Effective User Invitation Models for Online Survey Using Clustering Algorithm

Sen Shao, Shaochun Wu, Guobing Zou, Liang Chen
School of Computer Engineering and Science, Shanghai University
Shanghai 200444, China
shaolbs@163.com, scwu@shu.edu.cn, gbzou@shu.edu.cn, liangchenshu@gmail.com

Abstract—**With the continuous and rapid development of online questionnaire survey, the low response rate has plagued operating companies. To solve this problem, this paper proposed an effective user invitation model by our improved clustering algorithm, which analyzed large-scale historical user behavior characteristic data, including users' quality data, users' preferential data and users' similarity data. Extensive experiments with large-scale data from an online survey company have been conducted to validate the feasibility and effectiveness of our proposed approach. Experimental results demonstrate that the questionnaire response rate is increased and our approach can be easily deployed in real-world online survey application for effective personalized survey recommendation.**

Keywords-*online questionnaire survey; clustering algorithm; user invitation model; response rate*

I. INTRODUCTION

The history of online questionnaire survey could date back to 1994, initially conducted by Georgia Tech GVU Center. Online questionnaire survey holds many advantages over conventional survey. For example, it can complete a survey much faster and simpler, while consuming lower cost than traditional survey. Additionally, it is not limited by the time and space dimensions. Especially, sensitive topics or intimate issues will be intelligently analyzed and reflected. However, in the field of online questionnaire survey, it encounters varieties of application challenges. For example, some users respond to the wrong answers for the reward, which may not only lead to meaningless result itself, but analysis negatively influence the overall survey results. Some measures should be taken to detect and block those users as well as their questionnaires. Furthermore, the biggest issue that online questionnaire survey faces is low response rate, which directly affects the quality of survey results and increases the company's budget costs. In many cases, company has to send invitations a couple of times, because the number of respondents doesn't reach minimum standard. Especially, online survey doesn't come off because of lack of large-scale accumulated data before deadline.

To tackle above research challenge, our research motivation is based on large-scale survey data from a collaborative company and we propose an effective user invitation model for online survey via a clustering-based method. Our goal aims at decreasing low response rate problem. We conduct extensive experiments to validate the effectiveness of our proposed user invitation model. Compared with the existing random model, the experimental results demonstrate that our approach can be conveniently and effectively deployed for enterprise online survey.

The remainder of this paper is arranged as follows. Section II reviews the related work on sequential pattern mining, while user model and user behavior sequence is presented in Section III. We present user behavior analysis and its algorithm in Section IV. Experimental results are presented in Section V. Finally, we conclude the paper in Section VI.

II. RELATED WORK

Data mining has been widely applied in many applications fields, including SKICAT system [1] helping astronomers detect Earth-like planets, analyzing the basketball statistics and players data to help the coach develop tactics to inherent players' shortcomings.

The common ways of online survey include web, e-mail, and internet telephony and so on. But now, professional companies engaging in online survey have been the mainstream. Many large enterprises are willing to assign their investigation to SurveyMonkey which is the best company in the industry overseas. However, the online survey begins comparatively late in our country and competition is intense. How to apply data mining to the online survey is all the companies concern.

In view of clustering analysis, most of clustering algorithms are based on distance calculation, such as k-means clustering, k-medoid clustering and some other common clustering algorithms [2]. Thought clustering algorithm is relatively mature, it still difficult to directly apply existing clustering algorithms in questionnaire industry, due to its domain characteristics. The investigation in [3] demonstrated that response rate was very low in many research firms. For example, investigator Shaopeiji surveyed those companies who engaged in e-commerce in the Chinese mainland in 2005 and the response rate was only 2.1%, while investigator Judy Decou surveyed those companies' Yellow Pages who published in Canada in 2005 and the response rate was 2.5%.

Although online survey companies tried to figure out solutions to solve low response rate, most of solutions that improved response rate are from the angle of investigator. Moreover, the authors in [4] demonstrate that improving reward is just the supplementary role which still cannot play an important and decisive role. Even if you improve reward blind, results are just objective within a relative short period of time. Those known as professional respondents will answer questionnaire as the way what companies expect.

In order to reduce the impact of these factors, this paper integrates clustering algorithm into multi-characteristic data of all users to construct user invitation model. Based on the invitation model, we only send invitations by those expected users so as to improve the user response rate and save costs for company simultaneously.

III. USER INVITATION MODEL VIA CLUSTERING

The user information involved in online survey consists of multiple dimensions and each user has many attributes. If we directly carry on the raw data without any data preprocessing,

978-1-4673-7723-2/15 $31.00 © 2015 IEEE
DOI 10.1109/IMCCC.2015.307

it will lead to low efficiency as well as response data. So attribute selection is an important way for us to filter data before the application of our improved clustering algorithm. Attribute selection means to find out attribute subset from user's attributes to delete those irrelevant or redundant attributes that could help reduce complexity and run-time of clustering analysis when constructing user invitation model. Especially, "curse of dimensionality" will happen when it has many attributes, making the robustness of clustering analysis deterioration.

A. *Attributes Selection of Clustering Algorithm*

The process of attribute selection is shown in the following Fig. 1. Search algorithm usually performs heuristic search with the space reduction. This paper employs classical decision tree algorithm for attribute selection as follows.

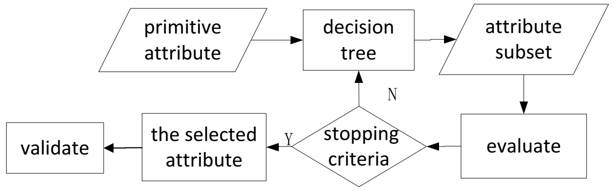

Figure 1. Process of clustering attribute selection

The major attributes of data set include user ID, user name, registration time, city, marital status, invitation times, login times, response times, screening times, give up times, success rate, give up rate, birthday, mobile phone verification, sex, personal income, education level, occupation, the response rate, SuccessCountType1, SuccessCountTypei and so on. Because many of the attributes are redundant or irrelevant, these attributes can be excluded before performing clustering analysis for user invitation model construction. We employ decision tree algorithm to make a construction similar to flowcharts, selecting the most important attributes in each tree node. Decision tree induction generally calculates the value of information gain for each candidate attribute as clarity-evaluation-function [5]. When we select the first attribute, it ranks with the greatest value of information gain among all of the attributes. As shown in Table 1, the education is the most important factor. Then recalculate the rest of attributes until decision tree is constructed completely.

Table 1. Selecting attribute from the highest information gain

Number	***Attributes***	***Information gain***
1	sex	0.0751
2	income	0.0182
3	education	0.0936
4	marital	0.022
5	Whether mobile phone through verification	0.0437
6	success rate	0.0076

After picking out attribute subset with top information gains, attributes are reduced from 33 to 12, which are shown in the following Table 2.

Table 2. Data set after attribute selection

number	identification	attribute
1	Id	userID
2	Age	age
3	Sex	sex
4	Mobile Valid	mobile valid
5	PersonIncome	income
6	EducationLevel	Educationlevel
7	SuccessCountType1	answer questionnaire type 1 times successful
8	SuccessCountTypei	answer questionnaire type i times successful
9	successRate	success rate
10	responseRate	response rate
11	giveUpRate	give up rate
12	screeningRate	screen rate

B. *Improved Clustering Algorithm*

Based on the idea of calculating the distance among initial clustering centers, the distance should be as far as possible. Through improving the way that selects initial cluster centers based on the simple cluster algorithm, Apache common data repository [6] realizes the improved algorithm.

The process of improved algorithm is mainly to determine how to select the initial cluster centers. First, we select a data randomly from data set as the first clustering center. Second, we calculate the distance between every data and the selected clustering center. Third, based on the strategy of selecting clustering center, we select new data as the clustering center. Fourth, we repeat the second and third processes until k clustering centers have been selected. Fifth, run the k-means clustering algorithm.

Actually, the key step of our improved k-means algorithm lies in the third step. The method apache common data repository is applied for finding initial k cluster centers as follows. We first calculate the distance by Euclidean distance $dist(x_i) = \sqrt{\sum_{k=1}^{n}(x_i - x_k)^2}$. Next, save the distances to an array and add distances all up, and thus we obtain the value $\sum_{i=1}^{n} dist(\mathrm{x}_i)$. Then, we select a value that can fall on $\sum_{i=1}^{n} dist(\mathrm{x}_i)$ randomly, and update the value by $Value -= \sum_{i=1}^{n} x_i$ until Value<=0. Now, Value is the next cluster center. The process of so-called k-means++ is shown in the following Fig. 2.

Input:

- k: the number of clusters of algorithm
- D: data set

Output:

- k clusters in collection of data objects D

Main Procedure:

(1) Select an object from the data set D randomly as the initial cluster center

(2) Repeat
(3) Execute the strategy of select the next cluster center, choose a new cluster center
(4) Until Selected k cluster centers
(5) Run k-means algorithm, output k clusters

Figure 2. Process of k-means++

C. *User Invitation Model Construction*

Since clustering analysis is carried out from different aspects to users who are divided into more than one type, it can provide a useful strategy for target survey invitations. First, clustering analysis based on user's quality is similar to partitioning of the club's membership, which divides the customers into gold, platinum and other types. By using this idea, we divide the users who participate in online survey into four types, including higher-quality, high-quality, medium-quality and low-quality. We construct a user invitation model based on users' quality level. Then, we combine clustering analysis based on user preferences with clustering analysis based on the basic characteristics of user for generating a comprehensive user invitation model.

a. User Invitation Model based on quality level

For online survey, the attitude of users to participate in the online survey determines the quality level of the user, so, response rate, success rate, give up rate and screening rate can well reflect the enthusiasm of users to questionnaire. However, the screening rate is passive operation for the users, so screening rate isn't considered as a evaluation index. According to the evaluation index, the high quality of the user should have a high response rate, high success rate and a low give up rate.

According to success rate, response rate and give up rate, we divide user's quality level into four types spanning from the highest quality to the low quality. Each of them has different value to an enterprise. Enterprises need to focus on high-quality users, because high-quality users mean that they could often successfully participate in online surveys and are important to enterprises.

For the clustering of user invitation model, all users will be divided into four different levels by improved k-means algorithm, through central point metrics including success rate, response rate, and give up rate to distinguish user's quality level. In most cases, the number of users with low-quality is much more than other three types, thus if enterprises only invite high-quality users, the quantity of high-quality users cannot meet enterprise's demands. Therefore, enterprises have to invite users from high-quality to low-quality to balance the tradeoff of invitation requirements. The constructed user invitation model is shown in the following Fig. 3.

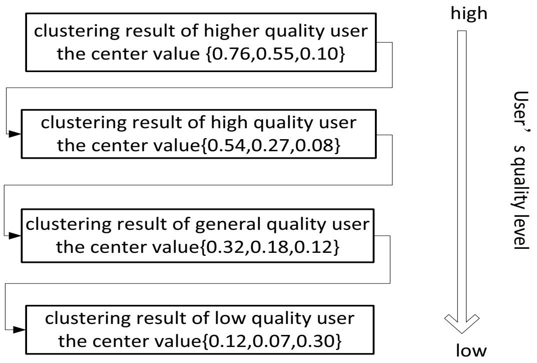
clustering result of higher quality user the center value {0.76,0.55,0.10}
clustering result of high quality user the center value{0.54,0.27,0.08}
clustering result of general quality user the center value{0.32,0.18,0.12}
clustering result of low quality user the center value{0.12,0.07,0.30}
high
User' s quality level
low

Figure 3. User invitation model based on quality level

b. User invitation model based on similar users and user preferences

To generate user invitation model, we first select users who possess their preferences among all of the users in data, and then eliminate the problem of data sparsity. By doing so, we obtain users' preferential set by k-means++ algorithm. Then, k-means++ algorithm is applied to the basic information of users, which clusters the same or similar features of the users. Thus, we obtain similar set of users. Finally, we combine the set of user preferences and the set of similar users to generate a candidate set of users who will be invited for online survey. User invitation model based on combination of similar users and user preferences is shown in Fig. 4. In real-world practical applications, user invitation model based on combination can improve response rate.

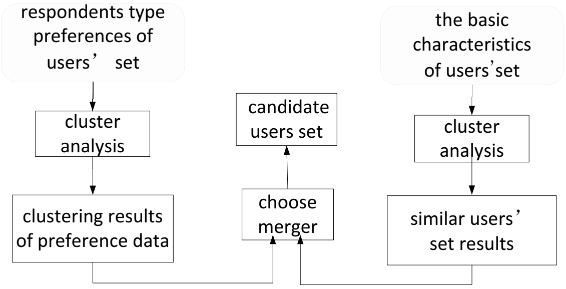
respondents type preferences of users' set
cluster analysis
clustering results of preference data
candidate users set
choose merger
the basic characteristics of users'set
cluster analysis
similar users' set results

Figure 4. User invitation model with the combination of similar users and user preferences

IV. Experimental Evaluation

A. *User Invitation Model Based on User's Quality Level*

a. Experimental design

After attribute selection, the data set is shown in Table 2. Experiment selects user id, response rate, success rate and give up rate as data set. Experiment was divided into two steps. The first step chooses the users data to execute clustering analysis using k-means++ algorithm, and then sets the value of k as 4. All users will be divided into four types, each type with a label. The second step repeats the experiment many times to compare response rate between randomly invited users and the one by user invitation model based on user's quality. The experimental results verify that clustering analysis can improve the response rate. First, we use a stochastic function to choose a questionnaire to send invitations, and then user invitation model based on user's quality level is applied to send another group of invitations.

The actual historical data comes from an online survey company, and the data set contains 534,879 records. Table 3 shows the results of the clustering algorithm based on user's quality.

Table 3. Results of clustering algorithm based on use's quality level

Id/user id	*category*	*Id/user id*	*category*
226070	A	7912	B
1267	B	106987	B
990511	B	30889	C
13857	D	643825	C
1746789	D	92029	C
985547	D	1744445	C
402	C	46622	D
… …	… …	… …	… …

We compare the response rate between these two invitation ways. Specifically, we select a questionnaire to send invitations. Then, we use two ways including randomly invite users randomly and inviting user based on user's quality level. Thus, we count up the number of users invited and responses after sending the questionnaires, and then calculate the response rate of each invitation way. Under the same experimental environment, we complete three experiments and the results of the experiment are shown in the following Table 4.

Table 4. Experimental results of comparison between two invitation ways for online survey

Statistics parameter *Invitation model*		Exp 1	Exp 2	Exp 3
Model of randomly invite users	count	26756	32679	4576
	response	100	10324	155
	response rate	***0.37%***	***31.59%***	***3.39%***
User invitation model via user's quality level	count	2229	6877	1288
	response	77	4988	141
	response rate	***3.45%***	***72.53%***	***10.95%***

b. Results analysis

In experiment 1 and 2, the response rate is improved many times. In experiment 3, although the number of response from invited users is almost the same by the two ways, the invitation number of randomly invited users is almost four times of the one by user invitation model based on user's quality level. At the same experimental condition, as the experimental data increases, the success rate will be a measurable reduction.

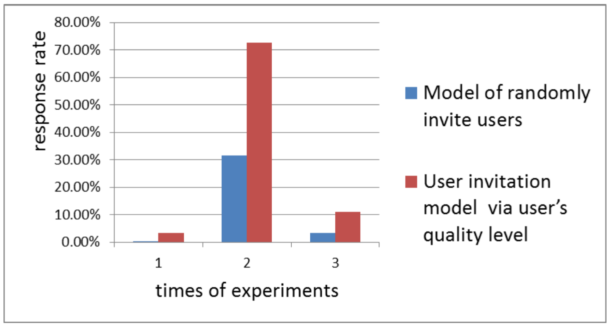

Figure 5. Experimental results for comparison between randomly invited users model and user invitation model based on user's quality

As shown in Fig. 5, the comparison results demonstrate that based on the clustering analysis of user quality can improve the response rate for online surveys, which can reduce the number of invitations as well as costs.

B. User Invitation Model with the Combination of Similar Users and User Preferences

a. Experimental design

Clustering analysis based on user preferences is mainly reflected in the type of questionnaire, by selecting the user's behavior features to represent each of users to cluster analysis. On the basis of selected attributes as shown in Table 2, experimental data include the number of successfully answering the first type of questionnaire, the number of successfully answering the second type of questionnaire, …, and the number of successfully answering the tenth type of questionnaire. The actual historical data comes from an online survey company, and the data set contains 329,744 records. Table 5 shows the results that are based on clustering algorithm of users' preferences.

Table 5. Clustering results based on user preferences

Id/user id	*The group belongs to the clustering results*
7912, 39922, 113938, 797783, …	1
402, 7958, 33898, 1666057, …	2
1976, 44544, 533535, 1479379, …	3
36699, 113109, 1091933, 1565479, …	4

Clustering analysis based on similar users chooses key features to represent different users. The same or similar characteristics among users indicate that they belong to the similar user group. The key attributes selected consist of age, sex, mobile phone verification, personal income, education level. The experiment results of clustering analysis based on similar users are shown in the following Table 6.

Table 6. Clustering results based on users' similarity calculation

Id/user id	*The group belongs to the clustering results*
121, 5737, 99622, 105346, 819051,	a
30452, 104732, 218202, 923057, 300230, …	b

7912, 94787, 195740, 1012363, 300230, …	c
1051, 89062, 111567, 1750053, …	d

In addition, we select users' preference questionnaire type 3 as the first category of users and select user's preference questionnaire 5 as the second category of users. The experimental results between randomly invited user model and user invitation model with the combination are shown in the following Table 7.

Table 7. Clustering results based on the combination of user invitation model

Statistics parameter ***Invitation model***		questionnaires of type 3	questionnaires of type 5
Random invitation model	total number	1180430	706240
	total number of success	135019	54242
	success rate	***11.44%***	***7.68%***
User invitation model with combination	total number	81659	42499
	total number of success	48488	15433
	success rate	***59.38%***	***36.31%***

b. Results analysis

From the results as shown in Table 7, we can find that through our proposed user invitation model with the combination of user preferences and similar users, we could develop appropriate sending strategies. First type users and second type users improve obviously in success rate, thus naturally all users' response rate also get improved. Additionally, several group experiments get success number similar, total invited number greatly reduced, reducing the costs of sending questionnaire for companies. At the same experimental condition, as the experimental data increases, the success rate will be a measurable reduction.

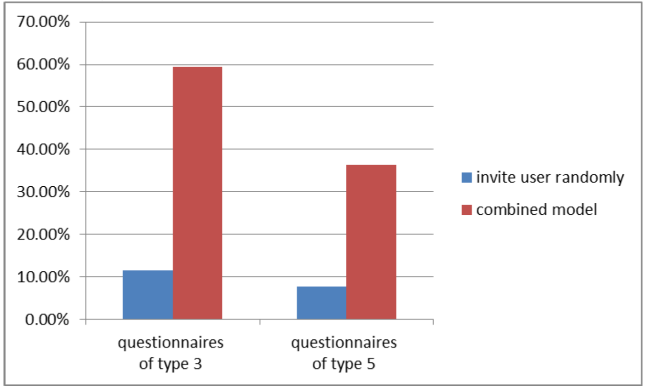

Figure 6. Comparison results of success rates between invite randomly invited user's model and user invitation model with the combination.

Experimental results reveal that the success rate of user invitation model based on combination is higher than the model of randomly invited users. The results verified the feasibility and effectiveness of our proposed invitation model with the combination. Thus, it could increase response rate for the online questionnaire survey companies.

V. CONCLUSIONS

To solve the problem of low response rate for online questionnaire survey, two user invitation models from user's quality level and combination are respectively proposed based on improved k-means clustering algorithm. The results demonstrate that both two invitation models improve user response rate. Therefore, the best model should be chosen to send invitations depending on the different application scenarios. When an online survey company needs to invite users to participate in a survey, they could rely on user invitation model based on user's quality level to send invitations to only higher-quality level and high-quality level users first. Second, they could further extract the type of the questionnaire and adopt user invitation model with the combination to send invitations. The steps above solve the problem of low response rate of online questionnaire survey.

In the future, our improved k-means++ algorithm only overcomes one drawback of the basic k-means algorithm, which strengthens the selection of the initial cluster centers by certain strategy. However, k-means++ algorithm does not solve the problem that the value of k still needs to be given. Thus, how to determine the value of k and whether the value of k is reasonable would be our concern for further work.

VI. ACKNOWLEDGMENT

This work was partially supported by a National Natural Science Foundation of China (61303096), a Shanghai Natural Science Foundation (13ZR1454600), an Innovation Program of Shanghai Municipal Education Commission (14YZ017), and a Specialized Research Fund for the Doctoral Program of Higher Education (20133108120029). We also would like to thank all of the anonymous reviewers for your insightful comments and useful suggestions.

REFERENCES

[1] Fayyad U. S., Djorgovski, and N. Weir. Automating the analysis and cataloging of sky surveys. Advances in Knowledge Discovery with Data Mining, pp.471-493, 1996.

[2] Verma M, Srivastava M, Chack N, et al. A comparative study of various clustering algorithms in data mining. International Journal of Engineering Research and Applications (IJERA), No. 2, pp.1379-1384, 2012.

[3] Shao Peiji, Fang Jamming, Decou Judy, et al. Research of the Web-based survey on the development of e-business in China, Canada, and Taiwan. In Proceedings of the 5th Asian e-business workshop. pp. 172-177, 2005.

[4] Jiaming F. The effect of incentives in Web surveys: results from three meta-analyses. Management Review, No. 2, 2013, pp.79~87.

[5] Duan W, Ma L, Xiang-Yang L U. An Improved Algorithm of Decision Tree Based on Information Gain and Minimum Distance Classification[J]. Science Technology & Engineering, 2013. Vol. 13, No. 6, 2013, pp.1671~1815

[6] Apache common data repository. http://commons.apache.org/, 2012.