

Portraying User Life Status from Microblogging Posts

Jiayu Tang*, Zhiyuan Liu, Maosong Sun, and Jiahua Liu

Abstract: Microblogging services provide a novel and popular communication scheme for Web users to share information and express opinions by publishing short posts, which usually reflect the users' daily life. We can thus model the users' daily status and interests according to their posts. Because of the high complexity and the large amount of the content of the microblog users' posts, it is necessary to provide a quick summary of the users' life status, both for personal users and commercial services. It is non-trivial to summarize the life status of microblog users, particularly when the summary is conducted over a long period. In this paper, we present a compact interactive visualization prototype, LifeCircle, as an efficient summary for exploring the long-term life status of microblog users. The radial visualization provides multiple views for a given microblog user, including annual topics, monthly keywords, monthly sentiments, and temporal trends of posts. We tightly integrate interactive visualization with novel and state-of-the-art microblogging analytics to maximize their advantages. We implement LifeCircle on Sina Weibo, the most popular microblogging service in China, and illustrate the effectiveness of our prototype with various case studies. Results show that our prototype makes users nostalgic and makes them reminiscent about past events, which helps them to better understand themselves and others.

Key words: text visualization; microblogging; topic model; sentiment analysis; keyword extraction

1 Introduction

In the Web 2.0 era, an increasing number of people use microblogs to share information and record daily life. The posts of a microblog user can thus usually reflect his/her interests, attributes, and life status. Some popular microblogging services, such as Twitter founded in 2006 and Sina Weibo founded in 2009, have been used by most of the Web users for years. These microblogging posts have become valuable records of users' individual life.

Because of the high complexity and the large number of microblogging posts, it is necessary to provide a quick summary to understand the life status of microblog users, including the subjective aspect (i.e., sentiments and opinions) and the objective aspect (i.e., interests and attributes). This is crucial for both user experience and commercial services: microblog users may want to quickly find what they have done in the past years, and commercial services may want to know the interests of a user. In fact, microblog users have shown a considerable interest in the summarization of microblogging posts. At the beginning of 2012, we developed an application to present the annual keywords of Sina Weibo users in 2011. The results have attracted a considerably large number of users to the site. Many Sina Weibo users have expressed their desire to automatically summarize and visualize their annual history. Moreover, Sina Weibo has recently released an official application to summarize some statistics of 2012 for a club user, including the number of posts

• Jiayu Tang, Zhiyuan Liu, Maosong Sun, and Jiahua Liu are with the State Key Laboratory of Intelligent Technology and Systems, Tsinghua National Laboratory for Information Science and Technology, Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China. Email: tjy430@gmail.com; lzy.thu@gmail.com; sms@tsinghua.edu.cn; alphaf52@gmail.com.

* To whom correspondence should be addressed.

Manuscript received: 2012-12-07; accepted: 2013-02-19

and the months in which the user published posts most frequently. A club user is an officially identified activist in Sina Weibo (<http://help.weibo.com/topic/club/>). The application has attracted more than 550 000 club users, approximately one-tenth of the club users in Sina Weibo, in a few days. However, the demonstration of the application is considerably simple, and the summarized information is limited to some simple statistical numbers. These phenomena indicate the importance and the necessity of the automatic summarization and visualization of microblogging posts over a long period of time from both the subjective aspect and the objective aspect. However, little work has been devoted to it, as the task is challenging for the following two reasons: (1) the number of posts posted by a user over a long period, e.g., a year, is usually large; and (2) the topics and keywords of these posts are highly diverse and dynamic.

To tackle the abovementioned challenge, we have developed LifeCircle, a multifaceted and interactive visualization prototype to help users to better understand themselves and others by exploring temporal interests, sentiments, and activities within an entire year. The prototype provides a multi-layered and time-aware design based on a radial visualization, which has the advantage of efficiently using the display space while conveying the hierarchical structure of time. The prototype consists of three visual primitives: (1) monthly sentiment summarization; (2) trends of microblogging posts; and (3) interest keywords and topics in various granularities of time. Each visual primitive encodes a particular sophisticated data-analysis technique and aesthetic visual treatment.

To the best of our knowledge, our work is the first to address the problem of visually and multi-dimensionally summarizing the life status of microblog users over time. Our work presents the following contributions: (1) We for the first time propose to summarize the life status and interests of microblog users over time. (2) We design a compact, multifaceted, and highly developed interactive visualization framework. It provides a comprehensive and informative summary to facilitate complex microblogging analyses. (3) We integrate multiple state-of-the-art techniques of natural language processing, including sentiment analysis^[1], keyword extraction^[2], and latent topic modeling^[3], and popular techniques of visualization, including tag cloud^[4] and radial visualization^[5], into the prototype system.

2 Related Work

2.1 Microblogging data analysis

At present, microblogs are almost the most popular social networking platforms and broadcast media. Some social scientists have analyzed the characteristic of microblogs, such as structure and relationships in social networks^[6-9].

Many researchers have investigated microblogs from the perspectives of computer science as well as social science, and most researches are related to text mining. In information retrieval, researchers study the differences between a traditional search of Web pages and a search in microblogs, such as real-time search and instant search^[10,11]. Twitter, the most popular microblogging service in the world, has been studied in relation to many traditional and new research topics, such as recommendations^[12], event detection and tracking^[13-15], sentiment analysis^[16-19], information propagation^[20,21], and authority identification^[22]. In natural language processing, many traditional tasks on microblogs, such as part-of-speech tagging^[23], named entity recognition^[24-26], and lexical normalization^[27], have become increasingly challenging because of the amount of noisy data.

However, these research achievements in text mining have not been explored well with respect to visualization, and this paper aims to bridge the two research fields in order to provide a comprehensive solution for microblog user life summarization.

2.2 Text visualization

LifeCircle is inspired in part by Themail^[28], a visualization that analyzes the interaction histories in email archives and shows stacked keyword lists over time to portray the relationships of users. Themail is faced with the problems such as an inadvertently high weight given to topical words in forwarded messages, unrepresentative keywords extracted from people's email signatures, and the granularity of the content parsing mechanism. We, in contrast, propose an efficient and effective framework to achieve a complex data analysis task in order to generate an intuitive circular visualization, which is more compact than Themail. Furthermore, LifeCircle shows more aspects as well as keywords to form better overall portraits for viewers.

In recent years, radial visualization has become a popular and common metaphor in information

visualization^[5]; it typically means the arrangement of data in compact concentric rings and has the ability to display multi-dimensional datasets. Typical examples include Hyperbolic Browser^[29], Sunburst^[30], InterRing^[31], Information Slices^[32], and DocuBurst^[33] for visualizing and manipulating large hierarchies, and Radial Traffic Analyzer^[34] and VisAlert^[35] for visualizing relationships among disparate entities. Some quantitative studies have confirmed its usability for visualizing hierarchical datasets^[5,36,37]. In this paper, the year-month-day data are arranged in a hierarchical structure, and we can use radial visualization to obtain a compact result.

Tag clouds are also a popular visual representation for text data and are typically used in the Web 2.0 era to visualize the frequency or popularity distribution of key terms that depict website content. Wordle (<http://www.wordle.net/>) is one of the most outstanding tag cloud visualization tools for generating aesthetic visual representation from a text corpus. It is significantly different from regular tag clouds because of its efficient use of the display space and is widely used in the social community and mainstream media^[38]. However, Wordle is not designed for interactive visualization, and the original algorithm for placing tags tends to be slow. We designed our keyword placement algorithm in a circular area on the basis of Wordle and proposed optimizations to obtain considerable computational acceleration.

3 Data Analysis

In this section, we describe the main pipeline for microblogging data processing. We show how to perform Chinese Word Segmentation (CWS) and Part-Of-Speech (POS) tagging and topic summarization. Then, we present a hybrid method of combining both the dictionary-based approach and the emoticon-based approach for the sentiment analysis. We then discuss how interest keywords are extracted by combining a frequency-based method and a translation-based method. Finally, we show how to build bridges between each topic and the related keywords. Figure 1 shows the entire framework for analyzing microblogging data.

3.1 CWS and POS tagging

In this paper, we take Sina Weibo, the largest microblogging service in China, as our research platform. Most posts in Sina Weibo are in

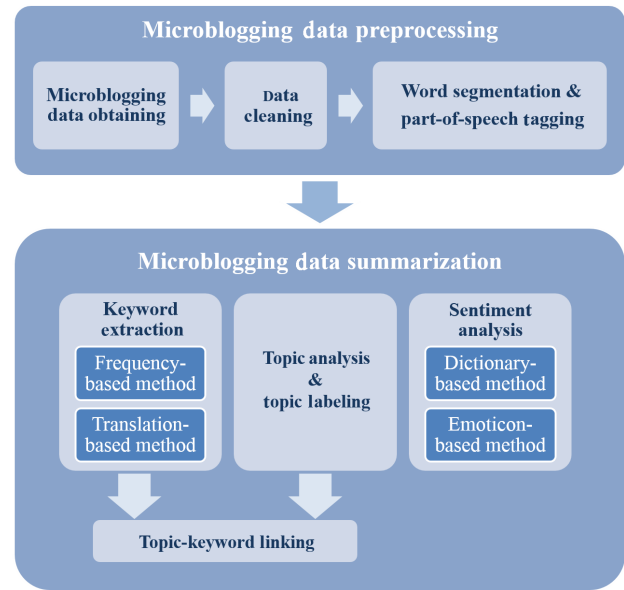


Fig. 1 Framework of microblogging data analysis.

Chinese. Hence, we perform CWS and POS tagging before topic summarization, keyword extraction, and sentiment analysis.

Before CWS and POS tagging, we first clean the microblog users' posts and only keep plain text data. Take Sina Weibo data for example. We remove the user names mentioned in posts, which are usually irrelevant to the themes of posts. If not removed, these names would rank high in the frequency-based keyword extraction. Further, we remove URLs and other non-text items from the posts.

After obtaining the clean data from the microblog users' posts, we perform CWS and POS tagging using THULAC, a practical system developed by the natural language processing group at Tsinghua University^[39]. Note that other POS taggers such as Stanford Log-linear Part-Of-Speech Tagger^[40] can also be easily embedded in our framework to support other languages.

Words used in microblogging posts are changing over time, and new words are generated every day. To make the extracted keywords in the visualization up-to-date, we have to take care of the new Out-Of-Vocabulary (OOV) words. To address the problem, we maintain a large external vocabulary for CWS and POS tagging. For building the external vocabulary, we collect new words from the following two sources, Hot Keyword List of Sina Weibo (<http://data.weibo.com/top/keyword/>)

and New Word Dictionary of Sogou Pinyin (<http://pinyin.sogou.com/dict/>). Sogou Pinyin is the most widely used Chinese Input Method Editor in China. Sina Weibo updates its list daily, and Sogou Pinyin updates its dictionary weekly.

3.2 Topic summarization

The themes or topics of microblogging posts are mostly reflected by nouns. Hence, we only take nouns for topic summarization and keyword extraction. Table 1 shows the POS tag set that we use for the two tasks.

In this paper, we use Latent Dirichlet Allocation (LDA)^[41] for topic summarization. LDA is a generative model assuming that documents are distributions over topics and that topics are distributions over words. As a representative and significant latent topic model, LDA has been widely used for learning latent topics from a large collection of documents^[42, 43].

Suppose that there is a set of large-scale microblogging posts, $P = \{p_1, p_2, \dots, p_N\}$, where N denotes the number of posts. We manually set the number of topics as K . In our prototype system for case studies, K is set 100. From these posts, we can learn K latent topics using LDA. Each topic is a multinomial distribution over a set of words. By manually inspecting these topics, we remove noisy topics with nonsense and group the leftover topics into 13 categories by taking the coverage and variance of these topics into consideration. The number of categories is determined by topic annotator. Each category is assigned a label, which is used for the visualization. The categories are discussed in Section 5.1.1.

Given a microblog user, we infer the related topic distribution over K topics according to his/her posts and then, select the top-6 categories for the visualization.

3.3 Sentiment analysis

For the sentiment analysis, the most straightforward approach is a dictionary-based method^[44]. This approach uses a sentiment term dictionary to determine

the sentiment of a text according to the frequencies of the sentiment words in the text. Because of its simplicity, this approach has been widely used in various applications of sentiment analysis. However, this approach faces the severe issue of OOV and may always adversely affect performance. Recently, researchers found that emoticons can be used for sentiment classification^[45] because these icons are widely used in microblogging posts to convey user emotions^[46]. In this paper, we design a hybrid approach for sentiment analysis by combining both the dictionary-based approach and the emoticon-based approach.

For the dictionary-based method, we maintain an emotion dictionary containing both positive and negative words, each of which has a score for recording its polarity. For each microblogging post, after CWS and POS tagging, we evaluate the emotions of each word according to the emotion dictionary. In this way, each post is assigned an emotion polarity score v_d by aggregating the scores of all the words in the post.

For the emoticon-based method, since each emoticon in the microblogging posts is automatically denoted within brackets, we can extract it by a simple string match. We treat these emoticons in the microblogging posts as sentiment labels and assign a positive or negative score to each. Each post is thus assigned with an emotion score v_e by aggregating all the scores assigned to the emoticons in the post.

Finally, the emotion score v_i of a post i is calculated as $v_i = (1 - \alpha) \cdot v_d + \alpha \cdot v_e$, where α is the smoothing factor that ranges from 0 to 1.

3.4 Keyword extraction

Most microblogging posts of a microblog user usually talk about decentralized but related subjects. For example, a user may talk about “javac” in one post, “JRE” in the second post, and “JDK” in the third post. These posts all talk about Java technologies, and it is better to identify “Java” as the keyword, even though the word “Java” itself may not be frequently mentioned by the user. To identify keywords from microblogging posts effectively and efficiently, we employ a novel keyword extraction method proposed by Liu et al.^[2], which combines a frequency-based approach and a translation-based approach. Each keyword i identified by this method is assigned an importance score s_i .

In this paper, we take the dynamics of keywords into consideration. The basic idea is as follows: If a keyword appears smoothly in every month, it will

Table 1 Selected POS tag set.

Tag	Description
n	Common nouns
ns	Place names
nz	Other proper nouns
np	Person names
ni	Institute names
i	Idioms

be less representative, and we set its font size to a relatively small value; while if the keyword only appears significantly in several months, we set its font size to a large value. This also confirms the idea of Term Frequency-Inverse Document Frequency (TF-IDF) that is widely used in natural language processing and information retrieval. We define the distinctiveness of a keyword over the entire year in order to measure its significance with respect of time. The distinctiveness score of keyword i in month j is calculated as follows:

$$s_{ij} = s_i / \sqrt{n} \quad (1)$$

where n is the number of the months in each of which i appears.

3.5 Topic-keyword linking

Topics and keywords represent the semantics and interests of users at different levels. Topics are at a coarse-grained level, while keywords are at a fine-grained level, and they are related to each other. To better demonstrate a user's life status, we build connections between topics and keywords to show their latent relationships. The construction process is as follows.

Each topic is assigned a probabilistic distribution over the word vocabulary, $W = \{w_1, w_2, \dots, w_M\}$, where M denotes the size of vocabulary. Keywords extracted using our method are a subset of the words derived by LDA; for instance, for a keyword k_i , $k_i \in W$. Thus, we can calculate which topic is most related to k_i according to the probabilities derived by LDA. In this way, we link each keyword and its most related topic to further visualize the relations between keywords and topics.

4 Visualization

In this section, we describe the visual and interactive design of LifeCircle. With respect to the fact that the metaphor compatibility has a significant effect on the accuracy of viewers' understanding^[47], we use common clock and tag cloud visual metaphors in our visualization. With the cyclic nature of the year compatible with the clock metaphor, a compact visualization is generated by taking advantage of the radial space-filling layouts in space efficiency.

We first present the user interface of our visualization, followed by the interaction design for viewers who wish to drill down to more detailed information. Then, we briefly introduce our efficient and aesthetic layout algorithms.

4.1 User interface

LifeCircle provides a comprehensive solution for analyzing and visualizing an individual's microblogging posts over time. The interface is designed in a circular fashion. From the center out, we arrange multiple layers of information including temporal topics, keywords, sentiments, and the trends of posts (shown in Fig. 2). The overall design is based on a metaphor of a clock and the time is tracked in a clockwise direction.

4.1.1 Annual and monthly topics

In order to present a general overview of a microblog user, we select to show annual topics in the center of the circle. This view presents up to 6 topics that the microblog user talks about most frequently over the entire year. The font size of these topic words is determined by their frequencies, and the color is randomly selected considering the aesthetics.

Adjacent to the innermost circle of the annual topic words, we set twelve small sectors with several colored dots in each to present the monthly topics talked about by the user in each month. We use the colors of dots to indicate the corresponding topic words in the innermost circle. Dots in each month are ordered by their corresponding topic frequencies. For example, if a certain topic is talked about most frequently in one month, its corresponding dot will be the biggest and placed above all the other dots.

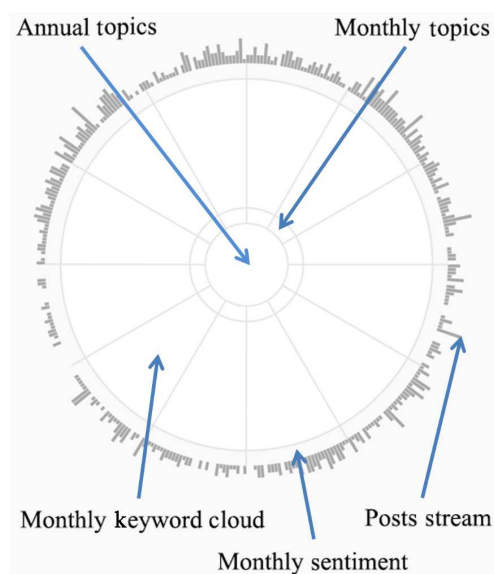


Fig. 2 User interface of LifeCircle.

4.1.2 Monthly keyword clouds

Moving outward from the annual and monthly topics, we place twelve keyword clouds to reveal a more detailed portrait of the microblog users' life over time.

The font size of keywords corresponds to the importance score computed by our keyword extraction method. We by default set all keywords in grey color with different shades. The larger the number of months that a keyword appears in, the darker will be the shade of grey that it will be denoted in.

4.1.3 Monthly sentiments

On the outline of the circle, we use a ring containing two colors to illustrate the sentiments of the microblog user in each month. The negative sentiment is shown in a cool hue, and the positive one is illustrated in a warm hue.

4.1.4 Time-based posts stream

On the basis of the design goal of providing viewers with a closer examination of facts, the time-based streaming of microblog users' posts is shown outside the circle. The histograms along the circle indicate the dynamics of the microblogging posts over the entire year, and each histogram represents a microblog user's daily posts. A viewer can quickly get an overall

comprehensive view of the post trends of the user.

4.2 User interactions

To make LifeCircle easy to use, we design our interaction model with simple operations of clicking and hovering.

Figure 3 shows an interacting screenshot of LifeCircle. The visualization panel shows the connection among a certain topic word, related keywords, and posts. In this figure, the annual topic word “旅游 (travel)” in green has been clicked. This causes the related monthly keywords and the histograms of posts related to this topic are colored in green. Here, some typical monthly keywords are “签证 (visa),” “旅行 (tour),” “旅游 (travel),” and “秦皇岛 (Chinwangtao)” (a resort of China).

A user may not quickly understand what a particular keyword is talking about; thus, additional detailed information might be helpful. In Fig. 3, after we click the keyword “吉隆坡 (Kuala Lumpur),” the corresponding posts are found and shown in the left post content box. The post content box shows the details of each related post, including the posting time-stamp, post content, and the display names of other microblog users who commented on the post. A

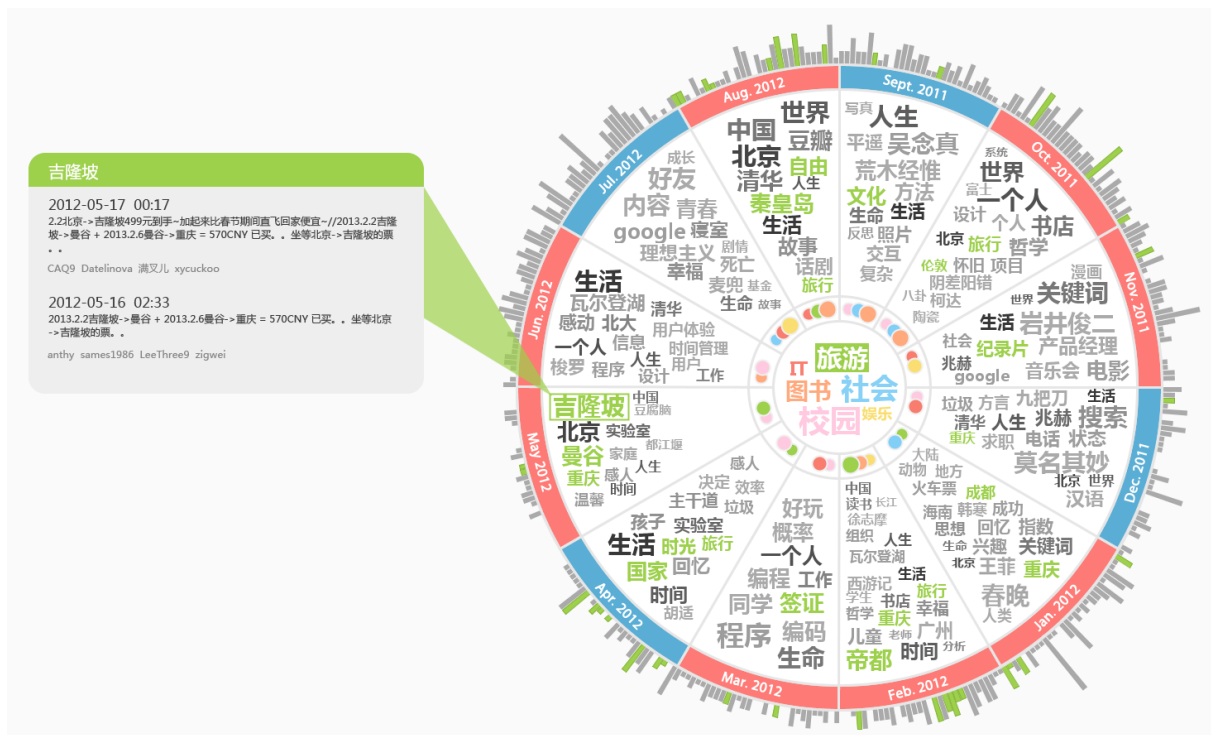


Fig. 3 Annual overview of a microblog user, including main topics, monthly keywords of his posts, and his sentiments. The keywords related to the specific topic “旅游 (travel)” are highlighted in green. The posts related to the selected keyword “吉隆坡 (Kuala Lumpur)” are shown in the left box.

viewer can also switch to another microblog user's visualization by clicking the user ID shown at the bottom of each post. Moreover, when the cursor hovers over one histogram of the outermost histogram stream, the corresponding posts will also be shown in the post content box.

To build a compact visualization, each area of the sectors for monthly keywords is limited by default, and a more detailed view for each month can be seen by double-clicking the corresponding sector. Animated transitions are employed for demonstrating additional keywords in the detailed view as such animation can significantly improve graphical perception^[48]. Figure 4 shows the transition process for a selected month. In the detailed view of monthly keywords, more keywords with a slightly small weight are displayed to provide a more comprehensive understanding for the selected month.

Throughout the design of these interaction methods, we hope to provide viewers with an intuitive and comprehensive review of the microblog users' life.

4.3 Algorithms

In this section, we describe the main algorithms designed for the visualization in our prototype.

4.3.1 Layout of topic words

LifeCircle follows the algorithm of Wordle^[31] for determining the locations of topic words in the innermost circle: for a given topic word, we select the center of the circle as the starting point and radially update its position on a spiral of increasing radius to perform collision detection between the word to be

placed and the words already placed. The topic words are placed one by one in the descending order of font size.

After the procedure, if there are any remaining words that cannot be positioned anymore, the font size of the topic words is decreased proportionally and the process is repeated. The layout algorithm will not be terminated until all the terms are positioned.

4.3.2 Layout of monthly keywords

LifeCircle mimics the rationale of Wordle and packs keywords as tightly as possible for a compact visualization result. As many keywords as possible are packed in each sector; however, the font size cannot be smaller than the minimum legible font size. For computational efficiency, further optimization is performed in the following two aspects.

(1) The most important keywords are placed close to the outer edge of the sectors. For a given keyword, we perform collision detection between the word to be placed and the words already placed by scanning sweep arcs from the one close to the outer edge of the sector to the one close to the inner edge.

(2) For a specified keyword, the interval of sweep arcs and the interval of candidate positions on each sweep arc are fixed during scanning. However, the intervals vary in proportion to the font size of the keywords (shown in Fig. 5). This idea is inspired by ManiWordle^[49] with the observation that if a word A is smaller than a word B, it is more likely that word B and not word A will collide with the words already on the canvas during collision detection.

The purpose of all these optimization methods is

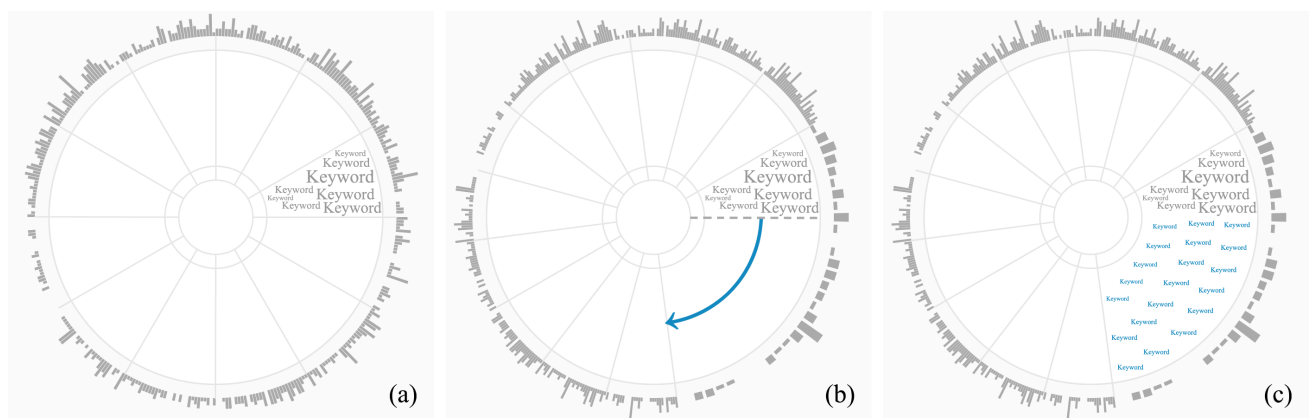


Fig. 4 Process for providing a detailed view of monthly keywords: (a) font size of keywords in rest of the months decreases proportionally; (b) sweep angles of sectors and widths of post histograms change smoothly; and (c) additional keywords are placed according to the algorithm described in Section 4.3.2.

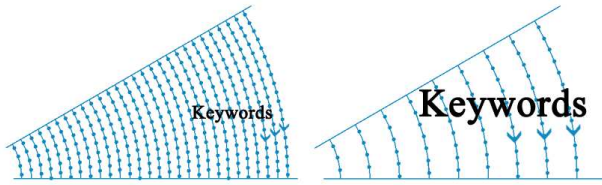


Fig. 5 Interval of sweep arcs and interval of candidate positions for collision detection on each sweep arc are larger when bigger keywords are placed. Each dot represents a candidate position.

to reduce the frequency of the collision detection task during the determination of keyword positions.

5 Experiments and Discussion

We developed a prototype system of LifeCircle using JSP and Processing.js (<http://processingjs.org/>). The system performance was evaluated using case studies.

5.1 Experiments

In the experiments, we select Sina Weibo as the platform to develop our prototype. We show the

overview of our prototype system in Fig. 6. The prototype system consists of two main parts for the data analysis: the offline module for preprocessing and the online module for real-time processing.

5.1.1 Offline module

The offline part mainly achieves preprocessing tasks: (1) trains the LDA model for inferring the topics of a microblog user; (2) pre-defines the emoticons for emoticon measurement; (3) trains the translation model for keyword extraction.

We use PLDA+^[50], a parallel implementation of LDA, to extract latent topics of microblog users from posts. The data for training the LDA model includes 8 million original posts of Weibo users. We label all the topics derived by LDA and classify them into 13 categories as the pre-defined candidate topics on the basis of the topic content coverage and the topic variance. Table 2 shows these pre-defined candidate topics for a given Weibo user and some typical related words for each topic.

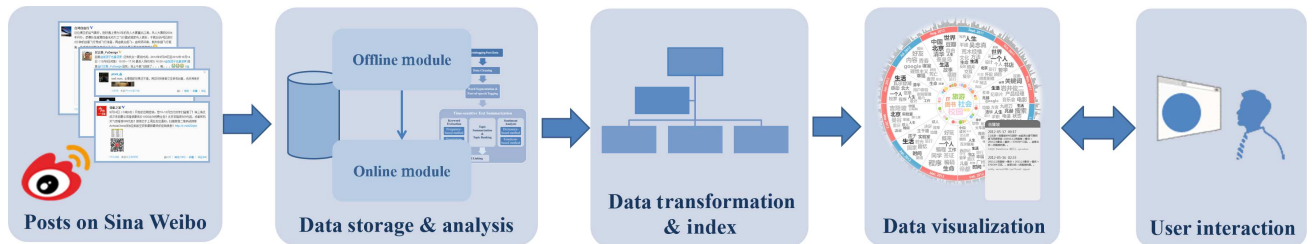


Fig. 6 Overview of LifeCircle implanted on Sina Weibo.

Table 2 Topic list and typical related words for each topic.

ID	Topics	Typical related words
1	历史 (history)	古代 (ancient), 文物 (cultural relic), 人类学 (anthropology), 博物馆 (museum), 祖先 (ancestor)
2	娱乐 (entertainment)	偶像 (idol), 还珠格格 (a teleplay), 星际争霸 (a game), 星座 (signs), 演出 (show)
3	IT (information technique)	邮件 (email), 鼠标 (mouse), 浏览器 (browser), 程序 (software), 代码 (code)
4	校园 (school life)	清华 (Tsinghua), 学术 (academics), 母校 (alma mater), 院士 (academician), 实验室 (laboratory)
5	情感 (emotion)	结婚 (marriage), 单身 (single), 约会 (date), 爱情 (love), 分手 (break up)
6	财经 (finance and economics)	资本 (capital), 股市 (stock market), 信用 (credit), 基金 (fund), 经济 (economy)
7	体育 (sports)	欧冠 (UEFA Champions League), 曼联 (Manchester United), 姚明 (Yao Ming, a basketball player), 总决赛 (Finals), 球迷 (a ballgame fan)
8	旅游 (travel)	酒店 (hotel), 机场 (airport), 三亚 (Sanya, a resort), 签证 (visa), 旅行 (tour)
9	社会 (society)	警方 (police), 群众 (crowd), 国务院 (the State Council), 火灾 (fire), 地震 (earthquake)
10	图书 (books)	阅读 (read), 出版社 (publisher), 书店 (bookstore), 小说 (novel), 洛丽塔 (Lolita, a book's name)
11	美食 (food)	食品 (food), 火锅 (hot pot), 味道 (taste), 葡萄 (grape), 晚餐 (dinner)
12	艺术 (art)	雕塑 (sculpture), 国画 (traditional Chinese painting), 意象 (image), 作品 (works), 油画 (painting)
13	购物 (shopping)	抢购 (snap up), 消费 (consume), 优惠 (discount), 营销 (marketing), 淘宝 (Taobao, a shopping site)

In order to perform the emoticon-based sentiment analysis, we select twenty-three emoticons as sentiment labels, according to their clear emotion representations and their frequencies of use. In all, these selected emoticons occur more than 83 million times in 1.9 billion posts of 1.4 million Weibo users. Table 3 shows the sentiment categories of the different selected emoticons.

To perform the translation-based keyword extraction, two resources are used for training the translation model: book descriptions with their tags on Douban as the translation pairs, and news articles and their titles as the translation pairs. Douban is the largest book review website in China (<http://book.douban.com/>).

5.1.2 Online module

The online part collects the Weibo users' microblogging data in real time to achieve the text summarization and visualization task.

Microblogging data, including the posting time, post contents, and users who have commented on each post, are available via the Application Programming Interfaces (APIs) provided by Sina Weibo. When a user authorizes LifeCircle to access his available data on Sina Weibo, LifeCircle quickly analyzes the data using the methods described in Section 3. After the data analysis and transformation for the visualization, the system forms a visual summarization for a Weibo user and pushes it to the web server for display on the user's browser.

5.2 Case studies

To verify the usability of our prototype, we performed case studies with the system.

5.2.1 Example

Figure 3 shows an example of the annual summarization of a Weibo user's life. The main topics of the user's posts are “校园 (school life)” (shown in pink), “社会 (society)” (shown in blue), “旅游 (travel)” (shown in green), “图书 (books)” (shown in orange), “IT”

(shown in red), and “娱乐 (entertainment)” (shown in yellow). For our initial exploration, we click on the topic word “旅游 (travel),” which is illustrated in green, and the monthly keywords related to it are highlighted in green, which can be treated as a corroboration of the user's interest in this topic.

On the scope of 2011 November, the yellow and red dots indicate that the main topics of this month are “娱乐 (entertainment)” and “IT.” Observing the keywords of this month, we find that distinctive keywords such as “岩井俊二 (Shunji Iwai)” (a Japanese director's name), “漫画 (comic),” “音乐会 (concert),” and “电影 (movie),” which are all related to entertainment, are assigned a high weight, clearly representing the important life status of the Weibo user in November. Furthermore, the outer sentiment arc illustrated in a warm hue indicates that the user was in general in a state of euphoria. The outermost histogram stream shows a peak on January 22, and the contents of the post box show that this phenomenon occurs because of the user's lively discussion on the 5-h-long Spring Festival Gala on CCTV.

The keywords in a relatively dark shade of grey, such as “生活 (life),” “中国 (China),” “北京 (Peking),” and “一个人 (single),” reflect the long-term status and characters of this microblog user. Varied keywords in light grey in each month expose the ever-changing nature of people's life records.

This user case demonstrates the capability of LifeCircle to revisit the annual life status of a microblog user.



5.2.2 Blind test

To further verify the usability and intuition of our prototype to form a quick overall view of a microblog user, we performed a blind test.

(1) Participants: We asked 4 students majoring in Computer Science to use our prototype system. They were all long-term users of Sina Weibo and had followed each other for more than one year.

Participant A posted few posts and just used the microblogging service to gain up-to-date status of his friends. Participant B liked to read funny short stories on the microblog and followed dozens of Weibo users who specially collected and shared such stories. Participants C and D were considerably similar. They both used the microblog to record their daily life and sometimes reposted information mainly about society, music, and movies.

Table 3 Emoticons in each sentiment category.

Sentiment	Number	Emoticons
Positive	13	
Negative	10	

(2) Setup: Using LifeCircle, we collected the participants' microblogging data for the previous year and formed four interactive visualizations without telling them the ground-truth owners. We gave the participants a 2-min explanation of LifeCircle and then asked them to use it for recognizing the owner of each visual representation in 5 min.

To avoid participants from just looking up every post's content to confirm the answer, we took the post content box out of the prototype system provided to the participants. Furthermore, participants are asked to comment freely about their impression of using our prototype system.

(3) Results: Successfully, all 4 participants distinguished these four visual representations. The visualization of participant A was easily identified, which was in our expectation because of the fact that the number of keywords in his case was the smallest. Participant B's visualization was not hard to tell either, as the sentiment values of eleven months were positive and keywords about fun, e.g., "joke" and "funny," were denoted by a big font and in dark grey, indicating their considerably frequent use. Although the behaviors of participants C and D were similar on Sina Weibo, participants were able to differentiate between them using different evidence. Participant A said he pointed out C's visualization by finding the keywords "wedding" and "bridegroom," which indicated a wedding C attended in May 2012. Participant B distinguished C's visualization by finding that the numbers of posts on several consecutive days were zero and realized that C was abroad and did not post any content on those days. Participant D distinguished the visualization of herself mainly by noticing a more frequent appearance of the topic "shopping" in each month as well as several small distinctive keywords, which came up in the detailed view of monthly keywords.

The result of the blind test demonstrates that our prototype system represents the status of microblog users well.

5.3 User feedback and discussion

After the blind test, we recruited 6 more participants to use our prototype system. LifeCircle was received well by all the 10 participants. Participants were interested in its ability to show an intuitive visualization and to allow them to interact with different elements.

While we expected to get feedback on the system's

usability, participants were more inclined to talk about the nostalgia caused by the visualization. One participant commented, "Seeing all the little tiny things that I have talked about in this way was a kind of euphoria. It seemed like a photo album of my online activities." Most of our participants said the contrast between everyday state and extraordinary events in their life was nicely illustrated in the visualization. One participant commented, "In the visualization area of November 2011, a musician's name came up with big size. I like him so much that I had wrote a number of messages talked about his performance in November. It was a memorable experience and the visualization also showed the positive sentiment of mine." These feedbacks indicate the ability of our prototype to help a viewer to easily and quickly look back on his life.

In addition to the pleasure felt when one has a tangible depiction of what he already knew, the participants were fascinated by the ability of LifeCircle to discover what they did not remember and bring unexpected surprises. One participant commented, "The keyword 'job hunting' came up in my visualization. I felt curious as I hardly paid any attention to subjects related to job or employment. Then I clicked on this keyword and found that a PhD student of my lab had been looking forward to a postdoctoral position during that time and I reposted this message with additional comment." Another participant commented that he observed a similar situation: "I saw an unusual pick of outermost histogram stream, I wondered why I made so many posts on that day. I clicked on the histogram and the posts shown made me remember that I had argued with a friend on Weibo by reposting every reply of mine. I felt ashamed and realized I should apologize to my friend." Eight participants described similar situations, when they focused on keywords that seemed out of place and finally found that they had forgotten certain events involving these words. Such situations demonstrate that viewers can easily and freely find interesting and memorable bits of data, even though LifeCircle was not designed for a query task.

As a result, the participants agreed on the necessity and usefulness of forming such an overall view of their life status. One participant pointed out that there is a need to review one's past time and then adjust somehow to a better lifestyle, because of the quickening pace of life and the information overload problem nowadays. He appreciated our prototype with

its ability to show an overview of the important aspects of his past life in an intuitive manner.

On the other hand, not all feedback was positive. One participant commented, “Sentiments are typically rather fluent and may vary a lot over the period of one month, so I don’t think a user’s affective state can be summarized on a monthly basis.” However, it can be improved by showing the sentiments on a daily or individual basis according to the emotion values that we have already calculated for each microblogging post.

Participants also suggested some new functional features, and the most frequently mentioned one were for an option to adjust the time scale. LifeCircle displays a variety of information over time; thus, temporal rhythms are an important aspect of consideration. To enable scalability with respect to time, the angle of sectors and the size of keywords are considered to vary in proportion to the time interval. Sectors can be expanded for obtaining detailed information as described in Section 4.2, when dealing with an extraordinarily long period of time. Further, some participants also wanted more related context, e.g., geographic information, to recall the exact memories.

6 Conclusions and Future Work

In this paper, we presented an interactive time-aware visualization prototype, LifeCircle, to help microblog users to visually retrospect their life statuses. By exploring and analyzing long-term microblogging post data, we summarized annual topics, monthly keywords, monthly sentiments, and microblog activity to generate a radial visualization for a certain microblog user. We implemented a prototype system for Sina Weibo, the largest Chinese microblogging service. Several use cases demonstrated the usefulness of our prototype. Satisfactory results showed its value for nostalgia based on reminiscence. The prototype system also enabled one to drill down to the details of certain events, which helped the microblog users to better understand themselves and others.

The positive feedback on the prototype has prompted us to explore novel visualization approaches of microblogging post contents even further. We consider the following work as our future research plan:

(1) We will make our prototype configurable on a certain time scale.

(2) In addition to monthly sentiments, we will show fine-grained sentiments with respect to keywords and posts to make our prototype more analytical and interactive.

(3) We will continue our work and deploy our system on the Web to make it available to the public and further evaluate the utility and get more feedback on helping portray the users’ life status.

(4) We plan to introduce more textual information, e.g., hierarchical topics, and social information, e.g., interactive behaviors of microblog users, into our visualization in a reasonable way.

Acknowledgements

This research is supported by the National Natural Science Foundation of China (Nos. 61170196 and 61202140), and by the Singapore National Research Foundation under its International Research Centre @ Singapore Funding Initiative and administered by the IDM Programme Office.

References

- [1] B. Liu, Sentiment analysis and opinion mining, *Synthesis Lectures on Human Language Technologies*, vol. 5, no. 1, pp. 1-90, 2012.
- [2] Z. Liu, X. Chen, and M. Sun, Mining the interests of Chinese microbloggers via keyword extraction, *Frontiers of Computer Science in China*, vol. 6, no. 1, pp. 76-87, 2012.
- [3] D. M. Blei, Probabilistic topic models, *Communications of the ACM*, vol. 55, no. 4, pp. 77-84, 2012.
- [4] F. B. Viegas and M. Wattenberg, TIMELINES: Tag clouds and the case for vernacular visualization, *Interactions*, vol. 15, no. 4, pp. 49-52, 2008.
- [5] G. M. Draper, Y. Livnat, and R. F. Riesenfeld, A survey of radial methods for information visualization, *IEEE Transactions on Visualization and Computer Graphics*, vol. 15, no. 5, pp. 759-776, 2009.
- [6] A. Java, X. Song, T. Finin, and B. Tseng, Why we twitter: Understanding microblogging usage and communities, in *Proc. 9th WebKDD and 1st SNA-KDD Workshop on Web Mining and Social Network Analysis*, New York, USA, 2007, pp. 56-65.
- [7] D. Zhao and M. B. Rosson, How and why people Twitter: The role that micro-blogging plays in informal communication at work, in *Proc. 15th ACM Int. Conf. on Supporting Group Work*, Sanibel Island, USA, 2009, pp. 243-252.
- [8] H. Kwak, C. Lee, H. Park, and S. Moon, What is Twitter, a social network or a news media? in *Proc. 19th Int. World Wide Web Conf.*, Raleigh, USA, 2010, pp. 591-600.

- [9] B. Krishnamurthy, P. Gill, and Arlitt M, A few chirps about twitter, in *Proc. 1st Workshop on Online Social Networks*, Seattle, USA, 2008, pp. 19-24.
- [10] M. Busch, K. Gade, B. Larson, P. Lok, S. Luckenbill, and J. Lin, Earlybird: Real-time search at Twitter, in *Proc. 28th Int. Conf. on Data Engineering*, Washington, USA, 2012, pp. 1360-1369.
- [11] J. Teevan, D. Ramage, and M. R. Morris, #TwitterSearch: A comparison of microblog search and web search, in *Proc. 4th Int. Conf. on Web Search and Data Mining*, New York, USA, 2011, pp. 35-44.
- [12] Z. Qu and Y. Liu, Interactive group suggesting for Twitter, in *Proc. 49th Annual Meeting of the Association for Computational Linguistics*, Portland, USA, 2011, pp. 519-523.
- [13] T. Sakaki, M. Okazaki, and Y. Matsuo, Earthquake shakes Twitter users: Real-time event detection by social sensors, in *Proc. 19th Int. World Wide Web Conf.*, New York, USA, 2010, pp. 851-860.
- [14] A. Culotta, Towards detecting influenza epidemics by analyzing Twitter messages, in *Proc. 1st Workshop on Social Media Analytics*, New York, USA, 2010, pp. 115-122.
- [15] S. Petrovic, M. Osborne, and V. Lavrenko, Streaming first story detection with application to Twitter, in *Proc. 11th Annual Conf. of the North American Chapter of the Association for Computational Linguistics*, Los Angeles, USA, 2010, pp. 181-189.
- [16] H. Saif, Y. He, and H. Alani, Alleviating data sparsity for Twitter sentiment analysis, in *Proc. 21st Int. World Wide Web Conf. Workshop on Making Sense of Microposts*, Lyon, France, 2012, pp. 2-9.
- [17] L. Jiang, M. Yu, M. Zhou, X. Liu, and T. Zhao, Target-dependent Twitter sentiment classification, in *Proc. 49th Annual Meeting of the Association for Computational Linguistics*, Portland, USA, 2011, pp. 151-160.
- [18] K. L. Liu, W. J. Li, and M. Guo, Emoticon smoothed language models for Twitter sentiment analysis, in *Proc. 26th AAAI Conf. on Artificial Intelligence*, Toronto, Ontario, Canada, 2012, pp. 1678-1684.
- [19] E. Kouloumpis, T. Wilson, and J. Moore, Twitter sentiment analysis: The good the bad and the OMG!, in *Proc. 5th Int. AAAI Conf. on Weblogs and Social Media*, Barcelona, Spain, 2011, pp. 538-541.
- [20] H. Chien-Tung, L. Cheng-Te, and L. Shou-De, Modeling and visualizing information propagation in a microblogging platform, in *Proc. 3rd Int. Conf. on Advances in Social Networks Analysis and Mining*, Kaohsiung, Taiwan, China, 2011, pp. 328-335.
- [21] C. Li, T. Kuo, C. Ho, S. Hong, W. Lin, and S. Lin, Modeling and evaluating information propagation in a microblogging social network, *Social Network Analysis and Mining*, doi:10.1007/s13278-012-0082-8.
- [22] A. Pal and S. Counts, Identifying topical authorities in microblogs, in *Proc. 4th Int. Conf. on Web Search and Data Mining*, New York, USA, 2011, pp. 45-54.
- [23] K. Gimpel, N. Schneider, B. O'Connor, D. Das, D. Mills, J. Eisenstein, M. Heilman, D. Yogatama, J. Flanigan, and N. A. Smith, Part-of-speech tagging for Twitter: annotation, features, and experiments, in *Proc. 49th Annual Meeting of the Association for Computational Linguistics*, Portland, USA, 2011, pp. 42-47.
- [24] T. Finin, W. Murnane, A. Karandikar, N. Keller, J. Martineau, and M. Dredze, Annotating named entities in Twitter data with crowdsourcing, in *Proc. 11th Annual Conf. of the North American Chapter of the Association for Computational Linguistics Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, Los Angeles, USA, 2010, pp. 80-88.
- [25] X. Liu, S. Zhang, F. Wei, and M. Zhou, Recognizing named entities in tweets, in *Proc. 49th Annual Meeting of the Association for Computational Linguistics*, Portland, USA, 2011, pp. 359-367.
- [26] A. Ritter, S. Clark, Mausam, and O. Etzioni, Named entity recognition in tweets: An experimental study, in *Proc. 2011 Conf. on Empirical Methods on Natural Language Processing*, Edinburgh, UK, 2011, pp. 1524-1534.
- [27] B. Han and T. Baldwin, Lexical normalisation of short text messages: makn sens a #twitter, in *Proc. 49th Annual Meeting of the Association for Computational Linguistics*, Portland, USA, 2011, pp. 368-378.
- [28] F. B. Viegas, S. Golder, and J. Donath, Visualizing email content: Portraying relationships from conversational histories, in *Proc. 24th SIGCHI Conf. on Human Factors in Computing Systems*, Montréal, Canada, 2006, pp. 979-988.
- [29] J. Lamping and R. Rao, The hyperbolic browser: A focus+context technique for visualizing large hierarchies, *Journal of Visual Languages & Computing*, vol. 7, no. 1, pp. 33-55, 1996.
- [30] J. Stasko and E. Zhang, Focus+context display and navigation techniques for enhancing radial, space-filling hierarchy visualizations, in *Proc. 6th IEEE Symposium on Information Visualization*, Salt Lake City, USA, 2000, pp. 57-65.
- [31] J. Yang, M. O. Ward, and E. A. Rundensteiner, InterRing: An interactive tool for visually navigating and manipulating hierarchical structures, in *Proc. 8th IEEE Symposium on Information Visualization*, Boston, USA, 2002, pp. 77-84.
- [32] K. Andrews and H. Heidegger, Information slices: Visualising and exploring large hierarchies using cascading, semi-circular discs, in *Proc. 4th IEEE Symposium on Information Visualization, Late Breaking*

- Hot Topics*, Research Triangle Park, NC, USA, 1998, pp. 9-12.
- [33] C. Collins, S. Carpendale, and G. Penn, DocuBurst: Visualizing document content using language structure, *Computer Graphics Forum*, vol. 28, no. 3, pp. 1039-1046, 2009.
- [34] D. A. Keim, F. Mansmann, J. Schneidewind, and T. Schreck, Monitoring network traffic with radial traffic analyzer, in *Proc. 1st IEEE Symposium on Visual Analytics Science and Technology*, Baltimore, USA, 2006, pp. 123-128.
- [35] L. Yarden, J. Agutter, S. Moon, R. F. Erbacher, and S. Foresti, A visualization paradigm for network intrusion detection, in *Proc. 5th IEEE Workshop on Information Assurance and Security*, New York, USA, 2002, pp. 30-37.
- [36] T. Barlow and P. Neville, A comparison of 2-D visualizations of hierarchies, in *Proc. 7th IEEE Symposium on Information Visualization*, San Diego, USA, 2001, pp. 131-138.
- [37] R. Vliegen, J. J. van Wijk, and E. J. van der Linden, Visualizing business data with generalized treemaps, *IEEE Transactions on Visualization and Computer Graphics*, vol. 12, no. 5, pp. 789-796, 2006.
- [38] F. B. Viegas, M. Wattenberg, and J. Feinberg, Participatory visualization with wordle, *IEEE Transactions on Visualization and Computer Graphics*, vol. 15, no. 6, pp. 1137-1144, 2009.
- [39] K. Zhang and M. Sun, A stacked model based on word lattice for Chinese word segmentation and part-of-speech tagging, <http://nlp.csai.tsinghua.edu.cn/thulac>, 2013.
- [40] K. Toutanova, D. Klein, C. D. Manning, and Y. Singer, Feature-rich part-of-speech tagging with a cyclic dependency network, in *Proc. 4th Annual Conf. of the North American Chapter of the Association for Computational Linguistics*, Edmonton, Canada, 2003, pp. 173-180.
- [41] D. M. Blei, A. Y. Ng, and M. I. Jordan, Latent dirichlet allocation, *Journal of Machine Learning Research*, vol. 3, pp. 993-1022, 2003.
- [42] T. K. Landauer, P. W. Foltz, and D. Laham, An introduction to latent semantic analysis, *Discourse Processes*, vol. 25, no. 2-3, pp. 259-284, 1998.
- [43] T. Hofmann, Probabilistic latent semantic indexing, in *Proc. 22nd Annual Int. ACM SIGIR Conf.*, New York, USA, 1999, pp. 50-57.
- [44] P. D. Turney, Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews, in *Proc. 40th Annual Meeting on Association for Computational Linguistics*, Denver, USA, 2002, pp. 417-424.
- [45] J. Read, Using emoticons to reduce dependency in machine learning techniques for sentiment classification, in *Proc. 43rd Annual Meeting on Association for Computational Linguistics Student Research Workshop*, Ann Arbor, Michigan, USA, 2005, pp. 43-48.
- [46] S. Aoki and O. Uchida, A method for automatically generating the emotional vectors of emoticons using weblog articles, in *Proc. 10th WSEAS Int. Conf. on Applied Computer and Applied Computational Science*, Stevens Point, Wisconsin, USA, 2011, pp. 132-136.
- [47] C. Ziemkiewicz and R. Kosara, Preconceptions and individual differences in understanding visual metaphors, *Computer Graphics Forum*, vol. 28, no. 3, pp. 911-918, 2009.
- [48] J. Heer and G. G. Robertson, Animated transitions in statistical data graphics, *IEEE Transactions on Visualization and Computer Graphics*, vol. 13, no. 6, pp. 1240-1247, 2007.
- [49] K. Koh, B. Lee, B. Kim, and J. Seo, ManiWordle: Providing flexible control over wordle, *IEEE Transactions on Visualization and Computer Graphics*, vol. 16, no. 6, pp. 1190-1197, 2010.
- [50] Z. Liu, Y. Zhang, E. Y. Chang, and M. Sun, PLDA+: Parallel latent dirichlet allocation with data placement and pipeline processing, *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, pp. 1-18, 2011.



Jiayu Tang is a master student of the Department of Computer Science and Technology, Tsinghua University. He got his BEng degree in 2010 from the Department of Computer Science and Technology, Tsinghua University. His research interests are information visualization and social computation.



Zhiyuan Liu is a postdoctoral fellow of Tsinghua University. He got his BEng degree in 2006 and his PhD in 2011 from the Department of Computer Science and Technology, Tsinghua University. His research interests are keyphrase extraction, social tag analysis, and social computation. He has participated in 5 research projects funded by NSFC and 863 as principal investigator or project member. He has published over 20 papers in international journals and conferences including ACM Transactions and EMNLP. He was awarded Tsinghua Excellent Doctoral Dissertation in 2011.



Maosong Sun is a professor of the Department of Computer Science and Technology, Tsinghua University. He got his BEng degree in 1986 and MEng degree in 1988 from Department of Computer Science and Technology, Tsinghua University. He got his PhD degree in 2004 from Department of Chinese, Translation and Linguistics, City University of Hong Kong. His research interests include natural language processing, Chinese computing, Web intelligence, and computational social sciences. He has published over 140 papers in academic journals and international conferences in the above fields. He serves as a vice president of the Chinese Information Processing Society, the council member of China Computer Federation, the member-at-large of ACM China Council, the member of

Expert Committee of National Language Resource Surveillance and Research Center, the co-director of Tsinghua University-National University of Singapore Joint Research Center on Next Generation Search Technologies, and the Editor-in-Chief of the *Journal of Chinese Information Processing*.



Jiahua Liu is a PhD student of the Department of Computer Science and Technology, Tsinghua University. He got his BEng degree in 2012 from the Department of Computer Science and Technology, Tsinghua University. His research interests are Chinese computing and social computation.