Research on Danmaku Knowledge Discovery Service Under Computational Communication

Li Wang[1,2,*] Zhihui Liu[1,2] Hongqi Han[1,2]

[1] *Key Laboratory of Rich-media Knowledge Organization and Service of Digital Publishing Content,China,100038*
[2] *Institute of Scientific & Technical Information of China,China,100038*
[*]*Corresponding author. Email: wl@istic.ac.cn*

ABSTRACT
Danmaku is a new type of user comment data with large amount of data, strong timeliness and close connection with video content. It contains a large amount of explicit and implicit knowledge used in computational communication to analysis user behaviour. This paper uses statistics and topic mining methods to analysis the relationship between the length and quantity of danmaku and the topic. This paper takes the danmaku of HUAWEI P30 mobile phone on BILIBILI website as an example for quantitative empirical research. The results shows that the danmaku is more superficial and popular. The audience is younger, and the culture contains more perceptual elements.
Keywords: *Danmaku, knowledge discovery, LDA, computational communication*

1. INTRODUCTION

As a new data form and source, time-stamped text data such as danmaku records a lot of information that cannot be obtained by traditional social science data acquisition methods. New types of data improve on the shortcomings of social science research.

Communication Research originates from the Computational social science proposed by LAZER [1]in 2009, which uses computing technologies (such as automatic classification, semantic modeling, natural language processing, simulation and statistical model, etc.) to analyse and discovery the knowledge. In 2013, WATTS [2] deepened the concept by introducing cross-platform data to analyse individual actions.

The rise of artificial intelligence has attracted the attention of Communication scholars. The collaborative development of big data and artificial intelligence technology has promoted the transformation of traditional Communication science, and then the concept of Computational Communication Research has emerged [3]. At the same time, artificial intelligence technology for the processing of unstructured data has a natural advantage, which is also a key element to promote computing communication advance with the times.

Along with the emergence of new types of data, a large amount of explicit and implicit knowledge emerges. It is an important subject in computational communication to store and discovery the knowledge, so as to analyse the patterns and rules behind human communication behaviours, as well as the generation mechanism and basic principles.

2. BACKGROUND

2.1. The study of Danmaku under communication

In recent years, with the rapid development of digital media technology, the use of danmaku system in network video has been on the rise. This kind of real-time comment data can be directly displayed on the video interface in the form of sliding subtitles. Danmaku originated from the Japanese video-sharing website "NICONICO". Since its establishment, danmaku system has played an important role in Japanese entertainment, politics and other fields and made remarkable achievements. Many Japanese politicians, including Abe Shinzo, have taken danmaku websites and contents as important positions to attract young people's votes. On August 26th, 2019, the ministry of science and technology, the publicity department of the CPC central committee and other six departments issued a document to promote the in-depth integration of culture and science and technology.

The advantage of danmaku is that the perspective of comment is more micro and more targeted. The disadvantage is that due to the limitation of time, the comments are usually covered by the following content quickly, so the expression is more superficial and popular. At the same time, the danmaku fully releases the users' scattered but compact desire for expression, so it is also full of banter, gags and other ways. And the biggest characteristic is that the comment and video content are more closely.

At present, danmaku is generally studied from the perspective of traditional communication studies. The

academic believes that the participants of danmaku have the characteristics of concealment and randomness. The danmaku reflect the public culture, while danmaku distributed by the participants has the nature of "carnival square" [4].Based on the theory of use and satisfaction, Wang [5] analysed the reasons for the popularity of danmaku from three aspects: emotion, entertainment and social contact, including weakening loneliness, relieving pressure, and satisfying self-identity and group identity. From the perspective of communication studies, Liao [6] analysed and studied the online ridicule culture in danmaku videos from the perspective of discourse system construction, symbol interaction in danmaku videos and communication media. At the same time, she found that there were serious polarization of audience groups, lack of management of danmaku comments and infringement of danmaku video websites. Jiang [7] found that, not like the multi-centre split propagation mode of traditional video websites, the danmaku propagation mode is mainly reflected in the infinite update and circulation. Based on field theory of Bourdieu, Tian [8] studied the formation of cultural field of BILIBILI by combining the methods of questionnaire and in-depth interview. The analysis results showed that BILIBILI, as a cultural field with limited production, had its autonomy from other fields in the development process.

From the above studies, it can be found that the current research on danmaku from the perspective of communication studies is still limited to the qualitative analysis under the traditional communication studies, and lack of in-depth quantitative analysis and research.

2.2. Knowledge discovery service research

Knowledge discovery is divided into generalized knowledge discovery and narrow knowledge discovery. Broadly, knowledge discovery refers to the discovery of new things, which may not need processing or refinement. In a narrow sense, knowledge discovery is also called database knowledge discovery, which is defined as "a non-trivial process to identify effective, novel, potentially useful and ultimately comprehensible patterns from data sets" [9]. The emergence of knowledge discovery is to solve the problem of information overload faced by people in the digital information age [10].

Knowledge discovery technology is a research hotspot in the field of mapping, and is widely used in medical information, business information, geographic and environmental information, bioinformatics and other fields containing a large number of complex data [11]. Knowledge discovery techniques include statistical analysis, decision tree, neural network, rule reasoning and information extraction. Compared with traditional concept, knowledge discovery which is under with the artificial intelligence focuses on viewpoint mining, standpoint detection and semantic analysis [12]. At present, a lot of research experience of knowledge discovery has been accumulated in communication research, mainly including automatic content analysis, semantic analysis and network analysis [13].

3. Research methods and experiments

It can be found that the application of knowledge discovery technology to danmaku data is a transition from the traditional communication to the computational communication. Based on this, this paper mainly from the statistics and subject analysis of the danmaku content of knowledge discovery processing.

This paper collected the data from website of BILIBILI, took "HUAWEI P30" as the index word, "most danmaku", "10-30 minutes" and "digital area" as the criteria, and selected the top 100 videos to crawl danmaku and video comments respectively. There were more than 92,000 barrage screens.

3.1. The relationship between barrage length and quantity

Length and quantity are important factors to discover the explicit and invisible knowledge contained in danmaku. Based on this, this paper first carries out a single analysis on the length and quantity of the danmaku. Secondly, the correlation between them is analysed quantitatively. Therefore, the corresponding conclusion can be drawn.

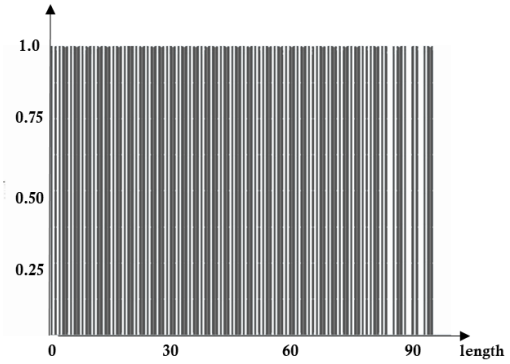

Figure 1 Histogram of danmaku length

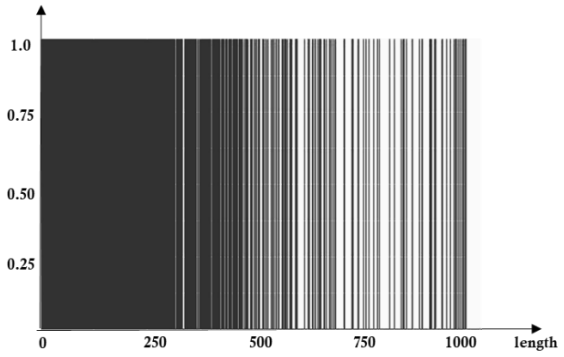

Figure 2 Histogram of comment length

Length and quantity are important factors to discover the explicit and invisible knowledge contained in danmaku. Based on this, this paper first carries out a single analysis on the length and quantity of the danmaku. Secondly, the correlation between them is analysed quantitatively. Therefore, the corresponding conclusion can be drawn.

This model firstly analyses the text length of danmaku and makes statistics on the text length of over 92,000 danmaku in 100 videos.

From the central trend, the median was 49.5, the mean is 49.775. The minimum value is 1 and the maximum value is 108.Skewness is 0.05, the degree of skewness is small. Kurtosis is -1.16, which is the distribution of flat peak.

In order to better illustrate the distribution characteristics of the length of danmaku, this paper crawled the corresponding comment data of all videos and drew the histogram corresponding to the number of comments.

In the length distribution of video comment data, the median is 262, the mean is 311.02, the minimum is 2, and the maximum is 1032.The skewness is 1.09 and kurtosis is 0.8. It is a highly skewness distribution and a peak distribution.

The length and quantity of danmaku are counted. The Pearson coefficient is calculated to be -0.6809813. It can be concluded from the above experimental results that the distribution of the length of the barrage is relatively stable, and it is mainly a short text . There is a negative correlation between the two attributes of danmaku.

Therefore, the comparison the length of video comment data and danmaku shows that there is a big difference between them. The video comment length interval has a larger span and a skewed distribution. The distribution of danmaku is smaller and more concentrated.

The research results have two meanings: 1. From the perspective of machine learning, short text analysis algorithm is more suitable for danmaku. 2. From the perspective of communication studies, existing qualitative studies have found that danmaku has the characteristics of more superficial expression, popularization and anonymity. Because of its objective length, the short text contains less users knowledge than the long text. The results of this experiment support the conclusion to some extent.

3.2. Danmaku topic mining

In the general sense, the topic refers to the meaning expressed by people in daily communication, the central idea expressed in the article and so on. Statistically, the document is the probability distribution of the topic, while the topic is the probability distribution of the word.

In machine learning and natural language processing, the topic model is an unsupervised statistical model for discovering abstract topics in documents. The topic model can be used to simulate the document generation process and obtain relevant information of each topic through parameter estimation. It can extract the low-dimensional mathematical representation from the high-dimensional word space and effectively cluster the words in the text.

Topic is a classical LDA model choice in this study, it is composed of Blei David M. Ng, Andrew Y. Jordan in 2003. It can be a document focus on the theme of each document in the form of probability distribution, and through the analysis of some document to extract their topic. It also can according to the subject clustering or text classification. At the same time, it is a typical word bag model, that is, a document is composed of a group of words with no sequential relationship between words. In addition, a document can contain multiple topics, and each word in the document is generated from one of those topics.

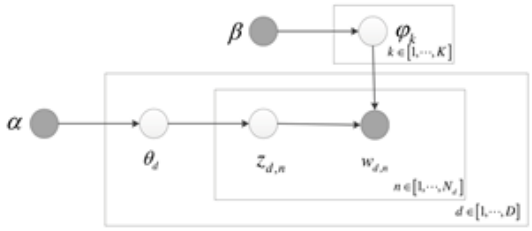

Figure 3 LDA model

The news data obtained from the Internet needs to be pre-processed by Chinese word segmentation, removing stopped words and other methods, and then the text data is further transformed into structured data. In this paper, JIEBA is used for word segmentation, and the text is screened by using the stop-word list of Harbin Institute of Technology.

(1) Chinese word segmentation Compared with English, Chinese has great difficulty in word segmentation and part of speech recognition due to the absence of Spaces between words and the large number of polysemy. Among them, the word segmentation method based on string matching, understanding and statistics 4 has been widely used in natural language processing. With the further research on Chinese word segmentation technology, the corresponding Chinese word segmentation tools have appeared one after another. At present relatively popular in the word segmentation tools have JIEBA, CTCLAS, JBosonNLP, IKAnalyzer, NLPIR and SCWS, including JIEBA participle has the largest number of word segmentation is applicable tools, it is based on the prefix dictionary to scan all the word, and then a statement of all the characters may be generated by the word structure of directed acyclic graph, and the path is based on dynamic programming algorithm to find the largest probability to get the maximum segmentation combination. In this paper, JIEBA word segmentation is used to segment news texts.

(2) eliminate stop words refer to certain words or words that are automatically filtered out before or after text data in order to improve the accuracy and efficiency of text processing. For example, news text contains a large number of words without clear meaning, such as mood auxiliary words, conjunctions, prepositions and punctuation, etc. These words do not have any effect on topic discovery, but they occur frequently. If they are ignored, the accuracy of subsequent topic discovery will be affected. Therefore, it is necessary to establish a stop list. Among them, "Baidu Stop Word List", "Harbin Institute of Technology stop word list" and "Sichuan University Machine Intelligence Laboratory Stop Word List" are the most common in text processing.

We will use the theme model to train the pre-processed data, and we will set the parameters toα=0.1, β=0.01, K=20. After model convergence, we selected five representative results:

Table 1 List of topics and terms

TOPIC	TOPIC WORDS
*Topic 1	[mobile] [camera] [single-lens reflex] [not] [shoot] [sure] [take a picture] [HUAWEI] [photo] [major]
*Topic 2	[suit] [really] [this] [feel] [like] [still] [sky] [not bad] [think] [a bit]
*Topic 3	[mobile] [audio] [what] [evaluating] [tastv] [comment] [experience] [not]
*Topic 4	[in case] [smile] [draw a lottery or raffle] [lottery] [what] [bring down] [win rate] [million] [Ha, ha, ha] [Ha, ha, ha, ha]
*Topic 5	[Samsung] [HUAWEI] [training studying] [Danmaku] [so] [how] [sales volume] [inland] [China] [like that]

Table2 List of High frequency word

WORD	FREQUENCEY
HUAWEI	6076
Samsung	4171
Say	3562
No	2916
Mobile	2881
Buy	2572
Really	2275
Apple	2020
Camera	1710
So	1663
Screen	1490
MI	1399
Pretty	1241
Ha,ha,ha	1154
Want	955
More	927
Awesome	892
Photo	858
Like	855
Money	852
Take a picture	837
Audio	813
P30	810
Danmaku	790
Black	770
◦ ◦ ◦	764
Two	763
Do	762

We can see from the results, the relationship between the word and the word is still relatively high by this topic model. Topic1 mainly related to photography. Topic2 is around the appearance of the mobile phone. Topic3 is a list of comment for the videos. Topic4 is users' feedback on the danmaku lottery, and Topic5 is a contest between fans of different brands. These five topics are basic covers the danmaku and comments of the topic, which have a certain generalization effect.

Combined with the high frequent danmaku words, it can be found that topic 1 and topic 5 have higher weight among all topics. From the perspective of computational

communication, the analysis of individual danmaku users shows that they pay more attention to the video content itself. On the other hand, previous studies on danmaku mentioned that danmaku originated from ACG (Anime, Comics, Games) culture in subculture. This kind of cultural audience has the characteristics of high enthusiasm, high activity, superiority and strong preference.
Further analysis, both from the results of subject word clustering and high-frequency word clustering, we found that the danmaku proportion of the ridicule class is relatively high. This also confirms the qualitative research conclusion of existing scholars, that is, danmaku audiences are younger. On the other hand, many scholars believe that although danmaku originated from ACG culture [14], it contains the characteristics of carnival square and participatory culture. Compared with other comments, the ridicule comments contain more emotional elements [15]. The results of this experiment also demonstrate the conclusion from a quantitative point of view.

4. CONCLUSION

With the development of Internet technology, the enthusiasm of users to participate in the discussion has increased, which is in direct proportion to the increase in the number of users' comments and the enrichment of comments. This paper selects danmaku, a new mode of criticism, as the research object, and through quantitative analysis from the perspective of computational communication, excavates danmaku text itself and the knowledge contained in danmaku respectively. This paper demonstrates the conclusion of the qualitative analysis of traditional communication. At the same time, the results of this experiment also found that machine learning method should be used to analyse danmaku, but the corresponding algorithm of short text analysis should be used for reference. In the future research, the semantic analysis of danmaku can be carried out to further explore the viewpoints of danmaku users, so as to optimize the results. In addition, in the topic mining algorithm selection, this paper adopts the most basic LDA topic mining algorithm. In the future, we can try to apply the improved LDA algorithm to danmaku text topic mining and get better results.

ACKNOWLEDGMENT

This research was financially supported by the youth fund of ISTIC “based on barrage comments and user reviews of sentiment analysis research(Grant No.QN2019-11)”.

REFERENCES

[1] LAZER D, PENTLAND A, ADAMIC L, et al. Computational Social Science[J].Science, 2009, 331(6018):719-723

[2] WATTS D J. Computational social science:Exciting progress and future directions[J].The Bridge on Frontiers of Engineering, 2013, 43(4): 5-10

[3] Chao Naipeng. Artificial intelligence and computing communication [J]. People's forum · academic frontiers,2019(20):20-31+107.

[4] Liu Xiaowei, the research on the micro blog carnival under the theory of carnival theory, is a case of sina weibo's "Spring Festival gala".

[5] Wang Lu, who is based on the theory of "use and satisfaction" theory, based on the theory of "use and satisfaction", is based on the theory of "use and satisfaction".

[6] Liao Xiaoying. Research on the culture of the network of communication from the perspective of communication. The effect of the b station is the study of the propagation of the b station.

[7] Jiang Hanxue. The video study of the video in the field of communication is the study of the dache. Central China normal university, 2014.

[8] Yang Tian. The cultural field construction of the video website of bilibili video website is a study of the study of the study of the study of the social sciences and humanises (icssh2018).

[9] Fayyad U M, Piatetskyshapiro G, Smyth P, et al. Knowledge Discovery and Data Mining: Towards a Unifying Framework[M].MA:MIT PRESS,1996:88-92

[10] Fayyad U M, Piatetskyshapiro G, Smyth P, et al. From data mining to knowledge discovery: an overview[C]. knowledge discovery and data mining, 1996: 1-34.

[11] Wang Min, Zhang Zhiqiang. A Content Analysis of Knowledge Discovery Papers in Information Science and Library Science. New Technology of Library and Information Service, DOI：10.11925/infotech.1003-3513.2008.02.12.

[12] Huang Lili. Knowledge discovery model and empirical research on text data of social media [D]. Jilin university,2016.

[13] Li Sumei. Knowledge map analysis of library knowledge discovery research in China in recent years [J]. Henan library journal, 2008,38(09):122-124.

[14] Peng Xixian , Yu Xiang, Chris Zhao, and Hockhai Teo. "Understanding Young People's Use of Danmaku Websites: the effect of Perceived Coolness and

subcultural Identi.." pacific asia conference on information systems (2016).

[15] Niu J, Li S, Mo S, et al. Affective Content Analysis of Online Video Clips with Live Comments in Chinese[C]. ubiquitous intelligence and computing, 2018: 849-856.