Understanding a Celebrity with His Salient Events

Shuangyong Song, Qiudan Li, and Nan Zheng

Laboratory of Complex Systems and Intelligence Science
Institute of Automation, Chinese Academy of Sciences
95 Zhongguancun East Road, Beijing, China 100190
{shuangyong.song,qiudan.li,nan.zheng}@ia.ac.cn

Abstract. Internet has become a resourceful platform for people to collect information. Specially, it becomes one of the main ways to understand a celebrity. However, the huge volume of information makes troubles for people to get what they really want. How to filter out needless information through numerous data and form a brief review of a celebrity become necessary for people to understand the person. In this paper, we propose a novel solution for understanding a celebrity by summarizing his most salient historical events, and a framework is outlined. The framework contains three main components: attention tracking, event mining from News, and event summarization. First, with the comparison of users' attention and media attention on a celebrity, News corpus is proved to be able to represent the users' attention. Second, keywords are extracted from the News according to different time periods for choosing summary sentences. Third, a final event description of the celebrity will be given. Finally, we will show the user interface of our system. Our experimental results show that the proposed solution can effectively process the news corpus and provide us with accurate description of the celebrity.

1 Introduction

With the development of the Web 2.0, people can easily find abundant information about a celebrity, but the complexity and redundancy of the information make it difficult to obtain the most necessary content. For example, about 152,000,000 web pages are returned to a user when he queries 'Obama' in Google.com. Those web pages may belong to different online Medias (i.e. News, Blog, Micro-Blog, etc.) and have various types of text formats. So it is hard for people to classify them by different topics or by different periods. Describing a celebrity with his most salient historical events, referred as character description, becomes important for users to fastly and conveniently understand the celebrity. A summarized text to present event evolution is necessary for general users to review events about a celebrity.

By analyzing the salient historical events about a celebrity and summarizing these events on a timeline with appropriate sentences, we are able to understand the celebrity's life facilely. Chieu and Lee proposed a query based event extraction model to summary events about the query along a timeline [4]. In this model, they rank the sentences which are queried from a news corpus with interesting and bursty feature to represent different events about the query. However, those sentences are too brief to describe events in detail, and some sentences have content relationship with each

A. An et al. (Eds.): AMT 2010, LNCS 6335, pp. 86–97, 2010.

other which sometimes makes the defined 'events' indistinguishable. On the other hand, Platakis et al. proposed a novel method of event summarization, in which an event was defined as a group of correlated terms [6]. They detected the frequent terms with similar temporal dynamics in a given period, and describe the events in this period with those terms. Through this method, events can be extracted more accurately, and some implicit association among those terms can be found. However, the result of this method is also oversimplified, and without a timeline, users could not fully understand the events of a celebrity.

In this paper, we design a new summarization method to describe a celebrity with his salient historical events. First, with the comparison of users' attention and media attention on a celebrity, News corpus is proved to be able to represent users' attention. Therefore, we track users' attention of a celebrity, which is based on the amount of News relevant to him, and detect his salient events by finding bursts in the stream of News corpus. Second, keywords with similar temporal dynamic are extracted from the News corpus according to different burst time periods, which can make the summary of each event more appropriate. Third, we extract sentences which contain the keywords we have obtained, and then delete redundant sentences by calculating their content similarity, helping to get a more accurate description of the celebrity.

Compared with the other two summarization methods we mentioned above, this form of summary extracts the most salient events about a celebrity from the related web pages, and provides a more detailed description of each event. 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 method was proposed in section 3. In section 4, some analysis of our experimental results and a screenshot of our system are given. Finally we make a conclusion and discuss our plans for future work in section 5.

2 Related Work

Our work is related to a series of work on character description, stream data mining, and natural language processing.

The work on character description aims to represent a character with his characteristics, societal attributes, and the events occurring to him, which is an important part of our work. Expert finding has aroused the interest of many researchers [7, 11, 1]. Balog et al. [1] proposed two general strategies in expert finding: one is to model an expert's knowledge based on the documents they are associated with, and the other is to locate documents on topic, and then find the associated expert. Both methods get good performance compared with other unsupervised techniques, indicating the importance of forming reliable associations in expert finding systems. Chen et al. [11] focused not only on the extensive knowledge about an expert but also the strong social links with him. They modeled the social network as a graph, in which the vertices indicate persons and the edges represent the relationships between persons. In this way, the problem of finding the "starring authors" in social network has turned to be detecting the vertices which have big weight and strong associations with others in a graph. Zhu et al. [7] used multiple levels of associations to solve this problem, but the basic idea is consistent with that of Chen et al [11]. In Chieu et al [4], a method of extracting the events about a person along a timeline was proposed. An event was

represented by a single sentence, and a series of sentences, having the biggest interest score, were ranked by time to describe the person with his historical events. Our work is more related to [4], but we focus on the person's most salient events and summarize them with a detailed description instead of a single sentence.

The work on stream data mining [6, 9, 10, 5] considers a single event but not the events which have causality with it or the events which have same attributes with it. Platakis et al. attempt to discover bursty terms and correlations between them during a time interval, and adequately describe an event with those terms [6]. Kumar et al. identify bursty communities from Weblog graphs by taking account of the considerable attention [9]. In [10], a topic evolution graph is built and used to trace topic transitions, i.e. changes in the cluster labels rather than the cluster themselves. Suhara et al. proposed a method to extract the action relation between two keywords from blog articles when two such related keywords are given [5], and represented evolution of event over time for a single concept or sentiment using those correlated words. In this paper, taking a celebrity as the center, we summarize the events about him along a timeline instead of considering a single event.

Finally, some work on Chinese text summarization has been done. Kuo et al. [2] proposed a sentence reduction algorithm by informative words, event words and temporal words to deal with both length constraints and information coverage. In [8], a kernel words based approach for sentence extraction in text summarization is proposed, which achieves a high accuracy rate. Lin et al. proposed a method to calculate the affix similarity between two sentences by identifying all the comment substrings between them [3]. The method in [3] is modified and introduced into our system in the event summarization part. The redundant sentences can be deleted effectively by this method to generate a more accurate description of the celebrity.

3 Proposed Algorithm

Figure 1 shows the system architecture of our proposed events summarizer (summary generation system). The input of the system is a celebrity's name and the output is the summaries of his most salient events. The system processes the summarization in three main steps as aforementioned: (1) tracking the public attention to a celebrity and detecting the bursts from the curve of attention; (2) mining the events from the News (or Weblogs) data in the period around the bursts detected in step 1; (3) summarizing the results. These steps are performed in multiple sub-steps.

We download the news or blogs related to the celebrity as our stream data, and paint them into curve by the change of time intensity. From the curve, we can intuitively understand what a burst means. "A sequence of events is considered bursty if the fraction of relevant events alternates between periods in which it is large and long periods in which is small" [9]. So whenever the popularity of a specific keyword dramatically and unexpectedly increases, a *burst* is marked [6]. Here, we assume that a burst arises as a result of a hot event or at least an event receiving obviously more attention than other events in the same period. The keywords for representing the bursts (or events) are first extracted using the statistical methods of frequency, and then chosen with the comparison of the similarity between them. Then those words are used to rank the sentences for summarizing events. We discuss each of the sub-steps in the following sections.

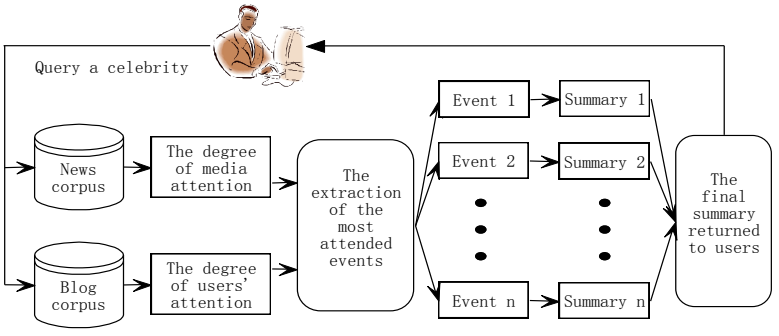

Fig. 1. System architecture of the events summarizer

3.1 Attention Tracking

The "users' attention" and "media attention" are defined as described in Baidu-Index[1].

> ***"The degree of users' attention*** *is evaluated on the statistic basis of millions of internet users' searching frequency in Baidu, targeted on keywords, analyzed and calculated by the weights of number of various words' searching frequency in Baidu search web, and finally illustrated by curve graph."*
>
> ***"The degree of media attention*** *is based on the amount of news most relevant to the keywords in Baidu news search in the last 30 days. After being weighted, the final data were obtained and displayed in the form of surface map."*

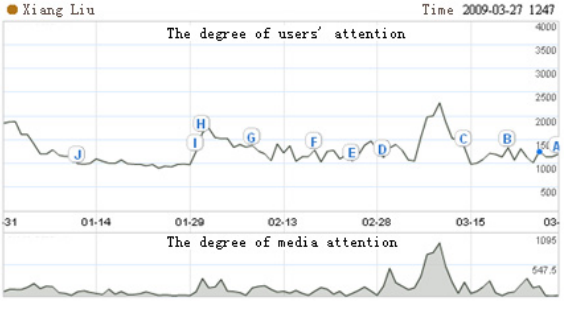

Fig. 2. The degree of users' attention and media attention in a quarter

The curves of users' attention and media attention we obtain from the Baidu-Index are given in Figure 2, which shows the degree of users' attention and media attention in a quarter. We can see that the media attention and the users' attention on a celebrity have almost followed the same trend. Therefore we can use one of them to characterize the whole changing trend. Here, we choose the news data downloaded from a popular Chinese news website Xinhua[2], since the degree of media attention is based

[1] http://index.baidu.com/
[2] http://www.xinhuanet.com/

on the News corpus and the News corpus has a structured format. Issued a query (a celebrity's name), the website will return all the news about it. We take the query *Liu Xiang* as an example. The pages returned according to him are downloaded as our corpus. This corpus contains more than twenty thousand sentences included in 1841 independent news articles.

Figure 3 shows the changing curve of numbers of News in Xinhua. It can be seen that there are 5 rapid increases in this curve, which are defined as bursts. The fourth one was around 12-03-2008, and the last one was around 03-09-2009. They are the same dates as shown in Figure 2. This similar trend has proved again that News data can characterize the whole changing trend of the attention degree about a celebrity.

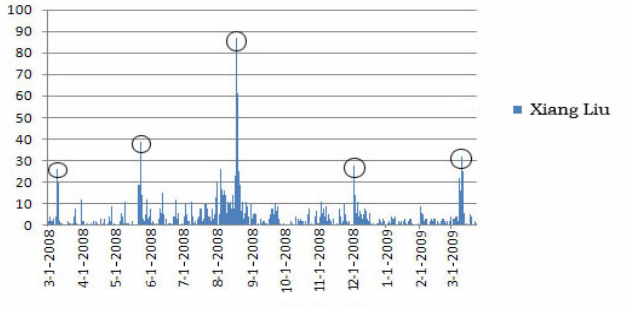

Fig. 3. The changing curve of News number

3.2 Data Preprocessing

The input raw texts need to be processed by a Chinese word segmenter and a part-of-speech (POS) tagger.

1) Word segmentation and annotation

Because of the flexibility of the grammar and the particular expression in Chinese sentences, we should first do some primary processing on data. We use a Chinese lexical analysis system ICTCLAS[3] as a lexical analyzer to process the input raw texts. ICTCLAS includes word segmentation, POS tagging and unknown words recognition. Its segmentation precision is 97.58%. The recalling rates of unknown words using roles tagging achieve more than 90% [11]. These works are prepared for the statistic of words and the redundant segments deletion in our experiments.

2) Resolving Chinese temporal expressions

The same temporal expression, for example 'yesterday', may denote different times in different documents, so the temporal expressions in documents should be converted into calendrical forms [2]. The most familiar two forms of date are '5/4/2004' and 'May 4th 2004'. We use the latter in this paper and convert all temporal expressions into this form.

[3] http://ictclas.org

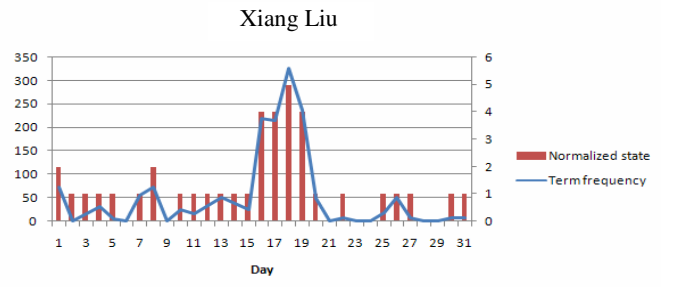

Fig. 4. The explanation of the 31-dimensional *vector* we defined

3.3 Mining Events from News

There are several methods to denote an event: 1) describing an event with its features like time, place and the people involved in [12]; 2) charactering an event using a small subset of keywords that are able to describe one or more real life events occurring during the period of study [6]; 3) representing each event with a sentence extracted from a collection of related documents [4].

We use the second method mentioned above. We find out all the bursts in our stream News data, and take every burst as a salient event of the celebrity. For each event, we choose the News in 31 days (somewhat equivalent to one month) around the burst time as its sub-corpus. Then we count the most frequent words in this sub-corpus, and represent them in the form of vector, in which every dimension reflects the number of times the word appearing in the corresponding day. In this way, every candidate keyword will be changed into a 31-dimensional vector. We give an example to explain the 31-dimensional *vector* we defined. In Figure 4, we can see that the term frequency of "Liu Xiang" was normalized into a 0 to 5 state series. Accordingly, every dimension in the *vector* of "Liu Xiang" was an integer between 0 and 5. This processing will enable us to calculate the time similarity between terms more conveniently.

$$Dis(V_1,V_2)=\sqrt{\sum_{n=1}^{31}(V_{1n}-V_{2n})^2} \tag{1}$$

Furthermore, we adopt a Euclidean-based distance metric in formula (1) to calculate the distance between two vectors. If the distance between each two vectors in a group of terms is less than 4 (the threshold we set here empirically), this group of terms is defined to be the similar timeline keywords, which will further be used to choose sentences for describing an event. In formula (1), $Dis(V_1,V_2)$ means the distance between two vectors V_1 and V_2, we calculate it with the Euclidean-based distance between those two 31-dimensional vectors.

3.4 Events Summarization

We divide this task into two subtasks: sentence extraction and redundant sentences deletion. In sentence extraction, the extracted keywords are used to rank the sentences as aforementioned. Yang et al. [8] proposed a method for ranking Chinese

sentences by keywords, which assumes that the more the in-degree and out-degree of an entity, the more important the sentences contain this entity. So we use the extracted keywords to choose the sentences.

In this paper, we delete redundant sentences by calculating their content similarity, so as to provide a concise description of events. The method in [3] is used to calculate the content similarity between two Chinese sentences, an English example is shown here for an easy understanding. Suppose two sentences are: S'1: "It will be fine tomorrow" and S'2: "Tomorrow will be sunny". First, we separate "Tomorrow will be sunny" into {"Tomorrow will be sunny", "will be sunny", "be sunny", "sunny"}, and then compare them with "It will be fine tomorrow".

As shown in Figure 5, "Tomorrow will be sunny" has a 1 word length common word with S'1, i.e., "tomorrow", "will be sunny" has a 2 word length common prefix with S'1, "be sunny" has a 1 word length common prefix with S'1, and "sunny" has a 0 word length common prefix with S'1. Finally, the value of the content similarity of the two sentences ConSim (S'1, S'2) is the sum of {1, 2, 1, 0}, which is 4 for the given example. Normalizing this value into a number between 0 and 1, we further define (ConSim (S'1, S'2) + ConSim (S'2, S'1)) / (ConSim (S'1, S'1) + ConSim (S'2, S'2)) as the final content similarity between them.

The content similarity between two Chinese sentences is finally defined as below:

$$\mathrm{Sim}(S1,S2)=\frac{\mathrm{ConSim}(S1,S2)+\mathrm{ConSim}(S2,S1)}{\mathrm{ConSim}(S1,S1)+\mathrm{ConSim}(S2,S2)} \tag{2}$$

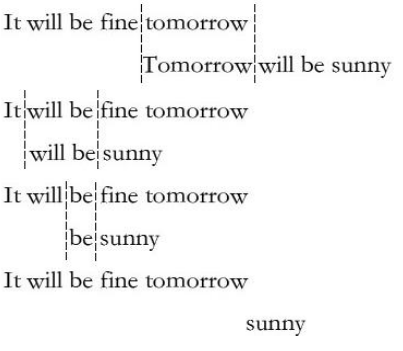

Fig. 5. The content similarity of the two sentences

4 Experimental Results

4.1 Experiments on Event Mining

We take the 5th burst in Figure 4 as an example to explain the whole process. Figure 6 presents the changing curve of numbers of News in a month around this burst.

After preprocessing the data, we count the number of every word in our corpus, and choose those whose amount is more than 62 (twice the number of days) as our keywords candidates. We denote them with 31-dimensional vectors, normalize every dimension into a 0 to 5 state, and calculate the similarity between two terms by a Euclidean-based distance metric as mentioned in section 3.3.

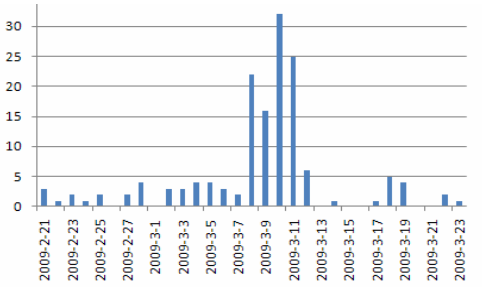

Fig. 6. The changing curve of numbers of news in a month around 2009-3-10

We follow two rules for choosing the final groups of keywords: 1) Every term in this group has a similar timeline as the changing numbers of News in the same period, and this similarity is also defined by the distance between the vectors in this group of terms and the vector of the news' number, whose dimensions are also normalized into a 0 to 5 state. 2) The distance between each two vectors in this group of terms is less than a threshold. Rule 1 is used to make sure that the keywords we extracted are related to the event we want to describe, and rule 2 indicates that those terms produce similar activity, which also means that they has a close connection with each other. The values of the thresholds in rule 1 and rule 2 are set between 4 and 5 empirically.

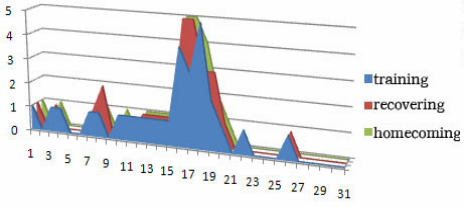

Fig. 7. The timeline-curve of the three words "training", "recovering" and "homecoming"

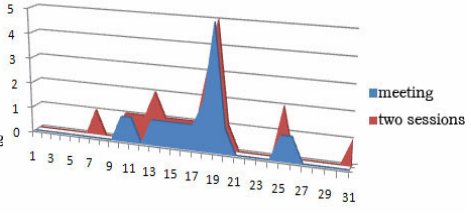

Fig. 8. The timeline-curve of the two words "meeting"and "two sessions"

As seen in Figure 7, the three words "training", "recovering" and "homecoming" have very similar time curves, as well as "meeting" and "two sessions" shown in Figure 8. These words all became bursty ones around March 10th 2009. Trying to evaluate the accuracy of this result, we found out that on March 8th 2009 Liu Xiang arrived at Shanghai Pudong International Airport after the surgery from the United States, and then participated in "two sessions" in China.

4.2 Sentence Extraction and Summary Formation

We first delete the same sentences in our database, which is abundant because of the reprints between News articles. Then we take sports star *Liu Xiang* as an example to explain how to choose the sentence candidates using the keywords we extracted in the above part. Algorithm 1 shows the procedure of choosing sentence candidates. S is

```
SENTENCE-CHOOSING(S)
Input: Set of sentences S.
1       for each s ∈ S
2       if s contains ("Liu Xiang")
3               and contains (("training" and "recovering")
4                       or ("homecoming" and "recovering")
5                       or ("training" and "homecoming")
6                       or ("two sessions")
7                       or ("meeting"))
8               and contains some time mark
9       preserve s.
10      else
11              delete s.
12      end
```

Algorithm 1. Sentence Choosing

the set of sentences in our one month News corpus about Liu Xiang. In line 8 of algorithm 1, we add a condition that the sentences we want must contain some time mark, for the purpose of deleting the News' titles. After such processing, the number of result sentences has been reduced to 16.

We then carry out a k-means cluster analysis on the 16 sentences, and set k to be 4. The distance between two sentences is the reciprocal of the similarity between them, which can be calculate by formula (3). Then we choose the 4 "center sentences" that are closest to the center point in every cluster as our ultimate summary sentences. The final summarization result of the last burst event in Figure 3 is given below:

When the "two sessions" opened, Liu Xiang was still in the United States, carrying out training for recovery after his operation. "Trapeze" once again absenting the "two sessions" has aroused some controversy.

Liu Xiang finished his training recovery in Houston, United States, and flew back home on March 7th.

On March 10th, Liu Xiang went to his first public training after a 3 months long training recovery in United States.

On the evening of March 10th, despite a little tired, he insisted on flying to Beijing to attend the meeting of the CPPCC National Committee, performing his duties.

The italicized part above is the summary of one event about Liu Xiang, which describes the event "Liu Xiang came back from United States after his training recovery".

4.3 Discussion of the Results

Empirically, our result is really able to character a person accurately and succinctly. In this section, we evaluate our experimental result from both objective and subjective point of view. We first perform a small statistical experiment to evaluate our system by examining whether the event we extracted can reflect the interest of the News viewers. We download 654 News articles, which are in the same way as we downloaded pages from Xinhua website, from another famous News website Sina[4] in the same period as in our previous experiments. Two annotators annotate those News

[4] http://news.sina.com.cn/

independently with "related to the event we extract" or not, then meet to compare those double annotated files. A final annotation result after their agreement show that 506 News are related to the event we extract, while the else 148 ones were not.

We believe that the number of comments of News can represent the interest of users on it, so we compared the number of comments of News in related part to unrelated one. As shown in table 1, most of the News in both parts has less than 100 comments. But the News with more than 100 comments has an obviously larger proportion in related part than in unrelated part. There is a News even has more than 1000 comments, of which the number is actually 2605. The average number of comments of News in related part is 171.21, while the number of unrelated part is 46.95. From this result, we can easily judge that our summary of events can heavily reflect the users' interest.

Table 1. Number of News related/ unrelated to the event we extracted (NC means "Number of Comments")

	Related	unrelated
NC<10	149	51
10<NC<100	177	68
100<NC<1000	179	29
1000<NC	1	0
total	506	148

To further evaluate our proposed method, we conduct a comparison against other two types of character description generated by Chieu's method [4] and Platakis's method [6]. About the same event mentioned above – "Liu Xiang came back from United Stated after his training recovery", we give the results of those two methods below as an example:

> **Chieu's method:** *On March 10, Liu Xiang went to his first public training after a 3 months long training recovery in United States.*
> **Platakis's method:** *training, recovering, homecoming, two sessions, meeting.*

50 celebrities are chosen as our queries, whose descriptions are summarized by Chieu's method, Platakis's method and our method, and five Ph.D. students are

Table 2. The score of each method

Scores / Methods / Questions	Chieu's method	Platakis's method	***Our method***
Q1: Can it describe the events accurately?	0.823	0.954	**0.952**
Q2: Can it describe the events succinctly?	0.722	0.720	0.612
Q3: Can it describe an event allsidedly?	0.845	0.612	**0.842**
Q4: Can it describe a celebrity allsidedly?	0.812	0.785	0.754
Q5: Can it emphasize the most salient events?	0.708	0.912	0.855
Q6: Does it have a good timeline?	0.901	0.284	**0.970**
Q7: Can it separate each event distinctly?	0.774	0.925	**0.934**

employed to score our experimental results with a number between 0 and 1 (0, 0.2, 0.4, 0.6, 0.8, 1.0) at seven aspects which are listed in Table 2. In general, a score close to 1 indicates that the automatically generated description is with good quality at this aspect. Table 2 shows the final average score of each method at every aspect. From table 2, we can see that our method performs well and evenly at all aspects discussed.

4.4 User Interface

Figure 9 represents our system interface, which allows obtaining web content on any given celebrity by sifting the News stream. Users are able to express the celebrity of his interest by a search string and the system will show the period of events occurring for the celebrity and output the summarization result about those events. By searching via such an interface, people can easily understand a celebrity with his salient events.

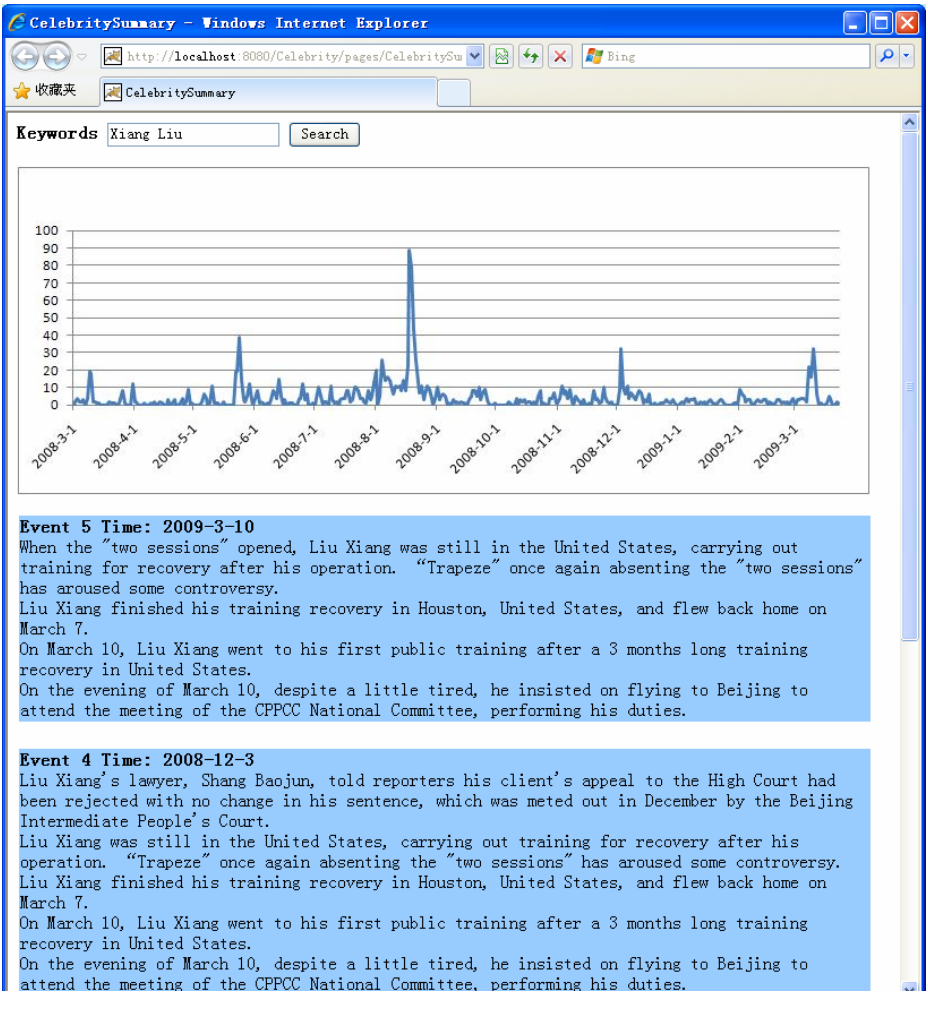

Fig. 9. An example of a celebrity's description

5 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. 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 on user queries. 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 partly supported by the projects 863 (No. 2006AA010106), 973 (No. 2007CB311007), NSFC (No. 60703085).

References

1. Balog, K., Azzopardi, L., Rijke, M.: Formal models for expert finding in enterprise corpora. In: International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 43–50 (2006)
2. Kuo, J.J., Chen, H.H.: Multi-document Summary Generation using Informative and Event Words. TALIP 7(1), 1–23 (2008)
3. Lin, K.H., Yang, C., Chen, H.H.: Emotion Classification of Online News Articles from the Reader's Perspective. In: Proceedings of International Conference on Web Intelligence, Institute of Electrical and Electronics Engineers, Sydney, AU, pp. 220–226 (2008)
4. Chieu, H.L., and Lee, Y.K., Query Based Event Extraction along a Timeline. In: International ACM SIGIR Conference on Research and development in Information Retrieval, Sheffield, UK, pp. 425–432 (2004).
5. Suhara, Y., Toda, H., Sakurai, A.: Event Mining from the Blogosphere Using Topic Words. In: Proceedings of the 1st International Conference on Weblogs and Social Media (ICWSM 2007), Boulder, Colorado, USA (2007)
6. Platakis, M., Kotsakos, D., Gunopulos, D.: Searching for Events in the Blogosphere. In: Proc. Int'l Conf. World Wide Web, WWW 2009, pp. 1225–1226 (2009)
7. Zhu, J., Song, D., Rüger, S.: Integrating Multiple Windows and Document Features for Expert Finding. JASIST 60(4), 694–715 (2009)
8. Yang, W., Dai, R., Cui, X.: A Novel Chinese Text Summarization Approach Using Sentence Extraction Based on Kernel Words Recognition. In: FSKD 2008, pp. 134–139 (2008)
9. Kumar, R., Novak, J., Raghavan, P., Tomkins, A.: On the Bursty Evolution of Blogspace. In: Proc. Int'l Conf. World Wide Web, WWW 2003, pp. 159–178 (2003)
10. Mei, Q., Zhai, C.: Discovering evolutionary theme patterns from text: an exploration of temporal text mining. In: Proceeding of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining, Chicago, USA, August 21-24, pp. 198–207 (2005)
11. Chen, D., Tang, J., Li, J., Zhou, L.: Discovering the Starring People from Social Networks. In: Proc. Int'l Conf. World Wide Web, WWW 2009, pp. 1219–1220 (2009)
12. Zhao, X., Qin, B., Che, W., Liu, T.: Research on Chinese Event Extraction. Journal of Chinese Information Processing 22(01), 3–8 (2008) (in Chinese)