Multiple User Characteristic Models for Online Survey based on FP-Tree Algorithm

Kaiqiang Guo, Shaochun Wu, Guobing Zou, Honghao Zhu
School of Computer Engineering and Science, Shanghai University
Shanghai 200444, China
lo_wolf@163.com, scwu@shu.edu.cn, gbzou@shu.edu.cn, wozhuhonghao@163.com

***Abstract*—Online survey recently has received many attractions and become an important way for enterprises to accumulate data and facilitate their business development. Under the analysis of user attributes, user behavior and survey attributes in the domain of online survey, this paper proposes five user characteristic models for online survey based on improved FP-Tree algorithm. First, by analyzing the attributes of user and online survey, the users are divided into different categories, using our proposed improved FP-Tree algorithm to mine frequent items of user characteristics. By doing so, we then discover association rules between the user and the survey. Based on generated association rules, multiple user characteristic models are built to support enterprise for online surveys. Experimental results show that the improved FP-Tree algorithm can significantly benefit the performance compared with the traditional algorithm. By the analysis of different user characteristics models, it is concluded that there are obvious characteristics of online survey user and strong association rules between user attributes and the survey type.**

Keywords—Online survey, FP-Tree, Frequent items, Association rules, User characteristics

I. INTRODUCTION

With the rapid development of Internet and its applications, online survey is gradually replacing the position of the traditional artificial way. The majority of enterprises prefer the former way to investigate markets. However, for the user, they hope to obtain surveys sent by the companies, which they are interested in. In fact, many companies cannot do this that pushes the targeted survey to the desirable users. As a result, receiving feedback surveys from users has become a huge challenge to most of the companies.

In order to solve the problems existing in this field, data mining technique has been widely taken. General application of data mining techniques in the online survey research can be divided into two main research directions, namely survey-oriented data mining and user-oriented data mining. The former research aims at building the knowledge base, and filling a vacancy, etc. The later one is to make use of analyzing the basic information of the user, and their answered surveys behavior data.

In this paper, we make use of the association rules algorithm—FP-Tree to mine users' common characteristics, in order to achieve the purpose of pushing the survey to its desirable user, and to provide enterprises with a reasonable online survey strategy. We conduct extensive experiments to validate the effectiveness of our proposed user characteristic model. Compared with the existing one, the experimental results demonstrate that our model can be conveniently deployed for enterprise online survey.

II. RELATED WORK

In the process of development of online survey, there are many problems, such as the authenticity of users participated in the online survey. Above all the potential risks and obstacles, the most challenging one belongs to its low response rate. Some work [1] [2] show that the successful response rate only ranges around 2%. To deal with this knottiness, many companies are always trying to explore some methods and improving the user's reward as their primary consideration. However, the authors in [3] find that increasing the reward is not decisive.

Although data mining techniques could be adopted, its application is still in a primary stage in this field. In the work [4], the authors point that how to ensure the safety of user privacy is a significant topic. In [5], it reveals processing data with huge dimension in the use of classification or cluster algorithm is extremely troublesome.

We bring in association rules algorithm, by extracting the basic information of the users in the online survey and cleaning data. Then, we use the improved FP-tree algorithm to build a user characteristic model that is applied to mine the potential users' characteristics (frequent items) and strong association rules. Finally we reach a target of pushing the survey in accordance with users' characteristics.

III. THE USER CHARACTERISTIC MODEL BASED ON IMPROVED FP-TREE ALGORITHM

A. Data Preparation

In online survey, the original data consists of user basic attributes, survey attributes, and answered survey data. In our study, the following data is collected. (1) User data, namely user basic attributes. (2) Survey data, namely survey basic attributes. (3) User behavior data, namely users' behavior attributes, including the type of the answered survey, state of the blacklist and whether exchange prizes or not. The last two values are in the state of Boolean, '1' is yes, '0' stands for no.

Except the first two kinds of data, users' behavior data must be calculated through the basic data. Here is the analysis of these two kinds of data particularly.

1) Calculation of User Attributes

Through the analysis of the online survey, the user basic attributes mainly contain "name", "gender", "age", "marital status" (1 for yes, 0 for no), "from" (registered channels—ali, web), "city", "province", "education level", "income level", "occupation", "panel type" (shop panel, car panel, mass panel, baby panel), "mobile valid" (1 for yes, 0 for no), "e-mail valid" (1 for yes, 0 for no) and so on. Properties of user behavior include screening rate, success response rate, whether or not in the blacklist, whether or not exchange prizes.

978-1-4673-7211-4/15 $31.00 © 2015 IEEE
DOI 10.1109/UIC-ATC-ScalCom-CBDCom-IoP.2015.324

Detailed information of user basic attributes and behavior attributes are illustrated as shown in Figure 1 and Figure 2.

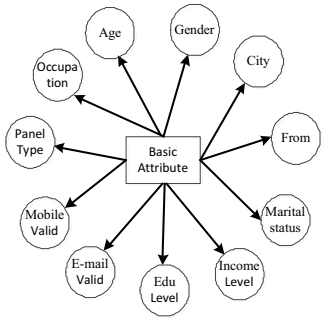

Figure 1. User basic attributes

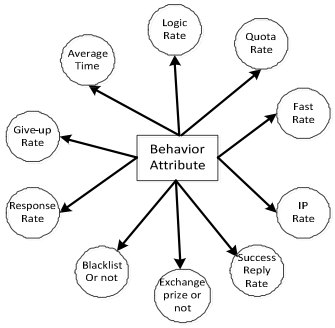

Figure 2. User behavior attributes

User basic attribute has been discussed, without any additional calculation. According to the field of online survey, properties of user behavior mainly are different screening rates, which means that the conditions of the users do not conform to the requirements of the survey and users must be screened, including the logic screening rate, quota screening rate, fast screening rate, IP screening rate, response rate, give-up rate. These properties need to be calculated with basic data. The specific calculation formulas for those user behavior attributes are as follows.

a) Logic Screening Rate

$$\text{logicRate} = \frac{\text{f\_surveyLogicScreeningCount}}{\text{f\_surveyResponseCount}} \quad (1)$$

f_surveyLogicScreeningCount represents the total count of the user screened by logic, while f_surveyResponseCount is the total count of the user response to the survey.

b) Quota Screening Rate

$$\text{quotaRate} = \frac{\text{f\_surveyQuotaScreeningCount}}{\text{f\_surveyResponseCount}} \quad (2)$$

f_surveyQuotaScreeningCount stands for the total count of the user screened by quota, and f_surveyResponseCount is the total count of the user response to the survey.

c) Fast Screening Rate

$$\text{fastRate} = \frac{\text{f\_surveyFastScreeningCount}}{\text{f\_surveyResponseCount}} \quad (3)$$

f_surveyFastScreeningCount is the total count of the user screened by fast answering, and f_surveyResponseCount represents the total count of the user response to the survey.

d) IP Screening Rate

$$\text{IPRate} = \frac{\text{f\_surveyIPScreeningCount}}{\text{f\_surveyResponseCount}} \quad (4)$$

f_surveyIPScreeningCount is the total count of the user screened by IP address, and f_surveyResponseCount stands for the total count of the user response to the survey.

e) Response Rate

$$\text{responseRate} = \frac{\text{f\_surveyReplyCount}}{\text{f\_surveyResponseCount}} \quad (5)$$

f_surveyReplyCount is the total count of the user replying to survey, while f_surveyResponseCount represents the total count of the user response to the survey.

f) Give-up Rate

$$\text{giveupRate} = \frac{\text{f\_surveyGiveupCount}}{\text{f\_surveyResponseCount}} \quad (6)$$

f_surveyGiveupCount is the total count of the user giving up the survey, and f_surveyResponseCount stands for the total count of the user response to the survey.

g) Success Reply Rate

$$\text{successReplyRate} = \frac{\text{f\_surveySuccessReplyCount}}{\text{f\_surveyInviteCount}} \quad (7)$$

f_surveySuccessReplyCount is the total count of the user successful reply to the survey, and f_surveyInviteCount stands for the total count of the user invited.

h) Tolerance--Average Time

The average time of the user taking in finishing all his or her surveys. It can be calculated as follows.

$$\text{averageTime} = \frac{\text{f\_surveySuccessTimeSum}}{\text{f\_sAmount}} \quad (8)$$

f_surveySuccessTimeSum represents the total time of the user taking in finishing all his or her surveys, while f_sAmount is the total count of the user successfully finishing the online survey.

2) Generalization of User Attributes

To be facilitated, there is a great need to generalize some attributes. User behavior attributes except the Boolean attributes, need to process generalization operations. The k-means++ clustering algorithm is taken into account to analyze attributes for user behavior. Generalization results of user behavior attributes are shown in the following TABLE I.

TABLE I GENERALIZATION OF USER BEHAVIOR ATTRIBUTES

Attribute	Cluster interval	Levels
Logic Rate	[0,0.25)	Lower
	[0.25,0.5)	Low
	[0.5,0.8)	High
	[0.8,1]	Higher
Quota Rate	[0,0.2)	Low
	[0.2,1]	High
Fast Rate	=0	Low
	>0	High
IP Rate	=0	Low
	>0	High
Response Rate	[0,0.3)	Low
	[0.3,0.5)	Middle
	[0.5,1]	High
Give-up Rate	=0	Low
	>0	High

Success Reply Rate	=0	Lower
	(0,0.2)	Low
	[0.2,0.5)	High
	[0.5,1]	Higher
Average Time	(0,100]	Lower
	(100,500]	Low
	(500,1000]	High
	(1000,+]	Higher

3) Analysis of Survey Attributes

Survey attributes are important factors for online survey. They describe the various features of online survey, mainly including "ID", "name", "type", "length", "survey description", "creation time", etc. They are shown in Figure 3.

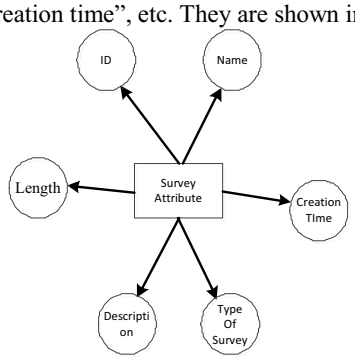

Figure 3. Analysis of survey attributes

In the process of creating the model, we utilize "type of survey" and "length". The first one is used to analyze the association between type of survey and user attributes. "length" is mainly calculated by the equation (8).

B. Improved FP-Tree Algorithm

Han, etc.[6] present FP-tree algorithm to generate frequent itemsets without candidate itemsets. However, one obvious weakness is that the algorithm has to scan the database twice. In such a case, when there are thousands of records, it is not allowed to scan the database for many times. Here are the improvements as follows.

(1) First, it scans database and save the records to a TXT file. The attributes are split by separator '#'.

(2) Second, TXT file is analyzed to build a FP-tree.

(3) While scanning the condition FP-tree, we mine frequent patterns according to the length of the path P, instead of calculating all the combinations of P. The improved FP-tree algorithm is described as below.

Input: D: Database Transaction
minSupp: minimum support
minConf: minimum confidence
Output: Frequent items and association rules in descending order by support and confidence, respectively.
Step 1: Scan all the data records in the database and save the records into a TXT file.
Step 2: Scan the TXT file line by line, save the attribute line into a List.
Step 3: Set parameters: minSupp and minConf.
Step 4: Generating frequent itemset 1 with the list. Then creating the FP-tree with list.
Step 5: Starting from the original suffix pattern (itemset 1), build its conditional pattern base, delete the item that is lower than the minSupp at the same time, then generate the conditional FP-tree. Traverse every single path on the tree, dealing with the path--P as following.
Building itemset 2, add each of the items on the P to itemset 2.
Building item set 3, P Length=2, add all the items on P to the item set 3; P Length>2,build all the combinations of all items on P (the size is 3 at least) to generate the itemset 4 or larger.
Step 6: Sort all the patterns in descending order by the support, and generate the association rules. Save them to TXT file.

C. Construction of User Characteristic Model

To make comprehensive analysis of the correlation between user characteristics and online survey, five kinds of characteristic models are constructed. The basic input/output are shown in TABLE II.

TABLE II BASIC INPUT/OUTPUT

Input	Parameters	Output
Set of user attributes Set of survey attributes	minSupp minConf	Frequent pattern P Rules R

Set is formally defined as S = {A1 = V1, A2 = V2, ..., An =Vn}, where Ai stands for an attribute name, and Vi is for the corresponding attribute value.

According, frequent pattern P is formalized as P = < I1, I2, ..., In, Sup>, where Ii represents item and Sup is used for the support degree.

Association rules are formally defined as R = {I1, I2, ..., In→ In+1, Con}, where Ii stands for an item, and Con represents the confidence degree.

The models built are specified as follows.

1) Characteristic model of tolerance

Accumulative time of completed surveys is initially calculated for each user to analyze the average effective time, namely the taken time used to complete every survey. Then we use improved FP-tree algorithm to offer the relationships between user attribute and the different tolerance levels as shown in TABLE I.

2) Characteristic model of health

Relationships between "user health", including all the rates, and user attributes are evaluated at all the levels shown in TABLE I.

3) Characteristic model between user and survey

The same user in response to different surveys will show some correlations. Meanwhile, different users in response to different surveys also have different characteristics. So the improved FP-tree algorithm is exploited to mine user common characteristics of different types of surveys, and excavate the correlation of the same user in response to different types of survey, such as to recommend the survey to a user.

4) Characteristic model of whether the user is in the blacklist or not

Previous work shows that the blacklist and non-blacklist user perform various characteristics. For example, the blacklist users often give up the survey or answer casually. On the contrary, non-blacklist is much more reliable to respond to a survey. Under this assumption, this model is applied to analyze the performance characteristics of these users to help an enterprise find those who may be in the blacklist in time to reduce their cost by reducing sending surveys to them.

5) Characteristic model of exchanging prizes

A conclusion is drawn by the feedback from the enterprise data that the prize stimulates the behavior of user replying to surveys to a certain degree. Besides, users have distinct behavior in exchanging prizes. For a further study of excavating those users' characteristics to verify the stimulatory effect, establishing this model will be effective to find the potential relationship.

IV. EXPERIMENTAL EVALUATION

A. Experimental Setup

We select 490,085 valid user records as the initial user sets from the actual operating database of a company and choose user attributes and survey attributes as described in section III as the input. Configuration of the experimental environment is as follows. Windows 7 is set up as OS and our experiment runs on Intel i3-2130 processor. The memory size is 8GB and local MySQL database is deployed to store and manage survey datasets.

1) Experiment 1

To illustrate the effectiveness of the improved FP-tree algorithm and compare with the traditional FP-tree one, 11696, 76387 and 467655 user records are chosen as test input data set. The output data contains all the frequent patterns and all the association rules produced by these two algorithms, respectively. The performance is measured by the total taken time.

2) Experiment 2

For the five models, the experiments are designed as follows.

(1) For model 1, input data is shown in TABLE III.

TABLE III INPUT DATA FOR CHARACTERISTIC MODEL 1

Level	Number of records
Lower	48265
Low	84954
High	26775
Higher	9623

(2) For model 2, input data is shown in TABLE IV.

TABLE IV INPUT DATA FOR CHARACTERISTIC MODEL 2

User Label	Number of records
Fast Rate	1093
IP Rate	19903
Logic Rate	230537
Quota Rate	57377
Response	66103
Give-up	26254
No Give-up	106552
Success Reply	53010
Failure Reply	211616

(3) For model 3, the 6 types of surveys and those who complete the surveys are shown in Figure 4.

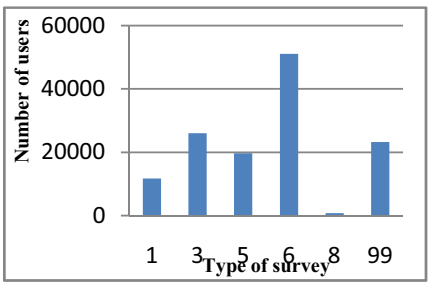

Figure 4. Number of users completing the survey

As shown in Figure 4, there are few users completing the type 8. As a result, the experiment chooses the type 1, 3, 5, 6, 99 and their associated users as input data for model 3.

(4) For model 4, input data is shown in TABLE V.

TABLE V INPUT DATA FOR CHARACTERISTIC MODEL 4

User Label	Number of records
Blacklist	22430
Non Blacklist	467655

(5) For model 5, input data is shown in TABLE VI.

TABLE VI INPUT DATA FOR CHARACTERISTIC MODEL 5

User Label	Number of records
Exchange Prizes	238736
Non Exchange Prizes	17394

B. Experimental Results and Analysis

1) Performance Comparison

The performance between our proposed improved FP-tree algorithm and the traditional one is shown in Figure 5 and Figure 6, respectively.

As shown in the figures: The former figure points that when the size of data set is small, the improved one is much faster and it has a good performance. And when the size is larger shown in figure 8, the improved one can still reduce the time in generating the frequent patterns.

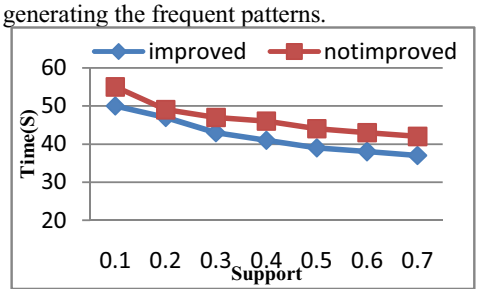

Figure 5. Performance comparisons (Data Set 76387)

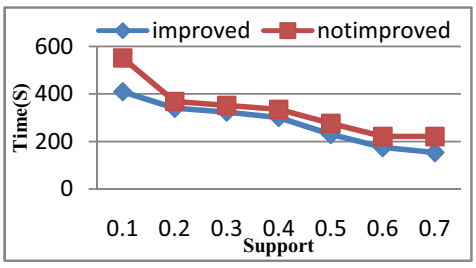

Figure 6. Performance comparisons (Data Set 467655)

2) Results of the Five Characteristic Models

(1) For model 1, the characteristics of tolerance in four levels are shown in TABLE VII. Here, the parameter of minSupp is assigned by 0.6.

TABLE VII RESULTS OF CHARACTERISTIC MODEL 1

Level	Characteristics
Lower	Mobile=1,From=ali Panel=shop &sass
Low	Mobile=1,From=ali Panel=shop
High	Gender=male, age=29~40
Higher	Gender=female, Mobile=1 From=ali, Panel=Shop

(2) For model 2, a part of characteristics of "user health" is shown in TABLE VIII. Here, we configure the parameter minSupp by 0.6.

TABLE VIII RESULTS OF CHARACTERISTIC MODEL 2

Health	Characteristics
High Response	Gender=male, E-mail=1
Low Response	Gender=male, Age=18~28
Give-up	Gender=female, Panel=shop
Non Give-up	Gender=male, Pane=shop

Under the above experimental results, we conclude that female user is more likely to give up a survey, while male users except the age between 18 and 40 have a relatively high response. Therefore, they should be considered to be sent surveys first when meeting the quota.

(3) For model 3, TABLE IX shows the users characteristics in different types, and association rules among the survey types are shown in TABLE X (minSupp=0.2,minConf=0.7). TABLE XI shows the association rules between user attributes and the survey types (minSupp=0.5, minConf=0.8).

TABLE IX USER CHARACTERISTICS IN SURVEY TYPES

Survey Type	Characteristics
1	Gender=female, Marry=1 Age=29~40, Panel=car
3	Gender=male, Age=29~40 Education=2
5	From=ali, Marry=1, Panel=car
6	From=ali, Education=2
99	Age=29~40, Marry=1, From=web

TABLE X ASSOCIATION RULES AMONG SURVEY TYPES

Rules	Interpretation
5,6→3:0.76	For those who answer the type 5 and 6, there is a possibility of 76 percent to answer type 3.
6,99→3:0.82	For those who answer the type 6 and 99, there is a possibility of 82 percent to answer type 3.
1,6→3:0.86	For those who answer the type 1 and 6, there is a possibility of 86 percent to answer type 3.

TABLE XI ASSOCIATION RULES BETWEEN USER ATTRIBUTES AND SURVEY TYPES

Rules	Interpretation
From=ali→Type=6 0.76	For those who are from ali, there is a possibility of 76 percent to answer type 6.
Panel=shop→Type=6 0.76	For those who are shop panel, there is a possibility of 76 percent to answer type 6.
Type=3, Mobile=1 →Type=99:0.82	For those whose mobile is valid and answer type 3, there is a possibility of 82 percent to answer type 99.

(4) For model 4, TABLE XII shows the users characteristics in blacklist and non-blacklist. Here, the parameter is adjusted to 0.6.

TABLE XII USER CHARACTERISTICS IN BLACKLIST OR NOT

User Label	Characteristics
Blacklist	Gender=male, Mobile=0, Marry=0, Panel=mass
Non-blacklist	Gender=male, Mobile=1, Age=18~28, Education=2

From the experimental results, we conclude that the attribute–"mobile valid" is of great significance for users. Company should try best to let the registered user to validate their mobile phones.

(5) For model 5, characteristics of users who exchange prizes and don't exchange prizes are shown in TABLE XIII.

TABLE XIII USER CHARACTERISTICS OF EXCHANGING PRIZES OR NOT

User Label	Characteristics
Exchange Prizes	Age=29~40, Mobile=1, E-mail=1
No Exchange Prizes	Gender=male, From=web, Panel=mass

From above results, the conclusion is that the company should encourage the "no exchange prizes" users to answer the survey by offering them more opportunities to exchange prizes.

V. CONCLUSION

In this paper, the early investigation indicates that the potential value of the data from online surveys is considerable. To uncover these characteristics between user and survey, we proposed an improved FP-tree algorithm to find association rules for effective online survey. According to the online surveys, user data and the answer data, we first clean the raw data and calculate the behavior attributes. Then, we take advantage of these basic data and improve the algorithm of FP-tree. Finally, we build five kinds of user characteristics models, and get better experimental results compared to the traditional one.

ACKNOWLEDGMENT

This work was partially supported by the National Natural Science Foundation of China (61303096) and Shanghai Natural Science Foundation (13ZR1454600).

REFERENCES

[1] Shao Peiji, Fang Jianming, Decou Judy, et al. Research of the Web-based survey on the Development of E-Business in China, Canada, and Taiwan [C]. In Proceedings of the 5th Asian e-business workshop. Jeju, Korea: Kaisat Press, 2005, pp.172~177.

[2] Nina, R., T. Sharon, Several factors affect e-research validity [J]. Marketing news, No. 15, 2003, pp.50~53.

[3] Fang Jiaming.The Effect of Incentives in Web Surveys: Results from Three Meta-Analyses [J]. Management Review, No. 2, 2013, pp.79~87.

[4] Torra V, Navarro-Arribas G. Data privacy[J]. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2014, 4(4):269–280

[5] Shinde A, Church G, Janakiram M. Feature extraction and classification models for high-dimensional profile data[J]. Quality and Reliability Engineering International, 2011, 27:885–893.

[6] Han J, Pei J, Yin Y. Mining frequent patterns without candidate generation [C]. In Proceedings of the Acm Sigmod Record. Acm, 2000, 29(2): 1-12