A context model for collecting diversity-aware data

Matteo Busso[1,*], Xiaoyue Li[1]

[1]Department of Information Engineering and Computer Science, University of Trento, Via Sommarive 9, 30123, Trento, Italy

Abstract

Diversity-aware data are essential for a robust modeling of human behavior in context. In addition, being the human behavior of interest for numerous applications, data must also be reusable across domain, to ensure diversity of interpretations. Current data collection techniques allow only a partial representation of the diversity of people and often generate data that is difficult to reuse. To fill this gap, we propose a data collection methodology, within a hybrid machine-artificial intelligence approach, and its related dataset, based on a comprehensive ontological notion of context which enables data reusability. The dataset has a sample of 158 participants and is collected via the iLog smartphone application. It contains more than 170 GB of subjective and objective data, which comes from 27 smartphone sensors that are associated with 168,095 self-reported annotations on the participants context. The dataset is highly reusable, as demonstrated by its diverse applications.

Keywords

Diversity, Big Thick Data, Situational Context, Data Collection, Methodology

1. Introduction

Diversity-aware data are essential for a robust modeling of human behavior in context. Nowadays it is common to associate people behavioral data with large data collections based on smartphone and smartwatch sensors, which allow to observe the person in her everyday life. However, as rich as these data collections are, many useful variables are unavailable, therefore people are "at best, thinly described" [1], since the granularity of the sensor data is essential but not enough to represent people's diversity in their context. Clearly, diversity lies not only within the person behavior but also in its interpretation. However, the lack of essential variables makes the data "often used 'out of context', which decrease the 'meaning and value'" [2].

To generate diversity-aware data, several hybrid techniques are applied, such as annotation through labels [3, 4, 5], aggregation, fusion or integration of data, for example in user profiling and record linkage [6]. A particularly important technique, which is closer to our approach, is blending, namely combining sensor data sources with high quality ethnographic data [7], with the aim of creating *Big Thick Data*. Thick data differs from Big (Thin) Data because it extends on many dimensions, gathering information that reveals the emotions and contexts of people.

HHAI-WS 2023: Workshops at the Second International Conference on Hybrid Human-Artificial Intelligence (HHAI), June 26—27, 2023, Munich, Germany

*Corresponding author.

matteo.busso@unitn.it (M. Busso); xiaoyue.li@unitn.it (X. Li)

http://knowdive.disi.unitn.it/matteo-busso-3/ (M. Busso); http://knowdive.disi.unitn.it/xiaoyue-li/ (X. Li)

0000-0002-3788-0203 (M. Busso); 0000-0002-0100-0016 (X. Li)

However, despite the obvious benefits of the techniques listed above and widely adopted, there are some weaknesses. First of all, most of the annotations are done by the researcher after the data collection, thus losing the immediacy and the wealth of information that derives from the confrontation with the subject who is providing the data [8]. This leads to a second problem, which affect the reuse of the collected data. The labelling process (whether they are the codings done from an anthropologist, or the integration work of a data scientist), although enriching the dataset content, reduces its reusability across disciplines (a well known issue within the F.A.I.R.[1] research field), which indirectly leads to a reduced diversity in data interpretation.

To fill this gap, we propose a state-of-the-art rich dataset, called SmartUnitn2 (SU2), for recognizing people context[2]. To generate a dataset that is both annotated and reusable at the same time, we followed a hybrid human-artificial intelligence approach, based on a comprehensive theory of context representation that integrates the person's point of view on the surrounding situation [10, 11] within the data collected by the smartphone sensors. The approach provides a related ontology, which improve the dataset interoperability. Furthermore, to enhance the cross-domain reuse, the dataset is based on interdisciplinary standards and it is built following guidelines from sociology, which has a strong tradition in data collection methodology (see, e.g., [12]).

The remainder of the paper is organized as follow. Section 2 presents the notion of context, while Section 3 describes its operationalization within the data collection process and the resulting dataset. Section 4 suggests how to extend the datasets and provides several use cases. Section 5 closes the paper.

2. The Situational Context

A situational context is a model that represents scenarios in the world from the person's point of view, whom we call `me`, which can be characterized by her *External* (e.g., age, gender, but also her activities) and *Internal* (e.g., personality and emotions) states. The *Situational context* of `me`, denoted as $C(me)$, is defined as follows:

$$C(me) = \langle L(C(me)), E(L(C(me)))\rangle. \tag{1}$$

where $L(C(me))$ is the *Location* recognized by `me`, while $E(L(C(me)))$ is the *Event* experienced by `me` within the location of the current scenario. The location and event are considered as priors of experience and delineate the general spatial and temporal boundaries of the current scenario from `me`'s perspective. This is predicated on the notion that a person must invariably occupy a physical space and engage in at least one activity at any given time. For instance, when a person reads a paper in her office, the office is the location, while the activity of reading is the event that defines the current context. Therefore, a change of context is concomitant with a change of location or event.

Within the spatio-temporal context, other objects can interact with each other. We define them as *Parts of a Context*, denoted as $P(C(me))$, as follows:

[1] https://www.go-fair.org/fair-principles/, see also [9]

[2] The dataset respects the General Data Protection Regulation (GDPR) and it is approved by the IRB00009280 with protocol n. 2016-027 "SmartUnitn".

$$P(C(me)) = \langle me, \{P\}, \{O\}, \{F\}, \{A\} \rangle \tag{2}$$

where $\{P\}$ and $\{O\}$ are, respectively, a set of *Person*s and *Object*s populating the context. $\{F\}$ is a set of *Function*s, representing the roles that `me`, persons or objects have towards one another (e.g., Mary is a friend of `me`). $\{A\}$ is a set of *Action*s involving `me`, persons and objects (e.g., `me` opens a computer). Further details regarding *Function* and *Action* can be found in [13].

Based on the situational context model, we define a *Life sequence* of `me`, denoted as $S(me)$, as a sequence of contexts during a certain period of time:

$$S(me) = \langle C_1(me), C_i(me), \ldots, C_n(me) \rangle; \quad 1 \leq i \leq n \tag{3}$$

where $C_i(me)$ is the i_{th} situational context of `me`. Further information on how the notion of context can be extended to involve a person's life sequences can be found in [14].

3. Collecting diversity-aware data

To observe the scenarios in the world from the person's point of view, as described above, while respecting the methodological criteria of social sciences, a hybrid data collection was conducted involving 158 students for a period of one month. The collection was held through an innovative smartphone application, called iLog [15][3], which allows both to interact with the participants (e.g., by sending questions) and to collect data from all the smartphone sensors.

We consider context as a 4-tuple, which can be observed through a set of 4 questions which were asked on a regular basis (i.e., every half hour for the first two weeks of data collection, and every hour for the second two weeks), which are:

1. WHAT - "What are you doing?" to annotate the ongoing *Event*s of the person.
2. WHERE - "Where are you?" to annotate the current *Location* of the person.
3. WHOM - "Who is with you?" to annotate the *Person* the participant was with.
4. WHITIN - "What is your mood?" to annotate the person *Internal state*.

These annotations are collected according to the time diaries methodology, a classic social science approach [20], that can be based on the HETUS[4] standard. To this standard we added a mood related question to document the *Internal* state of the person. In addition, and to consider other *Internal* and *External* states, a profiling questionnaire was collected, following the standards of other data collections (in particular [21]) and asking question based on reliable standardized scales, such as a short version of the Big Five Personality traits [22] among others.

The resulting dataset contains more than 170 GB of parquet data coming from 27 smartphone sensors, which are associated with 168.095 self-reported annotations. A detailed description of the data collection can be found in the technical report [23], while the set of data is described in the LiveoPeople Catalog[5].

[3] [16, 17, 18, 19] is a list of publications which describe the use of iLog and of iLog collected data in various studies.
[4] *Harmonized European Time Use Surveys*: https://ec.europa.eu/eurostat/web/time-use-survey
[5] LivePeople: https://datascientiafoundation.github.io/LivePeople/

4. Diversity-aware applications

This data set has already been used for a fair number of applications and it can be extended for further reuse, for instance via machine learning or integrating it with other datasets (in this latter case via a full exploitation of the ontological definition of the situational context). An example of a possible extension considering the OpenStreetMap data from Trentino (Italy) is provided in the Live Data Catalog[6]. We provide below a set of cross domain reuse of the dataset.

Mobile social media usage The work [24] was conducted with a previous version of the SU2 dataset, involving the sensor data called `Running Applications`, `Event` annotations, and questionnaire data. These variables were used for analysing the logs of social media apps and comparing them to students' credits and grades. The results show a negative pattern of social media usage that has a major impact on academic activities.

Predicting human behavior The study [25] investigates the role played by four contextual dimensions based on the data about `Events`, `Location`, and `Person`'s social ties, on the predictability of individuals' behaviors. The analysis shows how self-reported information has a substantial impact on predictability. Indeed, from the authors' example, the annotations of the location convey more information about activity and social ties than the information derived from GPS.

Complex Daily Activities, Country-Level Diversity, and Smartphone Sensing This study [26] is based on the context notion described in this paper, which lead to the collection of the Diversity 1 dataset [27], involving 8 different countries, allowing for cross-country diversity aware analysis. The study leverages data from multiple sensors and participant-reported `Events` to recognize complex daily activities within a diversity-aware model that shows how algorithms performs better when cross-country diversity is taken into account.

5. Conclusion

In this paper we considered how current data collection techniques allow only a partial representation of the diversity of people and often generate data that is difficult to reuse. Therefore, we proposed a data collection methodology, based on an ontological notion which considers the person point of view within her context and that guides a hybrid human-artificial intelligence approach that produces highly reusable data, and its related dataset. Finally, we showed how the dataset is highly reusable, as demonstrated by its diverse applications.

Acknowledgments

The work is funded by the *"WeNet - The Internet of Us"* Project, funded by the European Union (EU) Horizon 2020 programme under GA number 823783.

[6]LiveData: https://datascientiafoundation.github.io/LiveDataTrentino/

References

[1] N. G. Fielding, R. M. Lee, G. Blank, The SAGE handbook of online research methods, Sage, 2008.

[2] D. Boyd, K. Crawford, Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon, Information, communication & society 15 (2012) 662–679.

[3] B. Fu, N. Damer, F. Kirchbuchner, A. Kuijper, Sensing technology for human activity recognition: A comprehensive survey, IEEE Access 8 (2020) 83791–83820.

[4] X. Wu, L. Xiao, Y. Sun, J. Zhang, T. Ma, L. He, A survey of human-in-the-loop for machine learning, Future Generation Computer Systems (2022).

[5] Y. Vaizman, K. Ellis, G. Lanckriet, Recognizing detailed human context in the wild from smartphones and smartwatches, IEEE pervasive computing 16 (2017).

[6] K. Shu, S. Wang, J. Tang, R. Zafarani, H. Liu, User identity linkage across online social networks: A review, Acm Sigkdd Explorations Newsletter 18 (2017) 5–17.

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

[8] W. E. Pentland, A. S. Harvey, M. P. Lawton, M. A. McColl, Time use research in the social sciences, Springer, 1999.

[9] E. PwC, Cost of not having fair research data. cost-benefit analysis for fair research data, European Commission (2018).

[10] F. Giunchiglia, Contextual reasoning, Epistemologia, special issue on I Linguaggi e le Macchine 16 (1993) 345–364.

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

[12] H. F. Weisberg, The total survey error approach, Chicago Press, 2009.

[13] F. Giunchiglia, M. Fumagalli, Teleologies: Objects, actions and functions, in: ER- International Conference on Conceptual Modeling, ICCM, 2017, pp. 520–534.

[14] F. Giunchiglia, X. Li, M. Busso, M. Rodas-Britez, A context model for personal data streams, in: Web and Big Data: 6th International Joint Conference, APWeb-WAIM 2022, Nanjing, China, November 25–27, 2022, Proceedings, Part I, Springer, 2023, pp. 37–44.

[15] M. Zeni, I. Bison, B. Gauckler, F. Reis, F. Giunchiglia, Improving time use measurement with personal big collection - the experience of the european big data hackathon 2019., Journal of Official Statistics (2020).

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

[17] F. Giunchiglia, M. Zeni, E. Gobbi, E. Bignotti, I. Bison, Mobile social media and academic performance, in: International conference on social informatics, Springer, Cham, 2017, pp. 3–13.

[18] F. Giunchiglia, E. Bignotti, M. Zeni, Human-like context sensing for robot surveillance, International Journal of Semantic Computing 12 (2017) 129–148.

[19] F. Giunchiglia, M. Zeni, E. Big, Personal context recognition via reliable human-machine

collaboration, in: 2018 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops), IEEE, 2018, pp. 379–384. doi:10.1109/PERCOMW.2018.8480307.

[20] J. P. Robinson, The time-diary method, in: Time use research in the social sciences, Springer, 2002, pp. 47–89.

[21] R. Wang, F. Chen, Z. Chen, T. Li, G. Harari, S. Tignor, X. Zhou, D. Ben-Zeev, A. T. Campbell, Studentlife: assessing mental health, academic performance and behavioral trends of college students using smartphones, in: Proceedings of the 2014 ACM international joint conference on pervasive and ubiquitous computing, 2014, pp. 3–14.

[22] O. P. John, E. M. Donahue, R. L. Kentle, Big five inventory, Journal of Personality and Social Psychology (1991).

[23] F. Giunchiglia, M. Busso, M. Zeni, I. Bison, A survey on students' daily routines and academic performance at the university of trento (2022).

[24] F. Giunchiglia, M. Zeni, E. Gobbi, E. Bignotti, I. Bison, Mobile social media usage and academic performance, Computers in Human Behavior 82 (2018) 177–185.

[25] W. Zhang, Q. Shen, S. Teso, B. Lepri, A. Passerini, I. Bison, F. Giunchiglia, Putting human behavior predictability in context, EPJ Data Science 10 (2021) 42.

[26] K. Assi, L. Meegahapola, W. Droz, P. Kun, A. De Götzen, M. Bidoglia, S. Stares, G. Gaskell, A. Chagnaa, A. Ganbold, T. Zundui, C. Caprini, D. Miorandi, J. L. Zarza, A. Hume, L. Cernuzzi, I. Bison, M. D. Rodas Britez, M. Busso, R. Chenu-Abente, F. Giunchiglia, D. Gatica-Perez, Complex daily activities, country-level diversity, and smartphone sensing: A study in denmark, italy, mongolia, paraguay, and uk, in: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, CHI '23, Association for Computing Machinery, New York, NY, USA, 2023. URL: https://doi.org/10.1145/3544548.3581190. doi:10.1145/3544548.3581190.

[27] F. Giunchiglia, I. Bison, M. Busso, R. Chenu-Abente, M. Rodas, M. Zeni, C. Gunel, G. Veltri, A. De Götzen, P. Kun, et al., A worldwide diversity pilot on daily routines and social practices (2020) (2021).