Summer 5-28-2021

Behavior Analysis and Recognition of Hidden Populations in Online Social Network Based on Big Data Method

Minglei Li
School of Information Management and Statistics, Hubei University of Economics, China ; School of Public Administration, University of Electronic Science and Technology of China, China,
liminglei@hbue.edu.cn

Guoyin Jiang
School of Public Administration, University of Electronic Science and Technology of China, China

Wenping Liu
School of Information Management and Statistics, Hubei University of Economics, China

Junli Lei
School of Information Management and Statistics, Hubei University of Economics, China

Full Research Paper

Behavior Analysis and Recognition of Hidden Populations in Online Social Network Based on Big Data Method

Minglei Li[1,2], Guoyin Jiang[2], Wenping Liu[1], Junli Lei[1]*

[1]School of Information Management and Statistics, Hubei University of Economics, China

[2]School of Public Administration, University of Electronic Science and Technology of China, China

Abstract: Hidden populations refer to the minority groups that not well-known to the public. Traditional statistical survey methods are difficult to apply in the study of hidden populations because of that the hidden populations individuals are very troublesome to be found and they are not willing to share the inner opinion with the others. On the other hand, with the development of the Web 2.0, the hidden populations gather and share their views in online social networks due to the openness and anonymity of the Internet. So, this paper analyzes the behavioral characteristics of the hidden populations based on their data in online social networks. This paper uses the lesbian population as an example and analyzes the behavioral characteristics of lesbian by analyzing the data of the lesbian population in Douban Group. First, the activity data on lesbian are collected from Douban Group. Second, behavior characteristics of lesbian are analysed, the regional characteristic, temporal characteristic and text characteristic are mined out by big data method. Third, a lesbian recognition model is proposed based on the above analytical characteristics, and the effectiveness of the recognition model is varified by experiment study. The research of this paper is helpful to understand the behavioral characteristics of hidden populations deeply, and provides decision-making basis of management and service for hidden populations.

Keywords: Hidden population, Online social network, Lesbian, Behavior analysis, Lesbian recognition

1. INTRODUCTION

Hidden populations[1-3] refer to the minority groups that not well-known to the public. Generally speaking, hidden population is the population that hard to be contacted, and the people in these groups are unwilling to expose themselves for many various reasons, such as: HIV/AIDS, LGBT, patients with depression. Hidden populations are a small proportion of the overall population. However, there is a large population in China, so the absolute quantity of hidden population is big. Due to long misunderstanding, the hidden populations are very sensitive and self-protective, and they are unwilling to communicate with the mainstream society or express their true ideas in the real society [4]. At the same time, the people involved in hidden population are more complicated, and it is easy to generate events that affect social security such as extreme thoughts, spread of diseases,and fraud, which bring great hidden dangers to society[5]. Therefore, it is very meaningful for stable and harmonious development of society to improve the level of investigation and analysis of hidden populations and manage hidden populations reasonably and effectively.

The traditional analysis methods of hidden population are mainly based on statistical survey, such as snowball sampling[6,7] and respondent-driven sampling[8,9]. However, these traditional methods have many limitations. First, the hidden population is difficult to find, and sufficient survey samples cannot be obtained easily. Second, the hidden population is unwilling to participate in the survey. The last but not all, the cost of traditional method is high. With the development of the Web 2.0 , especially the online social networks, many people share their comments on the Internet. Due to the anonymity and openness of the Internet, many hidden populations are more willing to gather and share their ideas in online social networks. And the hidden populations leave a large amount

* Corresponding author. Email: liminglei@hbue.edu.cn(Minglei Li)

of data in online social networks. So we can understand the behavioral characteristics of the hidden population through the analysis of these data [10,12]. The amount of these data materials is very large, and the value density of them is low, so the traditional statistical analysis methods are difficult to find valuable information. Big data method [13,14] is a technology that specializes in processing large-scale data, and is widely used in online data mining, social management and other fields. In this paper, we use big data method to analyze the online social network data of the hidden population, and gain insight into the behavior characteristics of the hidden population.

This paper uses lesbian as an example. We collect the data of lesbian in Douban Group, and analyse the online behavior characteristics of lesbians based on big data method, then propose a recognition model to identify the lesbian population based on machine learning model. The paper is organized as follows. Section 2 presents the overall method of this paper. Section 3 shows the behavior characteristics analysis based on big data method, specifically, we aim to extract the text feature of lesbian based on text analysis technology. By using the behavior characteristics features, Section 4 designs lesbian recognition model based on machine learning model. In Section 5, experimental study is used to demonstrate the effectiveness and practicability of the recognition model. Finally, the conclusions and future works are given in the Section 6.

2. OVERALL METHOD

2.1 Overall structure of the method

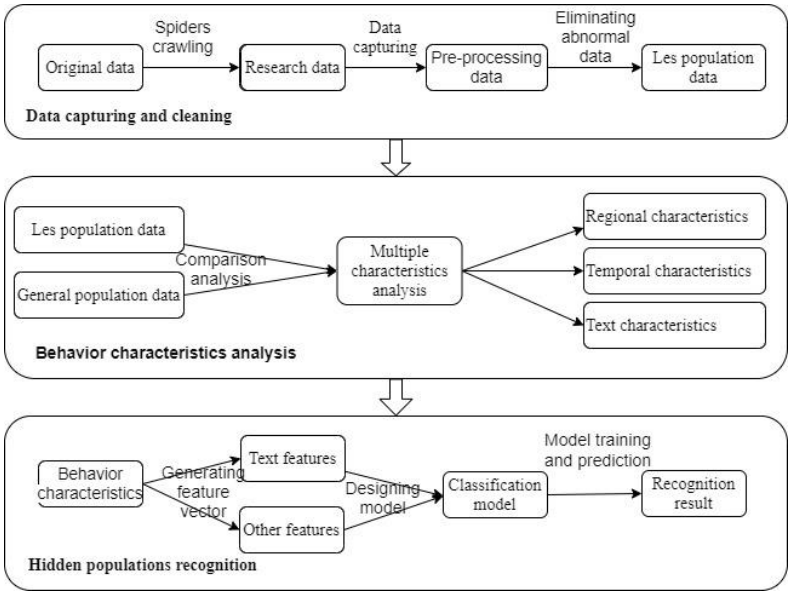

Figure 1. The overall structure of this method

Aiming to analyze the behavior characteristics of hidden populations and recognize the hidden populations based on their data in online social networks. The overall structure of the method proposed in this paper is shown as Figure1. There are three steps in the method, including data capturing and cleaning, behavior characteristics analysis and hidden populations recognition. Specifically, first, the data in online social networks are collected by Python crawler, data cleanning is used to improve the quality of the data, and irrelevant data, such as spammer and advertisement, are eliminated in this step. Second, the comparative analysis between hidden population and general population is used to study the behavior characteristics, we use the big data technology to analyze the regional characteristics and temporal characteristics, and use text analysis technology to analyze the test characteristics of hidden populations. And finally, we construct the feature vector based on the above characteristics, and design a recognition model based on machine learning model to distinguish the hidden population from general population.

2.2 Data sources

As one of the world's largest Chinese online social network, Douban Group is provided by Douban, and there are more than 400 000 douban groups. The Douban Group involves many fields such as film and television, reading, photography, society, and many hidden populations such as homosexuality and depression. In the Douban group, like-minded users open their hearts and express their opinions freely, so we can get a lot of real data from it.

In this paper, we use lesbian as an example of hidden population, and collect the hidden population data from "Les Sky" Group ('https://www.douban.com/group/lala/'). And a general population data from " Readingmania" Group ('https://www.douban.com/group/readingmania/') is collected as comparative analysis. The collected data involve user information, post and comment data as shown in Table 1. We collect 50 000 comments from "Les Sky" Group and 50 000 comments from " Readingmania" Group.

Table 1. Data fields from Douban Group

User_information	Comment_information	Reply_information
User_id	Comment_id	Reply_id
User_name	Comment_title	Comment_id
User_residence	Comment_text	Author_id
User_url	Author_id	Reply_to_whom
User_join_time	Creation_time	Reply_floor
	Reply_number	Reply_text
		Reply_time

3. BEHAVIOR CHARACTERISTICS ANALYSIS

In this Section, we will find that the hidden population's behavior patterns in online social networks, by comparing and analyzing the behavioral characteristics of the hidden population and the general population in online social networks.

3.1 Temporal characteristic

It can be found that when the different populations are active on the online social network through temporal characteristic analysis. The result is shown as Figure 2. It can be found that the active period of users in Douban Group is night, and it reaches the peak at night, especially around midnight, and the low period is 1:00 to 8:00. In addition, it can be clearly found that between 8:00 to 20:00, the number of comments posted by general population is higher than that of the hidden population. However, after 21:00, the number of comments posted by the hidden population is higher than that of the general population. It can be seen that lesbians prefer to be in online social networks at night.

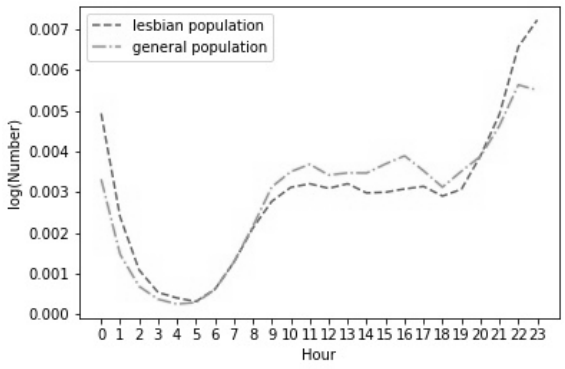

Figure 2. Comments submission time of different populations

3.2 Distribution of number of comments

In Figure 3, it shows the number distribution of comments under each post. As we can see from Figure 3, there are small number of comments under many posts of both the hidden population and general population. However, the hidden population has an obvious aggregation phenomenon, that is, there is a lot of discussion under certain posts.

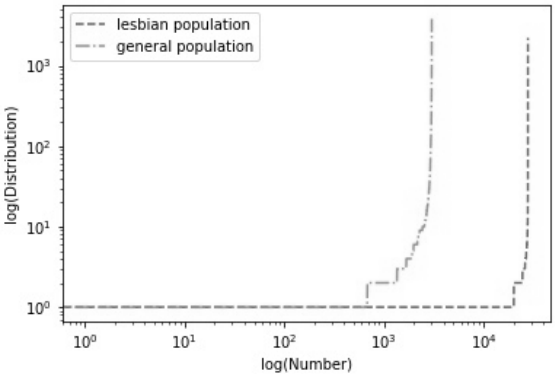

Figure 3. Number distribution of comments under each post of different populations

Figure 4 shows the number of comments posted by each user. It can be found from Figure 4, the number of comments posted by many users is small in Douban Group, and the hidden population is more likely to generate active users, who submit a large number of comments.

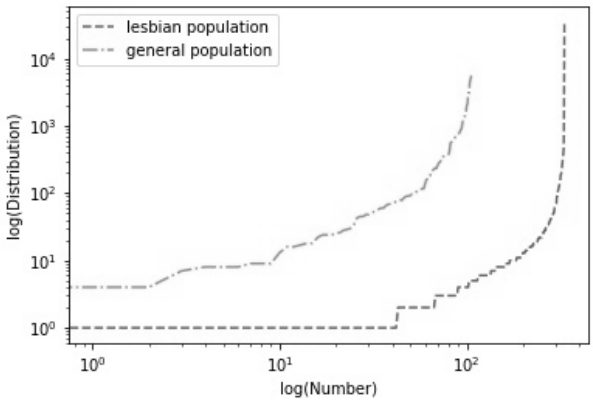

Figure 4. Number distribution of comments from each user of different populations

3.3 Textual characteristic

Figure 5. Comments text wordcloud of different populations

By textual characteristic analysis, we try to discover the topics that population care about and discuss generally. We generate the wordcloud of the comments text each from the lesbian and the general populations. The most of comments text in Douban Group are in Chinese, so we segment the comments text by Jieba module and remove irrelevant words, then the 100 most frequent words are used to generate the wordcloud. The wordcloud results are shown in Figure 5, the left wordcloud is from the lesbian population and the right is from the general population. It can be found from Figure 5, the part of high-frequency words from the hidden population and the general population are same, indicating that the lesbian population is also interested in the topic that the general population cares about. In addition, many unique words, such as "sister" and "feeling", have appeared in the high-frequency words from lesbian population.

4. HIDDEN POPULATION RECOGNITION

From the behavior characteristic analysis result in Section 3, it can be found that there are many differences between the hidden population and the general population. So, we can use these different characteristics to identify the hidden population.

4.1 Features for hidden population recognition

In this paper, the temporal characteristic, number of comments characteristic and textual characteristic are used as the feature vector of the hidden population recognition model. For temporal characteristic, based on the results in Section 3.1, the time is divided three periods: 8:00 to 20:00, 20:00 to 4:00 and the rest of time period. The model uses in which time period the user submits the most comments as the temporal feature value of this user. For number of comments characteristic, this model uses the number of comments submits by the user as the feature value of this author. For textual characteristic, the IF-IDF[15] is used to vectorize the comments text of the user. Specifically, based on the word segmentation results, we can calculate the TF-IDF value of each word by:

$$\mathrm{TF-IDF}(w,u) = \mathrm{TF}(w,u) \times I\mathrm{DF}(\mathrm{w}) \tag{1}$$

where, $\mathrm{TF}(\mathrm{w},\mathrm{u})$ is the number of word w appears in the comments text of the user u, $IDF(w) = log\frac{M}{N}$ (N is the number of word w appears in the text, and M is the number of all text). $TF-IDF(w,u)$ is the feature vector value of word w in comments text of the user u.

4.2 Recognition model

4.2.1 Model structure

The core of recognition model is XGboost[16]. The XGboost is an extension of the gradient boosting decision tree algorithm. XGboost belongs to ensemble learning model, and its basic idea is to combine multiple Classification And Regression Trees (also known as CART).

Given data set $D = \{(x_i, y_i)\}$ (x_i is the feature vector mentioned in Section 4.1 and y_i is label of user). The integrated model is expressed as

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), f_k \in \mathcal{F} \tag{2}$$

where K is the number of CARTs in the integrated model, $\mathcal{F} = \{f(x) = \omega_{q(x)}\}$ ($q : \mathcal{R}^m \to T, \omega \in \mathcal{R}^T$) is the space of CARTs. T represents the number of leaves in the decision tree. Each f_k corresponds to an independent tree structure q and leaf weights ω.

The objective function of XGboost is expressed as

$$argMin\mathcal{L}(\Theta) = \sum_i l(\hat{y}_i - y_i) + \sum_k \Omega(f_k) \tag{3}$$

The objective function consists of two parts, the first part $l(\cdot)$ is the training error. And the second part $\Omega(\cdot)$ is the regularization item that helps to smooth the final learnt weights to avoid over-fitting. $\Omega(\cdot)$ penalizes

the complexity of the model as

$$\Omega(f) = \gamma T + 1/2\lambda\|\omega\|^2 \tag{4}$$

where γ and λ are coefficients.

4.2.2 Training algorithm of the model

The training of XGboost is a boosting algorithm. This algorithm trains a base learner from the training data set firstly, and improves the model based on the base learner iteratively. That is to say, each iteration the model is retained, and a new item is added to the model as

$$\begin{aligned} &\hat{y_0}^{(0)} = 0, \\ &\cdots \\ &\hat{y_i}^{(t)} = \sum_{k=1}^{t} f_k(x_i) = \hat{y_i}^{(t-1)} + f_t(x_i). \end{aligned} \tag{5}$$

where $\hat{y_i}^{(t)}$ is the predicted value of the samples i in the round t. The choice of adding a new item in each iteration is to minimize the objective. And, the objective function Equation (3) can be turned as

$$\begin{aligned} \mathcal{L}^{(t)} &= \sum_{i=1}^{n} l(y_i, \hat{y_i}) + \sum_{k} \Omega(f_k) \\ &= \sum_{i=1}^{n} l(y_i, \hat{y_i}^{(t-1)}) + f_t(x_i) + \Omega(f_t) + C. \end{aligned} \tag{6}$$

The error function $l(\cdot)$ is squared error, and the objective function Equation (5) can be written as:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} [2(\hat{y_i}^{(t-1)} - y_i) f_t(x_i) + f_t(x_i)^2] + \Omega(f_i) + C. \tag{7}$$

Taylor expansion is used approximately and the objective function Equation (6) can be defined as

$$\tilde{L}^{(t)} \backsimeq \sum_{i=1}^{n} [l(y_i, \hat{y_i}^{(t-1)}) + g_i f_t(x_i) + 1/2 h_i f_t^2(x_i)] + \Omega(f_i) + C, \tag{8}$$

where $g_i = \partial_{\hat{y}^{(t-1)}} l(y_i, \hat{y}^{(t-1)})$and $h_i = \partial^2_{\hat{y}^{(t-1)}} l(y_i, \hat{y}^{(t-1)})$. The constant terms are removed and the simplified objective at step t is

$$\tilde{L}^{(t)} \backsimeq \sum_{i=1}^{n} [g_i f_t(x_i) + 1/2 h_i f_t^2(x_i)] + \gamma T_t + 1/2\lambda\|\omega_t\|^2. \tag{9}$$

XGboost generates and updates the decision tree，and being trained by the samples data set，until the objective function satisfies the condition.

5. EXPERIMENTAL RESULTS AND ANALYSIS

5.1 Experimental data

We collect 50 000 comments from "Les Sky" Group and 50 000 comments from " Readingmania" Group. After removing the irrelevant user such as spammer and advertisement, we mark the users who have submitted comments in "Les Sky" Group as lesbian, and mark the users who have submitted comments in " Readingmania " Group as general user. And we choose randomly 3000 users in "Les Sky" Group and 3000 users in " Readingmania " Group. We will use the behavior features to identify whether a user in Douban Group is lesbian.

5.2 Experimental results and analysis

The recognition model in this paper is developed based on XGboost library†, and we implement the recognition model by Python3 language.

In this paper, the data set is divided into two parts randomly. The one part, which is 80% of the total users, is the training set and is used to train the recognition model. The other one, which is the remaining 20% of total users, is the testing data and is used to verify the effectiveness of this recognition model.

† https://github.com/dmlc/xgboost

5.2.1 Performance comparison of different features

In this section, we study the impact of using different features on recognition model. We design two features sets: in the features set one (FS1), the features are all the features mentioned in Section 4.1; and in the features set two (FS2), the features are only the textual features. Each of FS1 and FS2 is used to train XGboost by the training data, and verify the performance of the recognition models by the testing data. The hidden population identification problem is a binary classifying problem, so the performance evaluation indicators for binary classifying model are used to measure the effectiveness of the recognition models, such as: accuracy, recall, and precision. The results is shown in Table 2.

Table 2. Performance comparison of different features

Features set	Accuracy	Precision	Recall
FS1	0.8503	0.8357	0.8713
FS2	0.7520	0.7618	0.7333

From the results in Table 2, it can be found that the recognition model using FS1 has better performance than using FS2. The result indicates that all the behavior characteristics analyzed in this paper are helpful to identify the hidden population. And it also reflects that more features and more implicit information. So, the features set FS1 is used in the next experiment.

5.2.2 Performance comparison of different models

In this section, we compare the recognition model XGboost to other familiar machine learning models: linear regression model (LR), decision tree model (DT), ANN model and SVM regression model. In ANN model, the number of hidden layer nodes is 120, the learning rate is 0.001, the maximum number of training time is 5000. In SVM, the kernel is gaussian function, which is vary popular in SVM. All the models are all implemented using Python's machine learning library sklearn. The comparison result is shown in Figure 6.

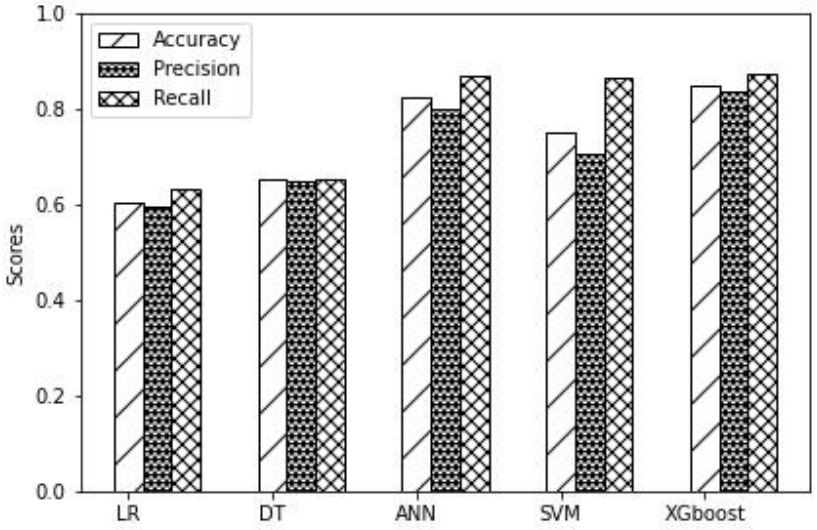

Figure 6. Performance comparison of different models

From the results in Figure 6, we can find that the recognition performance of XGboost is the best among all the models. The result illustrates the effectiveness of the method proposed in this paper.

6. CONCLUSIONS

In order to survey the hidden population deeply and widely, this paper analyzes the data in online social network of the hidden populations based on big data technology. This paper uses lesbian in Douban Group as an example. The main contributions of this paper are: (1) big data technology is used to analyze the data in online

social network of the hidden populations, so as to understand the behavior characteristics of these populations, (2) according to these behavior characteristics in online social network, a recognition model based on machine learning model is proposed to distinguish the hidden population from general population. Furthermore, the effectiveness and practicability of the recognition model are demonstrated with the data from Douban Group.

The research in this paper shows the potential and advantages of big data method in hidden population analysis. However, the methods and conclusions of this paper are still preliminary and limited, Our ongoing work is to consider the following aspects: (1) the more data and more types of hidden populations in online social network shoud be collected, and the methods of this paper will be applied for these data to get more information, (2) the group behaviors of hidden populations are also very important, we will use the social complex network analysis and other methods to analyze the group behaviors of the hidden populations, (3) this paper mainly analyzes the static characteristics of the hidden population, the next work is to study the dynamic characteristics of the hidden populations and design a model to predict the dynamic characteristics of the hidden populations.

ACKNOWLEDGEMENT

This work was partially supported by the National Natural Science Foundation of China under Grant 72071031 and Grant 62072163, and the Humanities and Social Sciences Research Youth Foundation of the Ministry of Education of China under Grant 20YJCZH072 ,the Fundamental Research Funds for the Central Universities of China (No. ZYGX2017KYQD185), and the Educational Commission of Hubei Province, China under Grant 19Q145.

REFERENCES

[1] Spreen M. Rare populations, hidden populations, and link-tracing designs: What and why?[J]. Bulletin of Sociological Methodology/Bulletin de Methodologie Sociologique, 1992, 36(1): 34-58.

[2] Magnani R, Sabin K, Saidel T, et al. Review of sampling hard-to-reach and hidden populations for HIV surveillance[J]. Aids, 2005, 19: S67-S72.

[3] Wesson P, Reingold A, McFarland W. Theoretical and empirical comparisons of methods to estimate the size of hard-to-reach populations: a systematic review[J]. AIDS and behavior, 2017, 21(7): 2188-2206.

[4] Griffiths P, Gossop M, Powis B, et al. Reaching hidden populations of drug users by privileged access interviewers: methodological and practical issues[J]. Addiction, 1993, 88(12): 1617-1626.

[5] Duncan D F, White J B, Nicholson T. Using Internet-based surveys to reach hidden populations: case of nonabusive illicit drug users[J]. American journal of health behavior, 2003, 27(3): 208-218.

[6] Frank O, Snijders T. Estimating the size of hidden populations using snowball sampling[J]. JOURNAL OF OFFICIAL STATISTICS-STOCKHOLM-, 1994, 10: 53-53.

[7] Rao A, Stahlman S, Hargreaves J, et al. Sampling key populations for HIV surveillance: results from eight cross-sectional studies using respondent-driven sampling and venue-based snowball sampling[J]. JMIR public health and surveillance, 2017, 3(4): e72.

[8] Lu X, Bengtsson L, Britton T, et al. The sensitivity of respondent□driven sampling[J]. Journal of the Royal Statistical Society: Series A (Statistics in Society), 2012, 175(1): 191-216.

[9] Handcock M S, Gile K J, Mar C M. Estimating hidden population size using respondent-driven sampling data[J]. Electronic journal of statistics, 2014, 8(1): 1491.

[10] Liu C, Lu X. Analyzing hidden populations online: topic, emotion, and social network of HIV-related users in the largest Chinese online community[J]. BMC medical informatics and decision making, 2018, 18(1): 1-10.

[11] Rice E, Tulbert E, Cederbaum J, et al. Mobilizing homeless youth for HIV prevention: a social network analysis of the acceptability of a face-to-face and online social networking intervention[J]. Health education research, 2012, 27(2): 226-

236.

[12] Young S D, Jaganath D. Online social networking for HIV education and prevention: A mixed methods analysis[J]. Sexually transmitted diseases, 2013, 40(2).

[13] Ghani N A, Hamid S, Hashem I A T, et al. Social media big data analytics: A survey[J]. Computers in Human Behavior, 2019, 101: 417-428.

[14] Oussous A, Benjelloun F Z, Lahcen A A, et al. Big Data technologies: A survey[J]. Journal of King Saud University-Computer and Information Sciences, 2018, 30(4): 431-448.

[15] Kim D, Seo D, Cho S, et al. Multi-co-training for document classification using various document representations: TF–IDF, LDA, and Doc2Vec[J]. Information Sciences, 2019, 477: 15-29.

[16] Chen T, Guestrin C. Xgboost: A scalable tree boosting system[C]//Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 2016: 785-794.