# Online Survey Prediction Model for High Response Rate via Decision Tree

Naibin Luo, Shaochun Wu, Guobing Zou, Xiang Shuai
School of Computer Engineering and Science, Shanghai University
Shanghai , China
jsluonaibin@163.com, scwu@shu.edu.cn, gbzou@shu.edu.cn, alxbeang@gmail.com

*Abstract*—**In recent years, online survey has received many attractions and become an important way for companies to promote their business development. According to the characteristics of the accumulated online survey data, we present a new approach to construct a prediction model which is based on ID3 algorithm to forecast users' response behavior. Especially, we introduce XML to store the prediction model on the basis of the structure property of decision tree. We collect lots of online survey data and validate the effectiveness of our proposed approach. According to analysis of experiment results, the main contributions of this paper are three-fold: (1) response rate based on prediction model of decision tree is higher than that of random way; (2) the improved prediction model of ID3 can improve the crawling rate of classic ID3 algorithm; (3) the cost of online survey can be reduced by the prediction model effectively.**

***Keywords-online survey; prediction model; decision tree; response rate***

## I. INTRODUCTION

The online survey is an essential method to collect data online in the field of social investigations. In the traditional case, questionnaires and interviews can be used to collect data. Manually releasing and recycling of paper questionnaire are needed by survey firm. This period of this way is too long to make analysis and statistics, while needing more cost as well. However, with the rapid development of Internet and its applications today, researchers could collect data needed from the Internet rather than via manually recycling of paper questionnaire. Nowadays, online survey is more popular which mainly consists of some websites. In addition, the way of online survey is simpler and faster than traditional survey while the cost has reduced obviously [1].

Although there are many advantages and benefits in online survey, the response rate in online survey is still low which exists commonly in the field of online survey, and has become a huge challenge to be solved. Thus, it is the hot-pot issue[2]. When the questionnaire is provided for the user sample, the investigation agency could not make sure that the questionnaire is replied availably, leading to loss of plenty of questionnaire resources wasted, besides while consuming adding additional the costs of for online survey. Thus, how to accurately recommend surveys to those desirable users and acquire high response from them is the core issue we tackle. Based on this research challenge background, we proposed a new and effective prediction model for online survey with high response rate. Firstly, we will classify the basic information of user samples and plenty of historical questionnaires which have been replied into different categories via the method of data mining algorithm, and then we build a prediction model in order to forecast response behavior of user samples. To validate the feasibility of the proposed model, we conduct extensive experiments. The results demonstrate the effectiveness of the prediction model compared to the traditional random way.

The remainder of this paper is arranged as follows. Section II reviews the related work, while prediction model via ID3 algorithm is presented in Section III. We present experimental results and analysis in Section IV. Finally, Section V concludes the paper.

## II. RELATED WORK

The research and application in online survey abroad developed in late 90's. GVU Center of Georgia Institute of Technology used online survey to conduct Internet usage at first time. This was the first time for human beings to use online survey to investigate and accumulate data. The research on online survey field abroad went early. It has made many empirical researches. Besides, JM Stanton from Syracuse University presented their findings about collecting data from online survey on the Internet. Recently, the research of online survey is mainly focusing on response rates, survey data quality problems and the certain issues around effective survey. MD Kaplowits demonstrated that the design way would significantly affect the response rate on online survey [3]. In addition, JS Laguilles raised lottery incentives that can improve the rate on questionnaire [4].

Network investigation in domestic went on lately. Some descriptive research work had been published gradually since 2005. With the fast development of Internet technology those years, online survey research has become a hot-pot issue. Huang et al proposed the method of questionnaire design based on social network [5]. Zhang et al presented a method of advertising effect on the basis of investigation [6], so the data mining techniques have become more and more important in online survey field. Chen et al presented a method of online survey system based on data mining technique [7]. It mainly discussed the issue of missing data on online survey questionnaire and took out missing data filling algorithm based on s-SM.

Recently, the research on data mining in online survey field is still worth conducting not only in abroad but also in domestic. This paper aims at investigating the prediction model for high response to online survey via classification algorithm in data mining.

978-1-4673-7211-4/15 $31.00 © 2015 IEEE
DOI 10.1109/UIC-ATC-ScalCom-CBDCom-IoP.2015.325

IEEE computer society

## III. CONSTRUCTION OF PREDICTION MODEL BASED ON ID3 ALGORITHM

### *A. Improved ID3 based on Samples Active*

There is a large amount of data about response records in online survey system. According to limited attributes of samples and questionnaires, it is hard to sort completely by training data. In order to make the majority vote in traditional ID3 algorithm work effectively, it can not only represent the subset of data classification, but also reflect the current subset classification objectively. In this paper, sample response activity weight is added to each record in majority vote. In other words, each record in response is taken by the weight for 1. In this case, majority vote cannot reflect different weight of each sample in each subclass. But in reality, the answer record's response can also apply probability theory to reflect the current probability. We define the sample activity which accounts for records replied in the total survey. As a result, we introduce sample response activity in ID3 algorithm as weight in the process of majority vote. So improved counting formulas in majority vote are as follows:

Votes of Reply:

$$V_a = \sum_i^n c_i \tag{1}$$

Votes of No Reply:

$$V_b = \sum_i^m 1 - c_i \tag{2}$$

Activity of Sample: $c_i$

$$c = \frac{R}{N} \tag{3}$$

R presents the number of reply and N is the total number of online survey. Under the equation 1, we can obtain $V_a$ which is vote for reply. In equation 2, $V_b$ can be got as voting for no reply. If $V_a > V_b$, return the reply, otherwise we return no reply. The results can not only reflect the recovery of the current sample, but also reflect the high active.

### *B. Construction of Predition Model*

Response quality is decided by attributes of users and questionnaires. Therefore, it can reach relational model of each attribute and response behavior by the analysis of all related attributes via decision tree. Analyzing attributes of questionnaires is adopted by the method of construction, and the response condition is appointed as the target attribute. So a decision tree model will be built and it can predict response behavior in accordance with ID3 algorithm. The model can make a highly accurate prediction by related attributes so that it could improve the response rate.

#### *1) The Input Data of Training Model*

Each related attribute in online survey needs to be analyzed and generalization of data collection for training should be calculated that is adopted in ID3. After passing through the property reduction and correlation analysis, it can get larger relevance of questionnaire collection as shown in the following Table 1.

TABLE 1 ATTRIBUTE REDUCTION AND INFORMATION GAIN COMPUTATION

| ID | Attribute | Attribute Values | Info Gain |
|---|---|---|---|
| 1 | **Gender** | Male; Female | 0.0751 |
| 2 | Occupation | Director; Manager; Staff: Other | 0.0008 |
| 3 | Income | Low; Mid;Upper; High | 0.0182 |
| 4 | **Education** | Pri; Mid; BSC; MSC;PHD | 0.0936 |
| 5 | **Marriage** | Married; Unmarried | 0.0220 |
| 6 | **Mobile** | Yes; No | 0.0437 |
| 7 | **Email** | Yes; No | 0.0481 |
| 8 | **From** | Inner; Outer | 0.0213 |
| 9 | AvgQues/Year | Few: Less; More; Multi | 0.0058 |
| 10 | Success Res Rate | Low; General; High | 0.0076 |
| 11 | Log Times/Year | Low; General; High | 0.0097 |
| 12 | **Ques Type** | Consumption; Product; Ad; Satisfaction; Rapid; Tracking; Contact; Other | 0.0811 |
| 13 | Ques Size | Short; Mid; Long | 0.0213 |

In Table1, it is still reached as many as 13 through the analysis of properties. If we adopt all 13 properties into the ID3 algorithm, the calculation is still of high complexity as we need to compute the information gain for every attribute, especially when facing millions of records of survey data. That will affect the prediction efficiency of response rate for online survey. As a result, it is still mandatory to select about 7 properties from reduced attributes for a better effect. As the input data, it can accelerate the computation speed and improve the prediction efficiency of online survey [8].

In information theory, the higher the entropy is, the more confused the system is. On the contrary, the lower the entropy is, the better the system for class recognition is. Shannon firstly proposed the formula of calculating the entropy of information by using the following equation.

$$H = \sum_{k=1}^n p_k \log_2 p_k \tag{4}$$

In formula 4, n represents the number of symbol of consisted information (i.e., the number of classes we need to differentiate), while $p_k$ stands for the probability of each class among all of the dataset.

All information gain by formula 4 as Table 1 shown can be calculated easily. From the table, the Top-K (K=7) attributes in terms of information gain are *Gender, Education, Marriage, Mobile, Email, From, and Ques Type.* In addition, considering in online survey there are only a few type of questionnaires. So we can train the certain type of questionnaire in decision tree model for improving the efficiency and prediction accuracy. The input sequence is as follows.

*<Gender, Education, Marriage, Mobile, Email, From, Ques Type>*

Among the sequence, the *Ques Type* is the target attribute in ID3. There are three types in *Ques Type,* which are response and succeed, response but not succeed and no response.

#### *2) Prediction Model of Decision Tree*

Repeated training for a certain type of questionnaire from training data, we can induce a decision tree model based on ID3 for prediction of user behavior. The Ques Type is as a target attribute in the prediction model. Every node in decision tree is an input attribute in algorithm. Each branch is a division of an attribute's possible value. For example, a simple decision tree prediction model is illustrated as Figure 1.

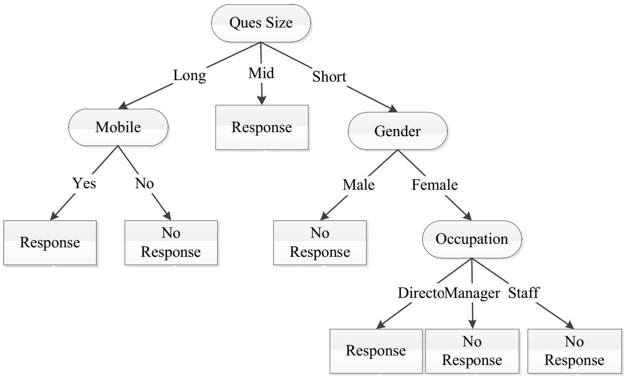


Figure 1 A decision tree model for response behavior

### C. *XML Storage Structure for Predicion Model*

XML is a markup language for information storage and propagation by containing structured information. It is a very flexible structure and can transmit across different platforms. For a specific data operation and management purpose, it stores data for a random structure under the condition of predefined syntax regulations. In realistic world, information could be used to describe objects, while XML could always be described that it transfers between different systems conveniently. It consists of all kinds of nodes which can be created and organized like a tree. Under these advantages, XML is suitable for describing a decision tree prediction model we build for user class recognition in online survey here.

Under the characteristics and regulations in XML, we use XML to depict a decision tree for Figure 1 as follows.

```
<?xml version="1.0"?>
<root>
<Ques Size = Long>
<Mobile = Yes>
Response
</Mobile>
<Mobile = No>
No Response
</Mobile>
</Ques Size>
<Ques Size = Mid>
Response
</Ques Size>
<Ques Size = Short>
<Gender = Male>
 No Response
</Gender>
<Gender = Female>
<Occupation = Director>
 Response
</Occupation>
<Occupation = Manager>
 No Response
</Occupation>
<Occupation = Staff>
No Response
</Occupation>
<Gender>
</Ques Size>
</root>
```

### D. *Application of Prediction Model*

After the construction of prediction model for response behavior, we need the model to predict user prediction behavior for online survey. As shown in Figure 1, the training test data below is adopted for prediction in Table 2, among the table, the last column is the prediction result and, the others are the inputs for the prediction model.

TABLE 2 EXAMPLE FOR PREDICTION USED TEST DATA

| **ID** | **Mobile** | **Gender** | **Occupation** | **Ques Size** | **Response** |
|---|---|---|---|---|---|
| 001 | Yes | Female | Director | Long | Yes |
| 002 | Yes | Male | Staff | Short | No |
| 003 | No | Female | Staff | Mid | Yes |
| 004 | No | Male | Manager | Short | No |
| 005 | No | Female | Director | Short | Yes |
| 006 | Yes | Male | Manager | Long | Yes |

The prediction process for each test data and its result are shown as follows.

001: Ques Size (Long) ->Mobile(Yes) ->Response

002: Ques Size (Short) ->Gender(Male) ->No Response

003: Ques Size (Mid) -> Response

004: Ques Size (Short) ->Gender(Male) ->No Response

005: QuesSize(Short)->Gender(Female)->Occupation (Director)->Response

006: Ques Size (Long) ->Mobile(Yes) ->Response

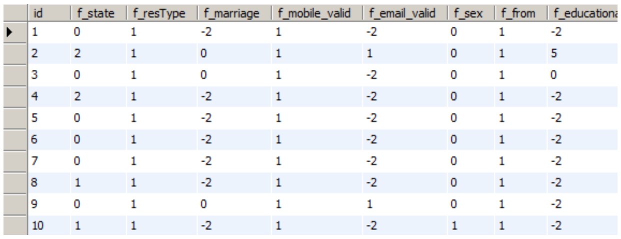
## IV. EXPERIMENTAL RESULTS AND ANALYSIS

### A. *Design of Experiment*

In current online survey system, random sending is used as invitation strategy for filling surveys. In this paper, prediction model is built for predicting response behavior based on ID3 algorithm, thus only those predicted users will receive questionnaires by the decision tree model. Therefore it will significantly improve the response rate with cost reduction.

#### *1) Experimental Data*

Experimental data for prediction of response behavior in online survey includes user information table, questionnaire table and questionnaire trace table. In order to gain the data easily, training data id3_data is constructed which consists of users' ID, questionnaire ID, Ques Type and all attributes shown in Table 1. The partial training data is shown in the following Figure 2.



| id | f_state | f_resType | f_marriage | f_mobile_valid | f_email_valid | f_sex | f_from | f_educationa |
|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 1 | -2 | 1 | -2 | 0 | 1 | -2 |
| 2 | 2 | 1 | 0 | 1 | 1 | 0 | 1 | 5 |
| 3 | 0 | 1 | 0 | 1 | -2 | 0 | 1 | 0 |
| 4 | 2 | 1 | -2 | 1 | -2 | 0 | 1 | -2 |
| 5 | 0 | 1 | -2 | 1 | -2 | 0 | 1 | -2 |
| 6 | 0 | 1 | -2 | 1 | -2 | 0 | 1 | -2 |
| 7 | 0 | 1 | -2 | 1 | -2 | 0 | 1 | -2 |
| 8 | 1 | 1 | -2 | 1 | -2 | 0 | 1 | -2 |
| 9 | 0 | 1 | 0 | 1 | 1 | 0 | 1 | -2 |
| 10 | 1 | 1 | -2 | 1 | -2 | 1 | 1 | -2 |

Figure 2 id3_data for training attributes selection

According to the information gain from Table 1, training attributes (id3_attr_value) could be achieved as shown in Table 3 below. Training attributes could be modified by altering table id3_attr_value and there are no efforts to change the model generation program. The values of attrValue in Table 3 have been converted to numeric type for efficient running with the model.

TABLE 3 ID3_ATTR_VALUE OF SELECTED TRAINING ATTRIBUTES

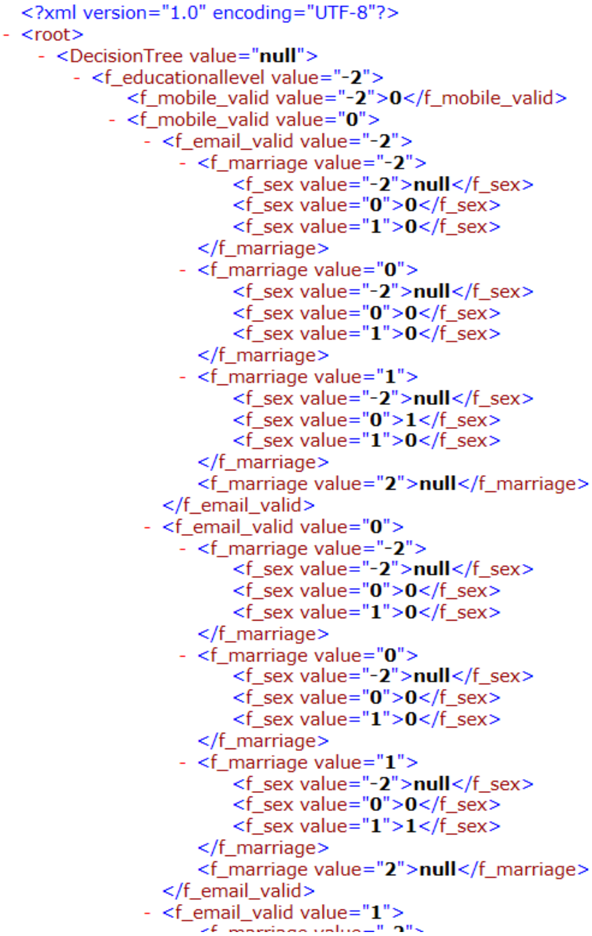
| id | *attrName* | *attrValue* |
|---|---|---|
| 1 | from | 0;1 |
| 2 | mobile_valid | 1;0 |
| 3 | email_valid | 1;0 |
| 4 | marriage | 0;1 |
| 5 | sex | 1;0 |
| 6 | educational | 0;3;4;5;6 |
| 7 | Type | 1;2;3;4;5;6;7;8;9 |
| 8 | state | 1;2;8;21 |

According to the structure of attributes in Table 3, approximate 200,000 sending records of questionnaire have been selected from source data as experimental data. Among them, 80% of records are used as training data and remaining 20% of records are taken as test data.

*2) Experimental Method*

At first we use the training data to train the decision tree model by ID3 algorithm. We mainly focus on offline validation by using test data, instead of test the response of users in the real network. Therefore, remaining data will be tested by the experiment. It is supposed that 20% data of questionnaire was not sent, and then these samples' response behavior can be predicted by the user behavior response model. Meanwhile, by using random sending prediction to each sample of users, experimental results can be compared. It is worthwhile to notice that questionnaire sending style is adopted by random access. Random sending test in our experiments is more convincing to compare with the results.

In the view of questionnaire of Type 1, prediction model which is described by XML is shown below as Figure 3.



```
 <?xml version="1.0" encoding="UTF-8"?>
- <root>
  - <DecisionTree value="null">
    - <f_educationallevel value="-2">
        <f_mobile_valid value="-2">0</f_mobile_valid>
      - <f_mobile_valid value="0">
        - <f_email_valid value="-2">
          - <f_marriage value="-2">
              <f_sex value="-2">null</f_sex>
              <f_sex value="0">0</f_sex>
              <f_sex value="1">0</f_sex>
            </f_marriage>
          - <f_marriage value="0">
              <f_sex value="-2">null</f_sex>
              <f_sex value="0">0</f_sex>
              <f_sex value="1">0</f_sex>
            </f_marriage>
          - <f_marriage value="1">
              <f_sex value="-2">null</f_sex>
              <f_sex value="0">1</f_sex>
              <f_sex value="1">0</f_sex>
            </f_marriage>
            <f_marriage value="2">null</f_marriage>
          </f_email_valid>
        - <f_email_valid value="0">
          - <f_marriage value="-2">
              <f_sex value="-2">null</f_sex>
              <f_sex value="0">0</f_sex>
              <f_sex value="1">0</f_sex>
            </f_marriage>
          - <f_marriage value="0">
              <f_sex value="-2">null</f_sex>
              <f_sex value="0">0</f_sex>
              <f_sex value="1">0</f_sex>
            </f_marriage>
          - <f_marriage value="1">
              <f_sex value="-2">null</f_sex>
              <f_sex value="0">0</f_sex>
              <f_sex value="1">1</f_sex>
            </f_marriage>
            <f_marriage value="2">null</f_marriage>
          </f_email_valid>
        - <f_email_valid value="1">
```

Figure 3 Prediction model is partially described by XML

In accordance with prediction model, the process of predictions and results are as follows.

Process 1: Type 3-->Mobile 1-->Gender 1-->0
Result 1: Forecast Value:0 Real Value:1

Process 2: Type3-->Mobile 1-->Gender 0-->1
Result 2: Forecast Value:1 Real Value:1

Process 3: Type 3--> Mobile 1--> Gender 1-->0
Result 3: Forecast Value:0 Real Value:0

Process 4: Type 3--> Mobile 1--> Gender 0-->1
Result 4: Forecast Value:1 Real Value:1

Process X means the Xth record which is produced by each branch of decision tree model, while Result X consists of the Forecast Value predicted by the model and Real Value provided in reality. Above records reflect the prediction process of response behavior by decision tree model in our experiment.

## *B. Experimental Parameters*

In the experiment, all types of questionnaires are corresponding to Type 1, Type 2, Type 3, Type 5 and Type 8. All experimental parameters are shown in Table 4.

TABLE 4 PARAMETERS SETTING IN THE EXPERIMENT

| **Parameter Name** | **Express Meaning** |
|---|---|
| **Type** | The type of questionnaire. |
| **Amount of Training** | Amount of training for prediction model |
| **Amount of Untested** | Amount for test in model and random way |
| **Response & Succeed** | Response and succeed, Set state 1 |
| **Response &Not Succeed** | Response but not succeed, set state 2 |
| **Not Reply** | No reply, set state 0 |
| **Real Response Rate** | $RealResRate = \frac{HitsofReply}{AmountofUntested}$ |
| **Random Response Rate** | $RandomResRate = \frac{HitsofRomdom}{AmountofRes}$ |
| **Model Response Rate** | $ResRateByModel = \frac{HitsbyModel}{PredictResbyModel}$ |
| **Number of Prediction** | Number of users who will reply by random or model |
| **Hits** | Number of users who reply in collection of prediction |
| **Amount of Real Response** | Number of users who reply in untested collection |
| **Hit Rate** | $HitRate = \frac{Hits}{NumberofPrediction}$ |
| **Capture Rate** | $CaptureRate = \frac{Hits}{AmountofRealResponse}$ |

## *C. Experimental Results*

1) *Experiment A (Type: 1, Amount of Training:213,072)*
2) *Experiment B (Type: 3, Amount of Training:201,288)*
3) *Experiment C (Type: 5, Amount of Training:197,048)*
4) *Experiment D (Type: 8, Amount of Training:133,993)*

TABLE 5 REFERENCE OF TEST DATA AND RESPONSE RATE IN FOUR EXPERIMENTS

| **Parameter** | **A** | **B** | **C** | **D** |
|---|---|---|---|---|
| Type | 1 | 3 | 5 | 8 |
| Amount of Training | 213072 | 201288 | 197048 | 133993 |
| Amount of Untested | 2632 | 16035 | 29219 | 3489 |
| Response & Succeed | 8 | 540 | 125 | 692 |
| Response & No Succeed | 72 | 1674 | 121 | 1680 |
| No Reply | 2552 | 13821 | 28973 | 1117 |
| Real Response Rate | 3.00% | 13.80% | 0.80% | 68.00% |
| Random Response Rate | 3.20% | 14.00% | 0.90% | 67.50% |
| ResRate of Decision Tree | 5.30% | 25.20% | 3.40% | 74.50% |
| ResRate of Improved Model | 5.90% | 26.30% | 3.30% | 74.00% |

In each experiment, hit rate and capture rate can be computed via different methods by the formula in Table 4. Forecast results in four experiments are in Table 6, and the comparisons between hit rate and capture rate are shown below in Figure 3-5.

TABLE 6 FOUR EXPERIMENT FORECAST RESULTS

| | **Methods** | **A** | **B** | **C** | **D** |
|---|---|---|---|---|---|
| Hit Rate | Random | 3.20% | 14.00% | 0.90% | 67.50% |
| | Common | 5.30% | 25.20% | 3.40% | 74.50% |
| | Improved | 5.90% | 26.30% | 3.30% | 74.00% |
| Capture Rate | Random | 53.80% | 50.40% | 54.90% | 50.30% |
| | Common | 20.00% | 29.40% | 30.50% | 36.30% |
| | Improved | 38.80% | 39.50% | 56.90% | 51.50% |

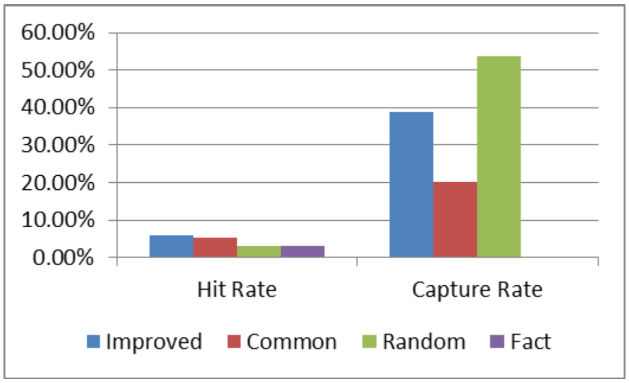


Figure 3 Prediction methods in Experiment A

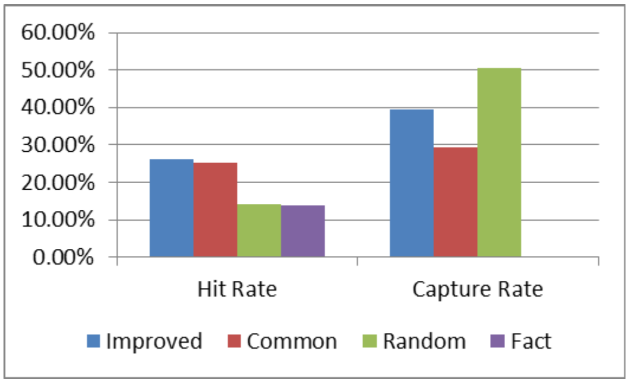


Figure 4 Prediction methods in Experiment B

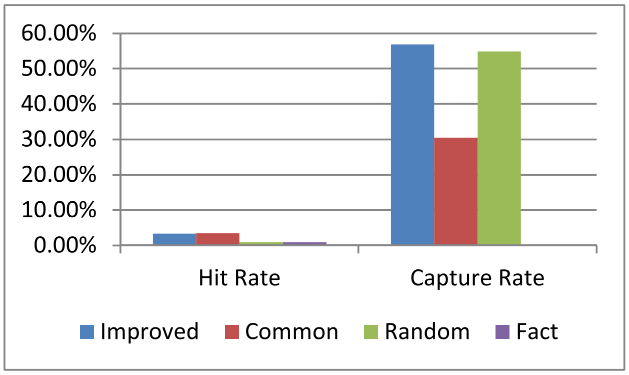


Figure 5 Prediction methods in Experiment C

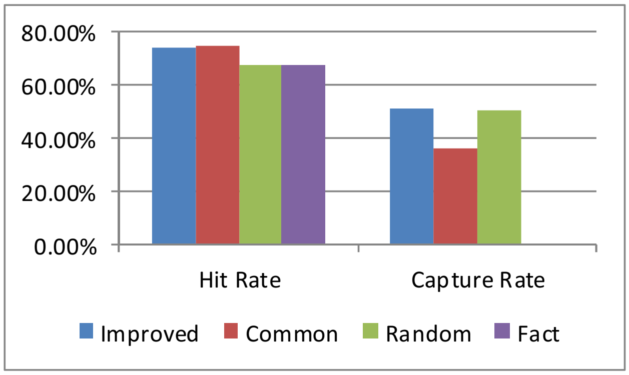


Figure 6 Prediction methods in Experiment D

### D. *Experimental Analysis*

From experimental results in Figures 3, 4, 5 and 6, we can indicate that response rate can be improved obviously by decision tree model. What is more, quantities of sending questionnaires are reduced more compared with random access which is a significant step to reduce costs on surveys. Capture rate is another important criterion to evaluate prediction model which refers to proportion in actual that response records grabbed successfully by the model or random access. From all four experiments, capture rate with decision tree is lower than that of random access. But improved model changes this situation. Besides, hits in total may reduce a little, but far less than the quantities of response reduced by prediction model. Compared with a large number of reduced sending questionnaires, a small amount of reduction hits can be acceptable while the cost reduced a lot.

## V. CONCLUSION

This paper proposes a prediction model for forecasting response rate based on ID3 algorithm in decision tree classification theory and store the generated prediction model using XML. Extensive experiments have been conducted on large amount of real online survey questionnaires. The results demonstrated the feasibility and effectiveness of our proposed prediction model with high response rate. In addition, improved decision tree model could achieve high capture rate compared with common decision tree. This leads to lower cost of online survey while using the prediction model.

## ACKNOWLEDGMENT

This work was partially supported by the National Natural Science Foundation of China (61303096) and Shanghai Natural Science Foundation (13ZR1454600). We also would like to thank all of the anonymous reviewers for their insightful suggestions and comments that will significantly improve the quality of our manuscript.

## REFERENCES

[1] Weible, Rick, and John Wallace. Cyber research: The impact of the Internet on data collection [J]. Marketing Research, 1998, 10:19-26.

[2] Fang Jiaming, Shao PJ , et al. Empirical Study of Factors Affecting the survey response rate based on network [J]. Management Review, 2006, 10:12-17.

[3] Kaplowitz M D, Lupi F, Couper M P, et al. The Effect of Invitation Design on Web Survey Response Rates [J]. Social Science Computer Review, 2012, 3:339-349.

[4] Laguilles, Jerold S., Elizabeth A. Williams, and Daniel B. Saunders.Can lottery incentives boost web survey response rates? Findings from four experiments [J]. Research in Higher Education, 2011, 52(5),:537-553.

[5] Huang Lu. Questionnaire design study based on social networks [D]. Shangdong University, 2011.

[6] Zhang Hua. Research onadvertising effectiveness based on web survey [D]. East China Normal University, 2012.

[7] Chang Yongtai. Research on web survey based on data mining [D]. Jiangsu University, 2007.

[8] Zhou M, Tao J. An Outlier Mining Algorithm Based on Attribute Entropy [J]. Procedia Environmental Sciences, 2011, 11:132–138.