A Model for Summarizing Celebrities with Microblogging Users' Interest

Shuangyong Song\textsuperscript{1}, Dong Sang\textsuperscript{2}, Hongyun Bao\textsuperscript{3}

\textsuperscript{1}Internet Application Laboratory, Information Technology Research Department, Fujitsu R&D Center Co., Ltd., Beijing 100027, China.
\textsuperscript{2}China Institute of Forensic Science (CIFS), Beijing 100038, China
\textsuperscript{3}State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China

Abstract—With the rapid development of Web 2.0, microblogging such as twitter is increasingly becoming an important source of up-to-date topics about what is happening in the world. Specially, it becomes one of the main ways for users to understand some celebrities. However, the huge volume of information makes troubles for people to get what they really want. How to filter out needless information through numerous microblogging data and form a brief review of a celebrity become necessary for microblogging users. In this paper, we propose a novel solution for understanding a celebrity by summarizing microblogging users’ interest on him/her, and a framework is outlined. Users’ interest are generally shown on the most salient historical events of the given celebrity, and usually users’ emotion can be reflected with their microblogs contents. Therefore, the proposed model first utilizes a burst detection algorithm to extract burst periods of the time series of given celebrity’s microblogs, and then conduct emotion analysis on the microblogs in burst event periods. Experimental results show that the proposed solution can effectively summarize celebrities with microblogging users’ interest.

Keywords—Microblogging; text summarization; celebrity summarization; user interest; burst detection; user emotion analysis.

I. INTRODUCTION

Over the last few years, microblogging is increasingly becoming an important platform for users to acquire information and publish some reviews and personal status. In microblogs, information about celebrities account for a very large proportion, and microblogging has become an important way for tracking celebrities \cite{3}. Therefore, summarizing a celebrity with microblogging information can help users quickly know him/her \cite{1}. However, mining useful information from massive microblogs is a very hard task, and it has become an important method for summarizing celebrities that filtering out content with lower importance.

In \cite{4}, we try to understand a celebrity with detecting his most salient events from web news dataset, and key sentences about different salient events are used to describe him/her. This proposed method was proved to be useful for describing celebrities. However, this method contains no effective temporal burst detection technique since web news data don’t have obvious temporal characteristics, and further microblog content analysis for burst events are not conducted because news content are edited by professional news writers and they cannot reflect real web users’ emotions on given celebrities. Therefore, we improve this method for adapting it to microblogging data.

In this paper, we propose a novel model for summarizing celebrities based on detecting his/her most concern events in microblogging platform. It contains three steps: 1) for each given celebrity, we collect the related microblogs about him/her and save them into a time series, then we utilize burst detection technique to detect periods with most microblogging users’ interest; 2) we summarize microblogs in each burst period for detecting the main cause events of this burst; 3) for main cause events in each burst period, we perform a dictionary based user emotion analysis. Based on the above steps, we can get a summarization result of a celebrity’s most concerning events in different time and the user emotions on those events.

Main technical points of the proposed model include: 1) Summarizing celebrities with considering microblogging users’ interest; 2) Detecting the periods with most microblogging users’ interest with the burst detection technique; 3) detecting the main cause events of each interest burst period with text clustering technique. The rest or this paper is organized as follows: In section 2, we provide a brief review of the related work. The introduction of our model was proposed in section 3. In section 4, experimental results and some related analysis are given. Finally, we make a conclusion and discuss our plans for future work in section 5.
II. RELATED WORK

In recent years, research work related to microblogging content analysis has attracted many researchers’ interest [8]. Our work is related to the work on automatic text summarization, user interest detection and friend recommendation. In this section, we discuss those related work respectively.

Liu and Yu [1] summarize a given user with extract three kinds of relation between him/her and other named entities: person to person relationship, person to organization relationship, and person list relationship. Wu et al. proposed a model to generate personalized annotation tags for twitter users [2], which perform a simple description form of users effectively. Purohit et al. locate the problem of automatic generation of informative expertise summary or taglines for Twitter experts in space constraint imposed by UI design. In our previous work, we propose a novel solution for understanding a celebrity by summarizing his most salient historical events [4]. However, this method contains no effective temporal burst detection technique since web news data don’t have obvious temporal characteristics, and further microblog content analysis for burst events are not conducted. In this paper, we improve the method in [4] for adapting it to microblogging data, with utilizing both burst detection technique and user emotion analysis technique.

User interest detection is a useful work for user behaviour analysis. Existing studies have focused on detecting users’ interest by user actions [9], user ontology [10], user navigation patterns [11] or user voting [12, 13]. Besides, classification model [14] and topic model [15] are common methods for user interest detection [16]. In this paper, we detect users’ interest from user created microblog data through the user voting technique in [12, 13], and then create user interest based time series for given celebrity [7].

Many friend recommendation works are based on user summarization and similarity calculation between summarized users’ feature vectors. Eirinaki et al. [17] proposed a trust-aware system for user recommendations which analysed the semantics and dynamics of the implicit and explicit connections between users via a discounting factor. Lo and Lin [18] propose a new recommendation algorithm named weighted minimum-message ratio (WMR) which generates a limited, ordered and personalized friend lists by the real message interaction number among web members. Zheng et al. [19] propose a temporal-topic model to analyse users’ possible behaviours and predict their potential friends in microblogging. Based on the idea in our paper, similarity calculation between celebrities’ salient events can be used to recommend celebrities to microblogging users as friends to follow.

![Fig. 1. An overview of the research design framework.](image-url)
III. PROPOSED MODEL

A. System Architecture of the Proposed Model

In this section, we present the architectural design of our proposed celebrity summarization model for microblogging content. The overview of the system architecture for our model is shown in Figure 1, which consists of three functional modules, namely, time series statistic & burst detection module, event detection & event description module, and emotion analysis module.

In time series statistic & burst detection module, based on a given celebrity, we first collect microblogs related to him/her, including microblogs mentioning or replying him/her, and then time series of his/her microblogs can be counted, on which we can perform a burst detection algorithm to detect the most famous events of the given celebrity. A ‘state level detection’ based burst detection model is adopted to handle this step. In event detection & event description module,

In emotion analysis module, a dictionary based user emotion classification technique has been utilized to get a detailed emotion analysis result for each detected burst event, and furthermore, we apply our previous work in [20], which can detect the relation between emoticons and emotions, and then emoticons can be utilized as emotion feature, similar with emotional words in our collected dictionary. The mechanism of each functional module in our proposed model will be discussed in detail in the following sections.

B. Time Series Statistic & Burst Detection

Time Series Statistic: for each given celebrity, we collect the related microblogs about him/her, which includes his/her real name or microblog name. Then we count the each day number of those microblogs and save them into a time series which can reflect the microblogging users’ interest on this celebrity. Inspired by the idea in paper [2], we define a set of microblogs which are related to a celebrity c as:

\[ M_c = \{ m_{c1}, m_{c2}, \ldots , m_{cn} \} \]  

(1)

Where n means the number of microblogs. Then we consider the temporal aspect of the microblogging time series of a celebrity c, and define it as the following state series:

\[ S_c = [t_{c1}, t_{c2}, \ldots , t_{ct}, \ldots , t_{ctf}] \]  

(2)

Where \( S_c \) represents the time series of c, and \( t_{ci} \) means the number of microblogs published by c in the \( i^{th} \) day in our dataset. T in the definition means the total number of days in our chosen period of time.

For above mentioned time series, we utilize burst detection technique to detect periods with most microblogging users’ interest.

Burst Detection: burst detection is used to discover topics’ event-based bursty periods. For the \( M_c \) of a given c, we accordingly define the gaps in time between those microblogs as:

\[ G_c = \{ g_{c1}, g_{c2}, \ldots , g_{c(n-1)} \} \]  

(3)

Where \( g_{cj} \) means the time interval between \( m_{ci} \) and \( m_{cj+1} \). Bursts of a celebrity are described as ‘grow in intensity for a period of time, and then fade away’ [5]. We define bursty periods of celebrity c as:

\[ B_c = \{ b_{c1}, b_{c2}, \ldots , b_{cm} \} \]  

(4)

Where \( b_{cj} \) means the \( j^{th} \) detected burst period of c. For an arbitrary c, we adopt the burst detection technique proposed in [5] to obtain \( B_c \) from \( G_c \). With the number of bursty intensity states z determined, we can calculate bursty intensity states of all gaps in \( G_c \), which are between 0 and z. \( C_j(i) \) is defined to be the minimum cost of \( g_{ji} \) ending with state j, and the state of \( g_{ji} \) is defined as:

\[ g_{ji} = \text{argmin}_j C_j(i) \]  

(5)

Where the formula of \( C_j(i) \) is defined in formula (2), with initial conditions \( C_0(0) = 0 \) and \( C_j(0) = \infty \) for \( j > 0 \).

\[ C_j(i) = -\sum f_j(g_k) + \max(C_j(i-1) + r(l, j)), 0 \leq l \leq z \]  

(6)

In formula (6), \( f_j(g_k) \) is a function representing the distribution rule of \( G_c \) and \( r(l, j) \) is a function of cost incurred by moving from state l to state j. In addition, burst of intensity v is defined to be a maximal interval over which states of index v or higher persist. We define the sequence of those intervals as \( B_c \). Finally, we delete the intervals less than one day, which are unlikely to be real bursty periods, and define \( B_c \) to save the left intervals.

C. Event Detection & Event Description

Event Detection: For each burst period detected in the previous subsection, we perform the following event detection technique: 1) for each microblog, we process the Chinese word segmentation [6] and create a word vector for it; 2) we cluster microblogs in each burst period with a self-adaptive threshold based clustering method; 3) Choose the clusters with biggest number of microblogs as the main cause events of this burst period. The mentioned ‘self-adaptive threshold based clustering method’ is detailedly explained below.
We design a clustering method with self-adaptive threshold for the cause event detection task in burst periods. In different burst periods, microblogs may have different content granularity, for example, in a burst period microblogs may contain contents on sports, movies, and games, and we need to detect events on different domains, while in another burst period microblogs may just contain contents on various sports, so we need to detect events about different kinds of sports. The clustering threshold $\delta$ for a burst period in our method is defined in formula (7) as ‘average Euclidean distance value of all microblog couples in this burst period’, which is based on the ‘content distribution granularity’. In formula (7), assuming that there are $x$ microblogs in a given burst period, $x^*(x-1)/2$ means the number of ‘microblog couple’, $w$ is a weight parameter, which we empirically set to be 0.9. $V(m_i)$ and $V(m_j)$ mean the topic vectors of microblogs $m_i$ and $m_j$, and $Ed(V(m_i),V(m_j))$ means Euclidean distance between $V(m_i)$ and $V(m_j)$. After calculating of $\delta$, we start the clustering with a random microblog, with assuming that it is a cluster, and if the Euclidean distance between a new microblog and it is smaller than $\delta$, we put the new microblog into this cluster, if not, we assign the new microblog to a new cluster. Then, all the microblogs will be assigned to a steady cluster in the end.

$$
\delta = w * \frac{\sum_{i=1}^{x} \sum_{j=i+1}^{x} Ed(V(m_i),V(m_j))}{x^*(x-1)/2}
$$

(7)

**Event Description:** For each cause event, we detect keyphrases as the description of it. Furthermore we exhibit those event descriptions together to be the general description of burst cause events. The keyphrase detection technique is detailedly explained below.

Keyphrases mean words or word groups which have high degree of diffusion and influence in special period. If a phrase rarely shows in microblogs, this phrase will not be thought as a candidate. Therefore, phrases whose frequency is lower than a threshold number will not be taken as candidate keyphrases. In the processing of candidate keyphrase extraction, keyphrases are always single word nouns or two-gram nouns empirically with lots of experiments [22]. In this paper, besides all single word nouns and two-gram nouns, we also extract trigram and four-gram nouns from part-of-speech results, and statistic them for ranking with frequency.

Then we process the ‘substring problem’, e.g. a trigram noun is the substring of a four-gram noun or a single word noun is the substring of a two-gram noun, with a rule of substring merging similar with the method in [13]: in formula (1), $T_{\text{length}}$ is the length of a phrase, which is the number of characters in it, and $T_{\text{frequency}}$ is the frequency of this phrase. $T_{\text{value}}$ means the ‘value for keeping’ of a phrase, which is determined by the above two factors $T_{\text{length}}$ and $T_{\text{frequency}}$.

$$
T_{\text{value}} = T_{\text{frequency}} \times T_{\text{length}}
$$

(8)

If a ‘substring noun’ has a larger $T_{\text{value}}$ than the longer noun which contains it, we keep the substring noun. Otherwise, we keep the longer noun and delete the substring noun. Finally, all the saved phrases whose frequencies are larger than a threshold number will be taken as candidate keyphrases, which in this paper is set to be 5 empirically.

**D. Emotion Analysis**

There are some different methods for user emotion analysis, and we can roughly divide them into content based analysis and emoticon based analysis. For content based analysis, many studies have made contributions, such as research on dictionary based approaches, statistical corpus based approaches and machine learning approaches [23]. For emoticon based analysis, in our previous work [20] we proposed a modified topic-supervised biterm topic model to detect relationship between emoticons and emotions. In this paper, we consider content based analysis and emoticon based analysis together, for completely and accurately detect users’ emotion expression.

<table>
<thead>
<tr>
<th>Emotions</th>
<th>Chinese emotion words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Esteem</td>
<td>喜悦(jingpei), 敬仰(jingyang), 敬重(jingzhong)</td>
</tr>
<tr>
<td>Sadness</td>
<td>哭(ku), 悲痛(beitong), 叹息(tanxi), 悲哀(beiai)</td>
</tr>
<tr>
<td>Sympathy</td>
<td>同情(tongqing), 可怜(kelian), 可惜(kexi)</td>
</tr>
<tr>
<td>Happiness</td>
<td>惊喜(jingxi), 高兴(gaoxing), 欣幸(xinxing)</td>
</tr>
<tr>
<td>Anger</td>
<td>愤恨(fenhai), 愤慨(fenkai), 愤怒(fennu)</td>
</tr>
<tr>
<td>Fear</td>
<td>恐惧(youju), 恐怕(haipai), 胆怯(danqie)</td>
</tr>
<tr>
<td>Inferiority</td>
<td>自卑(zibei), 自惭形秽(zijianxinghui), 自侮(zinei)</td>
</tr>
<tr>
<td>Surprise</td>
<td>惊奇(jingqi), 吃惊(chijing), 惊讶(jingya), 震惊(zhenjing)</td>
</tr>
</tbody>
</table>
Emotional words are employed in many emotion classification tasks. There are two main approaches in order to compile or collect the emotional word list. Manual approach is very time consuming and it is not used alone. It is usually combined with the other automated approach as a final check to avoid the mistakes that resulted from automated methods. In this paper, we utilize a dictionary-based approach to realize emotion analysis task. A small set of emotional words is collected manually with known orientations. Then, this set is grown by searching in the well known corpora WordNet for their synonyms and antonyms. The newly found words are added to the seed list then the next iteration starts. The iterative process stops when no new words are found. After the process is completed, manual inspection can be carried out to remove or correct errors. Some examples of emotional words in emotion dictionary are given in TABLE I. Finally, we search arisen emotional words in microblogging contents and classify them into corresponding emotional categories based on this dictionary.

### TABLE II

**Examples of Emoticons on Sina Weibo**

<table>
<thead>
<tr>
<th>Emoticon</th>
<th>Meaning</th>
<th>Emoticon</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>😊</td>
<td>[laugh]</td>
<td>😞</td>
<td>[sad]</td>
</tr>
<tr>
<td>🧡</td>
<td>[handshaking]</td>
<td>👍</td>
<td>[candle]</td>
</tr>
<tr>
<td>❤️</td>
<td>[heart]</td>
<td>😷</td>
<td>[scold]</td>
</tr>
<tr>
<td>🦄</td>
<td>[sick]</td>
<td>😴</td>
<td>[yawn]</td>
</tr>
<tr>
<td>😍</td>
<td>[lovely]</td>
<td>😊</td>
<td>[OK]</td>
</tr>
</tbody>
</table>

Emoticon is a portmanteau word formed from 'emotion' and 'icon'. As new visual or nonverbal communication cues used in digital interaction, emoticons can realize pictorial representation of authors’ facial expressions or behavioural states, which contain rich emotional information of them and make important effects on web user emotion analysis. *Sina Weibo* is a well-known Chinese micro-blogging web service offering a fun and interactive way to discover and discuss information, and several examples of emoticons on *Sina Weibo* and their intuitive meanings are given in TABLE II. We utilize the result in our previous work on ‘emoticon-emotion’ relation detection [20] to detect microblogging users’ emotion expression as additional information for above dictionary based approach.

In [20], a modified topic-supervised biterm topic model was proposed to detect probabilistic relation between emoticons and emotions, some examples are given in TABLE III. Emoticons are ranked by classification probability for each emotion. We can see that there are some emoticons show in different emotional classifications, such as [fist], [jostle] and [lovely]. Actually, [fist] has some probability to express esteem and has some probability to express fear, so are the other two emoticons.

### TABLE III

**‘Emotion-Emoticon’ Relationship Examples**

<table>
<thead>
<tr>
<th>Emotions</th>
<th>Emoticons</th>
</tr>
</thead>
<tbody>
<tr>
<td>Esteem</td>
<td>[lovely], [fist], [breeze], [jostle], [adore]</td>
</tr>
<tr>
<td>Sadness</td>
<td>[scold], [anger], [shocked], [wave], [orz]</td>
</tr>
<tr>
<td>Sympathy</td>
<td>[sad], [candle], [handshaking], [scared], [sick]</td>
</tr>
<tr>
<td>Happiness</td>
<td>[rabbit], [nothing], [OK], [powerful], [lovely]</td>
</tr>
<tr>
<td>Anger</td>
<td>[rainy], [mania], [evil laugh], [interrogative]</td>
</tr>
<tr>
<td>Fear</td>
<td>[footprint], [grieved], [fist], [black line]</td>
</tr>
<tr>
<td>Inferiority</td>
<td>[humph], [fall asleep], [cry], [grimace]</td>
</tr>
<tr>
<td>Surprise</td>
<td>[what], [jostle], [not care], [tragedy], [Altman]</td>
</tr>
</tbody>
</table>

For each detected burst period, we detect the existing emotional words and emoticons, and count them for analyse the emotional distribution of microblogging users. From user emotion analysis results of different events, we can get the general emotional distribution of users on the given celebrity, which is helpful for the reputation building and crisis management.

### IV. EXPERIMENTAL RESULTS AND DISCUSSION

We conduct preliminary case-study experiments of user interest based celebrity summarization, for roughly evaluate the performance of the proposed model.

#### A. Dataset and Experimental Design

We utilize ‘microblog search function’ in *Sina Weibo* to collect microblogs about two celebrities: 1) Xiang Liu, one of China’s most commercially successful athletes and has emerged as a cultural icon, whose 2004 Olympic gold medal was the first in a men’s track and field event for China; 2) Taylor Swift, an American singer-songwriter and actress, who signed with the independent label Big Machine Records and became the youngest songwriter ever hired by the Sony/ATV Music publishing house.
Finally, we collect related microblogs to those two celebrities from April 01, 2015 to June 30, 2015, which contain 13,501 microblogs and 150,312 microblogs respectively.

B. Experimental Results and Discussions

Fig. 2. The $F$-value results of different values of $v$ for our model on user interest burst detection.

In [4], local highest points of web news statistical curve are detected and 31 day long period around each local highest point is defined as a bursty period. This burst detection method has obvious shortcoming and in this paper we replace it with automatic method in [5]. We still set the number of bursty intensity states $z$ as 8, which was used in the original Burst Detection paper, and what we need to additionally set is the burst of intensity $v$. For detecting the optimal $v$, we use $F$-value as an indicator. For burst detection results with each $v$, we compare them with manually tagging results to get Precision and Recall value, and then further obtain $F$-value. In particular, Precision and Recall means day-level calculated results. Figure 2 shows the $F$-value results when varying $v$ as eight different integer values from 1 to 8. As shown, the curve peaks at $v=3$, which indicates that it can achieve the best performance. Therefore, we set $v=3$ in the following experiments.

Figure 3 and Figure 4 show the celebrity summarization results. Compared to results in [4], detected burst periods have more flexible length, which can reflect the changes of users’ interests more accurately. Besides, more than one events can be detected as cause events of a same burst period [21]. Furthermore, user emotion analysis result can enhance users’ understanding of celebrities.
V. CONCLUSION AND FUTURE WORK

This paper proposes a novel method to extract and summarize the most salient events of a celebrity from Chinese News corpus. With this method, we first extract keywords, which describe an event, and then rank the sentences and remove redundant sentences according to these keywords. The experimental results show that our summary can concisely and accurately describe a celebrity. Innovations of this paper are given below: 1) we propose a method to detect a celebrity’s most concerning events based on microblogging users’ interest burst detection; 2) we propose a method to detect the main cause events of microblogging users’ each interest burst.

Currently, this system works independently out of any search engine. It is our intention to integrate it with a search engine so that it can work in real time. Based on this work, we are going to address the issue of how to find the associated rules between events, and event prediction is also a key point of our future research work.

Acknowledgments

This research is supported by the NSFC project 61501463, the Youth Innovation Promotion Association of the Chinese Academy of Sciences, and the Open Project of Key Laboratory of Big Data Mining and Knowledge Management, Chinese Academy of Sciences.

REFERENCES


1 http://wordnet.princeton.edu/
2 http://weibo.com
3 http://js.weibo.com/