SEA: A System for Event Analysis on Chinese Tweets

Yaqiong Wang[1], Hongfu Liu[1], Hao Lin[1], Junjie Wu[1*], Zhiang Wu[2], Jie Cao[2]
[1]School of Economics and Management, Beihang University, China
[2]Jiangsu Provincial Key Laboratory of E-Business, Nanjing University of Finance and Economics, China
[*]Corresponding Author: wujj@buaa.edu.cn

ABSTRACT

Recent years have witnessed the explosive growth of online social media. Weibo, a famous "Chinese Twitter", has attracted over 0.5 billion users in less than four years, with more than 1000 tweets generated in every second. These tweets are informative but very fragmented, and thus would be better archived from an *event* perspective, as done by Weibo itself in the "Micro-Topic" program. This effort, however, is yet far from satisfaction for not providing enough analytical power to events. In light of this, in this demo paper, we propose SEA, a System for Event Analysis on Chinese tweets. In general, SEA is an event-centric, multi-functional platform that conducts panoramic analysis on Weibo events from various aspects, including the semantic information of the events, the temporal and spatial trends, the public sentiments, the hidden sub-events, the key users in the event diffusion and their preferences, etc. These functions are enabled by the integration of various analytical models and by the NoSQL techniques adopted purposefully for massive tweets management. Finally, a case study on the "Spring Festival" event demonstrates the effectiveness of SEA. To our best knowledge, SEA is the first third-party system that provides panoramic analysis to Weibo events.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications—*Data Mining*; H.3.3 [**Information Search and Retrieval**]: Text Mining—*complexity measures, performance measures*; J.4 [**Social and Behavioral Sciences**]: Miscellaneous

Keywords

Chinese Tweets, Event Analysis, Public Sentiments, Weibo

1. INTRODUCTION

The development of online social networks has attracted enormous Internet users in this decade. They are becoming the mainstream online social media for information sharing.

KDD'13, August 11–14, 2013, Chicago, Illinois, USA.

Twitter (`www.twitter.com`), a microblogging site launched in 2006, has over 300 million registered users, with over 140 million posts, known as tweets, being published every day. In China, Weibo (`www.weibo.com`), a Twitter-like service launched in 2009, has accumulated more than 500 million users in less than four years. Every second, more than 1000 Chinese tweets are posted on Weibo.

The massive tweets have conveyed abundant and timely information to Weibo users, but also brought the serious information-overload problem. Moreover, the tweets are often too short and scattered on the Weibo system, which prevents users from knowing complete pictures. One way to alleviate this is to archive related tweets into *events*, as done by Weibo itself in its "Micro-Topic" program. But this is yet an initial step, and further study is in great need to enable *event analysis* on archived tweets. For instance, people would like to know the semantic information of an event, the temporal and spatial trends, and the public sentiments towards the event, from a *macro-scope view*. Moreover, some people might have further interests in the *micro-scope view* of the event, such as the key users in propagating the event and their preferences, the important tweets related to the event and their retweeting history, and the interesting sub-events hidden inside the big event.

We therefore propose SEA: a **S**ystem for **E**vent **A**nalysis on Chinese tweets to meet this challenge. In general, SEA is an event-centric, multi-functional, and big-data-oriented analytical platform. It is capable of performing panoramic analysis on Weibo events from both the macro-scope and micro-scope views. These multi-functions are based on the integration of various analytical models and on the adoption of NoSQL techniques designed purposefully for massive document management. Finally, a case study on the "Spring Festival" event demonstrates the effectiveness of SEA. To our best knowledge, SEA is the first third-party system that provides panoramic analysis to Weibo events.

We finally briefly review some related systems and research. In industry, "Twitter Analysis" (`analytics.twitter.com`) showcases the deep user information, and "Micro-Topic" (`topic.weibo.com`) collects the relevant tweets. But neither of them can perform panoramic analysis on Weibo events. In the literature, there have been a great deal of related research, such as the ones for event detection [8, 7], sentimental analysis [10, 2], and information diffusion [6, 9]. These studies, however, are still fragmented and need to be further integrated and adjusted upon practical demands on big-data analytics.

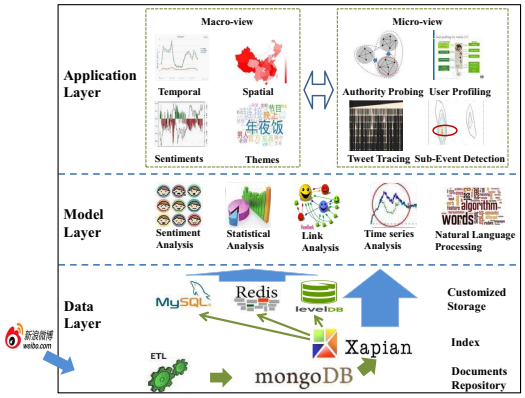

Figure 1: The framework of SEA.

2. SYSTEM OVERVIEW

In this section, we briefly introduce the framework of SEA. As shown in Fig. 1, SEA consists of three layers, namely the Application, the Model and the Data layers.

The "Application Layer" contains SEA's analytical modules for a user-specified event, with a user-friendly searchbox as the entry. In general, SEA provides two views for an event. The "Macro-View" summarizes an event using statistics in four aspects, i.e., the temporal and spatial distributions of relevant tweets, the public sentiments towards the event, and the semantic information of the event. The "Micro-View" provides a closer look at the event, by probing the whole authorities in event diffusion, by profiling them with the Area-of-Interest and Topic-of-Interest information, by exploring important tweets using vivid retweeting trees, and by detecting important sub-events with semantic descriptions. The macro- and micro-views together provide panoramic information about an event, which, in turn, reflects the event-centric and multi-functional features of SEA.

The "Model Layer" integrates various important models to support the analytical functions of SEA. For instance, *Link Analysis* is adopted for identifying key users in information diffusion and for finding the Area-of-Interest of the users. *Natural Language Processing* is frequently used whenever semantic information is needed. *Time Series Analysis* can help to identify important sub-events hidden within a big event. More details about the usage of these models will be given in Section 3 below.

The "Data Layer" is designed purposefully for handling massive tweets. Specifically, SEA adopts MongDB, a well-known NoSQL database, to store unstructured raw tweets. This enables the future expansion of SEA to the distributed environment, as more tweets keep coming in. SEA also provides two main functions for data manipulation. The first one is to extract, transform and load tweets from Weibo to the database using a parallelled ETL module. The second one is to retrieve tweets and preprocess the data to feed analytical models in an efficient manner, via the indexing structure specially designed to speed up the information retrieval. Due to the page limit, we omit the technical details of the data layer.

3. METHODS

In this section, we introduce in detail some key models employed by SEA.

3.1 Whole-Authorities Probing

It is always very important to identify the whole set of authorities in the diffusion of an event. The problem is, we can only maintain a sub-network of about 80 thousand users in SEA, which is indeed a *tiny* portion compared to the over 0.5 billion total Weibo users. To deal with this, we propose a Whole-Authorities Probing (WAP) scheme as follows.

WAP rests itself on the assumption that given an event (1) each authority should be involved in at least one popular tweet about that event (publish or retweet that tweet) , and (2) each popular tweet should be captured by at least one user stored in SEA. In this way, we can take the users in SEA as social sensors for popular tweets, and then use these tweets to probe the whole authorities even outside SEA. In detail, we take a three-step procedure as follows: *i*) Search tweets related to some event published by users in SEA; *ii*) Trace the retweeting tree for every tweet, and integrate multiple trees into a directed graph with all the stakeholders; *iii*) Identify all the authorities from the graph.

The trace of the retweeting tree in Step *ii* is actually non-trivial. To reduce the number of requests through API, we first locate the root user of the tree based on the meta-information of the tweet, and request API once for all the retweets. These retweets are then sorted in an increasing order of the publishing time, and the retweeting tree is finally constructed from the root to the leaves according to the user information led by the first "//@" mark in the title of each retweet. For the retweets missing any "//@" mark in the title, we simply link the users to the root. The integration of multiple retweeting trees in Step *ii* is straightforward by pooling the users and drawing the directed edges weighted by the sum of the retweeting times.

The generated graph might still be too huge to handle given a popular event. We employ three strategies to alleviate this. First, only the top-K most popular tweets (with highest retweeted times) will be traced in Step *ii*, which, in many cases, already cover most users. Second, only the retweets published in the first T hours will be used for building the retweeting trees, since authorities usually involve in an event in its early stage and they are the driving force for the diffusion. We can set K and T specifically according to the event context. Finally, the Hadoop framework is adopted to parallel matrix computations when using Pagerank to find authorities. Note that $K = 20\%$ and $T = 24$ were employed as the default settings in SEA.

3.2 User Profiling

User profiling is for further understanding the preferences and behavioral patterns of Weibo users, especially the authorities found above. In SEA, Weibo users are profiled in four aspects (D.B.I.P.): Demographics, Behaviors, Influences, and Preferences. Demographic information can be obtained directly via API requests. Behavioral information is concerned with the publishing history of a user, described by various statistics such as *#tweets in total*, *#tweets in a certain period*, and *the ratio of retweets to total tweets*, which generally indicate the vitality of a user in information diffusion. Influential information is characterized by features such as #followers, #followees, and the importance of followers, which could help to estimate the potential influences a user can cast by his voice. Preferential information, however, cannot be computed directly as above, and thus is the key challenge in user profiling.

We first tag a user by his *Area of Interest* (AoI). A method based on link analysis is proposed here. The basic idea is a user with an obvious AoI is more likely to follow or be followed by users with the same AoI. Along this line, we first categorize AoI into eight major classes including Culture, Education, Entertainment, Fashion, Finance, Media, Sport, and Technology. Each class i is then assigned with a *Seed Set* S_i containing 100 users with obvious AoI in class i, manually selected from the Hall of Fame maintained by Weibo. We then boost S_i into a larger *Prototypical Set* P_i by adding some followees of S_i [3]. Suppose a user f is followed by at least one user in S_i. Let $n(f, S_i)$ denote the number of *followers* of f from S_i. Then the membership score of f with respect to class i can be computed as

$$M(f, S_i) = n(f, S_i) / \sum_{i=1}^{8} n(f, S_i), \quad i = 1, \cdots, 8. \qquad (1)$$

The followees with top-200 $M(f, S_i)$ are then combined with the users in S_i to form P_i, $i = 1, \cdots, 8$. Now we can tag a user according to his outside links to the users in the prototypical sets. Given a user u, let $n(u, P_i)$ denote the number of *followees* of u in P_i, $i = 1, \cdots, 8$. Then similar to Eq. (2), the membership score of u w.r.t. class i is

$$M(u, S_i) = n(u, P_i) / \sum_{i=1}^{8} n(u, P_i), \quad i = 1, \cdots, 8. \qquad (2)$$

We finally tag u by the two AoIs with the highest membership scores.

We then tag a user by his *Topic of Interest* (ToI). The LDA model with a fast implementation [1] is adopted for this purpose. To train LDA, we collect 4 million tweets published by 0.15 million users in Feb. 2013. All these tweets are preprocessed to form the corpus, based on which a mixture of topics are finally assigned to the users.

3.3 Sentimental Analysis

The public sentiments toward an event are often the concern of parties of various interests. SEA employs an emoticon-based strategy to build a Naïve Bayes classifier for sentimental analysis.

In our previous work [10], we have discussed the reasons for modeling sentiments on emoticons, including the wide adoption of emoticons in Chinese tweets (nearly 85.5% of users used emoticons in 2011), the unavailability of authoritative Chinese dictionaries, the rapid emergence of new words, and the concept drift of some popular words. The choose of Naïve Bayes classifier is due to its simplicity and high-efficiency, which actually performs better than the Support Vector Machines in [10]. Nevertheless, we have made some progresses in SEA as follows.

First, we map the emoticons to only two types of sentiments: positive and negative, which helps to improve the mapping accuracy. In detail, we first extract all the emoticons from over 8 million tweets published by around 0.2 million users in 2011; we then filter out "cold" or ambiguous emoticons and manually map the remaining emoticons to the two sentiments; we further relate each emoticon to the semantics of the hosted tweets, to validate the mapping accuracy. As a result, we finally obtain 141 emoticons on the positive side and 96 on the negative side.

Some special treatments are as follows. First, some defects of Chinese-word segmentation tools, such as filtering out some important Chinese separate words (e.g., love, praise, cool) and keeping the irrelevant star names led by "@", are carefully addressed in SEA. Second, the captions of the emoticons are added to the feature set of the training corpus. Third, the original tweets are used instead when the retweets are blank. These treatments indeed improve the quality of the sentimental classifier in SEA.

Table 1: Classification validation.

	Precision	Recall	F-measure
Positive	0.810	0.788	0.799
Negative	0.793	0.815	0.804
Average	0.802	0.801	0.801

Experiments on 8.23 million tweets (P/N=3/1) using 10-fold cross validation demonstrate that the Naïve Bayes classifier built on the under-sampling strategy yields satisfactory results, as shown in Table 1.

3.4 Sub-Event Detection

It is not unusual that an event is driven by a series of interesting sub-events and thus has a longer life cycle. The difficulty in detecting sub-events is that one must distinguish between sub-events that could have occurred by chance and sub-events that are statistically significant. We here employ a rank-based statistic method [5] as follows.

We define a time series of length N by $T = [v_1, v_2, \cdots, v_N]$, where v_t is the value, e.g., #tweets in total, at time t. To reduce the noise interference, we first transform the quantities into the rank values. That is, we rank each point in T, denoted as r_t $(1 \le t \le N)$, from 1 to N, where 1 is assigned to the smallest value and N to the largest value. Let $Q(s, w)$ be the sum of the rank values within the window $[s, s + w - 1]$:

$$Q(s, w) = r_s + r_{s+1} + \cdots + r_{s+w-1}, \qquad (3)$$

where $1 \le s \le s + w - 1 \le N$. By changing s and w in $Q(s, w)$, we can obtain all the possible values of $Q(s, w)$. We then use Monte Carlo simulation in SEA to determine the p-value of each $Q(s, w)$, and the one with $p < \alpha$ indicates a significant sub-event, where α is the significance level. Note that the number of sampling is at least $1/\alpha\epsilon^2$[4], where ϵ is the expected error. We set $\alpha = 0.05$ and $\epsilon = 0.05$ by default in SEA.

4. CASE STUDY

In this section, we show a case study of SEA on the "2013 Spring Festival" event. To this end, we search tweets in SEA using "spring" (in Chinese). Altogether 51.9 thousand tweets were hit, which were published by 37.3 thousand users.

4.1 The Macro-View

The macro-view of SEA consists of four parts: the temporal trend, the spatial distribution, the semantics, and the sentiments of an event. These parts for the Spring Festival event are summarized in Fig. 2, from which several interesting findings can be drawn. First, the number of tweets sharply drops at Feb. 9th and does not break out of the low valley until Feb. 15th. This valley corresponds to the Spring Festival holidays exactly, which indicates that the life on Weibo steps aside in the Chinese traditional festivals. In addition, the semantics and sentiments are also closely related to the happiness theme of Spring Festival. For instance, the most staring words in the cloud are "family re-

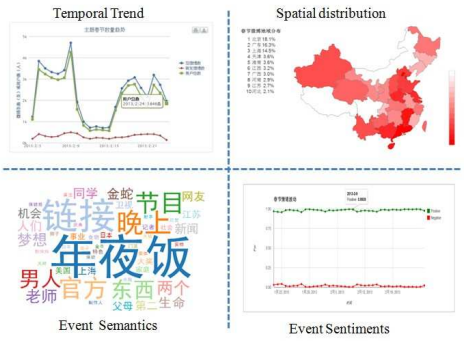

Figure 2: Macro-view by SEA.

Outside SEA	Inside SEA
李云迪YUNDI	Vista看天下
刘谦	人民日报
途牛旅游网	李开复
王旋akaWX	薛蛮子
三星手机官网	2013央视春节联欢晚会
#users: 1180573	#users: 37280

Figure 3: Sample authority-probing results.

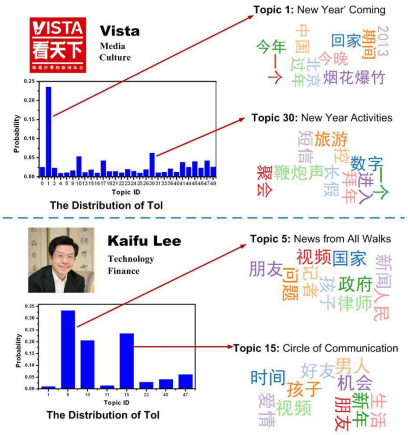

Figure 4: Sample user-profiling results.

union dinner" and "New Year's Eve entertainment", and the sentiments reflect the delighted celebration of the festival.

4.2 The Micro-View

Here, we take a deeper look at the microscopic results of SEA from four aspects: authority probing, user profiling, tweet tracing and event detection.

Authority Probing. By using WAP illustrated in Sect. 3.1, we expand the network scale from 37280 to 1217853 (1180033 outside SEA). Fig. 3 showcases several key authorities inside and outside SEA. For instance, Kaifu Lee and Manzi Xue are active opinion leaders on Weibo, and Qian Liu and Yundi Lee are partners of the magic show on CCTV Spring Festival Gala. In this way, not only the authorities inside SEA, but also those outside SEA can be detected effectively.

User Profiling. Fig. 4 shows the preferential information of two users: Vista and Kaifu Lee. AoI tags assigned to Vista are "Media" and "Culture", which are consistent to the fact that Vista is a famous news magazine in China. Also, "Technology" and "Finance" are accurate for Kaifu Lee since he used to work in IT industry and now is a venture capitalist. Moreover, the ToIs by LDA for the two users also make sense. We can observe that the ToIs of Vista ranges widely, since Vista aims to discuss various topics about "New Year". The top-2 topics of Kaifu Lee tell us that he is more interested in topics about the external environment for entrepreneurship (indicated by "Government", "Issues", etc.), and some personal understanding of life (indicated by "life", "friends", etc.).

Figure 5: Visualization of retweeting tree.

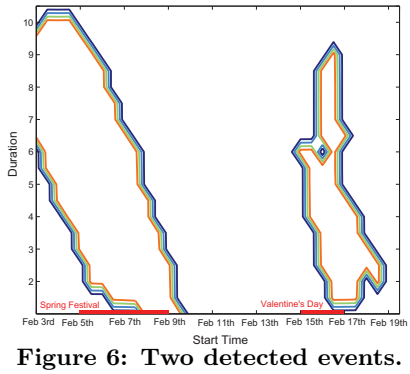

Figure 6: Two detected events.

Tweet Tracing. Fig. 5 visualizes the retweeting tree containing 1341 retweets about "CCTV Spring Festival Gala" (`http://e.weibo.com/2210168325/zirqa6MjY`). As can be seen, although the depth of this tree is 4, the majority of retweets happen in the first hop. By moving the cursor on the nodes, we can browse the basic information of retweeting users and drill down for more details about them.

Event Detection. By using the sub-event detection method introduced in Sect. 3.4, we find two prominent events in February, as shown in Fig. 6. These two events relate respectively to the two adjacent festivals in 2013, i.e., the Spring Festival (Feb. 10th) and the Valentine's Day (Feb. 14th). Note that the events might happen in the dates around the festival days. For instance, people may do shopping and send greetings before the Spring Festival, and prepare flowers and/or gifts for his/her lovers before the Valentine's Day.

5. CONCLUSIONS

In this paper, we propose a System for Event Analysis (SEA) on Chinese tweets. In general, SEA is an event-centric, multi-functional platform designed purposefully for massive tweets analysis. SEA successfully integrates the spatio-temporal, textual, and networked information extracted from tweets and users, to provide extensive analysis to Weibo events. A case study on the "Spring Festival" event demonstrates the effectiveness of SEA. To our best knowledge, SEA is the first third-party system that provides panoramic analysis to Weibo events.

6. ACKNOWLEDGE

This work was partially supported by NSFC under Grants 71171007, 70901002, and 71031001. Dr. Jie Cao's work was supported in part by NSFC under Grant 71072172.

Due to the page limit, we omit the reference list here. Full paper is at http://idec.buaa.edu.cn/weiboSEA/KDD13.pdf.