On Design of Problem Token Questions in Quality of Experience Surveys

Jayant Gupchup*, Ebrahim Beyrami*, Martin Ellis*, Yasaman Hosseinkashi*, Sam Johnson*†, Ross Cutler*
*Microsoft Corporation, {jayagup, ebbeyram, maellis, yahossei, sajohnso, rcutler} @microsoft.com
†currently affiliated with Facebook, sam.johnson@fb.com

***Abstract*—User surveys for Quality of Experience (QoE) are a critical source of information for application developers. In addition to the common "star rating" used to estimate Mean Opinion Score (MOS), more detailed survey questions (problem tokens) about specific areas provide valuable insight into the factors impacting QoE. This paper explores two aspects of problem token questionnaire design. First, we study the bias introduced by fixed question order, and second, we provide a methodology to manage the size of the survey while keeping it informative. Based on 900,000 calls gathered using a randomized controlled experiment from Skype, we find that token selections can be strongly biased due to token positions and display design. This selection bias can be significantly reduced by randomizing the display order of tokens. It is worth noting that users respond to the randomized-order variant at levels that are comparable to the fixed-order variant. The effective selection of a subset of tokens is achieved by extracting tokens that provide the highest information gain over user ratings. This selection is known to be in the class of NP-hard problems. We apply a well-known greedy submodular maximization method on our dataset to capture 94% of the information using just 30% of the questions.**

***Index Terms*—QoE; Survey design; VoIP; data analysis**

I. Introduction

Several Internet telephony applications employ end-of-call user surveys to gather data on in-call QoE [1]. In addition to the five star rating (MOS [2]), the percentage of calls rated 1 or 2 (poor call rate, or PCR) is often tracked as a measure of media quality. Previous studies have shown the value of combining PCR with an additional problem token questionnaire (PTQ) to gather detailed insights [3]. The UI design of the PTQ used in this study is provided in [3].

The range of questions (tokens) used to capture these detailed problem areas has been studied in depth [4], [5]. However, to the best of our knowledge, the impact of presentation order on token selections has not been studied in a live, deployed system. Our work is motivated by the practical challenges faced in analyzing questionnaire data. For example, one analysis showed that the contribution of one-way audio (one side can hear, but the other side cannot) to PCR dwarfed the other areas by a factor of two on mobile platforms. However, further investigation showed that the most important factor for this gap was the display order of questions; users were 40% more likely to select the 'no sound' token purely due to its position at the top of the survey. While one-way audio remains an important problem area, we found that after randomization of the order, other impediments such as audio distortion and poor image quality occur at comparable levels.

In mobile environments, the screen size is limited, so the number of questions needs to be kept small to avoid the need for a scrollbar. The key question is - how to minimize the number of questions while maximizing their power in explaining PCR. Moreover, studies have shown the benefit of shortening surveys (without losing information) for improving response rate and data quality [6]. Identifying this subset of questions belongs to a class of NP-hard problems [7]. In order to solve this, we follow the lead of Krause *et al.*, leveraging the fact that information gain is a submodular function, and can be optimized using provable greedy approaches [8], [9]. As shown in Section IV, this approach maps well to our problem. The main contributions of this paper are: 1) Results of a large scale randomized, controlled experiment in a live VoIP system to measure the bias introduced by fixed order questions; and 2) An efficient solution to select a subset of tokens that maximizes information and minimizes correlation.

II. Related Work

There is a rich area of research and practice in general survey design, validation and question order [10], [11]. Factor analysis is commonly employed to analyze surveys with the number of factors being smaller than the number of questions [12]. There are many standards for subjective audio and video quality surveys, such as the ITU standards [13], [4], [5]. In this paper, our goal is not to replace these surveys, but instead to improve their utility by providing recommendations for presentation order, and a methodology to select a subset of informative questions. The selection of a subset of correlated random variables for maximizing the information gain has been studied in detail [8], [9]; this paper focuses on the application of these methods for QoE problem area surveys.

III. Impact of Question Display Order

We study the impact of question order on our PTQ using a randomized controlled experiment: 1) To learn whether randomization induces a change in the percentage of responses; and 2) To measure the change in selection rate of the individual audio and video tokens. The control population was shown the original questionnaire with fixed token order; for video calls, the audio tokens were always shown on the left while the video tokens were always shown on the right. The treatment population was shown the questions in randomized order; for video calls in the treatment population, the position of the audio and video panels (left/right) were selected at random. The details of the experimental design are below:

TABLE I: The difference in overall response rate in tokens between control group and treatment group

Population	Relative Delta	p-value
Audio-only calls	-1.38%	0.072
Video calls	-1.62%	0.001

- Responses were collected from one-to-one Skype calls on desktop platforms, including audio-only and video calls.
- The PTQ for audio calls contained only audio tokens, whereas for video calls it included audio and video tokens.
- The control and treatment groups each contained 450,000 calls, spanning over 100,000 unique users.

Note that we present all results using relative measures to preserve commercial confidentiality.

A. Overall Questionnaire Response Rate

First, we wanted to understand if randomizing the question order impacts the overall response rate of the PTQ. A user is said to respond to a PTQ if any token selection is made. The differences in overall token response rate between the control and treatment groups for audio-only and video segments is shown in Table I. There is no statistically significant difference in the audio population, but there is a change in the video population at the 99% significance level (i.e., $p < 0.01$). However, we find that the relative difference of 1.6% for video surveys is small enough that the benefits (Section III-B) of the randomized questionnaire outweigh the minor reduction in response rate.

B. Selection Rates of Individual Questions

The difference in selection rates of individual tokens between fixed and randomized ordering for desktop video calls is shown in Table II. A negative sign (red) indicates that selection rate decreased, while a positive sign (green) indicates selection rate increased. Although a change in the selection rate was expected, there are some significant insights from these results:

1) The selection rate of the top token is dramatically impacted for audio and video tokens. The decrease in selection rate between the two variants is greater than 20%. This shows the propensity for selecting the top token.
2) For audio calls, selection rates of the top four tokens decreased dramatically. However, for video, selection rates of the bottom four tokens increased. This is likely due to the bias of reading from top to bottom in most languages.

The impact of panel position is even more pronounced in mobile environments. We found the average selection rate for tokens requiring users to scroll was 49% lower. These results motivate the need for showing a small set of informative tokens to ensure that we do not lose the user's attention while responding to the questionnaire.

IV. Token Subset Selection

The question we are trying to address is the following: Given a limited budget of questions, k, is there a systematic process of selecting the questions to maximize information? In this paper, we propose the selection of tokens by applying the algorithm described by Nushi *et al.* [9].

TABLE II: The difference in selection rate of individual tokens for fixed vs. randomized display order in desktop video calls

Audio problem Token	**Relative delta**	**p-value**
I could not hear any sound	-26.7%	$\leq 2e^{-29}$
The other side could not hear any sound	-12.5%	$2e^{-23}$
I heard echo in the call	-12.7%	$6e^{-24}$
I heard noise in the call	-9.5%	$3e^{-18}$
Volume was low	-3.8%	0.01
The call ended unexpectedly	-2.4%	0.10
Speech was not natural or sounded distorted	$+3.6\%$	0.00
We kept interrupting each other	-1.7%	0.22
Video problem Token	**Relative delta**	**p-value**
I could not see any video	-20.4%	$\leq 2e^{-29}$
The other side could not see my video	-1.9%	0.39
Image quality was poor	-2.2%	0.06
Video kept freezing	$+10.1\%$	$3e^{-16}$
Video stopped unexpectedly	$+28.0\%$	$4e^{-45}$
The other side was too dark	$+25.2\%$	$8e^{-21}$
Video was ahead or behind audio	$+25.2\%$	$5e^{-39}$

A. Information Gain and Submodular Function Optimization

Information gain (IG) captures the amount of information "shared" between two random variables [14]. Mathematically, IG between variables, Y and X, is defined as $IG[Y;X] = H[Y] - H[Y \mid X]$; where H represents the entropy (uncertainty) [14] of a random variable. In our setting, Y denotes a poor call label and X denote the multivariate random variable covering all tokens. It should be clear that IG has the property of monotonicity [15]. We can easily see that for any two sets of random variables, $X_1, X_2 \colon X_1 \subset X_2 \subset X, IG[Y;X_2] \geq IG[Y;X_1]$. Building on the monotonicity property and borrowing notation from [8], IG also has a "diminishing returns" property. The incremental information gain obtained by adding a new element, $t : t \notin X_1$ and $t \notin X_2$, to a subset is higher than the incremental information gain obtained by adding the same new element to its superset. Mathematically, the principle of diminishing returns property is shown in the equation below:

$$\begin{aligned} IG[Y;\{X_1 \cup \{t\}\}] - IG[Y;X_1] \geq \\ IG[Y;\{X_2 \cup \{t\}\}] - IG[Y;X_2] \end{aligned} \tag{1}$$

In other words, the marginal benefit of reducing the uncertainty in Y by adding a new token to a smaller set is higher than any superset; this property is known as *submodularity* [8]. Krause *et al.* [8] show that a certain class of interactions (e.g., mutual exclusion) between variables can result in IG to not be strictly submodular. However, we do not see such interaction effects for our token dataset, and therefore no violation of submodularity of information gain.

The optimization of submodular functions is a known NP-hard problem; however, Nemhauser *et al.* [16] provide a greedy algorithm to solve this problem that is 63% of the computationally expensive and exhaustive solution. The algorithm to construct the minimal token set is iterative and fairly straightforward. In iteration 0, we start with an empty set, X_0. At every iteration i, we add the token, t, that maximizes the discrete derivative of information gain. Set X_i is constructed using the equation:

$$X_i = X_{i-1} \cup \arg\max_t IG[Y;(X_{i-1} \cup t)] \tag{2}$$

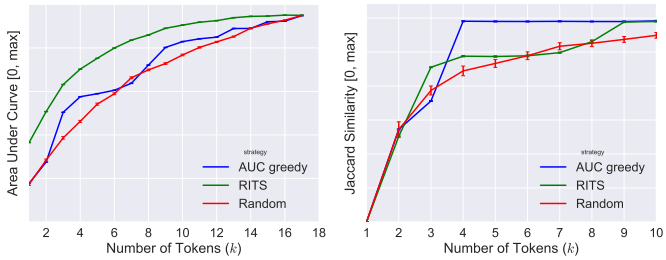

Fig. 1: Relative AUC performance of different strategies in selecting tokens is shown on left while Jaccard similarity scores are shown on right. Note: Scales removed for confidentiality.

where $t \in X \setminus X_{i-1}$ and $IG[Y; (X_{i-1} \cup t)]$ represents the information gain of Y by jointly considering the tokens $(X_{i-1} \cup t)$ for all candidate tokens t. Note that iteration 1 picks the token that provides the highest univariate information gain. In subsequent iterations, this method selects tokens that provide the information not already captured by the existing token set. By design, this results in selection of the least correlated tokens at every iteration. We view this method of selecting tokens as maximizing the return of information for tokens shown (hereafter referred to as *RITS*).

B. Evaluation of RITS

Data gathered from the treatment (randomized PTQ) population of the experiment was used for evaluation. The RITS method was evaluated using two quantitative metrics [14]: area under an ROC curve (AUC), and Jaccard Similarity (JS). While AUC measures the ability of the token set to discriminate between a good call and a poor call, Jaccard Similarity measures the pair-wise degree of overlap between the tokens. The ideal token set has an AUC close to 1 and a JS score of 0. Uncorrelatedness (i.e., a low JS score) is important when breaking down an overall quality metric into distinct factors. For a given token count, k, we studied and compared the performance of RITS with the following approaches:

1) Random: selecting a random subset of tokens.
2) AUC-Greedy: selecting tokens sorted in descending order of their univariate AUC.

In our evaluation, we used the random forest implementation from the Python scikit-learn library [17] to obtain the classification boundary with default settings. The error bars were obtained using 100 independent runs of train/test splits.

C. Results using RITS

The AUC and JS scores of the different token subset selection strategies are shown in Figure 1. Note that we represent the scale in terms of the maximum values obtained for our dataset, and hide the labels for commerical confidentiality. This allows for relative comparison between the selection strategies. The RITS method significantly outperforms AUC-greedy and random method for all k in terms of the AUC criterion. The shape of the RITS curve highlights the diminishing returns property. RITS also has a significantly lower JS score compared to the AUC greedy method for $k > 4$; this is because it is designed to find the tokens that provide information not already covered by the existing set of tokens. Since the AUC-greedy method does not consider correlation, it performs poorly on the Jaccard similarity measure. Using RITS, the first five tokens capture 94% of the total information content in our dataset.

V. Summary

In this paper, we studied two aspects of the problem token questionnaire design for VoIP applications: display order and token subset selection. Based on over 900,000 calls gathered from a randomized controlled experiment in Skype, we showed that there is strong bias in selection rates due to the presentation order of questions. The most dramatic impact is experienced by the top-most token. In mobile environments, scrolling can lead to a reduction in selection rate by as much as 49%. Motivated by these observations, we studied the problem of selecting a subset of tokens that maximize information while minimizing correlation. We achieved this by mapping it to the problem of submodularity maximization. By doing so, we were able to retain 94% of the information using just 30% of the questions. Finally, we would like to emphasize that these methods and results can vastly benefit the community as they significantly improve the quality of data gathered from any QoE survey.

References

[1] J. Jiang *et al.*, "Via: Improving internet telephony call quality using predictive relay selection," in *Proc. ACM SIGCOMM*, 2016.
[2] ITU-T, "Mean Opinion Score (MOS) terminology," 1996, Rec. ITU-T P.800.1.
[3] J. Gupchup *et al.*, "Analysis of problem tokens to rank factors impacting quality in voip applications," in *Proc. QoMEX*, 2017.
[4] ITU-T, "Methods for subjective determination of transmission quality," 1996, Rec. ITU-T P.800.
[5] ——, "Subjective video quality assessment methods for multimedia applications," 2008, Rec. ITU-T P.910.
[6] D. S. Allen, "The impact of shortening a long survey on response rate and response quality," Ph.D. dissertation, Brigham Young University, 2016.
[7] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of machine learning research*, vol. 3, no. Mar, pp. 1157–1182, 2003.
[8] A. Krause and D. Golovin, "Submodular function maximization." 2014.
[9] B. Nushi *et al.*, "Learning and feature selection under budget constraints in crowdsourcing," in *Fourth AAAI Conference on Human Computation and Crowdsourcing*, 2016.
[10] R. M. Groves *et al.*, *Survey methodology*. John Wiley & Sons, 2011, vol. 561.
[11] S. G. McFarland, "Effects of question order on survey responses," *Public Opinion Quarterly*, vol. 45, no. 2, pp. 208–215, 1981.
[12] D. Weintraub *et al.*, "Validation of the questionnaire for impulsive-compulsive disorders in parkinson's disease," *Movement Disorders*, vol. 24, no. 10, pp. 1461–1467, 2009.
[13] ITU-T, "Subjective performance evaluation of network echo cancellers," 1998, Rec. ITU-T P.831.
[14] P.-N. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining*. Addison-Wesley, 2006.
[15] T. M. Cover and J. A. Thomas, *Elements of information theory*. John Wiley & Sons, 2012.
[16] G. L. Nemhauser *et al.*, "Best algorithms for approximating the maximum of a submodular set function," *Mathematics of operations research*, vol. 3, no. 3, pp. 177–188, 1978.
[17] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in python," *Journal of machine learning research*, vol. 12, no. Oct, pp. 2825–2830, 2011.