Improving Accuracy in Mobile Human Contributions: An Overview

Niels van Berkel
The University of Melbourne
Melbourne, Australia
niels.van@unimelb.edu.au

Matthias Budde
Karlsruhe Institute of Technology
TECO / Pervasive Computing
Karlsruhe, Germany
budde@teco.edu

Senuri Wijenayake
The University of Melbourne
Melbourne, Australia
swijenayake@student.
unimelb.edu.au

Jorge Goncalves
The University of Melbourne
Melbourne, Australia
jorge.goncalves@unimelb.edu.au

UbiComp/ISWC'18 Adjunct, October 8–12, 2018, Singapore, Singapore
© 2018 Copyright is held by the owner/author(s). Publication rights licensed to ACM. ACM ISBN 978-1-4503-5966-5/18/10…$15.00
https://doi.org/10.1145/3267305.3267541

Abstract

The collection of human contributions through mobile devices is increasingly common across a range of methodologies. However, possible quality issues of these contributions are often overlooked. As the quality of human data has a direct impact on study reliability, more should be done to improve the accuracy of these contributions. We identify and categorise solutions aimed at increasing the accuracy of contributions prior, during, and following data collection. Our categorisation assists in the positioning of future work in this area and fosters the usage of cross-methodological practises.

Author Keywords

Mobile sensing; citizen science; self-report; Experience Sampling Method; Ecological Momentary Assessment; crowdsourcing; data quality; verification.

ACM Classification Keywords

H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

Introduction

With the increased availability of smartphones, in-the-wild collection of human-labelled data has become a feasible and increasingly popular solution for researchers [2,12]. This form of smartphone-based data collection has proven useful for a variety of

Challenges faced prior to data collection

- Decide on study design
- Determine questions / tasks
- Identify appropriate question / task scheduling technique
- Develop and design task / questionnaire software
- Find and select suitable participants
- Provide appropriate training / instructions to participants

Table 1: Summarized challenges related to data quality prior to data collection.

methodologies. For example, smartphones allow crowdsourcing tasks to be completed in the outside world rather than at a desk, and self-report researchers have the possibility to collect contextual information in addition to human-labelled self-reports.

It is interesting to note that although several of these methodologies have developed method-specific approaches to increase data quality, few of these consider the possibilities and limitations introduced by mobile devices. For example, situated crowdsourcing has made use of public displays to increase human accuracy by tapping into local knowledge [10]. Similarly, citizen science has seen an increase towards Open Data, enabling citizens to verify existing data and contribute to any data gaps. In self-report studies, researchers have embraced mobile devices to present questions, although the use of sensors or novel display options to improve data quality remains underused [2].

A common thread throughout the various applications of smartphones in human contribution tasks is the longitudinal nature of the studies. Rather than focusing on a single moment of data collection, these studies typically run for multiple days, weeks, or even months on end. As participants are requested to provide frequent contributions, their motivation to offer high-quality input decreases over time.

In this paper we identify the challenges faced by researchers when collecting longitudinal human contributions using mobile devices. We group these challenges in temporal relation to the data collection; challenges faced prior to, during, and following data collection. Then, we categorise and describe promising solutions as proposed throughout the literature on mobile human contributions. This categorisation offers not only an overview of current accuracy improvement methods, but can also be used as a framework for future contributions in this developing research area.

Prior to Data Collection

The accuracy of contributions can be affected long before data collection (Table 1). We identify a total of four areas with strategies related to the accuracy of human contributions.

Study Design

Perhaps the most influential decision made by the researcher is the design of a study. Decisions relevant to the quality of human submissions require the researcher to decide when to ask for data, how often to ask for data, and the desired completion time for such requests. Previous work has therefore investigated scheduling mechanisms for data contribution requests. Mehrotra et al. [14] propose that the quality of submissions can be negatively affected if participants are interrupted. To circumvent this problem, they suggest the use of interruptibility prediction based on contextual inference. The use of smartphones allows for the triggering of data requests following a certain event (*e.g.*, change in location) [2]. However, previous work shows that the use of such contextual triggers may bias the collection of data towards certain timeframes [13].

Participant Selection

Song et al. [16] state that the common objective of participant selection algorithms is to determine an optimal subset of participants which yields the expected quality of information given a set of constraints (*e.g.*, incentives and sensing capabilities). The use of human participants inherently introduces the risk of sampling

Challenges faced during data collection

- Loss of participant motivation over time
- Prevent / identify fraudulent input by participants
- Continuous increase in participant mistakes
- Potential for participant dropout
- Monitor data collection / collect meta data

Table 2: Summarized challenges related to data quality during data collection.

bias. This is especially true for projects for which compensation is limited, attracting participants with an above average interest in the studied phenomena. Attempts to recruit participants from a wider population can however also be detrimental to data quality. Stone et al. [17] offered a $250 reward in a self-report study and encountered poor data quality. They hypothesise that this is a result of attracting those interested mostly in the monetary reward rather than the study itself.

Participant Training / Instructions
Budde et al. [6] note that "*the most intuitive approach to ensure that users perform a task correctly is training*". The use of (individual) intake sessions is a common approach to provide one-on-one instructions to participants. While this typically works well, the downsides of this approach are inherit problems with scalability and the difficulty of reaching participants outside of your geographic area. Furthermore, in the case of longitudinal studies or a complicated set of tasks, it is likely that participants will (partially) forget given instructions – resulting in reduced accuracy.

Task / Questionnaire Design
The design of mobile questionnaires is constrained by the limited physical space and functionalities offered by smartphones. Consolvo & Walker [7] state that text readability (*e.g.*, font-size, contrast) and appropriated use of modalities (*e.g.*, text-based, audio-based) are important elements to consider when designing a mobile questionnaire.

An interesting approach in the design of mobile questionnaires is the use of a 'explicitly verifiable question' [11]. These are questions for which the researcher has the correct answer, and which are typically straightforward in nature. Verifiable questions have been shown to improve answer quality, as participants realise that their answers can be used to identify those who provide intentionally fraudulent input [11]. Their effect has not yet been verified in mobile based human contributions.

Hosio et al. [9] present a mobile application in which participants indicate their personal priorities in choosing a healthcare solution (*e.g.*, costs, duration). The content presented to participants is adjusted based on the indicated priorities. This approach, in which content is tailored to the interest of the participant, could be used in a variety of human-labelling studies to reduce the length and space required for task content.

During Data Collection

Longitudinal data collection introduces a variety of challenges related to the quality of data (Table 2). Given the often repetitive nature of the given tasks, it is not uncommon for participants to lose interest.

Intrinsic Motivation
A participant's intrinsic motivation describes an interest or willingness to contribute to a project based on internal rewards (rather than external rewards such as monetary compensation). Land-Zandstra et al. [12] summarise various reason that may stimulate intrinsic motivation among participants; ability to contribute to scientific research or the environment, genuine interest in the project or scientific topic, enjoyable to participate, and an interest to get involved with other people with similar interests. Intrinsic motivators are highly individual-dependent and may not be applicable to all studies involving human data contributions.

Extrinsic Motivation

Extrinsic motivators, in contrast to the aforementioned intrinsic motivators, are motivators external to the participant and often created by the researcher. The use of extrinsic motivators has been successfully applied in previous work to increase motivation. However, the literature also reveals some caveats related to the usage of extrinsic motivators.

The use of gamification elements is a fruitful area of exploration. Budde et al. [5] argue that 'Sensified Gaming' can be used to ensure sensing is carried out correctly. In an ESM field study, Van Berkel et al. [3] demonstrate that gamification resulted in an increase in both the number and quality of proactive participant contributions, as assessed by peer participants. However, the authors also note that gamification can have unintended side-effects, such as the pressure that can be experienced by a countdown element.

The use of micro-payments per individual data contribution is another extrinsic motivator. Talasila et al. [18] explored the use of micro-incentives in mobile crowdsensing and found that the quality of completed tasks with a higher monetary reward was higher in relation to comparable tasks with a lower financial reward. A downside to the use of extrinsic motivations is that they can quickly replace any intrinsic motivations a participant may have had.

Monitoring Contributions

The continuous online connection of mobile devices allows researchers to monitor continuously monitor contributions. Van Berkel et al. [2] suggest to contact participants if data contribution has come to an unexpected halt. However, this approach is not scalable for large-scale studies. Alternatively, automated systems can be developed that continuously assess participant accuracy and intervene when required. Even if no intervention is possible, contributions should still be monitored and meta data collected (*e.g.*, response rates), in order to assess data quality after-the-fact.

Pro-active Instructions

In-app instructions can be used to offer practical guidelines to participants as they are about to collect data. In a participatory sensing study by Budde et al. [6], the use of in-app instructions led to a significant reduction in the number of user errors. These instructions provided participants with the correct data collection procedure *in situ*. Not only mode and timing of instructions are important, so is the content. Wording should be precise, and its understanding should be double-checked as to prevent undesired results or side-effects on motivation [4].

Feedback

Budde et al. [6] show how mobile sensing can be employed to detect whether a sensing task is completed correctly. If an error is detected (e*.g.*, placement of the device), the software refuses to proceed and guides the participant to improve the taken sensing approach.

Work by Dow et al. [8] shows that feedback from external experts can be used to improve data quality in crowdsourcing. Furthermore, encouraging participants to assess their own work resulted in significant improvements of data quality. Atreja et al. [1] use NLP for the automated classification of human reports as well as the answering of citizen queries with virtual agents. The authors claim that the effective communication keeps users involved and motivated.

Following Data Collection

Although study data has already been collected at this stage, the researcher still faces a number of challenges (Table 3) in relation to the reliability of the dataset.

Challenges faced following data collection

- Identify fraudulent participants
- Identify incorrect responses
- Remove outliers
- Response shift (change in participant baseline for data collection)

Table 3: Summarized challenges related to data quality following data collection.

Data Filtering & Cleaning

Kittur et al. [11] propose that the time taken to complete a task can be used as an indicator of participants gaming the system (*i.e.*, a task which is completed suspiciously fast). While helpful, determining a suitable cut-off threshold remains difficult. The aforementioned explicitly verifiable question can be used to detect fraudulent input and subsequently remove subversive participants from the dataset.

Budde et al. [4] report on the use of automated logfiles or clickstreams to identify fraudulent participants, as well as inconsistent internal validity scores when using standardized questionnaires.

Response Shift

A challenge faced in the collection of reflective data is a shift in the meaning assigned to response scales by participants. For example, a participant reporting on the air quality on a 5-point Likert scale may discover that their assumed maximum level of polluted air is higher than anticipated – leading to a shift in the used baseline for air quality. This is known as the response shift phenomenon. A then-test can be used to measure the participant's recalibration through a retrospective pretest–posttest design (see *e.g.*, [15]).

Discussion

The presented overview of accuracy improvement techniques shows a variety of methods employed both prior, during, and following data collection. The choice for an appropriate improvement technique depends on study details (*e.g.*, intrinsic vs. extrinsic motivation of participants, face-to-face or online interaction with study participants, etc.) as well as the possibilities to evaluate participant contributions (*e.g.*, availability of ground truth answers, personal opinions or generalisable statements).

Assessing Accuracy

Key to all scientific efforts towards improving human accuracy is the ability to assess participant accuracy. Previous studies have employed various techniques to assess their participants' accuracy and therefore the effect of certain interventions. This includes comparison of participant data to ground-truth sensor data, analysis techniques to identify outliers and suspicious contributions, or utilising the knowledge of the crowd to compare participant answers. Lessons learned from studies containing ground-truth data can inform researcher decisions in studies which lack ground-truth, for example in the collection of emotional states.

Conclusion

Despite the increased reliance on human contributions across a variety of methodologies, current work on accuracy in mobile human sensing remains both limited and fails to cross methodological boundaries. Our overview provides a starting point for future accuracy improvement research in the community.

Acknowledgements

This work is partially funded by the Academy of Finland (Grants 286386-CPDSS, 285459-iSCIENCE, 304925-CARE), the European Commission (Grant 6AIKA-A71143-AKAI), and Marie Skłodowska-Curie Actions (645706-GRAGE).

References

[1] Shubham Atreja, et al. 2018. Citicafe: An Interactive Interface for Citizen Engagement. *IUI*, 617-628.

[2] Niels van Berkel, Denzil Ferreira and Vassilis Kostakos. 2017. The Experience Sampling Methods on Mobile Devices. *ACM Computing Surveys*, 50 (6). 93:91-93:40.

[3] Niels van Berkel, et al. 2017. Gamification of Mobile Experience Sampling Improves Data Quality and Quantity. *IMWUT*, 1 (3). 107:101-107:121.

[4] Matthias Budde, et al. 2017. Lessons from Failures in Designing and Conducting Experimental Studies – a Brief Anectodal Tutorial. *UbiComp Adj.*, 992-999.

[5] Matthias Budde, et al. 2016. Sensified Gaming: Design Patterns and Game Design Elements for Gameful Environmental Sensing. *Advances in Computer Entertainment Technology*, 1-8.

[6] Matthias Budde, et al. 2017. Participatory Sensing or Participatory Nonsense?: Mitigating the Effect of Human Error on Data Quality in Citizen Science. *IMWUT*, 1 (3). 1-23.

[7] Sunny Consolvo and Miriam Walker. 2003. Using the Experience Sampling Method to Evaluate Ubicomp Applications. *IEEE Pervasive Computing*, 2 (2). 24-31.

[8] Steven Dow, et al. 2012. Shepherding the Crowd Yields Better Work. *CSCW*, 1013-1022.

[9] Simo Hosio, et al. 2018. Mobile Decision Support and Data Provisioning for Low Back Pain. *IEEE Computer*.

[10] Simo Hosio, et al. 2018. Facilitating Collocated Crowdsourcing on Situated Displays. *Human–Computer Interaction*. 1-37.

[11] Aniket Kittur, Ed H. Chi and Bongwon Suh. 2008. Crowdsourcing user studies with Mechanical Turk. *CHI*, 453-456.

[12] Anne M. Land-Zandstra, et al. 2016. Citizen science on a smartphone: Participants' motivations and learning. *Public Understanding of Science*, 25 (1). 45-60.

[13] Neal Lathia, et al. 2013. Contextual Dissonance: Design Bias in Sensor-based Experience Sampling Methods. *UbiComp*, 183-192.

[14] Abhinav Mehrotra, et al. 2015. Ask, But Don't Interrupt: The Case for Interruptibility-Aware Mobile Experience Sampling. *UbiComp Adj.*, 723-732.

[15] Carolyn E. Schwartz, et al. 2004. Exploring response shift in longitudinal data. *Psychology & Health*, 19 (1). 51-69.

[16] Z. Song, et al. 2014. QoI-Aware Multitask-Oriented Dynamic Participant Selection With Budget Constraints. *IEEE Transactions on Vehicular Technology*, 63 (9). 4618-4632.

[17] Arthur Stone, Ronald Kessler and Jennifer Haythomthwatte. 1991. Measuring Daily Events and Experiences: Decisions for the Researcher. *Journal of Personality*, 59 (3). 575-607.

[18] M. Talasila, R. Curtmola and C. Borcea. 2016. Crowdsensing in the Wild with Aliens and Micropayments. *IEEE Pervasive Computing*, 15 (1). 68-77.