Visualization of Sina Weibo Propagation and Sentiment Analysis

Yunzhi Ye*, Ying Qian
School of Software Engineering, Chongqing University of Posts and Telecommunications
Graphic and multimedia laboratory, Chongqing, Chinaa
charles_yz@163.com

Abstract. Propagation and sentiment get more attention in social network research, but there is no intuitive way for users to get information directly and quickly. In this paper, combing data analysis and sentiment analysis, we build a system to analyze the Sina Weibo data and provide the results in a visual way. The system is divided into six parts, mainly leading users to find the regulation between time and reposts, the key players in the process of the propagation, the main phrases of the topic and the change in sentiment. Moreover, we provide improved text visualization and tree map to represent the data after text processing. The system contributes to Weibo data analysis and public opinion monitoring.

Keywords: Sina Weibo, visualization, propagation, sentiment.

1. Introduction

As the network goes into every household, people spend a lot of time on social software. They record and share the drops of life, contact with friends, focus on current events and participate in hot topics by using micro-blog. According to incomplete statistics, micro-blog such as Facebook, Twitter, Sina Weibo active users has reached 300 million monthly and data show TB even PB level growth, which means social network platforms occupy an important part in people's daily life. Users have become the recipients of the news, and become the source of the news in We-media era.

In recent years, researchers have setting off a wave of micro-blog data analysis, including content mining[1], user behavior analysis[2], sentiment analysis[3], geographical situational awareness[4]]and some other domains. However, none of them provides an intuitive way to present the results after analyzing. Great information visualization help people easily find the value or potential relation between data.

Here this paper presents a visual system, mainly analyzing the propagation and sentiment properties of Weibo data. The system consists of five parts, including reposts-time visualization, propagation-path visualization, tipping point visualization, content analysis visualization and sentiment visualization. And the system also provides visual analytic operation in the visualization model mentioned above. It enables users to have a better understanding of Weibo and it will help Weibo administrator to monitor public opinion.

The rest of this paper is organized as follows. Section II mentions some related works. Section III introduces method and visual models. Section IV shows the overview of the system. Section V discusses the further work and Section VI concludes the whole work.

2. Related work

2.1 Correlated Visualization Method

Qi Y et al studied the visualization method of media data combined with the media information and introduces text visualization and visualization method by using geographic information[5]. Chien-Tung Ho et al designed system to show the propagation path and social graph through analyzing the degree of propagation[6]. There are still many methods put forward by other scholars, such as real-time scalable analysis and localization of text visualization combined with clustering method[7], text visualization with unique layout[8] and so on.

2.2 Correlated Visualization Application

To get better analysis, good visualization applications are made recently. There is no standard definition for visualizing data, some focus on propagation of Weibo and visual analytic[9], some focus on problems and fictitious information[10] and some focus on the synergistic reaction of visualization[11].

2.3 Weibo Sentiment Analysis

Combined with natural language processing, visualization is popular in sentiment analysis. It helps decision making and opinion monitoring. Vu Dung Nguyen et al developed a framework to see the public sentiment in a geographical region[12]. Zhitao Wang et al used visual technique to study the relationship between sentiment and real life, the public emotions and events[13].

3. Method

3.1 Data Acquisition

Sina Weibo is one of the most popular social platform in China and provides official API for users to get some common data. It is a complex way to get official authorization, so this project changed mind to get data by using crawler technology. Scrapy(a crawler framework), beautifulsoup(an excellent parsing library) and regular expression are used to get and format all the data. Data are stored into mysql database.

One blog is chosen as an example. The content is about that one student was late for one minute of the Chinese GaoKao and was stopped out of the gate. The Fig. 1 shows the original blog.

Fig. 1 Original weibo

After collecting and structuring data, all the data are stored into three tables of the database. Their structures are as follows:

Table 1 Comment table

Comment				
UID(pk)	**content**	**time**	**com_num**	**pra_num**
user id	comment list	the time user remarks	the number of the comment	the number of the praise

Table 2 Repost table

Repost					
UID(pk)	**rep_hie**	**time**	**LUID**	**content**	**pra_num**
user id	repost hierarchy	the time user repost	user id of last hierarchy	repost comment	the number of the praise

Table 3 User table

User					
UID(pk)	**nickname**	**tVIP**	**gVIP**	**head_portrait**	**fans_num**
user id	user's nickname	user's vip type	user's vip grade	User's head portrait picture	the number of fans

In the three tables, this first raw is the name of the table and the second raw represents the properties of the entity with description at the third raw.

3.2 Propagation Analysis and Visualization

Propagation is a main property in social network. We suppose to find a common way or pattern to explain the regulation in the process of the propagation. There are two factors in the process, time and people. We try to figure out how reposts change from time and who are key players.

We built three visual models to analyze the information we got. In model 1, we established x-axis to represent the whole repost period and y-axis to represent the reposts amount per second. Then we outline each point with lines and form an area map, as shown in Fig. 2.

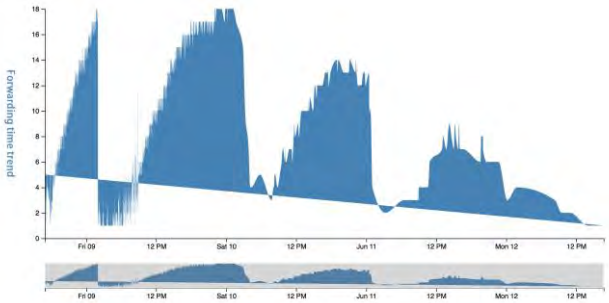

Fig. 2 Repost-time area map

In Fig. 3 and Fig. 4, we zoomed in Fig. 2, and we see more details. Repost decreased sharply during midnight and repost continued in the morning next day. The overall spread showed a downward trend until the heat disappeared.

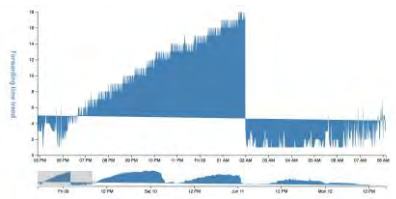

Fig. 3 Repost map1 (zoom in)

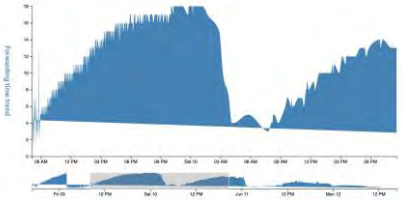

Fig. 4 Repost map2 (zoom in)

Weibo post can be depicted as a repost tree, where every node in the tree corresponds to one user. The original user is considered as the root and the parent node of a certain user is the node that represents the source user it directly reposts from. In Fig. 5, the red node represents the source user, the yellow node represents the repost user and links represent one reposts directly from the other. Considering the problem of layout, we extract some nodes in each layer, but we will take priority to extract nodes with child nodes and the node whose praise amount and repost amount are big enough. In the graph, we select at most 20 nodes and at most 5 nodes for every other level. In Fig. 5, we can clearly see a blog's largest spread level and main users in every level. And we provide drag method for every node for users to have a better view. Meanwhile, we use circular and bar graphs to further analyze each level of data, shown in Fig. 6 and Fig. 7.

Fig. 5 Repost tree map

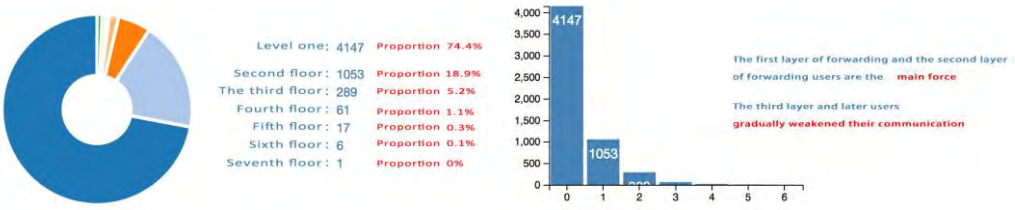

Fig. 6 Circular graph Fig. 7 Bar chart

About the key player model, we call it tipping point. The tipping point is related to the factors of adhesion, environment and character in journalism. In this paper, our focus is on character. User plays a major role when they repost or comment the source Weibo. Once a blog is posted, the fans of the source user are the audience, and people who comment, repost or praise the blog are audience, too. We built a model to calculate the weight of every player. The model expression is:

$$w_i = \begin{cases} \dfrac{r_i + p_i + f_i}{\sum_{i=1}^{g}(r_i + p_i + f_i)}, (t_i = "repost") \\ \dfrac{c_i + p_i}{\sum_{i=1}^{g}(c_i + p_i)}, (t_i = "comment") \end{cases} \quad (1)$$

w_i is the weight of a player in the process of propagation, t_i is the type of content user i has sent. r_i, p_i, f_i, c_i represent the amount of reposts, the amount of praise, the amount of fans and the amount of comments respectively. Then we selected the top 10 players to show in the system.

Fig. 8 Key players

In Fig. 8, we will see who are key players in the propagation clearly. When we move mouse on th e picture of different user, we can get the user’ s information in the center and we can distinguish be tween ordinary user and special user who has a VIP tag.

3.3 Sentiment Detection and Visualization

Sentiment resources, especially sentiment lexicons play a critical role in sentiment analysis. The e xisting Chinese lexicons, such as How-Net, are insufficient and useless for micro-blog. So How-Net are used as basic lexicon and PMI-IR(Point-wise Mutual Information-Information Retrieval) are add ed to enlarge the lexicon[15].

Step1: we use jieba(a technique to cut words) to cut all the text and remove stop words and words which have been already in lexicon, then we get a new text collection and we call it corpus.

Step2: we regard every phrase as a statistical mutual information point, and compute the PMI bet ween words from corpus and lexicon. The expression is:

$$PMI(word_1, word_2) = \log_2\left(\frac{p(word_1 \& word_2)}{p(word_1)p(word_2)}\right) \quad (2)$$

$p(word_1)$ is the frequency of word$_1$ appearance, $p(word_2)$ is the frequency of word$_2$ appearance, $p(word_1 \& word_2)$ is the the frequency of the common occurrence of word$_1$ and word$_2$.

Step3: we compute the similarity between words and positive words or negative words-Semantic Orientation (SO). The expression is:

$$SO(phrase) = PMI(phrase, pword) - PMI(phrase, nword) \quad (3)$$

pword, nword represents positive words and negative words respectively.

Step4: Add reference phrase to a positive dictionary or a negative dictionary.

Then we adapt sentiment analysis based on lexicon to compute every sentence’s value. The algorithm is described as follows:

Step1: read every sentence and cut them into clauses.

Step2: search sematic phrase from positive and negative lexicon and record the polarity and position.

Step3: scan before the sematic phrase to find degree words. If there exist degree words, sematic value multiple the weight of degree words.

Step4: scan before sematic phrase to find negative words such as no, never and record the number of negative words. If it is an odd number, sematic value multiple -1. If it is an even number, sematic value remains unchanged.

Step5: calculate every completed sentence’ s positive variance and negative variance.

Now, we have not only basic data, but also sematic data. For this part, we extract the main phrases from the original Weibo, count the high-frequency words, analyze the change trend of sentiment in the way of propagation and analyze the distribution of sentiment. About main phrases extraction, we used TextRank algorithm. We compute every node’ s weight and select the top 5 phrases. The algorithm is:

$$WS(V_i) = (1-d) * \sum_{V_i \in In(V_i)} \frac{w_{ji}}{\sum_{V_k \in Out(V_j)} w_{jk}} WS(V_j) \tag{4}$$

Then we got the result as the table4 shows.

Table 4 Key phrases

phrase	weight
child	1.0
don’t cry	0.6558805506053819
approach into the gate	0.6114534117801246
go to take an exam	0.6014580074699839
examinee	0.5998032470244555

Meanwhile, we count the high frequency phrases that appear in the text and display by text visualization. We improved the original wordle layout algorithm to present more words for the reason of that there are some blank space in the Chinese character. We choose the rectangular to present and the text size is calculated according to the weight. The text locates the center expanding to sides. Every time we put a new text into the center of the rectangle, then we determine whether there is a position conflict. If there is space, the text will be placed. If there is any conflict, the text will be placed around the scope of the conflict. Our result is shown in Fig. 9.

Fig. 9 Text visualization

About the key phrases, we did another thing. We select the top 10 high-frequency phrases, and find their most similar words in the corpus. At first, we trained the corpus to a model by using word2vec.

Then we call the most similar method in the word2vec to find the most similar phrases in the corpus by loading the model. Then we use bubble chart to present these data. In the Fig. 10 the same color bubbles stand for the same kind. The size of the word also depends on the word frequency.

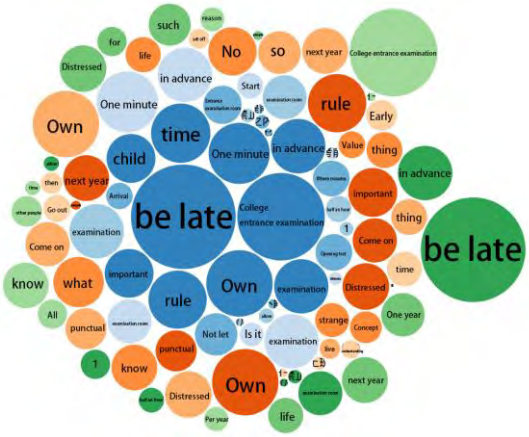

Fig. 10 Bubble chart

As for the sematic visual model, we designed three models to display all the data. In the first model, we want to represent that what is the sentiment in the crowd after the micro-blog is sent. We regard a point as a user, and the color of the point represents the sentiment value. We map the positive and negative sentiment to the red and blue channels of the RGB color space respectively. Animation helps user clearly know how many for and against in the crowd, shown in Fig. 11.

Fig. 11 Sematic analysis visualization

Because there are too many points and, point is too small and the difference of sentiment value is small, it is not easy to see the sematic differences. For the purpose of show the differences in sentiment, we built block map. In the block map, we count the frequency of different sentiment and lay the high-frequency sentiment in the center expanding to sides, shown in Fig. 12. Then we will clearly see the emotional level.

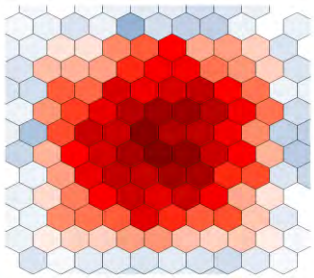

Fig. 12 Block map

We also pay attention to the sematic changes of micro-blog in the process of propagation. On the foundation of tree map, we add color channel to every node. The node's sentiment is opposite to its parent node and all its child nodes' sentiments are in the same trend, and we call it sentiment mutational point (SMP). Every SMP is an object to get focused and monitored because it will lead to a turn of public opinion. In the Fig.13, we can find two SMPs. we hide the nickname of the users and stroke the color only in red and blue color. When we move mouse on the node, and we will see the nickname.

Fig. 13 Sematic tree map

4. System Overview

Our system has six major components: data load modular, repost-time visual analysis, propagatio n-path visual analysis, tipping point analysis, content visual analysis and sentiment visual analysis.

The online interface, based on HTML and D3.js, is an intuitive way for users to get information e asily. Interactive operation provided in each part help user understand deeply. The system overview i s shown in Fig. 14.

Fig. 14 System overview

5. Further Work

To further complete the system, we look forward to getting more data of a blog and analyzing We ibo data more comprehensively. Also, we are going to complete the expert system by collecting user s action in the system to provide better visualization.

6. Conclusions

In this work, combined with nature language processing, we analyzed how a blog propagate to the public and what the sentiment is in the crowd. And we built a visual system to present the analysis results online. The system is formed in five parts, mainly analyze the propagation and sentiment properties by using visual technique.

References

[1]. Carter S, Weerkamp W, Tsagkias M. Microblog language identification: overcoming the limitations of short, unedited and idiomatic text[J]. Language Resources & Evaluation, 2013, 47(1):195-215.

[2]. Oulasvirta A,Lehtonen E，Kurvinen E，et al． Making the Ordinary Visible in Microblogs［J］. Personal and ubiquitous computing，2010，14(3) : 237 －249.

[3]. Zhang L, Jia Y, Zhou B, et al. Microblogging sentiment analysis using emotional vector[C]//Cloud and Green Computing (CGC), 2012 Second International Conference on. IEEE, 2012: 430-433.

[4]. A. M. MacEachren, A. Jaiswal, A. C. Robinson, S. Pezanowski, A. Savelyev, P. Mitra, X. Zhang, and J. Blanford. SensePlace2: GeoTwitter analytics support for situational awareness. In Proceed- ings of IEEE Conference on Visual Analytics Science and Technology, pages 181 –190, 2011.

[5]. Qi Y, Shi G, Yu X, et al. Visualization in media big data analysis[C]//Computer and Information Science (ICIS), 2015 IEEE/ACIS 14th International Conference on. IEEE, 2015: 571-574.

[6]. Ho C T, Li C T, Lin S D. Modeling and visualizing information propagation in a micro-blogging platform[C]//Advances in Social Networks Analysis and Mining (ASONAM), 2011 International Conference on. IEEE, 2011: 328-335.

[7]. Thom D, Bosch H, Koch S, et al. Spatiotemporal anomaly detection through visual analysis of geolocated twitter messages[C]//Pacific visualization symposium (PacificVis), 2012 IEEE. IEEE, 2012: 41-48.

[8]. Viegas F B, Wattenberg M, Feinberg J. Participatory visualization with wordle[J]. IEEE transactions on visualization and computer graphics, 2009, 15(6).

[9]. Ren D, Zhang X, Wang Z, et al. Weiboevents: A crowd sourcing weibo visual analytic system[C]//Visualization Symposium (PacificVis), 2014 IEEE Pacific. IEEE, 2014: 330-334.

[10]. Yu L, Asur S, Huberman B A. Artificial Inflation: The True Story of Trends in Sina Weibo. arXiv preprint[J]. arXiv, 2012, 1202.

[11]. Willett W, Heer J, Hellerstein J, et al. CommentSpace: structured support for collaborative visual analysis[C]//Proceedings of the SIGCHI conference on Human Factors in Computing Systems. ACM, 2011: 3131-3140.

[12]. Nguyen V D, Varghese B, Barker A. The royal birth of 2013: Analysing and visualising public sentiment in the uk using twitter[C]//Big Data, 2013 IEEE International Conference on. IEEE, 2013: 46-54.

[13]. Wang Z, Yu Z, Chen L, et al. Sentiment detection and visualization of Chinese micro-blog[C]//Data Science and Advanced Analytics (DSAA), 2014 International Conference on. IEEE, 2014: 251-257.

[14]. Turney, Peter D. Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL[C]// Proc. European Conference on Machine Lerning. 2001:491-502.