Attitudes of online users towards personal information leakage: based on Sina Weibo database

Wang Min[1,*], Zhilong You[1]

[1] College of Economics and Management, Shandong University of Science and Technology, Qingdao, China

Abstract. With the rapid development of internet, people pay more attention to personal information security. Drawing upon three components of attitude, this study was designed to realize online users' attitudes toward personal information leakage. Web crawl program was used to get the blog data from Sina Weibo. Results show that the main media for personal information leak include mobile phones, telephone, media and networks. People who have verification published blog more than no verification people. People pay more attention to account number. Pioneer and No verification people have more negative affection. Personal account, VIP and Organization people have more positive affection. If the blog has higher interaction, positive affect will also rise. Media's blogs exert an imperceptible influence on people's behavior.

1 Introduction

In recent years, with the rapid development of Internet, the number of Internet users has increased dramatically. According to the 45th China Statistical Report on Internet Development, by the end of April 30 , 2020. The number of Chinese Internet users had reached 904 million, Internet penetration rate was 64.5%[1]. With the rapid development of Internet, the problem of personal information leakage has become the social focus[2].

There are lots of research on person information security. It mainly focuses on the definition of privacy information and privacy concerns, exploring the factors and consequences of user privacy concerns for a special industry. The mainly research method is empirical study by questionnaire[2]. Information security has become an important issue which is related to people's life. However, the users' attitude to personal information leakage, which has not been addressed adequately thus far.

Recently, social media platforms have become a source of people to record their attitude about happenings in their life[3]. The wide use of the Internet has led to the rapid development of social media such as blogs, forums and social networks. Sina Weibo is a leading blog platform in China. According to the Sina Weibo Reports Second Quarter 2020. Unaudited Financial Results, monthly active users had a net addition of approximately 70 million users[4]. Micro-blog plays an important role in interaction and content distribution. The social media represented by micro-blog has produced a large number of users' subjective texts. It is called "user generated content"[5]. This kind of content contains the user's sentiments and attitudes.

Therefore, this paper based on Sina Weibo. Web crawl program was used to get user's blog content. We try to solve the following questions:

Q1. What are the users' 'attitude to personal information leakage?

Q2. Is there any difference between different kind of user's attitude?

2 Figures and tables

With the development of Internet, users' information protection has become a research hotspot. Researchers pay more attention to the user's behavior and attitude about information security[6]. Researcher suggests that YAHOO, WeChat, Viber or any other communications tools should protect users' information[7]. Ku and Yc found that users' privacy information concerns will affect their willingness to continue to use social networks[8].

Personal information, including name, age, address, family[9] and other demographics[10]. In big data era, concern is no longer limited to traditional items of personal information not only demographics, but also increasingly about behavior. In the process of mobile payment or online shopping, we provide personal information to get service or products. Browsing history, shopping data, or purchase tracking could be leaked[6]. In many cases, users are willing to send personal information to exchange for useful services. Malware will attack mobile devices to get personal information[7].

However, different people have different cognitive about personal information leakage[8]. Even since human get into big data era, great progress has been made in scientific research area. Scholars have published articles on Science, showing the use of Internet data. Internet data can help us to study human social behavior and the rules of social operation[7]. So, we use Internet data to analysis people's attitude about personal information leakage.

Attitudes have three components. The ABC of attitudes is made up of affect (feelings), behavior (responses) and

* Corresponding author: youzhilong23@163.com

cognitions (thoughts)[6].

Cognition (thoughts) including the fact, understanding, and evaluation of an individual's perception of attitude objects. Cognitions is a basic component.

Affect (feelings), it refers to an emotional experience held by an individual towards an object. Such as respect and contempt, likes and dislikes, sympathy and sarcasm. Behavior(responses), it refers to an internal reaction tendency held by an individual towards the attitudinal object.

It is coherent between the three components of attitude. The results show that the correlation between emotion and behavioral is higher than that between cognition and emotion, cognition and behavior. Cognitive is more independent and have less interaction with the other two components[7].

In this research, we use web crawl program to get the blog for cognition analysis. Together with the affect expressed by the blogger, we are also able to characterize the degree of the depression expressed in the text as negative or positive. Then, we do regression analysis to realize the relationship between affect and behavior.

3 Method

3.1 Data collection

This research uses "Sina Weibo" (PC terminal) as the data source platform. We used the keywords "personal information leak" to search the public blog information. We got 757 relative blog information. The data collection time is July 12, 2020. The publication time of blog is from June 2, 2020 to July 12, 2020.

3.2 Attitude analysis

Analyzing the words frequency in the material can help us understand the important words quickly. And realize the users' thoughts and understanding about information leak. We use word segmentation software to segment blog information we collected from Sina Weibo. This paper used Nvivo11 and Atlas. Ti software to analyze the word frequency of blog data.

Affect analysis refers to the analysis the emotion of the texts. Affect analysis mainly includes:

a. Tendency analysis. That is, the user emotion expressed in the text is positive, negative or neutral.

b. Degree analysis. That is, the intensity of the emotion expressed in the text, such as "hate" is deeper than "dislike".

c. Subjective and objective analysis. That is, whether the users' text is reflects real events objectively or not. Most micro-blog users publish subjective comments.

This paper focuses on tendency analysis. There are two main methods of affective analysis:

a. Dictionary-based method. It mainly analysis data through a series of emotional dictionaries and rules. Splitting paragraph to analysis the grammar. Then use encodings and algorithms to calculate the emotional value. However, this method is only suitable for a small number of text analysis.

b. Machine-based learning. Machine learning mainly relies on the classification method in machine learning to analyze the emotions in the text, which mainly includes two main steps: the first step is to extract the emotional information of subjective text through feature construction technology; the second step is to use classification technology to dig the emotional information contained in these texts.

This method is suitable for a large number of text analysis[8].

Sina Weibo is a platform for users to publish micro blog. The maximum number of blog words is up to 2000. It belongs to short text analysis. The number of our sample is 645 blog text, so it is more appropriate to choose a dictionary method to analysis affect. Procedure is as follows:

a. Constructed emotional dictionaries firstly. The BosonNLP sentiment dictionary is widely used in micro-blog research papers. So we choose BosonNLP to analyze the text data after word segmentation.

b. Find out affective words, such as "like, hate, boring, interesting…", negative words and adverbs. Then classify the results.

c. Calculate the score.

We define:

W= weight=1

A= the affective value of the affective word.

S=W*A= affective score

N= the degree value of the adverb

a. Firstly, we define W as a weight. The initial weight W is set to 1, starting from the first affective word.

b. Then find out if there are negative word or adverb before the next affective word. If there is a negative word, then W * - 1. If there is a adverb, W *N.

c. Do cycle calculation, and summarize the score S.

User's behavior is the process of identifying and responding to blog content, then implement the action. In Sina Weibo, user can click the button such as "Like", "Comment" or "Forward". We define interaction as the sum of "Like + Comments+ Forwards". Then calculate interaction to do further analysis.

4 Results

4.1 Quantitative statistics

After crawling the data, it was cleaned by removing unnecessary words like modal words in Chinese and irrelevant blogs, we got 643 useful blogs for further analyze. The sample of blogs is shown in Table 1.

4.2 Cognitive analysis

According to the cognitive characteristics of Internet users, we will focus on four dimensions: main body, behavior, media and content. This article attempts to explore the following questions: who is the subject of information leak? Which behavior will cause personal information leak? What is the main way to the leakage of information? What is the main content of information leak? The results are shown in Table 2 by cleaning out irrelevant words such as

mood words, adverbs, and concept words such as "personal information leakage".

Table 1. The number of useful blogs

Verification	Members	Percent	Blogs	Percent
Pioneer	5	0.86	5	0.78
Personal account	93	15.92	96	14.93
VIP	73	12.5	73	11.35
Organization	186	31.85	238	37.01
No verification	227	38.87	231	35.93
Sum	584	--	643	--

Table 2. Word frequency

No.	Words	Frequency	Percent%
1	Mobile phone	218	0．82
2	Social	193	0.73
3	Phone	146	0.55

From Table 2, we can know that:

a. There are two main bodies of personal information leak, that is, users and users' privacy infringer.

b. The keywords about users' behavior include "Replace" and "Conveniently". It shows that users' awareness of protecting personal information is not strong enough. Infringer can get information easier.

c. The main media for personal information leak includes mobile phones, telephone, media and networks. The industry is mainly related to social and bank.

d. In many kinds of information， people pay more attention to account number.

4.3 Affect analysis

Affect including emotion, is a component of attitude. Emotion is the result of the inner mental activities of the users. In this study, we have mined and analyzed the user's emotional to solve two problems.

What is the user's affect to the personal information leak? Is there any difference between different kind of verification user? Result is shown in Table 3.

In order to study the different affect between different kinds of verification users. In this paper, we get the following nuclear density maps to show affective value clearly.

Table 3. Affective score

Verification	Obs	Mean	Std. Dev.
Pioneer	5	-.050	0.333
Personal account	96	0.096	0.366
VIP	73	0.036	0.345

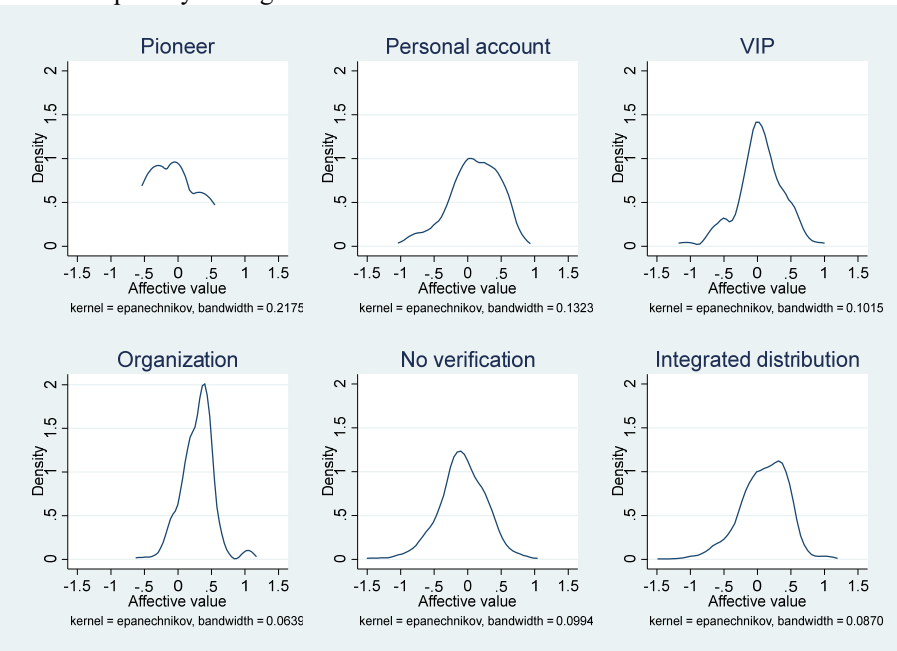

Fig. 1. Nuclear density maps of affective value

From the Figure 1, we can know that:

a. People who have verification published blog more than no verification people.

b. The standard deviation of Organization's blog is small, indicating that the attitude of Organization is more centralized and unified.

c. The mean affective value of Pioneer and No verification is negative. From Figure1(a and e) the area of negative is bigger. It means Pioneer and No verification people have more negative affection. The mean affective value of Personal account, VIP and Organization is positive. From Figure1(b, c and d), the area of positive is bigger. It means Personal account, VIP and Organization people have more positive affection. From Figure1(f), we can realize a integrated distribution. The area of positive is bigger than negative. The visual processing of object shape showing the right lateralization. The affective value of the whole blog is positive and rational.

4.4 Behavior analysis

We define that the act of forwarding or commenting is defined as interactive behavior.

This study will make a statistical analysis of users' interaction behavior. Solve the following questions: Is there any differences in the interaction behavior between different kind of users? Is there a correlation between affect and interactive behavior?

We define interaction as the sum of "Like + Comments+ Forwards". Table 4 is the top 10 interaction.

Table 4. Top 10 interaction

No.	Verification	Main contents	Interaction
1	Organization	please pay attention to change information if you change the phone number.	9202
2	Organization	you can login to the appoint website or make a call to report.	7959
3	Organization	please pay attention to change information if you change the phone number.	7856
4	Personal account	A famous TV stars' personal information was leaked.	5515
5	Personal account	A famous TV stars' personal information was leaked.	5179
6	Personal account	A famous TV stars' personal information was leaked.	4680
7	Organization	A famous TV stars' personal information was leaked.	4451
8	Organization	People who have bought the famous TV stars' concert tickets should pay attention to protecting personal information.	3627
9	Personal account	A famous TV stars' personal information was leaked.	2509
10	VIP	A famous TV stars' personal information was leaked.	2377

From Table 4, Organizations are the mainstream official media. Their blog ranks the top 3. The content of the blog describes the possible ways to cause personal information leakage. Meanwhile, they remind users to improve their awareness of protecting personal information. Give ways to report illegal activities.

The 4th to 10th blog is all about the information leakage about TV stars and ask to protect the privacy of the TV stars.

The media should form positive guidance to public opinion and disseminate rational information. At the same time, we should pay attention to the solution and solution of "personal information leakage".

At the same time, we should focus on current situation of star effect and its value.

5 Conclusion

This article presented a method which is able to realize people's attitude about personal information leakage[11]. We use the Web Crawler program to get the information about "personal information leakage" from Sina Weibo. For user-generated content, we can further determine the cognitive, affect and behavior[12].

The presented results clearly illustrated the key world to help us know the cognitive. Affect analysis showing that different kind of users has different affect tendency. Mainstream media should play a positive role to help users pay attention to personal information leakage. For users who express negative affect, it is important to focus on the reasons. Find out method to protect users' privacy information. Through the behavior analysis, we find that the interaction number of organizations, especially mainstream media, is bigger than other users. Media's blogs exert an imperceptible influence on people's behavior.

References

1. The 45th China Statistical Report on Internet Development. 2020
2. ITRC's data leak summary report. 2020
3. Gruzd, A., Hernández-García, Á. Privacy Concerns and Self-Disclosure in Private and Public Uses of Social Media. Cyberpsychology, Behavior, and Social Networking, **21**, 7(2017)
4. Syn S Y, Oh S. Why do social network site users share information on Facebook and Twitter?. Journal of

Information Science, **41** (2015)

5. Sina Weibo Reports. 2020
6. Fatima, I., Mukhtar, H., Ahmad, H. F., & Rajpoot, K. Analysis of user-generated content from online social communities to characterise and predict depression degree. Journal of Information Science, **44**, 5(2018)
7. Smith, E. R., Mackie, D. M., & Claypool, H. M. *Social psychology*. (Psychology Press, 2014)
8. Ku Y C , Chen R , Zhang H . Why do users continue using social networking sites? An exploratory study of members in the United States and Taiwan[J]. Information & Management, **50**, 7(2013)
9. Preibusch, S., Peetz, T., Acar, G., & Berendt, B. Shopping for privacy: Purchase details leaked to PayPal. Electronic Commerce Research and Applications, **15**, (2016)
10. Kim, Y., Oh, T., & Kim, J. (2015). Analyzing user awareness of privacy data leak in mobile applications. Mobile Information Systems, **20**, 15(2015).
11. Lazer D, Pentland A, Adamic L, et al. Social science. Computational social science. Science, **323** (2009)
12. Symeonidis, S., Effrosynidis, D., & Arampatzis, A. A comparative evaluation of pre-processing techniques and their interactions for twitter sentiment analysis. Expert Systems with Applications, **110** (2018)