A Methodology and System For Big-Thick Data Collection

Ivan Kayongo [1], Haonan Zhao [1], Leonardo Malcotti [1] und Fausto Giunchiglia [1]

Abstract: Pervasive sensors have become essential in research for gathering real-world data. However, current studies often focus solely on objective data, neglecting subjective human contributions. We introduce an approach and system for collecting big-thick data, combining extensive sensor data (big data) with qualitative human feedback (thick data). This fusion enables effective collaboration between humans and machines, allowing machine learning to benefit from human behavior and interpretations. Emphasizing data quality, our system incorporates continuous monitoring and adaptive learning mechanisms to optimize data collection timing and context, ensuring relevance, accuracy, and reliability. The system comprises three key components: a) a tool for collecting sensor data and user feedback, b) components for experiment planning and execution monitoring, and c) a machine-learning component that enhances human-machine interaction.

Keywords: Personal data collection, Human-aware AI, Big-thick data, Context, data quality

1 Introduction

Pervasive sensors are extensively utilised for data collection, which subsequently provides features that facilitate the understanding of the world [Yü14]. This is often referred to as Big Data [DK13]. Although Big Data offers an objective perspective of reality, it is unable to elucidate the subjective motivations that drive an individual's actions. On the other hand is Thick Data, a category of data sources that are consistent with ethnographically aligned and carefully analysed observational data [BD18]. Big data offers extensive quantitative insights, whereas thick data provides qualitative insights into human behaviour, experiences, and motives. When both are combined, to form Big-thick data, they provide a more comprehensive perspective of human needs and preferences to machines. One of the key elements of Big-Thick data is *context* [GBZ17; In03; Ru13], which is the situational setting of an individual that encompasses their internal condition within a common reference environment with others.

Current research on context data collection primarily focuses on context interruptions [Mi17], capturing of user attention [Me16], and enhancement of question response rates [SRV21]. However, the current data quality does not meet the demands of big-thick data. A significant challenge arises from machines needing to pose a high volume of questions to humans, often resulting in low-quality responses [Me15]. Certain methods also rely on fixed or random schedules that may not align with participants' availability or willingness to respond [Gi21b], thereby affecting the quality of the data.

[1] University of trento, Italy, ivan.kayongo@unitn.it, https://orcid.org/0009-0007-4429-7335;
haonan.zhao@unitn.it, https://orcid.org/0000-0003-0825-7188;
leonardo.malcotti@studenti.unitn.it, https://orcid.org/0009-0000-0589-6393;
Fausto.giunchiglia@unitn.it, https://orcid.org/0000-0002-5903-6150

doi:10.18420/inf2024_33

In light of the above, our research objective in this paper is *to develop a system that enables collection of high quality Big-Thick data, while ensuring minimal disturbance from the AI system*. Our aim is to collect quality Big-Thick data from both an objective and a subjective perspective, while monitoring this process. We offer participants flexibility by incorporating a Machine Learning component which adaptively schedules context questions based on participants' availability and willingness to respond as a means to improve data quality. For monitoring, our system enables visualizing the collected data via a dashboard, thereby providing real-time feedback on the quality for example, missing sensor and question data.

The system developed according to the methodology in this paper is an improvement and extension of iLog [ZZG14], which has been widely used in various data collection experiments and studies on human behavior. The first experiment, SU2 (*Smart University Two*), gathered data on students at the University of Trento, Italy. This data was used in [Gi21b] to predict individuals' behaviors, and in [Gi18] to examine the impact of social media usage on students' academic performance. Another major data collection involved two experiments, DIV1 (*Diversity One*) and DIV2 (*Diversity Two*) [Gi21a], which collected data on student behavior, mood, and food habits across eight countries. These extensions are motivated by the goal of having a fully integrated human-in-the-loop human machine interaction [Gi22a; Gi22b].

The organization of this paper is as follows: Section 2 provides an overview of the related work. Section 3 presents our overall architecture. Section 4 delineates the context model. Sections 5 and 6 elucidate the monitoring and scheduling methods, respectively. Finally, Section 7 offers the conclusion of the paper.

2 Related Work

The collection of big-thick data is mostly through; utilization of smartphone sensors and direct acquisition of information from participants. The former involves the use of sensors [Yü14], while the latter employs Time Use Diaries (TUDs) [Bo39]. In social sciences, the Experience Sampling Method (ESM) [LC14] is one of the principal methods for gathering information from participants to document their behaviors and this helps provide a ground truth as the data's meaning is directly provided by the user. However it is beset with issues pertaining to the quality of responses [BZ23; BZG24]. A significant challenge resides in accurately determining the optimal timing for responses, a task complicated by the difficulties in observing respondent behavior when questions are scheduled at fixed intervals thus impacting its quality.

A number of platforms exist that collect and process big-thick data from smart devices. For instance, DemaWare2 [SMK17] is a context-aware fusion system that employs OWL2 as its knowledge representation and reasoning language. The contextual information provided is used to identify links between observations that signify the presence of complex activities. Aware [FKD15], built on top of Beiwe [On21], is an open-source mobile platform

that generates user contexts from sensor data from smart devices and human digital questionnaires.

To better understand the collected data, visualization can be employed. It effectively supports data exploration, insight communication, and data model improvement / understanding, for quality data. For example in [SWM17], predictions of a deep learning model are explained through visualization We leverage visualization not only to gain insights into the collected Big-Thick data but also to monitor the progress of the experiment for quality data collection.

3 Logical Architecture

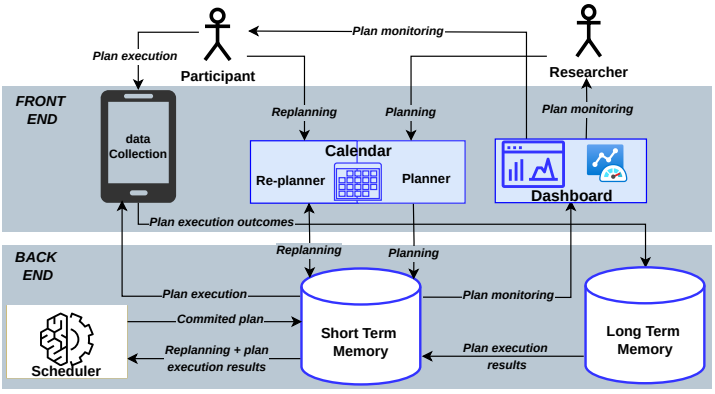

Fig. 1: Logical architecture

Fig. 1 depicts the overall system architecture; The light blue rectangles denote the primary components, the calendar and dashboard, with which individuals interact. Dark blue cylinders represent two databases, the first, termed the short term memory (STM) database, stores the plans from both researchers and participants, while the second, named Long Term Memory, the plan execution outcomes from participants. The orange rectangle represents the *scheduler*, a machine learning module that modifies plans based on participants' data. The humanoid figures symbolize two roles, i.e., researchers and participants. A researcher is one seeking to gather context information from participants, while a participant is one that imparts their knowledge through the data collection app. The arrow lines illustrate the interactions between the different components, and the mobile phone serves as the tool for collecting data and enables the participant to receive context questions.

The Big-Thick data collection process begins with the researcher's planning. Using the Planner component within the Calendar module, the researcher schedules context questions and sensor data collection. This plan is stored in the STM database as scheduled actions, which are then sent to the data collection application at the designated times for participants to provide responses. Participants can modify the plan through the re-planner component

in the Calendar module. Data collected during the execution of the plan is archived in the Long Term Memory database, while plan execution results are sent to the STM database. Furthermore, the scheduler component learns the optimal times to ask context questions based on the personal context and re-plan modifications from participants. It then sends these predicted times to the STM database to inform the planning process, ensuring the system adapts and responds effectively to changes.

Subsequently, the dashboard provides a means for both the researcher and participant to view and interact with the collected data. They can navigate through and filter the data, gaining valuable insights. For participants, this awareness of their lifestyle, as reflected in the data, can lead to positive behavioral improvements. Participants can also see how their data compares to that of other participants or to the limits set by the researcher.The dashboard provides researchers with a comprehensive overview of the experimental data, including any encountered problems. This enables them to promptly address these difficulties to ensure the seamless execution of the experiment for accurate data collection.

4 Representing Context

The goal of asking context questions or collecting sensor data is to understand the personal context annotation from a participant. The annotated context used in this study is defined in [GBZ17; Gi93]. As an example, let's consider an afternoon at a student's place where he is in a discussion with his friend. Fig. 2 shows this scenario as a knowledge graph, representing the personal context of the student. Each node represents an entity, e.g., person and room, with their respective attributes with values; for instance, attributes of *ME (whose context we are describing)* are $Class$, $Name$, $Mood$, *Notification time* and *Answer time* with the corresponding values as shown in Fig. 2. Edges represent relations between entities, e.g., *Sitting room* is $PartOf\ Home$, whereas; *Peter (person)* and the *dining table (Table)* which are both *in* the *sitting room* which $HasActivity$ of a $discussion$ taking place.

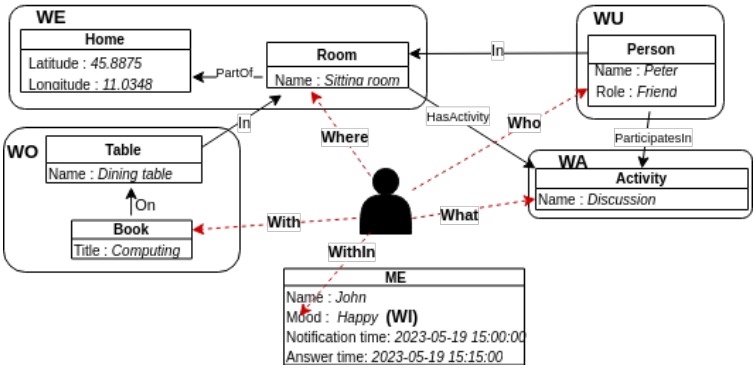

Fig. 2: A motivating example of the situational context model

The relation between context as a partial representation of the real world and the subject is represented in Fig. 2 as:

$$MyWorld = <Cxt, me> \quad (1)$$

where; *Cxt* is the context (real world) of the subject, including elements in his local immediate surrounding and *me* is the subject, represented as an entity with attributes and relations. The red arrows in Fig. 2 show the relation between *me* and *Cxt* thus;

$$Cxt = < WA , WE , WI , WO , WU> \quad (2)$$

where:

- *WA* is the temporal context from answering the question; "**W**h**A**t are you doing?". In Fig. 2, *WA* is the main activity taking place, i.e. the *discussion*.
- *WE* is the spatial context generated from the question "**W**h**E**re are you?". In Fig. 2, *WE* shows the most relevant location, i.e. the *sitting room*.
- *WI* is the internal context generated from the question "**W**hat mood are you **I**n?". As shown in Fig. 2, it is the emotional state (mood) of *me*.
- *WO* is the object context generated from the question "**W**hich **O**bject are you with?". In Fig. 2, *WO* includes a few objects e.g., the *dining table*, and the *book* from which they are *discussing*.
- *WU* is the social context generated from the question "**W**ho is with you (**U**)?". As shown in Fig. 2, *WU* focuses on *me*'s friend *Peter*.

The sensor data collected is used to add attributes to the context. For example in Fig. 2, the *latitude* and *longitude* attribute values in the *Home* entity of the *WE context* are sensor readings. Other collected sensor data includes; accelerometer, social media apps, Bluetooth devices among others, as detailed in [Gi21a].

5 Monitoring the Plan Execution

The dashboard, as a visualization component offers flexibility to both the researcher and the participants in monitoring the execution of the experiment plan. It allows them to view and explore the collected big-thick data at any given time. The dashboard helps them gain insights from the data as explained below.

Researcher: The researcher designs an experiment to determine which data to collect and when and how to collect the data. Through the dashboard, he is able to view each participant's data including the number of questions answered and sensor data collected per participant. He can also compare the data collected between different participants and even filter it using options to get those with the least or most contributions. Not only does it

facilitate him to gain insights in the collected data, but it also helps in the monitoring of the experiment progress to ensure that everything goes as planned. Details of the plan execution process including questions sent and data collected (answers and sensors) are displayed on graphs which in turn help in understanding if the plan is progressing as expected or if mitigation is needed, ensuring the collection of high-quality data.

Participant: The participant registers to take part in an experiment and installs the data collection tool on a smart device. Through the dashboard, participants can navigate and explore their data to gain insights about themselves. For example, depending on the experiment's nature and questions, they can learn about their lifestyle, such as if they eat a lot of snacks or spend excessive time in a particular location. Such insights can highlight negative behaviors and prompt positive lifestyle changes. Participants can also compare their data against other participants to see how they fare in terms of data collected in the experiment. Fig. 3 shows such an example, where one participant (green line) is compared to three others (orange, red and blue lines). The y-axis shows the number of questions answered whereas the x-axis shows the different experiment days. Such comparison prompts the participants improve on their contribution in the experiment.

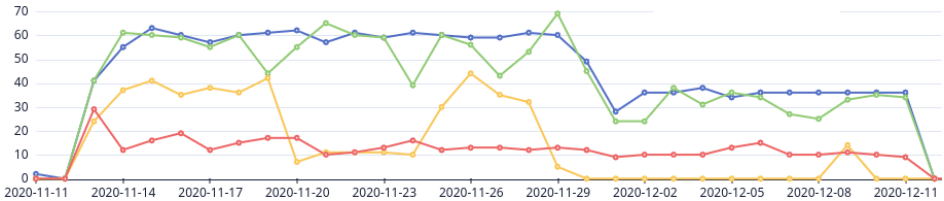

Fig. 3: Comparison of a participant's answers (green line) with three other participants (blue, red and orange lines)

The two functionalities explained above are some examples of the dashboard's functionalities; other key features include:

- Alerting / notifying the researcher or the participant in case of inconsistencies in the collected data;
- Viewing participant response time and answer completion time and how these are affected by other factors for example social media, mood or location (sensor data);
- Enabling the participants to set personal goals and keep track of progress by using the collected Big-Thick data.

6 Scheduling

The scheduler is a machine learning component which learns and modifies the experiment plan by using a combination of the answers given by a participant and the modifications

they make in the re-planner. Let's explain this with an example: Given that our participants are students, we try to refrain from sending questions while they are studying. To address this, we use the scheduler component to predict the periods when the participants are likely to be in class, thereby avoiding any disruption to their academic activities.

Using data from the WeNet experiment described in [Gi21a], we trained Random Forests, Decision Trees, Artificial Neural Networks, Logistic Regression, and Gaussian Naive Bayes on comprehensive training and testing sets encompassing all participants (170 students). We chose to binary encode study activities. Specifically, studying alone or with others and attending a classroom lecture are encoded as 1, and all other activities as 0. This information was collected from time diaries questionnaire with the question: "What are you doing?". The results are shown in Tab. 1 with the Random Forest classifier showing the highest prediction accuracy of 75.1%. Other features that can be used for training include; social demographics (e.g gender, department), time (e.g hour, weekday), situational context (e.g location, interacting individuals), personality traits [Do06], and mood.

Classifier	Accuracy	Kappa	Precision	Recall	F1 score	AUC
Random Forest	0.7510	0.3832	0.6903	0.4491	0.5442	0.8146
Decision Tree	0.7234	0.3338	0.6091	0.4587	0.5233	0.6821
Artificial Neural Networks	0.7209	0.3702	0.5781	0.5799	0.5790	0.7762
Logistic Regression	0.4557	0.3536	0.4892	0.0740	0.1286	0.6705
Gaussian Naive Bayes	0.6295	0.2388	0.4571	0.6366	0.5321	0.6676

Tab. 1: Prediction results of different machine learning classifiers on Wenet Data

7 Conclusion

In this paper, we have presented a novel system for quality Big-Thick data collection, designed to enhance user experience and improve the quality and quantity of collected data. We have fostered a meaningful collaboration between humans and AI by granting participants the flexibility to choose their response times for context questions and sensor collections. Our system incorporates a visualization component, allowing both researchers and participants to monitor the experiment's progress and gain insights from the data. Future work will focus on evaluating the system in real-world scenarios, comparing it with existing methods, and addressing potential ethical and privacy implications.

Acknowledgement

The research by Fausto, Ivan, and Leonardo were funded by the European Union's Horizon 2020 FET Proactive project "WeNet – The Internet of us", grant agreement No 823783. The work by Haonan received funding from the China Scholarships Council (No.202107820038).

References

[BD18] Bornakke, T.; Due, B. L.: Big–Thick Blending: A method for mixing analytical insights from big and thick data sources. Big Data & Society 5 (1), S. 2053951718765026, 2018.

[Bo39] Bowers, R. V.: Time Budgets of Human Behavior. 1939.

[BZ23] Bison, I.; Zhao, H.: Factors Impacting the Quality of User Answers on Smartphones. In: Proceedings of the Second International Conference on Hybrid Human-Machine Intelligence (HHAI 23). Bd. 3456, S. 208–213, 2023.

[BZG24] Bison, I.; Zhao, H.; Giunchiglia, F.: What Impacts the Quality of the User Answers when Asked about the Current Context? arXiv preprint arXiv:2405.04054, 2024.

[DK13] Das, T. K.; Kumar, P. M.: Big data analytics: A framework for unstructured data analysis. International Journal of Engineering Science & Technology 5 (1), S. 153, 2013.

[Do06] Donnellan, M. B. et al.: The mini-IPIP scales: tiny-yet-effective measures of the Big Five factors of personality. Psychological assessment 18 (2), S. 192, 2006.

[FKD15] Ferreira, D.; Kostakos, V.; Dey, A. K.: AWARE: mobile context instrumentation framework. Frontiers in ICT 2, S. 6, 2015.

[GBZ17] Giunchiglia, F.; Bignotti, E.; Zeni, M.: Personal context modelling and annotation. In: 2017 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops). IEEE, S. 117–122, 2017.

[Gi18] Giunchiglia, F. et al.: Mobile social media usage and academic performance. Computers in Human Behavior 82, S. 177–185, 2018.

[Gi21a] Giunchiglia, F. et al.: A worldwide diversity pilot on daily routines and social practices (2020). University of Trento, Technical Report. No.# DISI-2001-DS-01, 2021.

[Gi21b] Giunchiglia, F. et al.: Putting human behavior predictability in context. EPJ Data Science 10 (1), S. 42, 2021.

[Gi22a] Giunchiglia, F. et al.: A context model for personal data streams. In: Asia-Pacific Web (APWeb) and Web-Age Information Management (WAIM) Joint International Conference on Web and Big Data. Springer, S. 37–44, 2022.

[Gi22b] Giunchiglia, F. et al.: Lifelong Personal Context Recognition. arXiv preprint arXiv:2205.10123, 2022.

[Gi93] Giunchiglia, F.: Contextual reasoning. Epistemologia, special issue: I Linguaggi e le Macchine 16, S. 345–364, 1993.

[In03] Intille, S. S.; Rondoni, J.; Kukla, C.; Ancona, I.; Bao, L.: A context-aware experience sampling tool. In: CHI'03 extended abstracts on Human factors in computing systems. S. 972–973, 2003.

[LC14] Larson, R.; Csikszentmihalyi, M.: The experience sampling method. In: Flow and the foundations of positive psychology. Springer, S. 21–34, 2014.

[Me15] Mehrotra, A. et al.: Ask, but don't interrupt: the case for interruptibility-aware mobile experience sampling. In: Adjunct Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2015 ACM International Symposium on Wearable Computers. S. 723–732, 2015.

[Me16] Mehrotra, A. et al.: My phone and me: understanding people's receptivity to mobile notifications. In: Proceedings of the 2016 CHI conference on human factors in computing systems. S. 1021–1032, 2016.

[Mi17] Mishra, V. et al.: Investigating contextual cues as indicators for EMA delivery. In: Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers. S. 935–940, 2017.

[On21] Onnela, J.-P. et al.: Beiwe: A data collection platform for high-throughput digital phenotyping. Journal of Open Source Software 6 (68), S. 3417, 2021.

[Ru13] Runyan, J. D. et al.: A smartphone ecological momentary assessment/intervention "app" for collecting real-time data and promoting self-awareness. PloS one 8 (8), e71325, 2013.

[SMK17] Stavropoulos, T. G.; Meditskos, G.; Kompatsiaris, I.: DemaWare2: Integrating sensors, multimedia and semantic analysis for the ambient care of dementia. Pervasive and Mobile Computing 34, S. 126–145, 2017.

[SRV21] Sun, J.; Rhemtulla, M.; Vazire, S.: Eavesdropping on missing data: What are university students doing when they miss experience sampling reports? Personality and Social Psychology Bulletin 47 (11), S. 1535–1549, 2021.

[SWM17] Samek, W.; Wiegand, T.; Müller, K.-R.: Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. arXiv preprint arXiv:1708.08296, 2017.

[Yü14] Yürür, Ö. et al.: Context-awareness for mobile sensing: A survey and future directions. IEEE Communications Surveys & Tutorials 18 (1), S. 68–93, 2014.

[ZZG14] Zeni, M.; Zaihrayeu, I.; Giunchiglia, F.: Multi-device activity logging. In: Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication. S. 299–302, 2014.