A Neoteric Data Preprocessing Technique for Online Surveys

Akshay R.
Department of MCA
PESIT – Bangalore South Campus
Bangalore, India

Arti Arya
Department of MCA
PESIT – Bangalore South Campus
Bangalore, India

ABSTRACT
Online surveys is an essential research tool that are being applied in variety of research fields, including marketing, social and official statistics research and hence are one of the most popular data collection technique. Some people fill it genuinely and some do it randomly. Data collected through samples that are not filled genuinely may affect the analysis of the collected samples considerably. This paper proposes a preprocessing technique to select the samples that have genuine responses in order to make sure the final data collected from the survey is more precise and accurate. For this purpose the time duration an individual takes to provide his/her opinion to each question in questionnaire is captured. This captured time is used to check the percentage of questions that fall between the time ranges computed for each question using the proposed algorithm to indicate if the sample was filled genuinely. In doing so the samples that are found to be genuinely responded to, can be given more weight-age while analyzing the survey or randomly filled samples can be eliminated.

Keywords
Surveys, Questionnaire

1. INTRODUCTION
Surveys are a set of questionnaires distributed to be filled by a target set of people in-order to get feedback and opinions. These information can be analyzed to uncover the answers, make important decisions, understand the expectations of the clients etc.

A questionnaire [2] is a set of questions along with other prompts for the purpose of gathering information from respondents

An online survey [3] is a questionnaire that the target audience can complete over the Internet. Online surveys are usually created as Web forms with a database to store the answers and statistical software to provide analytics. Some of the examples for online survey are product review survey, customer satisfaction survey etc. [3]

Genuinely filled samples are those that are filled by providing instinctive and honest opinions to the questionnaire by reading and understanding the question correctly.

In survey terminologies the word population refers to the entire set of people on whom the survey is intended towards but at times it is impractical to survey the entire population. Therefore a set of samples from this population is selected and survey is conducted only on these samples and infer the information about the population [4]. Questionnaire data has been preprocessed and analyzed using fuzzy association mining concepts in [9].

1.1. Types of sampling techniques [4]

1.1.1 Probability-based sampling
Is one which makes use of probability theory to select the respondents [4,5]. In these cases the probability of each possible respondent in the population being selected to fill the survey is know in advance.

1.1.2 Non-probability based sampling
occur when either the probability of the respondent included in the sample cannot be determined, or it is left up to each individual respondent to choose to participate in the survey [6, 7, 8].

1.2.Sources of error in surveys[4]

1.2.1 Non-response error
Occurs when the respondent either doesn't give his response to the entire survey or to a part of the survey.

1.2.2 Measurement error
Arises when there is a difference in the survey response and the 'true' response. It happens when the respondents may not have answered questions honestly, or respondents may have misinterpreted the questions or may have made errors in answering the questions for a variety of reasons.

1.2.3. Error of coverage
Occurs when the entire part of the population is not included in the sample.

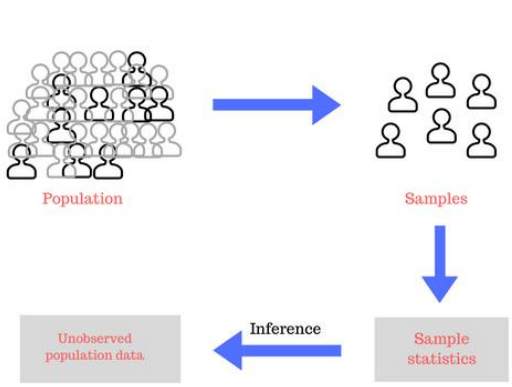

Fig 1. Description of the sampling procedure [4]

Either of the above mentioned sampling techniques can be used to select samples from a population. It is important to make sure that the sampling technique that is being implemented for a given survey helps in finding out those samples which approximately represent the data of the entire population to which the samples belong to. [4]

Since every sample represents a population it is picked from, it is necessary to ensure each of these samples are completely and genuinely filled. In case if they fail to do so it results in fetching incorrect data and hence fail to represent the population they belong to, which may impact the final analysis of the survey.

The present online survey systems have the ability to avoid non-response errors by ensuring that a respondent can submit the response only after he has answered all the questions in the survey before, but presently there exists no mechanism to check if the person filling the survey is filling it genuinely. Here in this paper an algorithm is proposed to check if a sample is genuine, keeping time as a parameter to overcome the measurement error that occurs in survey.

2. MOTIVATION

The online survey tool is being used by a wide range of people from different domains to get information ranging from simple feedback data to critical data that might be used for research purposes. For example online surveys are being used by a large number of startups currently to understand the market for their product. The data being fetched can make or break a product depending on how genuinely the surveys are being filled.

With technological evolution there are plenty of tools available to facilitate the surveyor to create surveys and publish them but unfortunately there is very less work happening currently on ensuring the data fetched from surveys are genuine in-spite of acknowledging the importance of data being fetched is very high. So in order to filter out such samples from the population that are not filled genuinely, an algorithm is proposed in this paper.

3. OVERVIEW OF SURVEY SYSTEM

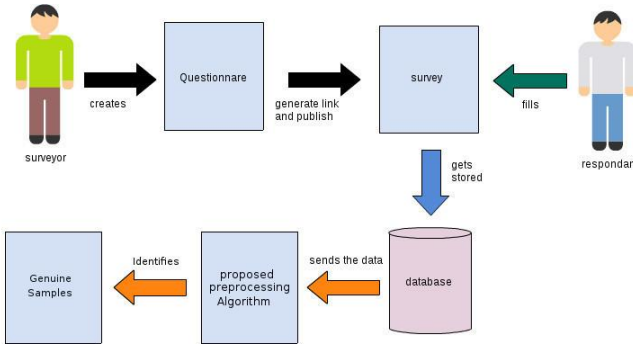

Fig 2. Flow diagram of collecting all the samples and selecting the genuine samples

Surveyor is either an individual or a group of individuals who roll out the survey containing a set of questions. In online surveys the questions are uploaded to any of the online survey creator platform upon which a unique link is provided for each survey in order to enable the surveyor to publish the survey to the respondent and collect the responses.

Once the samples are collected they are fed to the algorithm proposed in this paper to select the genuine samples among all the samples.

4. PROPOSED ALGORITHM

Fig 3.1 to fig 3.6 explains the complete idea behind the proposed system. A question from questionnaire is considered and the proposed algorithm is applied on it.

If the minimum time taken to answer the question is 10 seconds (lower limit) and the maximum time is 110 seconds (upper limit), there exists a time range of 100 seconds for that question, which is divided into 10 equal buckets with each buckets having 10 seconds range. After breaking down the time range into buckets the below chart is obtained (Fig. 3.1) where the percentage value inside each bucket represents the percentage of questions in each bucket.

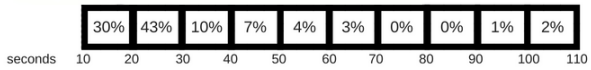

Fig 3.1 Percentage of samples in each bucket

Next, from the left and right extreme sides the buckets which has a percentage value less than a threshold value (4%) which is determined heuristically are eliminated. By doing so for the chart shown in fig 3.1 the five buckets from the left side is eliminated as they have a percentage value less than threshold value. Similarly the right most bucket already has a percentage value greater than threshold no buckets from the right side are eliminated. After this step a new lower limit (10 seconds) and upper limit (60 seconds) is obtained as shown in Fig 3.2.

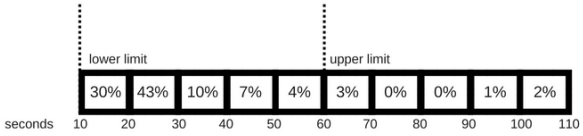

Fig 3. 2 After eliminating buckets with less than the threshold value

The same procedure is repeated with a the newly computed bucket size (5 seconds and 2.5 seconds each) such that the percentage at each extreme ends are greater than the threshold value (4%) as shown Fig 3.3 and Fig 3.4

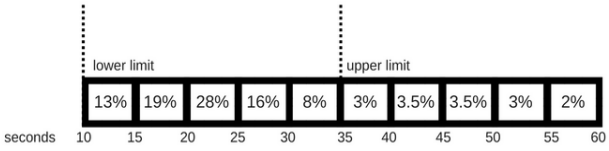

Fig 3. 3

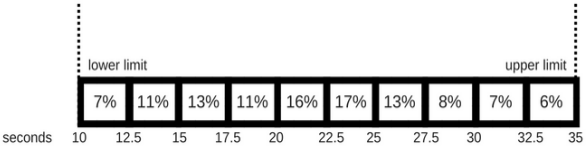

Fig 3. 4

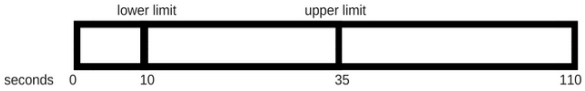

Fig 3. 5

After performing the above steps, a conclusion can be that the time taken to answer this question genuinely is between 10 seconds to 35 seconds.

Next, further divide the time between the lower_limit and upper_limit into three halves and assign them a weight based on the maximum percentage value in each bucket as show below.

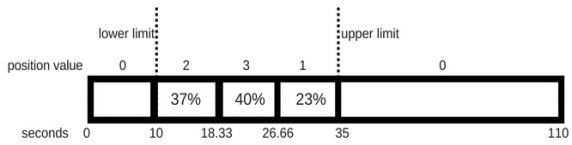

Fig 3. 6

Once the ranking is done as shown in Fig 3.6, each question from a sample is given a rank value depending upon which bucket, the time taken to answer the question falls. For example, if the ranking in Fig 3.6 is considered then the question answered in 12.34 seconds is assigned a weight value 2.

Similarly, the weight value for all the question of each sample in the survey is computed and sum it up, which gives us the cumulative weight value of each sample.

If this cumulative weight value is greater than the average cumulative value of all the samples then the sample is considered to be genuine. For example, consider the Fig 4 where each bar represents one sample and the number on the bar indicates the cumulative weight value of that sample. The first bar indicates the maximum cumulative value a sample can hold, that is 21 for this questionnaire (7 questions * 3). The samples with a cumulative weight greater than the average cumulative weight value of all the samples (9 for this survey) is considered as genuine.

5. PSEUDO CODE

```
1. Creating the buckets initially
Range ← upper_limit – lower_limit
bucket_size ← range / 100;
temp← lower_limit
For i← 1 to 10
        bucket[i][low] ← temp
        bucket[i][high] ← temp + bucket_size
        temp← temp + bucket_size
end for

2. Eliminating the extreme buckets with a percentage less than the threshold

for i← 1 to 10
        if ( bucket[i][percentage] < threshold)
                lower_limit ← bucket[i][high]
        else
                break
        end if
end for

for i← 10 down to 1
        if ( bucket[i][percentage] < threshold )
                upper_limit ← bucket[i][low]
        else
                break
        end if
end for
```

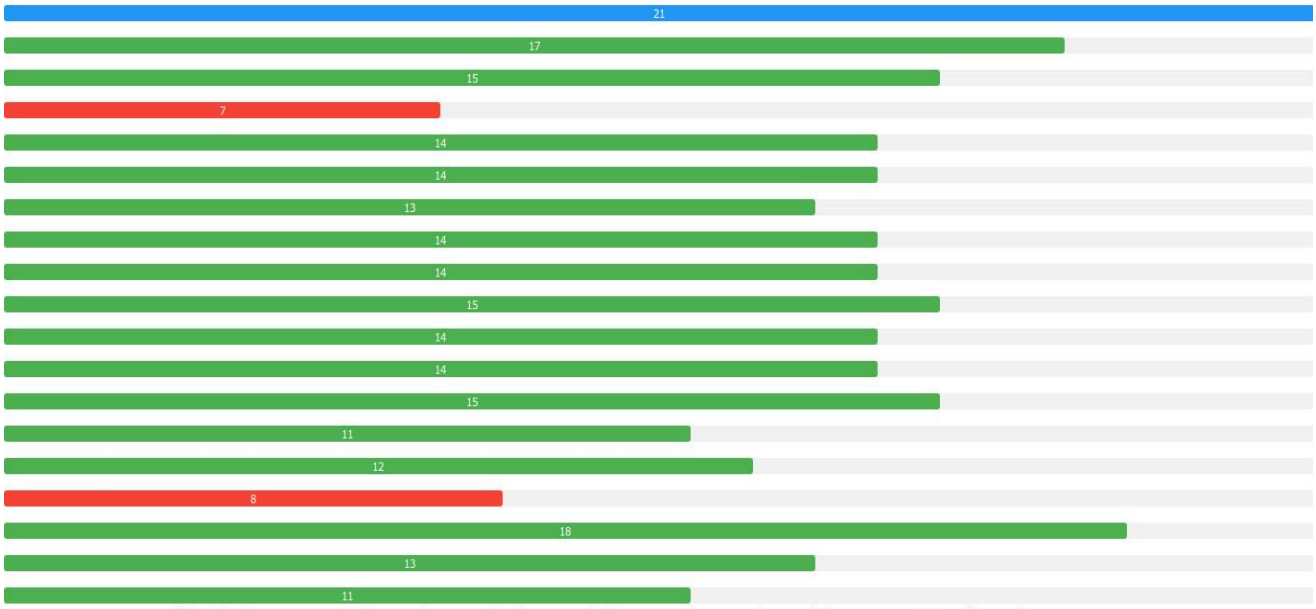

Fig 4 Representation of cumulative weightage of samples with respect to closed survey

6. RESULTS AND DISCUSSION

An online survey with seven questions was created and published to collect samples. The above Fig 4 shows the results obtained after applying the algorithm proposed in this paper.

The survey was conducted in two phases

6.1. Open survey

Here a link to the survey was created and posted on various online platforms to collect random samples.

6.2. Closed survey

Here a selected set of participants were asked to fill the survey in presence of the surveyor in order to validate if the survey was being filled genuinely.

From the above Fig 4 representing the samples filled in closed survey, it can be observed that out of the 20 samples filled in

closed survey 18 (bars colored in green) are found to be genuine. Therefore the proposed algorithm is able to select the genuine samples with an accuracy of 90%.

7. CONCLUSION

In this paper, a novel algorithm is proposed for preprocessing the data collected from online surveys which is capable of identifying genuine and not so genuine responses from the respondents. This algorithm has given an approximate results of up to 90 %. To best of our knowledge, no such algorithm is proposed so far to identify genuine responses from an online survey. Hence this algorithm can be used to select the genuine responses and perform analysis on them to infer the information about the population. This algorithm can be a very

Useful tool when there is a need to analyze critical information from survey data.

8. REFERENCES

[1] https://en.wikipedia.org/wiki/Survey_data_collection

[2] Questionnaire: https://en.wikipedia.org/wiki/Questionnaire

[3]Onlinesurvey:https://www.techopedia.com/definition/27866 /online-survey

[4] Fricker, Ronald D. "Sampling methods for web and e-mail surveys." N. Fielding (2008): 195-216.

[5] Erik Volz, Douglas D. Heckathon, Probability based estimation theory for respondent driven sampling", In J. of Official Statistics, Vol. 24, no.1, 2008, pp.79-97.

[6] Tansey, Oisín. "Process Tracing and Elite Interviewing: A Case for Non-Probability Sampling." PS: Political Science &Amp; Politics, vol. 40, no. 4, 2007, pp. 765–772.

[7] Schillewaert, Niels, Fred Langerak, and Tim Duhamel. "Non-probability sampling for WWW surveys: a comparison of methods." International Journal of Market Research 40.4 (1998): 307.

[8] Feild, Lucy, et al. "Using probability vs. nonprobability sampling to identify hard-to-access participants for health-related research: costs and contrasts." Journal of Aging and Health 18.4 (2006): 565-583.

[9] Chen, Yen-Liang, and Cheng-Hsiung Weng. "Mining fuzzy association rules from questionnaire data." Knowledge-Based Systems 22.1 (2009): 46-56.

9. APPENDIX

Sample Questionnaire:

1. Your gender?
   - Male
   - Female

2. Select your age group?
   - below 16 years
   - 17 years - 25 years
   - 26 years - 35 years
   - 36 years and above

3. Select your profession?
   - Student
   - Home maker
   - Salaried Employee
   - Self Employed
   - others

4. Select your location?
   - Bangalore
   - Mumbai
   - Chennai
   - Delhi
   - Other

5. How often do you play sports?
   - everyday
   - never
   - weekly once
   - weekly twice
   - monthly once
   - monthly twice
   - rarely

6. Are you happy with the amount of time you spend playing? If no, how often do you wish to play?
   - yes, I am happy with the time I spend on sports
   - no, I would like to play once in a week
   - no, I would like to play once in a month

7. What is your motivation to play sports?
   - time pass
   - passion
   - fitness
   - socialize