Reading-Weibo: A Sina Weibo Oriented Data Mining System

Bin Liu[a], Jingyuan Zhang[b], Qiang Liu[c], Han Li[d], Mingliang Zhang[e], Rui Qiu[f], Jingyang Zhao[g]

School of Economics and Management, Hebei University of Science and Technology, Shijiazhuang, 050018, China

[a]email: triba@126.com, [b]email: jinlri248@163.com, [c]email: lqdarg@163.com, [d]email: viplihan@163.com, [e]email: zmldarg@163.com, [f]email: qrdarg@163.com, [g]email: zjzdarg@163.com

Keywords: Sina Weibo; Social Network; Data Mining; Reading-Weibo

Abstract. Sina Weibo, as the most popular Chinese microblogging social network, has huge amounts of information that reflects all aspects of our social life. For discovering the hidden patterns or rules, this paper presents a Sina Weibo oriented data mining system named as Reading-Weibo. The system can conduct an on-line, multi-dimensional (nine dimensions), interactive and dynamic analysis on the transmission of a single microblog. The workflow of Reading-Weibo and its sentiment analysis module, the regional distribution module and propagation tree module are introduced with examples. Finally, we put forward our views on the future work in microblog mining field.

Introduction

Nowadays, the social media (e.g. microblog, social network sites, the open encyclopaedia and video sharing sites) provide information interchanging platforms that deeply affects people's work and life [1]. Until June 2014, the number of monthly active users of Twitter and Sina Weibo has reached 271 million and 143 million, respectively. Microblog has accumulated considerable user generated content (UGC), which needs effective data mining methods to discover the hidden knowledge and improve the efficiency of social production and life. For example, microblog based communication can improve the product image [2], and ultimately affect the consumption behaviour [3] (e.g. the prediction of box office [4]); microblog based customer relationship management (CRM) can improve the service and profit [5].

In this pare, we present our self-developed Weibo mining system named as Reading-Weibo, which puts focus on mining the transmission of a single microblog. Compared with other similar systems (e.g. Doodod, WeiboReach, Yeezhao and SocialDrip), Reading-Weibo can help users to analyze the microblog transmission in an on-line, multi-dimensional (nine dimensions), interactive and dynamic way. The coming content is structured as follows: the hot issues of current microblog mining field are firstly introduced. Then, Reading-Weibo and its main modules: sentiment analysis, regional distribution and propagation tree are given with examples. Finally, we summarize the challenging problems from the applications aspect.

Hot issues of microblog mining

Microblog user classification, which can be conducted according to the following two conditions: 1) the microblog user list [6]. The users are roughly divided into two categories: Elite users, who follow a small number of users but have the huge number of fans; Ordinary users, who follow and his fans are generally acquaintances in real life, hence the number of fans is limited. 2) The intent of using microblog [7]. The users are commonly divided into three categories: Information sources, who has a large number of followers; Friends, who may have friends, family and colleagues on their friend lists. Unfamiliar users may add someone as a friend at sometimes; Information Seeker, who is an information seeker might post rarely, but browses other users regularly.

Sentiment analysis, which is effective for guiding public opinion [8]. The text content of microblog often contains some emoticons, so we can analysis microblog emotional tendencies by using these emoticons. Ref. [9] calculated microblog sentiment index by defining the attitudinal words and establishing weighting dictionary, negative dictionary and interjection dictionary. However, the machine learning methods are not restricted by the specification and update of the dictionary, so its application is more widely.

Reading-Weibo, a microblog mining system

In this section, we propose a system for microblog dissemination analysis called Reading-Weibo (http://115.29.47.48/analysis_wb). It analyzes the dissemination of a single microblog from time, space, emotion and sex by diverse means. As Figure 1 shows, the first act is information collection for a complete system of microblog analysis, and we obtain information data through the application programming interface (API). Next step is preprocessing, including: using Ansj for word segmentation and removing stopwords. Ansj is an open source tool of Java Chinese word segmentation base on ICTCLAS of Chinese academy of sciences [10]. The processed data is stored in the Redis database, and the visualization is shown through the front-end web page after format transformation. The main modules include original microblog, sentiment analysis, and tag clouds, forwarding time analysis, statistical analysis, regional distribution, propagation graph and hierarchical analysis. The following will discuss the sentiment analysis module, regional distribution module and propagation graph module in detail.

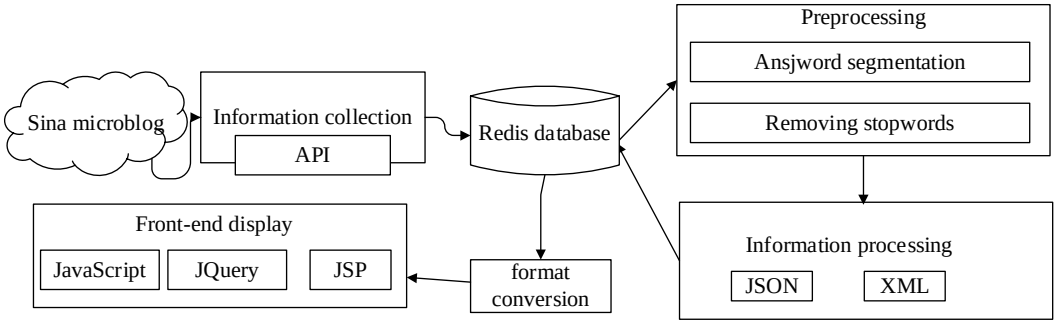

Fig.1. Workflow of Reading-Weibo.

Sentiment analysis module. Sentiment analysis module analyses forward and direct comments of users, thereby grasp user's emotional tendency. This system first uses API provided by Sina Weibo to access information resources. Next step is word segmentation and removing stopwords. And then a series of processed phrases is classified by sentiment lexicon and divided into positive words and negative words. Neutral words are not positive words or negative words. Besides, this system makes a statistic of the proportion of three kinds of emotional words in all the words, and finally gets the statistical chart of emotion. As shown in Figure 2, the ring on the left side is the statistics of emotional tendency，and the right side is the photo wall which shows 20 users containing neutral words.

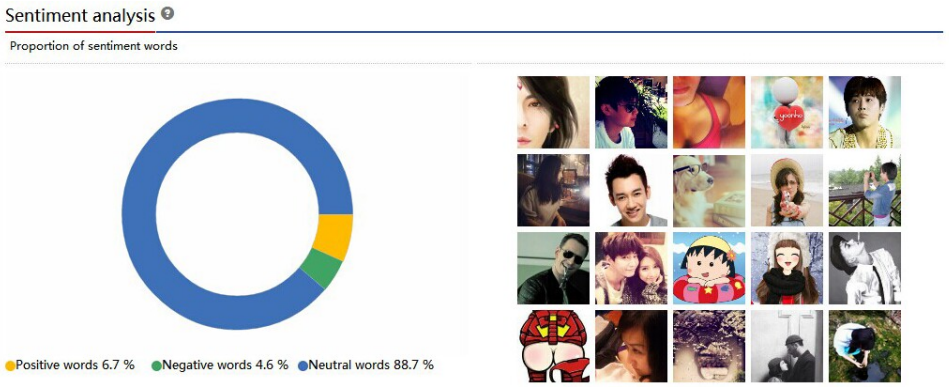

Fig.2. Sentiment analysis module.

Regional distribution module. Regional distribution module shows the distribution of forwarding in nationwide, and reflecting each forwarding user's contribution to the microblog information diffusion. This system analyzes regional information which extract from database after receiving the request of the regional distribution module, and transforms analysis results into JavaScript object notation (JSON). The JSON files are parsed and returned to the system user. As shown in Figure 3, each block on the map represents a province, and the deeper the color the more forwarding users the block represents. The default setting of the right side is the Weibo user who brings the largest number of forwarding.

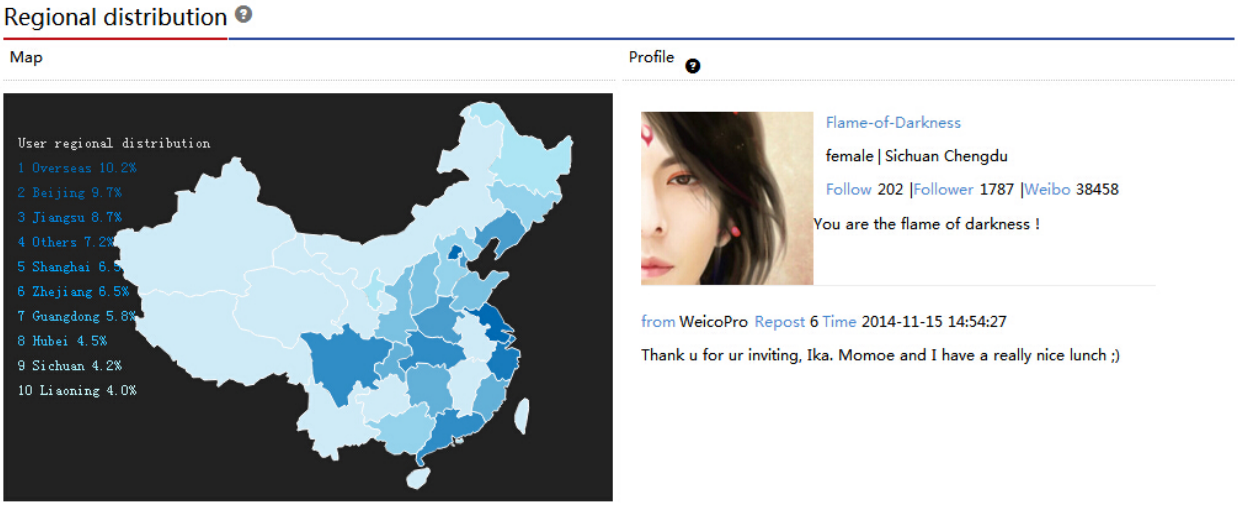

Fig.3. Regional distribution module.

Propagation graph module. This module occupies the important position in this system. It clearly displays the people's complex transmission relationships with the data visualization technique. The process is as follows: extracting the transmission relationships among microblog users and storing them into a XML file, then using the front-end visualization technology to extract and display the transmission relationships to the web page. As shown in Figure 4, each point of the graph represents a microblog forwarding user, and the lines represent the forwarding relationships between users. The biggest circular cluster, which is closest to original point, represents the direct forwarding users and other points are indirect forwarding users. We can clearly observe the process of information diffusion through layer upon layer forwarding.

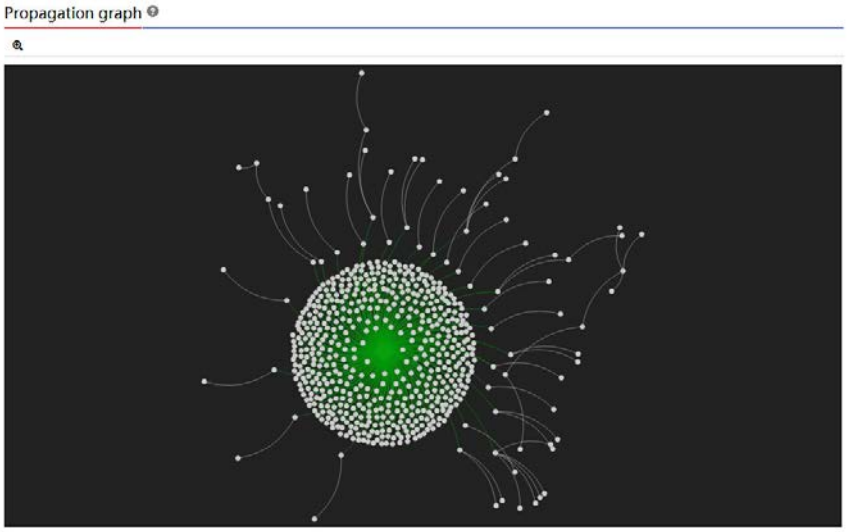

Fig.4. Propagation graph module.

Conclusion

This paper gives the hot issues of microblog mining and proposes Reading-Weibo, our self-developed microblog mining system. The future work of this field can be predicted from the

applicable aspect: 1) event prediction and management. Event prediction is still the major problem, but the accuracy of results remains to be improved. By research, event prediction of microblog can further strengthen individual or government's monitoring, and then improve the initiative and controllability; 2) commercial application. Microblog contains lots of business information because of a mass of users. Such as finding users' purchase intention, we can recommend products to them; by analyzing the comments, the decision makers can improve the market decision-making.

Acknowledgement

This work is supported by the National Natural Science Foundation of China (71271076); Statistical Scientific Research Projects of Hebei Province (2013H210); the Open Foundation of Five Platforms of Hebei University of Science and Technology (WH03); the National College Students' Innovative Entrepreneurial Training Program (201410082001); the Key Support Project for Graduate Education Teaching Reform of Hebei University of Science and Technology.

References

[1] Kaplan M. Kaplan, Haenlein Michael. The early bird catches the news: nine things you should know about micro-blogging [J]. Business Horizons, 2011 54 (2) 105-113.

[2] Hsu Chien-Lung, Liu Chia-Chang, Lee Yuan-Duen. Effect of commitment and trust towards microblogs on consumer intention [J]. International Journal of Electronic Business Management, 2010 8 (4) 292-303.

[3] Jansen Bernard J., Zhang Mimi, ET al. Twitter power tweets as electronic word of mouth [J]. Journal of the American Society for Information Science and Technology, 2009 60 (11) 2169-2188.

[4] Du Jingfei, Xu Hua, Huang Xiaoqiu. Box office prediction based on microblog [J]. Expert Systems with Applications, 2014 41 (4) 1680-1689.

[5] Malthouse Edward C., Haenlein Michael, Skiera Bernd, ET al. Managing customer relationships in the social media era: Introducing the social CRM house [J]. Journal of Interactive Marketing, 2013 27 (4) 270-280.

[6] Wu Shaomei, Hofman Jake M., MasonWinter A., ET al. Who Says What to whom on twitter. Proc. of the 20th Int. World Wide Web (WWW) [C]. New York: ACM Press, 2011. 705-714.

[7] Java Akshay, Song Xiaodan, Finin Tim, ET al. Why we twitter: Understanding microblogging usage and communities. Proc. of the 9th WebKDD and 1st SNA-KDD workshop on Web mining and social network analysis [C]. New York: ACM Press, 2007 56-65.

[8] Bollen Johan, Pepe Alberto, Mao Huina. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. Proc. of the 5th Int. AAAI Conf. on Weblogs and Social Media [C]. Washington DC: AAAI Press, 2011 450-453.

[9] Shen Yang, Li Shuchen, Zheng Ling, ET al. Emotion mining research on micro-blog. 1st IEEE Symposium on Web Society (SWS) [C]. Washington DC: IEEE Computer Society, 2009 71-75.

[10] Li Xiangdong, Zhang Cheng. Research on enhancing the effectiveness of the Chinese text automatic categorization based on ICTCLAS segmentation method. 4th IEEE International Conference on Software Engineering and Service Science (ICSESS) [C]. Washington DC: IEEE Computer Society, 2013 267-270.