Event Early Embedding: Predicting Event Volume Dynamics at Early Stage

Zhiwei Liu[1], Yang Yang[1*], Zi Huang[2], Fumin Shen[1], Dongxiang Zhang[1], Heng Tao Shen[1],
[1]Center for Future Media and School of Computer Science and Engineering, University of Electronic Science and Technology of China [2] The University of Queensland

ABSTRACT

Social media has become one of the most credible sources for delivering messages, breaking news, as well as events. Predicting the future dynamics of an event at a very early stage is significantly valuable, e.g, helping company anticipate marketing trends before the event becomes mature. However, this prediction is non-trivial because a) social events always stay with "noise" under the same topic and b) the information obtained at its early stage is too sparse and limited to support an accurate prediction. In order to overcome these two problems, in this paper, we design an event early embedding model (EEEM) that can 1) extract social events from noise, 2) find the previous similar events, and 3) predict future dynamics of a new event. Extensive experiments conducted on a large-scale dataset of Twitter data demonstrate the capacity of our model on extract events and the promising performance of prediction by considering both volume information as well as content information.

CCS CONCEPTS

•**Information systems** → **Information retrieval;** *Web mining;* Information systems applications;

KEYWORDS

social events; volume dynamics; content information; early prediction

ACM Reference format:
Zhiwei Liu[1], Yang Yang[1*], Zi Huang[2], Fumin Shen[1], Dongxiang Zhang[1], Heng Tao Shen[1],. 2017. Event Early Embedding: Predicting Event
Volume Dynamics at Early Stage. In *Proceedings of SIGIR '17, August 07-11, 2017, Shinjuku, Tokyo, Japan,* , 4 pages.
DOI: http://dx.doi.org/10.1145/3077136.3080700

1 INTRODUCTION

Recent years have witnessed the tremendous power of social media reshaping the ways of generating, distributing and consuming information, such as breaking news, topics and events. Numerous research endeavors have been dedicated to characterizing social messages [3] and events. For instance, on Twitter[2], tweets are attached with timestamps, which can assist detecting the information flow [12] and depicting the growth and decay of certain events [7, 8].

In this paper, we study the problem of predicting about time-series volume of events. Different from majority of existing work [1, 6] that build mathematic models to fit certain types of mature events, we utilise the limited information to foresee their future volume dynamics.

To facilitate the understanding of the organization of social data, we first clarify several concepts, including topic, event, volume and social noise. If a set of messages are related to some common subject [2], we define the set of messages as a **topic**. Generally, each topic comprises various underlying constituent parts which lie in different periods. Each constituent part reveals some important aspects of the topic. Hence, we define the constituent part as **event**.

We define **volume** illustrating the total number of messages in a predefined time window. Particularly, in Twitter, topic volume denotes the number of tweets with the same hashtags published in a certain time window (e.g., daily and hourly). Similarly, the searching interest volume is another type of topic volume provided in Google Trends[3]. **Event volume** is the number of messages in a constituent part of a topic. For instance, Figure 1 shows the topic volume dynamics of *Apple* in Google Trends. As seen, this topic is composed of several events, such as *Swift for iOS*, *Apple Special Event*.

The volume of an event rises from the emergence of the event and decays to zero when the event ends. However, events are not always readily intelligible within a topic in that there also exists social noise. In our context, we define **social noise** as the ever-lasting irrelevant part to the events (analogous to white noise in signal processing).

To make a prediction of event volume, most existing work only consider the volume feature. However, the content information is also useful because events with the same content information, which implies people's attitude, often have similar dynamics. Thus, we predict the future dynamics of events at early stage with both volume and content feature. Our predicting method is based on locally linear embedding algorithm [10]. For a new coming event, we try to find its neighbors and construct its future dynamics from previous

[*] Corresponding author: Yang Yang

[2] https://twitter.com/

[3] https://www.google.com/trends/

SIGIR '17, August 07-11, 2017, Shinjuku, Tokyo, Japan
© 2017 ACM. 978-1-4503-5022-8/17/08. . . $15.00
DOI: http://dx.doi.org/10.1145/3077136.3080700

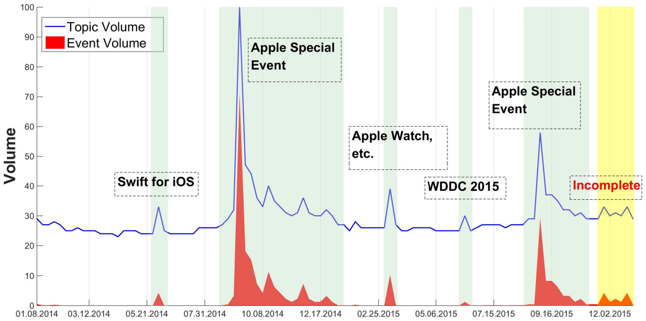


Figure 1: Search Interest for *Apple*. The blue line is the weekly topic volume of Apple, which is the hybrid of 5 events and random social noise volume. After denoising with our method, we can extract the 5 events volume shown in the red area.

events volume. The major contributions of our model are summarized as follows:

- **Early Prediction:** We propose a novel event early embedding model to predict event volume trend given limited data at a very early stage.
- **Multi-Feature Fusion:** We construct a new event using both the volume feature and content feature.
- **Novel Evaluation Metric:** We define a novel divergence function evaluating the difference of our prediction and ground truth.

The rest of the paper is organized as follows. We review the related work in Section 2. In Section 3, we present some preliminaries and the details of the proposed model. Experiments are reported and analyzed in Section 4, followed by the conclusion in Section 5.

2 RELATED WORK

In this section, we discuss some existing works about the social event prediction.

One of the most important task in event prediction is to predict the future popularity of tweets [5, 9]. The volume of event is also the value of popularity of the event though different from our definition of event. Previous work claim that modeling the collective behavior of users of a social media site allows the prediction of popularity of items from the users' early reaction [9]. After that, another generative model of predicting the final popularity of tweet is proposed [11]. Different from previous work, in this work, we study how to predict the future volume of a new event based on limited information (e.g., 24 hours). We try to predict the numerical value at every time point so that the dynamic trends of events can be observed clearly.

3 PRELIMINARY AND PROBLEM DEFINITION

In this section, we present some preliminary information of this work.

Given a set of n tweets corresponding to a certain topic $\mathbf{T}$, denoted as $\mathcal{T} = \{\mathbf{t}_i\}_{i=1}^{n}$. The i-th tweet $\mathbf{t}_i$ is represented as a triplet (tp_i, c_i, ts_i), where tp_i is the topic-word $\mathbf{t}_i$ belonging to, c_i indicates $\mathbf{t}_i$'s content and ts_i represents the timestamp. We further define $V = (V(t))|_{t=1}^{l}$ and $C = (C(t))|_{t=1}^{l}$ as the volume sequence and content sequence, respectively. Here, $V(t)$ is the number of the tweets in the topic $\mathbf{T}$ during the t-th time interval (e.g., 1 hour), $C(t)$ is the corresponding collective contents, and l is the length of $\mathbf{T}$'s life cycle.

Suppose $\mathbf{T}$ comprises of m events, denoted as $\{\mathbf{E}_j\}|_{j=1}^{m}$, where $\mathbf{E}_j = (V_j, C_j, s_j, q_j)$, $V_j = (V_j(t))|_{t=1}^{l}$ is the sequence of tweet volume of $\mathbf{E}_j$ and C_j is the collective content of $\mathbf{E}_j$. Let s_j and q_j be the start time and end time of $\mathbf{E}_j$, respectively. Here we have $1 \leq s_j < q_j \leq l$. At any time before the starting point or after the ending point, the event volume is 0. By defining the volume of social noise in the topic as $\xi = (\xi(t))|_{t=1}^{l}$, we model the topic volume as below:

$$V(t) = \sum_{j=1}^{m} V_j(t) + \xi(t), \ t = 1, 2, \ldots, l. \tag{1}$$

4 EVENT EARLY EMBEDDING MODEL

In this section, we elaborate the proposed **E**vent **E**arly **E**mbedding **M**odel (EEEM). Our model has two important parts. The first part is the collection of event corpus so than the new events can be matched with some previous events. The second part is prediction part, where the future volume dynamics of a event can be predicted by applying our event early embedding algorithm.

4.1 Social Denoising and Event Extraction

As illustrated in Figure 1, compared to the fast fluctuation of the volume of social events, the slight variation of noise volume hints us to make the assumption that the volume of social noise is an time-invariant constant. Thus, the topic volume model is simplified as

$$V(t) = \sum_{j=1}^{m} V_j(t) + \xi. \tag{2}$$

Performing integral with time-average for an infinite interval, we arrive at

$$\lim_{T\to\infty} \frac{1}{T} \int_0^T V(t)dt = \lim_{T\to\infty} \frac{1}{T} \int_0^T \sum_{j=1}^{m} V_j(t)dt + \xi. \tag{3}$$

It is observable that the sharp variation in topic volume normally corresponds to the emergence of events, which is similar to the impulse in a signal. Hence, we may conclude that most power of topic volume is from social noise. Another assumption is that under the same topic, different events do not overlap with each other in view of occurring time. Thus, as $T \to \infty$, we have

$$\lim_{T\to\infty} \frac{1}{T} \int_0^T \sum_{j=1}^{m} V_j(t)dt = \lim_{T\to\infty} \frac{1}{T} \sum_{j=1}^{m} \int_{s_j}^{q_j} V_j(t)dt. \tag{4}$$

Since in Eq. (4), the denominator is infinite while the numerator is finite, thus the value is 0. Substituting Eq. (4) into Eq. (3), we can obtain

$$\xi = \lim_{T \to \infty} \frac{1}{T} \int_0^T V(t)dt. \tag{5}$$

Now, we can extract the volume of social events by subtracting the volume of social noise from the volume of topic:

$$\sum_{j=1}^{m} V_j(t) = V(t) - \xi. \tag{6}$$

Then, we smooth the value such that all the event volume is greater than zero. Now, we can detect events as *the longest sub-sequences of consecutive non-zero members* in the sequence of the denoised topic volume $\{V(t) - \xi\}|_{j=1}^{l}$. Figure 1 presents an illustration of the denoising results on "Apple", and the red area parts is the extracted events.

4.2 Event Dynamics Prediction

In this part, we try to predict the new event volume future dynamics given only the early stage information. The basic idea is that we want to reconstruct the new events volume dynamics by a linear combination of its neighbors.

First we will use the early information to find the neighbors. However, previous events $\{\mathbf{E}_j\}|_{j=1}^{N}$ consist of the whole information of the event from the beginning to the end. Thus, in case of confusion, we separate the event early information denoting as $\{\mathbf{Ee}_j = (V_{ej}, C_{ej}, s_j, s_j + T_e)\}$, in which V_{ej} is the event early volume, C_{ej} is the event early content, and T_e is the early time duration.

Given the new event's early information $\mathbf{Ee}^{(q)}$, we represent the two types of early stage features vectors as $\mathbf{x}_v^{(q)}$ and $\mathbf{x}_c^{(q)}$. We propose the following similarity-level merge to facilitate $k$nn search:

$$S^{(e)}(\mathbf{Ee}^{(q)}, \mathbf{Ee}_j) = S^{(v)}(\mathbf{x}_{vj}^{(q)}, \mathbf{x}_v) \cdot S^{(c)}(\mathbf{x}_c^{(q)}, \mathbf{x}_{cj}), \tag{7}$$

where $S^{(e)}(\cdot)$ is the similarity factor of the new event and previous event, which is a product of $S^{(v)}(\cdot)$ and $S^{(c)}(\cdot)$, the volume similarity and content similarity respectively. $S^{(v)}(\mathbf{x}_v^{(q)}, \mathbf{x}_v)$ and $S^{(c)}(\mathbf{x}_c^{(q)}, \mathbf{x}_c)$ are defined as

$$\begin{cases} S^{(v)}(\mathbf{x}_v^{(q)}, \mathbf{x}_v) = \dfrac{\|\mathbf{x}_v^{(q)} - \mathbf{x}_v\|}{\mathbf{max}(\|\mathbf{x}_v^{(q)} - \mathbf{x}_v\|)}, \\ S^{(c)}(\mathbf{x}_c^{(q)}, \mathbf{x}_c) = \dfrac{\mathbf{x}_c^{\top}\mathbf{x}_c^{(q)}}{\|\mathbf{x}_c\|\|\mathbf{x}_c^{(q)}\|}, \end{cases} \tag{8}$$

where $\|\cdot\|$ is the ℓ_2 norm and $\mathbf{max}(\cdot)$ find the maximum value. Note that in order to make $S^{(v)}(\mathbf{x}_v^{(q)}, \mathbf{x}_v)$ and $S^{(c)}(\mathbf{x}_c^{(q)}, \mathbf{x}_c)$ comparable, we project volume feature and content feature to the scale of [0,1] by considering a maximum as denominator and cosine similarity respectively.

From the similarity of new event and previous events, we can find k neighbors which are k most similar events. To find the reconstruction coefficient vector $\mathbf{w}$, inspired by LLE algorithm [10], we try to minimize the following early reconstruction error:

$$\varepsilon(\mathbf{w}) = \frac{1}{2T_e} \left[\sum_{t=1}^{T_e} |V_e^{(q)}(t) - \sum_{j=1}^{k} \mathbf{w}_j V_{ej}(t)|^2 + \gamma ||\mathbf{w}||^2 \right] \tag{9}$$

where V_{ej} is the early volume of corresponding neighbor to $\mathbf{w}_j$ and γ is the regularization factor. The weight $\mathbf{w}_j$ summarizes the contribution of the jth event at early stage.

Finally, we construct the future volume dynamics, in particular, given a new event at its early stage, the predictive volume dynamics $V^{(q)}(t)$ of the new event is

$$V^{(q)}(t) = \sum_{j=1}^{k} \mathbf{w}_j V_j(t), \tag{10}$$

where $V^{(q)}$ is the predictive volume of the new event $\mathbf{E}^{(q)}$, $V_j(t)$ is volume of the neighbors corresponding with the weight $\mathbf{w}_i$. The underlying principle of the model (10) is that the early volume dynamics of a event is an early embedding of the future dynamics thus we learn the weights from the early stage and the future dynamics of new event can be predicted by reconstruct the dynamics from the neighbors.

5 EXPERIMENTAL EVALUATION

In this section, we evaluate the proposed EEEM for predicting the future volumes of social events at very early stage.

5.1 Data

For evaluation, we employed the Twitter dataset published by [4]. The dataset contains 10, 681, 232 tweets posted from 2013-08-01 to 2013-11-30. We regarded the trending hashtags (e.g., #iPad) as topics, which results in 18, 399 topics. We sorted all the topics in descending order of the topic volume and kept the top 5, 000 topics as our experimental data.

5.2 Event Volume Dynamics Prediction

We applied social noise reduction in 5000 topics and selected the events lasting more than 48 hours so that the event volume dynamics is long enough, which gives us 16707 samples in total. We sorted these events in ascending order of their start time, and chose the top 16507 events to form the historical event corpus and the rest latest 200 events as new events samples.

5.2.1 Evaluation Metric.

Here we define a *Divergence* $\mathcal{D}(V^*, V^g)$ to characterize the difference between our predicted sequence V^* and the true volume sequence V^g of the given event:

$$\mathcal{D}(V^*, V^g) = \frac{\text{Dist}(V^*, V^g)}{\text{Sim}(V^*, V^g)}, \tag{11}$$

where Dist$(\cdot, \cdot)$ and Sim$(\cdot, \cdot)$ are defined as

$$\begin{cases} \text{Dist}(V^*, V^g) = \displaystyle\sum_{t=1}^{l} \frac{(V^*(t) - V^g(t))^2}{\sqrt{V^g(t) + 1}}, \\ \text{Sim}(V^*, V^g) = \dfrac{\vec{V}^* \cdot \vec{V}^g}{\|\vec{V}^*\|\|\vec{V}^g\|}, \end{cases} \tag{12}$$

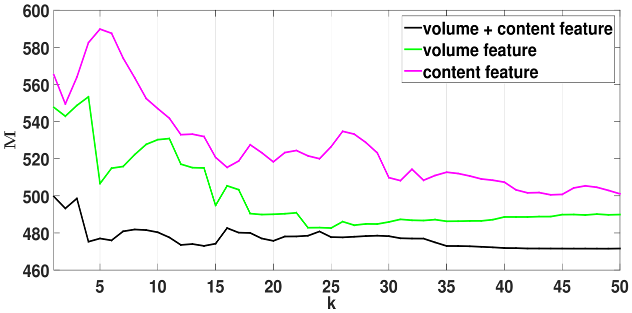


Figure 2: Measurement (sum of logarithm of divergence) of k w.r.t. three types of feature.

where $\vec{V}^*$ and $\vec{V}^g$ are the vector versions[4] of V^* and V^g, respectively.

While $\text{Dist}(\cdot,\cdot)$ measures the absolute difference between the two volume dynamics, $\text{Sim}(\cdot,\cdot)$ is the cosine similarity, which guarantees that even if the absolute distance is somehow far, but we still regard it as a better prediction since the prediction has a similar variant shape of dynamics to the real dynamics.

5.2.2 Optimization of Parameter. In this part, we optimizing our parameter based on the evaluation result. Given a test event, we use Eq. (10) to predict the volume and then exploit Eq.(11) to compute the divergence value. We utilize the sum of the logarithm of the divergence values of all the testing samples as evaluation measurement, denoted as $\mathbf{M}$, as shown in Eq.(13):

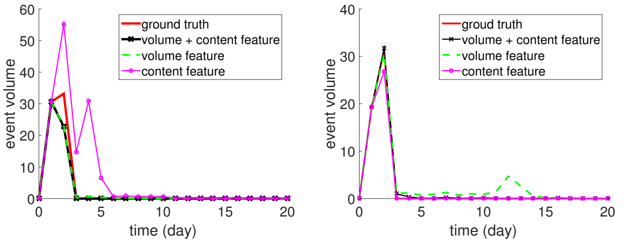
$$\mathbf{M} = \sum log(\mathcal{D}). \tag{13}$$

We set γ fixed with value 0.1. Additionally, we set the early time duration Te as 24 hours. We tune k in the range of $\{1, 2, \ldots, 50\}$. The experimental results are shown in Figure 2.

5.2.3 Individual Prediction Study. The number of neighbors is set to 35 as before and the early duration Te is set to 24. In Figure 3, We show two illustrative predictions of using differ feature. The divergence of each individual prediction is denoted as $\mathcal{D}^{(v+c)}$, $\mathcal{D}^{(v)}$ and $\mathcal{D}^{(c)}$, respectively.

6 CONCLUSION

In this work, we studied the problem of predicting event volume with limited early information. In the context of social media, we formally defined the concepts of topic, event, volume and social noise. Furthermore, we view the future dynamics as a high dimensional embedding generating from the early low dimensional event dynamics. We proposed a novel prediction model, termed event early embedding model (EEEM), to reconstruct a new event from its k neighbors based on both volume and content features. Additionally, we provide a novel evaluation metric. Extensive experiments on a large-scale Twitter dataset demonstrated the effectiveness of our methods.



(a) $\mathcal{D}^{(v+c)} = 18.5$,$\mathcal{D}^{(v)} = 18.8$, $\mathcal{D}^{(c)} = 1526$

(b) $\mathcal{D}^{(v+c)} = 1.30$,$\mathcal{D}^{(v)} = 40.3$, $\mathcal{D}^{(c)} = 3.49$

Figure 3: Fusion of feature v.s. Single feature.

ACKNOWLEDGMENTS

This work was supported in part by the National Natural Science Foundation of Chinaational Natural Science Foundation of China under Project 61572108, Project 61632007 and Project 61502081, the National Thousand-Young-Talents Program of China, and the Fundamental Research Funds for the Central Universities under Project ZYGX2014Z007, Project ZYGX2015J055 and Project ZYGX2016J080.

REFERENCES

[1] C. Bauckhage, K. Kersting, and F. Hadiji. Mathematical models of fads explain the temporal dynamics of internet memes. In *Proceedings of the ICWSM 2013*, 2013.

[2] J. Bian, Y. Yang, and T. Chua. Multimedia summarization for trending topics in microblogs. In *CIKM*, pages 1807–1812, 2013.

[3] J. Bian, Y. Yang, and T. Chua. Predicting trending messages and diffusion participants in microblogging network. In *The 37th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '14, Gold Coast , QLD, Australia - July 06 - 11, 2014*, pages 537–546, 2014.

[4] H. Cai, Z. Tang, Y. Yang, and Z. Huang. Eventeye: Monitoring evolving events from tweet streams. In *Proceedings of the ACM MM'14*, pages 747–748, 2014.

[5] H. Cai, Y. Yang, X. Li, and Z. Huang. What are popular: Exploring twitter features for event detection, tracking and visualization. In *ACM MM*, pages 89–98, 2015.

[6] X. He, M. Gao, M. Kan, Y. Liu, and K. Sugiyama. Predicting the popularity of web 2.0 items based on user comments. In *SIGIR*, pages 233–242, 2014.

[7] K. Y. Kamath and J. Caverlee. Discovering trending phrases on information streams. In *Proceedings of the 20th CIKM 2011*, pages 2245–2248, 2011.

[8] N. Kanhabua and W. Nejdl. Understanding the diversity of tweets in the time of outbreaks. In *22nd WWW '13*, pages 1335–1342, 2013.

[9] K. Lerman and T. Hogg. Using a model of social dynamics to predict popularity of news. In *Proceedings of the 19th WWW'2010*, pages 621–630, 2010.

[10] S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

[11] Q. Zhao, M. A. Erdogdu, H. Y. He, A. Rajaraman, and J. Leskovec. SEISMIC: A self-exciting point process model for predicting tweet popularity. In *Proceedings of the 21th ACM SIGKDD*, pages 1513–1522, 2015.

[12] Q. Zhao, P. Mitra, and B. Chen. Temporal and information flow based event detection from social text streams. In *Proceedings of the Twenty-Second AAAI*, pages 1501–1506, 2007.

[4] In practice, we find that $\vec{V}^*$ and $\vec{V}^g$ may have different lengths. To make them comparable, we simply expand the shorter one with value 0 to meet the length of the longer one.