# *Effective User Recommendation Model for Online Survey via BP Neural Network*

Honghao Zhu, Shaochun Wu, Guobing Zou, Naibin Luo
School of Computer Engineering and Science, Shanghai University
Shanghai 200444, China
wozhuhonghao@163.com, scwu@shu.edu.cn, guobingzou@gmail.com

***Abstract*—In order to increase the survey response rate in the online survey industry, this paper analyzes the data from a company in the perspective of information gain, and then proposes the user recommendation model which is based on improved BP neural network. Through this model users can be classified into different categories according to the user's quality. After that, by sending surveys invitation to high quality users, the response rate can be improved. Experimental results show that the proposed user recommendation model can significantly improve the survey response rate.**

***Keywords-Online survey, response rate, BP neural network, information gain***

## I. Introduction

Compared with the traditional survey industry, the online survey industry has a lot of advantages, such as wide dissemination scope of information, lower costs and quick response. It has received many attractions from the survey industry and gradually adopted by researchers. It leads to the trend of development in the future.

However, there is no doubt that online survey still has huge development space, especially in its theoretical research. At present, the research in this field is still lack of theoretical supports [1]. So aiming at theoretical research of online survey is very meaningful to promote the potential real applications. As we know, the success rate of online survey industry is about 3%. If the success rate or response rate could be improved, that would promote the economic benefits.

Based on above analysis on online survey industry and challenges to be tackled, we propose a user recommendation model for online survey industry via BP neural networks. Firstly, the information gain of all attributes is calculated. Then, those attributes which are highly related with the influence on response rates intensively are chosen. It obtains the mapping relationships between attributes and response rate by the improved BP neural network model. Finally, invitations will be sent to recommended users by the model, so that the response rate can be improved. It is less difficult to know that success rate will be improved by the advancement of response rate. To validate the effectiveness of our proposed model, experiments are conducted and the results demonstrate its feasibility.

## II. Related Work

According to the relevant literatures in recent years, the research on online survey industry can be divided into three directions [1], including the research of applications in online survey, the method of survey quality improvement, and the development of online survey system.

Nowadays, there are mainly four methods to improve the response rate. That is, attractive research topics, authoritative research institutions, material incentive mechanism, and letter remind [2]. But there are no theoretical solutions. The study of online survey is full of overall researches whereas lack of specific solutions to real survey applications [1]. In addition, there are plenty of papers containing the words, such as simple analysis, pre-test. We can conclude that currently the most domestic research deserves in-depth research on this field. So we need some intensive study in online survey.

Nowadays, there are many types of recommendation techniques, and these techniques have their own advantages or disadvantages for different application circumstances. According to the different methods, recommendation algorithms can be divided into model-based recommendation and memory-based one. Although the development of recommended techniques is thriving, it is still hard to pick an effective method and integrate it into actual data which is of great importance for survey industry.

## III. The USER recommendation model based on BP neural network

### *A. The basic idea of user recommendation model*

Based on the large number of user data and survey data, we adopt the idea of big data to improve the response rate. Approaches can be divided into two kinds. One is recommending the surveys that users might be interested in. The other is selecting high quality users, and then invitations will be sent to selected users. Considering the actual data of survey is coarser, we take the latter one.

Initially, a prediction model is trained by machine learning algorithm. This model will study the existing data and find the intrinsic characteristics of high quality users. Besides, we use this model to predict the quality of new users who has no experiences to answer the questionnaires. The structure of model is shown in Figure 1.

User can be divided into two groups: old users and new users. As to the old users, the response rate and success rate can be directly calculated by the existing data. To the new users their quality will be evaluated and predicted with the generated model via neural network. Afterwards, all high quality users are searched and added into a recommended user set. If online survey company sends invitations to these predicted high quality users, the response rate will be remarkably improved.

The user quality is evaluated by three items, which are composed by response rate, success rate, and blacklist. Taking into account correlative expert knowledge, we have definitions of these three items as table 1.

In the above, response rate and success rate are divided into three categories: Class 1, Class 2 and Class 3. Among these levels, Class2 and Class 3 are defined as high quality users. As the definition above, it is easy to know that success rate has

978-1-4673-7211-4/15 $31.00 © 2015 IEEE
DOI 10.1109/UIC-ATC-ScalCom-CBDCom-IoP.2015.326

intimate connection with response rate. There is little difference between success rate and response rate in marking user quality.

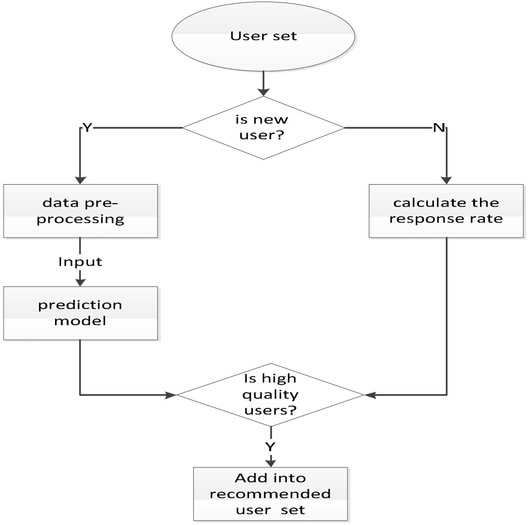

Figure 1. The overall framework of user recommendation model

TABLE 1 THE STANDARD OF USER QUALITY AND ITS DEFINITION

| Name | Definition | Class | Description |
|---|---|---|---|
| Response rate | Response times<br>/<br>Invitation times | Class 1：<br>[0-10%]<br>Class 2：<br>(10%-40%]<br>Class 3：<br>(40%-100%] | The sign of user vivid degree |
| Success rate | Success times<br>/<br>Response times | Class 1：<br>[0-30%]<br>Class 2：<br>(30%-60%]<br>Class 3：<br>(60%-100%] | means the quality of response |
| Blacklist | 1 meanings belongs to blacklist, 0 meanings not | Class 1 : 0<br>Class 2 : 1 | The user should never be invited |

The pseudocode which can recognize whether a sample belongs to high quality set or not shown as follows.

```
//A is the user that belongs to users set, S is recommend set
if(A is not in blacklist)
    then if(A.responseRate== Class 2 OR Class 3)
        then if(A.successRate==Class 2 OR Class 3)
            then add A into S
```

Because of a large amount of sample data and high dimensionality, the BP neural network is chosen as prediction model. After research of the data we find that response rate is in accordance with Pareto Principle, which means most of the surveys are replied by a small number of users. So we directly consider that user who has high response rate will respond to next survey with a great probability. And the result shows that this hypothesis is desirable and reasonable.

### *B. Data Pre-processing*

#### *1) Data Cleaning*

In online survey industry, raw data can be divided into user information, questionnaire information and behavior data. There are 32 useful attributes in raw data, such as education background, income, short message identification and so on in user information. These attributes contain some null value, noisy data and outlier. In view of the fact that sample data is surplus, noisy data and outliers are removed directly, and then assign zero for those places with null value.

#### *2) Data Generalization*

A large amount of data exists in raw data such as income, group and age, which is unfit for BP neural network. So these data have to be generalized. One part of generalization is processed as follows.

- Short message identification. This attribute only has two kinds of value “YES” and “NO”. So this paper just set “YES” as 1 and set “NO” as 0.
- Group. It is not only discrete attribute but also categorical feature. So it needs to be encoded at first, after that we transform it into numerical attributes.
- Personal income. It is a numeric kind of data that needs to be generalized to a higher level. In order to generalize it the expert knowledge is needed. And last, we generalize it as follows. (0, 1000), (1000, 4000), (4000, 10000), (10000, 20000), (20000, +∞).

#### *3) Attribute Reduction*

Feature selection is one of the crucial steps of data processing in the field of data mining and pattern discovery. With the deepening of the data mining, the research object is more and more complicated. As we know, a large number of high dimensional feature objects contain some redundant and noisy features. Therefore, when it comes to high dimensional feature objects, the algorithm of feature selection is required to find the sub-feature space [3]. Considering the high correlation between attributes and online survey, this paper applies feature reduction based on information gain to select those highly related attributes and then build the prediction model via BP neural network. Attributes with TOP-K (K=7) information gain are as follows:

TABLE 2 TOP-K INFORMATION GAIN (K=7)

| NO | Attribute | AttributeValue | Information Gain |
|---|---|---|---|
| 1 | Occupation | Manager; Staff; etc | 0.021 |
| 2 | Income | Low-income; middle-income; upper middle-income; high-income | 0.055 |
| 3 | education background | Primary school; secondaryschool; Bachelor; master; doctor | 0.093 |
| 4 | Marriage Status | Married; single; divorced | 0.022 |
| 5 | Short Message Identification | YES; NO | 0.029 |
| 6 | Email Identification | YES; NO | 0.043 |
| 7 | User Group | Online Shopping Group; Car Group; and so on | 0.081 |

*4) The selection of training set and testing set*

After data preprocessing, it remains 332,005 records. The distribution of the data is shown below.

TABLE 3 THE DISTRIBUTION OF THE DATA

| Class | Section | Data Size |
|---|---|---|
| 1 | [0-0.1] | 215,298 |
| 2 | (0.1-0.4] | 75,241 |
| 3 | (0.4-1.0] | 41,298 |

In order to avoid data offset, the training set is constructed by the computation of the following criterion.

$$S_{\text{Training Set}} = \left(\frac{n_1}{N} + \frac{n_2}{N} + \frac{n_3}{N}\right) * P \qquad (1)$$

Among them, $n_i$ is the number of class i, N stands for the sum of $n_i$, and P is the proportion of training set. The size of training set will affect the final result directly. Along with extensive experiments, 57 thousand training set and 50 thousand testing set are finally selected for model learning and performance checking.

## *C. The Building of Prediction Model*

Determining the structure of BP neural network model is crucially important to prediction model. Related research shows that the more number of hidden layer nodes of BP neural network has, the better approximation ability of model is. However, concerning the efficiency of neural network learning, the large number of hidden layer nodes is unfavorable. So how to determine the structure of BP neural network is difficult and important to decide in a real application [4]. In this paper, there are seven input nodes and three output nodes. And then the following equation is taken to calculate the number of hidden layer nodes.

$$n_{\text{hidden layer nodes}} = \sqrt{n+m} + a \qquad (2)$$

Among them, n represents the number of input nodes, and m is the number of output nodes. Here we set a=8. So the number of hidden layer nodes is 11, the structure of BP neural network is illustrated as follows.

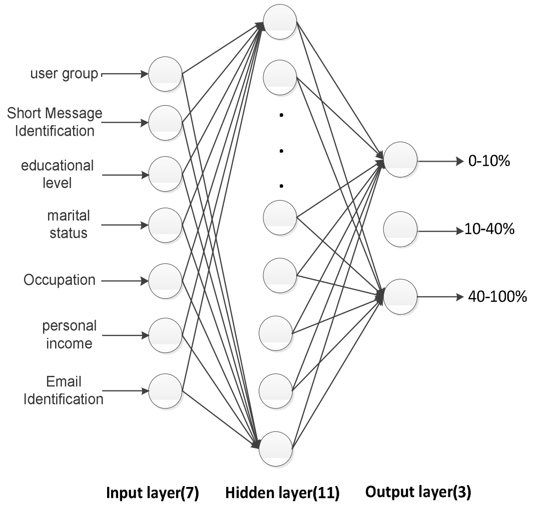

Figure 2. The structure of built BP neural network

## *D. The Improvement of Model*

Although the classic BP neural network model is adopted, the result is unsatisfactory in this data set. There is about 50% accuracy in predicting response rate. That is why we analyze the reason and further polish the built prediction model.

The classic BP neural network model is based on gradient descent algorithm. As we know, error function is a complex function with high-dimension, especially when have to face large-scale problems or network structure cannot be fully expressed. In these scenarios, BP neural network has some defects that cannot be avoidable. For example, trap into local minimum, slow in convergence and oscillation phenomenon [5]. Error function is the basis of model learning. If the error function can be improved according to the actual data, the learning effect might also be increased. In this paper, based on the realistic situation of attribute association, the error function in the perspective of information gain is introduced. New error function is more suitable for the data. The error function is formalized as follows.

$$E_A = \frac{1}{2} * \sum_{p=1}^{p} \sum_{l=0}^{m-1} [(d_l^{(p)} - y_l^{(P)})^2 * (\frac{p_l}{p_1+p_2+p_3+\dots+p_m} + 1)(3)$$

Among them, P represents sum of user, m means the number of output nodes, $d_l^{(p)}$ means the desired model output, $y_l^{(P)}$ stands for the model output. $p_l$ stands for the information gain of attribute. The intention of new error function is stressing the importance of attributes on the basis of the original function.

## *E. The Storage of Model*

Database is used to store the weights and threshold of each node of BP neural network, but the structure of BP neural network is complex relatively, which leads to be complicated in table design and slow in preservation model. In this paper these problems can be avoided by adopting the serialization and deserialization methods in the Java language.

First of all, the trained model is serialized into the document. After that, this model can be desterilized from the document whenever it is needed. So we can get the model easily. The model storage and application process is shown in the figure as below.

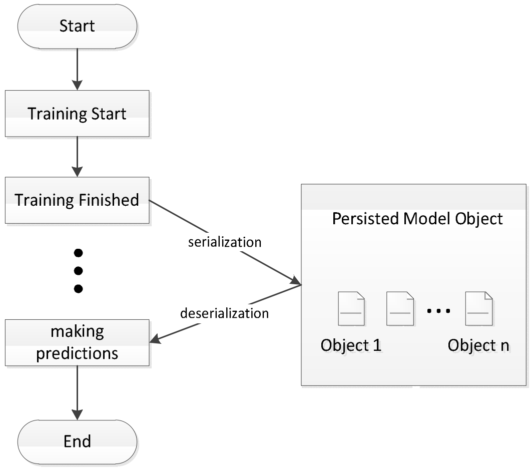

Figure 3. The storage of model by deserialization strategy

## IV. Experimental Evaluation

The most common strategy in online survey invitation is sending surveys randomly. This strategy is simple and feasible. But it is blind obviously, which leads to the rising cost of the online survey. The fundamental reason of this problem lies in the fact that online survey industry is blind to select users without any consideration on user quality filtering. For this reason, this paper proposes a new strategy that only high quality users should be invited. High quality users are predicted by BP neural network model. Experimental results show that new strategy is better than the existing commonly used one.

There is a hypothesis that all users are new registration in the following experiments. As a result, only the basic attributes of users are taken into consideration, for example, personal income and marital status. At the same time, the attributes that are associated with the survey are abandoned, for example, response time and success time. So we have no ideal in user classification before using classification. The goal of this model is to find the relation between user's attribution and user's loyalty.

### A. The Design of the Experiment

The questionnaire can be roughly divided into 10 types in survey industry, for example, Product Testing, TEATM, U&A etc. There are many experiments to each kind of surveys. Each experiment is divided into two groups, A and B. Group A is old strategy which invite randomly. Group B is new strategy which invite by recommendation model. The user prediction model is used for group B. Afterward the results of two groups are compared to validate the effectiveness of our proposed recommendation model for predicting high quality users.

### B. Experimental Results

The result of a questionnaire that belongs to type 1 is shown in the following table.

TABLE 4 THE EXPERIMENTAL RESULTS OF TYPE 1 ONLINE SURVEY

| Class | invitation number | Success number | Success rate |
|---|---|---|---|
| old strategy | 3,458 | 596 | 17.2% |
| new strategy | 2,746 | 591 | 21.5% |

The result of a questionnaire that belongs to type 2 is shown in the following table.

TABLE 5 THE EXPERIMENTAL RESULTS OF TYPE 2 ONLINE SURVEY

| Class | invitation number | Success number | Success rate |
|---|---|---|---|
| old strategy | 26,756 | 100 | 0.37% |
| new strategy | 6,589 | 92 | 1.39% |

The result of a questionnaire that belongs to type 3 is shown in the following table.

TABLE 6 THE EXPERIMENTAL RESULTS OF TYPE 3 ONLINE SURVEY

| Class | invitation number | Success number | Success rate |
|---|---|---|---|
| old strategy | 32,679 | 10,324 | 31.5% |
| new strategy | 15,781 | 9,985 | 63.2% |

The results of 574 surveys that can be found in existing data are shown in the following table.

TABLE 7 THE EXPERIMENTAL RESULTS OF EXPERIMENT FOUR

| Class | invitation number | Success number | Success rate |
|---|---|---|---|
| old strategy | 4,096,894 | 603,807 | 14.7% |
| new strategy | 2,287,442 | 560,733 | 24.5% |
| changes in total | 1,809,452 | 43,074 | |

Comparison results between two user recommendation models for online survey are shown in figures 4 and 5.

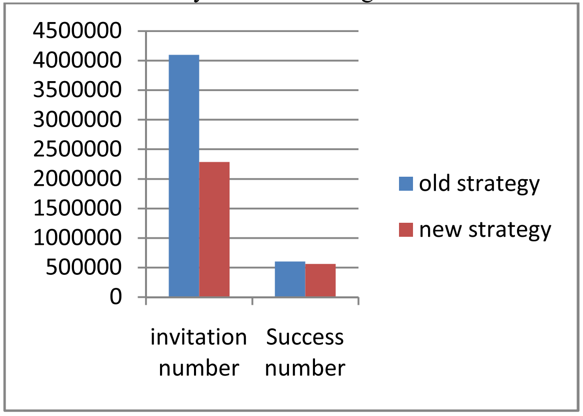

Figure 4. Comparison results on success number between to user recommendation strategies for online survey

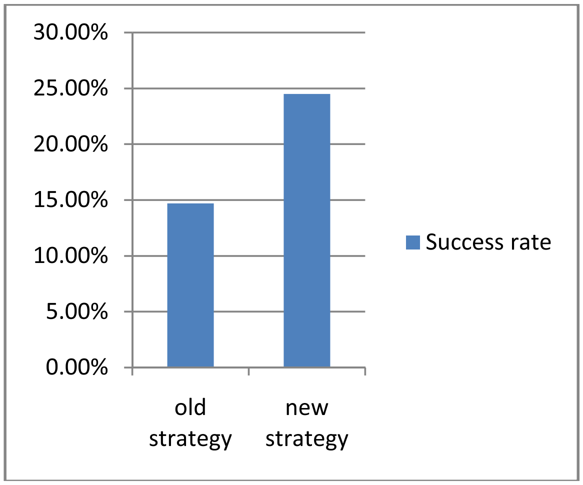

Figure 5. Comparison results on success rate between two user recommendation models for online survey

### C. Analysis of Experimental Results

These experimental results show that the success rate of online survey can be improved by high quality user recommendation model which is based on BP neural network. The result of experiment three is the most obvious among them. Success rate can be improved from 31.5% to 63.2%. At the same time, experiment one also obtains a good result. On the other hand, experiment two shows that the success rate is improved a little, but the rate is also improved twice as large as before. At last, this paper carries out experiment on 574 questionnaires which can be found in the raw data. It turns out that the invitation number is reduced to half the number of before while the success number is nearly the same. So the success rate is improved from 14.7% to 24.5%. The great potential economic benefits would be generated while this user recommendation model is used for online survey industry.

## V. Conclusion

To solve the research challenge of low success rate in current online survey industry, we propose a user recommendation model via BP neural network to improve the success rate.

At first, the data is preprocessed. Then, the information gain which is related to target attribute is calculated. Then this paper chooses seven attributes which have heavy information gain as the input of neural network model. At the same time, based on the features of data, we improve the error function of BP neural network, which raises the success rate. Finally, there are some experiments on recommendation model. The experimental results validate its feasibility. Our proposed approach to online survey can significantly improve the success rate.

## ACKNOWLEDGEMENT

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

## References

[1] Wan Chong. Network Survey System Analysis and Design[D]. Beijing Jiaotong University, 2014.

[2] Fang Jiaming, Shao Peiji, Su jie, etc. An Empirical Study on the Incentive Influencing the Response Rate of the Web-Based Survey[J]. Management Review, 2006, 18(10):12-17.

[3] Ren Jiang Tao, Sun Jing Hao, Huang Huan Yu, etc. Feature Selection Based on Information Gain and GA [J]. Computer Science, 2006, 33(10):193-195.

[4] Zhang Xiaoming, Wang Fang, Jin Yuxue, etc. Hidden Layer Structure Optimization of BP Network Based on Grey Incidence Degree and Sensitivity Degree[J]. Computer Measurement & Control, 2014, 22(9).

[5] Liang Jiu zhen, He Xin gui, Huang De shuang. Simple Conjugation-Gradient BP Algorithm for Feedforward Neural Networks[J]. Journal of Beijing University of Aeronautics and Astronautics, 2000, 26(5):596-599.