eng.* 27, 11 (2015), 3016–3028.

Jiawei Han, Jian Pei, and Yiwen Yin. 2000. Mining frequent patterns without candidate generation. In *Proceedings of the 19th ACM SIGMOD International Conference on Management of Data (SIGMOD'00)*. 1–12.

Guoliang Ji, Kang Liu, Shizhu He, and Jun Zhao. 2016. Knowledge graph completion with adaptive sparse transfer matrix. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence (AAAI'16)*. 985–991.

Abhishek Kumar and Hal DauméIII. 2011. A co-training approach for multi-view spectral clustering. In *Proceedings of the 28th International Conference on Machine Learning (ICML'11)*. 393–400.

Chenliang Li, Aixin Sun, and Anwitaman Datta. 2012. Twevent: segment-based event detection from tweets. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management (CIKM'12)*. 155–164.

Shaohua Li, Tat-Seng Chua, Jun Zhu, and Chunyan Miao. 2016. Generative topic embedding: a continuous representation of documents. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL'16)*. 666–675.

Tianyi Lin, Wentao Tian, Qiaozhu Mei, and Hong Cheng. 2014. The dual-sparse topic model: mining focused topics and focused terms in short text. In *Proceedings of the 23rd International Conference on World Wide Web (WWW'14)*. 539–550.

Yankai Lin, Zhiyuan Liu, Maosong Sun, Yang Liu, and Xuan Zhu. 2015. Learning entity and relation embeddings for knowledge graph completion. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence (AAAI'15)*. 2181–2187.

Shixia Liu, Xiting Wang, Yangqiu Song, and Baining Guo. 2015. Evolutionary bayesian rose trees. *IEEE Trans. Knowl. Data Eng.* 27, 6 (2015), 1533–1546.

Ting Liu, Andrew W. Moore, Alexander Gray, and Ke Yang. 2005. An investigation of practical approximate nearest neighbor algorithms. In *Proceedings of the 19th Annual Conference on Neural Information Processing Systems (NIPS'05)*. 825–832.

Xueqing Liu, Yangqiu Song, Shixia Liu, and Haixun Wang. 2012. Automatic taxonomy construction from keywords. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'12)*. 1433–1441.

Carlos Lorenzetti, Ana Maguitman, David Leake, Filippo Menczer, and Thomas Reichherzer. 2016. Mining for topics to suggest knowledge model extensions. *ACM Trans. Knowl. Discov. Data* 11, 2 (2016), Article 23.

Tiancheng Lou, Jie Tang, John Hopcroft, Zhanpeng Fang, and Xiaowen Ding. 2013. Learning to predict reciprocity and triadic closure in social networks. *ACM Trans. Knowl. Discov. Data* 7, 2 (7 2013), Article 5.

Andrew J. McMinn, Yashar Moshfeghi, and Joemon M. Jose. 2013. Building a large-scale corpus for evaluating event detection on twitter. In *Proceedings of the 22nd ACM International Conference on Information and Knowledge Management (CIKM'13)*. 409–418.

David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. 2010. Automatic evaluation of topic coherence. In *Proceedings of the 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL'10)*.

Min Peng, Jiajia Huang, Hui Fu, Jiahui Zhu, Li Zhou, Yanxiang He, and Fei Li. 2013. High quality microblog extraction based on multiple features fusion and time-frequency transformation. In *Proceedings of the 14th International Conference on Web Information Systems Engineering (WISE'13)*. 188–201.

Min Peng, Jiahui Zhu, Xuhui Li, Jiajia Huang, Hua Wang, and Yanchun Zhang. 2015. Central topic model for event-oriented topics mining in microblog stream. In *Proceedings of the 24th ACM International Conference on Information and Knowledge Management (CIKM'15)*. 1611–1620.

Michael Röder, Andreas Both, and Alexander Hinneburg. 2015. Exploring the space of topic coherence measures. In *Proceedings of the 8th ACM International Conference on Web Search and Data Mining (WSDM'15)*. 399–408.

Lidan Shou, Zhenhua Wang, Ke Chen, and Gang Chen. 2013. Sumblr: continuous summarization of evolving tweet streams. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'13)*. 533–542.

Jian Tang, Zhaoshi Meng, XuanLong Nguyen, Qiaozhu Mei, and Ming Zhang. 2014. Understanding the limiting factors of topic modeling via posterior contraction analysis. In *Proceedings of the 31st International Conference on Machine Learning (ICML'14)*. 190–198.

Jie Tang, Limin Yao, Duo Zhang, and Jing Zhang. 2010. A combination approach to web user profiling. *ACM Trans. Knowl. Discov. Data* 5, 1 (12 2010), Article 2.

Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei. 2004. Sharing clusters among related groups: hierarchical dirichlet processes. In *Proceedings of the 18th Annual Conference on Neural Information Processing Systems (NIPS'04)*. 1385–1392.

Dingding Wang, Shenghuo Zhu, Tao Li, and Yihong Gong. 2012. Comparative document summarization via discriminative sentence selection. *ACM Trans. Knowl. Discov. Data* 6, 3 (2012), Article 12.

Xiting Wang, Shixia Liu, Yangqiu Song, and Baining Guo. 2013. Mining evolutionary multi-branch trees from text streams. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'13)*. 722–730.

Xuerui Wang and Andrew McCallum. 2006. Topics over time: a non-Markov continuous-time model of topical trends. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'06)*. 424–433.

Zhenhua Wang, Lidan Shou, Ke Chen, Gang Chen, and Sharad Mehrotra. 2015. On summarization and timeline generation for evolutionary tweet streams. *IEEE Trans. Knowl. Data Eng.* 27, 5 (2015), 1301–1315.

Hui Yang. 2015. Browsing hierarchy construction by minimum evolution. *ACM Trans. Inf. Syst.* 33, 3 (2015), Article 13.

Jaewon Yang, Julian McAuley, Jure Leskovec, Paea LePenduz, and Nigam Shah. 2014. Finding progression stages in time-evolving event sequences. In *Proceedings of the 23rd International Conference on World Wide Web (WWW'14)*. 783–794.

Xintian Yang, Amol Ghoting, and Yiye Ruan. 2012. A framework for summarizing and analyzing twitter feeds. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'12)*. 370–378.

Zhiwen Yu, Zhu Wang, Huilei He, Jilei Tian, Xinjiang Lu, and Bin Guo. 2015. Discovering information propagation patterns in microblogging services. *ACM Trans. Knowl. Discov. Data* 10, 1 (2015), Article 7.

Hao Zhang, Gunhee Kim, and Eric P. Xing. 2015. Dynamic topic modeling for monitoring market competition from online text and image data. In *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'15)*. 1425–1434.

Jianwen Zhang, Yangqiu Song, Changshui Zhang, and Shixia Liu. 2010. Evolutionary hierarchical dirichlet processes for multiple correlated time-varying corpora. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'10)*. 1079–1088.

Jing Zhang, Jie Tang, Juanzi Li, Yang Liu, and Chunxiao Xing. 2015. Who influenced you? predicting retweet via social influence locality. *ACM Trans. Knowl. Discov. Data* 9, 3 (4 2015), Article 25.

Li Zheng, Tao Li, and Chris Ding. 2014. A framework for hierarchical ensemble clustering. *ACM Trans. Knowl. Discov. Data* 9, 2 (2014), Article 9.

Jun Zhu and Eric P. Xing. 2011. Sparse topical coding. In *Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence (UAI'11)*. arXiv: 1202.3778.

Xingwei Zhu, Zhao-Yan Ming, Yu Hao, Xiaoyan Zhu, and Tat-Seng Chua. 2014. Customized organization of social media contents using focused topic hierarchy. In *Proceedings of the 23rd ACM International Conference on Information and Knowledge Management (CIKM'14)*. 1509–1518.

Received XXXXX 2017; revised XXXXX 2017; accepted XXXXX 2017