Synergistic user ↔ context analytics

Citation for the original published paper (version of record):
Hossmann-Picu, A., Li, Z., Zhao, Z. et al (2016). Synergistic user ↔ context analytics. Advances in Intelligent Systems and Computing, 399: 163-172. http://dx.doi.org/10.1007/978-3-319-25733-4_17

(article starts on next page)

Synergistic User $\leftrightarrow$ Context Analytics

Andreea Hossmann-Picu[1], Zan Li[1], Zhongliang Zhao[1], Torsten Braun[1], Constantinos Marios Angelopoulos[2], Orestis Evangelatos[2], José Rolim[2], Michela Papandrea[3], Kamini Garg[3], Silvia Giordano[3], Aristide C. Y. Tossou[4], Christos Dimitrakakis[4], and Aikaterini Mitrokotsa[4]

[1] University of Bern, Switzerland
[2] University of Geneva, Switzerland
[3] SUPSI, Switzerland
[4] Chalmers University, Sweden

Abstract. Various flavours of a new research field on *(socio-)physical* or *personal analytics* have emerged, with the goal of deriving semantically-rich insights from people's low-level physical sensing combined with their (online) social interactions. In this paper, we argue for more comprehensive data sources, including environmental and application-specific data, to better capture the interactions between users and their context, in addition to those among users. We provide some example use cases and present our ongoing work towards a synergistic analytics platform: a testbed based on mobile crowdsensing and IoT, a data model for representing the different sources of data and their connections, and a prediction engine for analyzing the data and producing insights.

Keywords: crowd-sensing; information fusion; crowd-sensing analytics

1 Introduction

The goal of *(socio-)physical* or *personal analytics* [4,15,14,13] is to derive semantically rich insights about people (high-level activity, preferences, intentions) from low level measurements (e.g., location, type of activity), from their (online) social interactions, or from a combination of these. The results of such analytics could be used to improve customer engagement for businesses, provide space and event planning that accounts for the self-organising phenomena in crowds, and create higher value location-based services for users. People's behavior is influenced by their environment e.g., weather, infrastructure, air quality. For example: on a *rainy day*, one may take the bus rather than cycle. In some application scenarios, specialized information may also be useful. For example, when analyzing shopper behavior, if a *big sale* is announced, someone may reschedule her regular shopping to attend the sale. We argue here for building a much more comprehensive user context. We propose the concept of *synergistic user $\leftrightarrow$ context analytics*, illustrated in Fig. 1, as a way to promote the generalizability of an analytic initiative. Synergistic Analytics (SA) is a modular construction, consisting of the above-cited personal analytics core (based on smartphone and online

S. Loshkovska and S. Koceski (eds.), *ICT Innovations 2015*,
Advances in Intelligent Systems and Computing 399,
DOI: 10.1007/978-3-319-25733-4_17

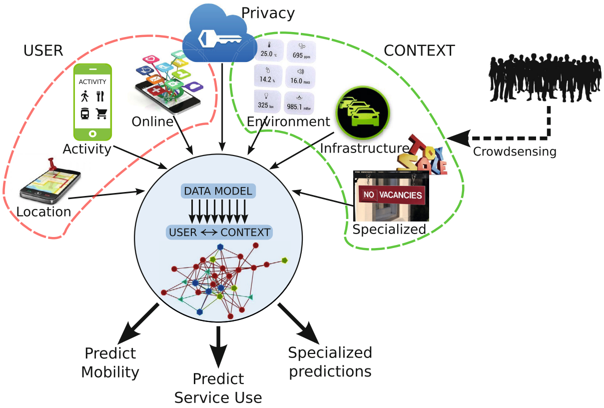

Fig. 1. Synergistic User ↔ Context Analytics

media data), enriched with extra layers of additional information, such as environmental, infrastructure-related or specialized data (e.g., retail). The results of Synergsitic Analytics will be much more than the sum of its parts: instead of isolated predictions of limited scope, deeper, semantically richer inferences are possible. Privacy-protection must be developed alongside and in full synergy with other system's components. Our platform for providing privacy-preserving, location- and context-based services to users aims to support a variety of applications, as discussed in Section 3. The scenarios for synergistic analytics underscore several scientific challenges, to be addressed by relying on the following research pillars, cf. Fig. 1: (i) location and activity prediction; (ii) context (environment, infrastructure etc) awareness via crowdsensing analytics; (iii) social profile and behavioral analytics; and (iv) privacy-preservation methods for each of the above. We present our ongoing work on three main aspects of the platform: a testbed with two units that we aim at integrating (a crowdsensing unit with smartphones and an Internet of Things (IoT) unit with sensors/actuators) in Section 4.1; a data model and storage solution, for efficiently representing and processing the highly heterogeneous information collected from the smartphones and from the sensors in Section 4.2; and a predictive analytics engine in Section 4.3.

2 Related Work

(Socio-)physical analytics. [14] presents a system that can integrate mobile sensing data with data from online social networks, to provide insights into user mobility and their interactions (both online and physical). *SocialFusion* [4] focuses on the immediate context of individuals, rather than on their interactions. In [13] a personal analytics engine generates high-level user states (e.g., emotions, preferences), which can be used to intervene in user actions. In [15] authors recognize the importance of a more comprehensive user context.

Mobile crowdsensing. A user's context consists of many variables: immediate (e.g., location), personal (e.g., activity, heart rate) and of a broader nature (e.g., weather, pollution). Traditionally, these variables are measured via standalone specialized sensors. Through *mobile crowdsensing (MCS)* [8], smart mobile devices can be used to infer or measure the above variables. This is achieved via the devices' sensing capabilities. While this solves the sensing problem, it creates new challenges: resource limitations (energy, bandwidth, computation), privacy and security issues, and the lack of a unifying architecture. The latter is important not only for making the best use of sensor data across applications and devices, but also for guaranteeing privacy and security. A common framework will allow seamless integration of both sensory and security information. We already defined the fundamental components of a mobile crowdsensing system – *crowd*, *server*, *task*) – and their interactions [2], and we have identified incentives for engaging the *crowd*. Our proposed synergistic analytics platform will tackle the additionally raised issues by integrating and jointly analyzing data from different MCS sensors to extract comprehensive patterns and predictions about user behavior and/or their context.

Privacy and security. Mobile crowdsensing (including location and activity sensing) raises many privacy and security concerns. The *crowd* provides sensed data to a *server*, which may or may not be trusted. If the server is not trusted, computation must be performed on encrypted data, which can be achieved via homomorphic encryption [10] or through secure multi-party computation [11]. Even if the server is trusted, private information may still leak, e.g., when a third party constructs clever queries that, if answered truthfully, cause the server to divulge private information. A characterisation of resistance to this is given by the concept of differential privacy [6]. These issues have not yet been addressed in the context of mobile crowdsensing, and it is our goal to design efficient algorithms, fitted for these cases. Our platform integrates privacy and security seamlessly, by embedding privacy and security guarantees within the graph that describes the relations between measured and inferred variables. For privacy, a simple solution is to utilise Bayesian posterior sampling for message passing [5], which allows us to trade off communication costs with privacy and accuracy.

3 Synergistic Analytics Use Cases

Highly Personalized Navigation. Current navigation applications are typically limited to a few transportation modes and miss complex context and user related data. Exploiting data on user preferences, transportation modes, and the environment, allows a more effective user-oriented navigation and recommender system. The data may include real-time traffic data, public transportation, rental vehicles, air quality, weather conditions, safety ratings and user habits. The system suggests places to visit, transportation modes, as well as important traffic and environmental data to city officials. Users benefit by improved social interactions, handling mobility more sustainably and efficiently. Security and privacy issues may arise, such as untruthful users and non-trusted local infrastructure.

User-Optimized Coupon Dispensing. An empirical study [19] found that *proximity* drives coupon redemption. It considered the behavior of people, while moving into the proximity of Subway restaurants: the authors showed that the distance to a restaurant is inversely proportional to the amount of monetary incentive needed to prompt people to redeem the restaurant coupons. However, the physical distance to a shop is not the only driving factor for an optimized coupon distribution. In fact, a better insight into potential customers' profile would allow a more effective dispensing. Along with proximity, other user-related information may be important driving factors, for example: *personal preferences* (i.e., a user who likes Italian food is most likely to visit nearby Italian restaurants) and *social network* (i.e., a user tends to go where their friends have already been). Consequently, a coupon distribution service could optimize the process of customer selection and coupon distribution, by exploiting the our synergistic platform, for retrieving location- and context-related user information.

Recommendation systems. Synergistic analytics could be exploited to make recommendations to users, according to their location and social profile similarity. For example, in a crowded touristic city, the dissemination of localized recommendations (i.e., interesting events and places in the city) among users would be more effective than static provider-based data distribution, in terms of both resource usage (downlink) and time for the recommendations to reach the target users [9]. Such an environment is usually populated by people with various social profiles and interests. The availability of rich information about users may improve the dissemination of localized recommendations by identifying the people and/or communities with similar profiles and interests.

4 Synergistic Analytics: Early Experiences

We present our efforts on three main aspects of the proposed privacy-preserving location- and context-based platform: a testbed with two units that we aim at integrating; a data model and storage solution, for efficiently representing and processing the highly heterogeneous information collected from the smartphones and from the sensors; and a predictive analytics engine.

4.1 Data Collection

For our generic platform for location- and context-based services, we need access to real(istic) data and to be able to easily develop, deploy and debug software on real(istic) end devices. We are building VIVO, a novel human- and sensor-based testbed with volunteers.

The VIVO volunteer testbed The VIVO testbed is based on the concept of *enrolled crowdsourcing*, which allows the deployment of several experiments, as opposed to the traditional usage of crowd-sourcing for a single experiment. VIVO provides a secure and privacy-respecting platform for *testbed users* to

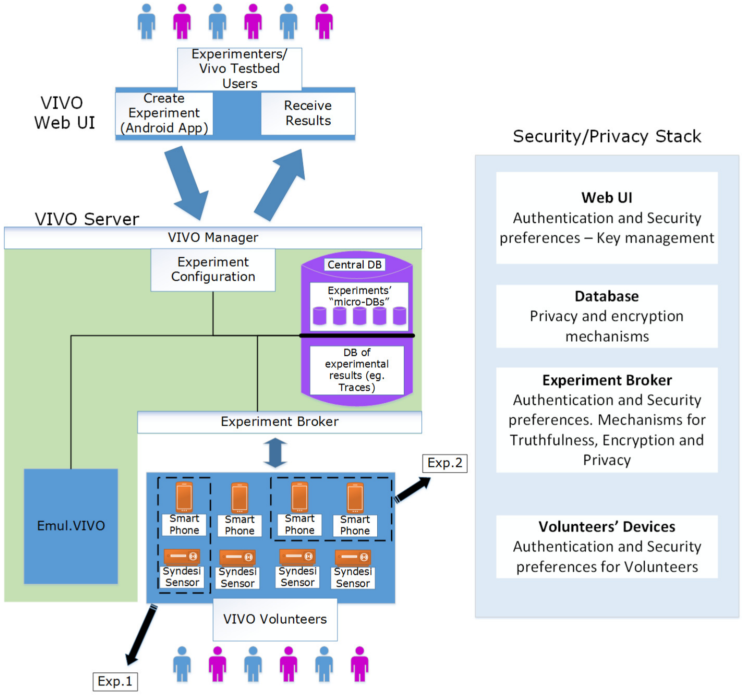

Fig. 2. VIVO Testbed Architecture

collect social, physical and environmental information. The information can be accessed remotely, as in traditional testbeds. However, VIVO differs from traditional testbeds in that it allows testing algorithms and solutions by scheduling and running them *in real time* on real mobile phones of people participating in the testbed (also called *volunteers*, not to be confused with *users*). Further, VIVO also provides an emulation environment for *users* to run and test experiments on already existing data, stored in the VIVO database. Unlike LiveLab [3] and SmartLab [12] (where a single specific and static application is installed on each smartphone to constantly save the data collected from the sensors), VIVO aims to offer more flexibility. More precisely, VIVO *testbed users* can dynamically deploy their own application on each *volunteer*'s device, as in PhoneLab[16]. However, while PhoneLab requires *volunteers* to run a modified version of the Android OS on their mobile (thus limiting the set of potential participants), VIVO applications run on normal Android versions without extra hardware requirements. In addition, VIVO promotes reproducibility of the experiments via its emulation environment. Fig. 2 depicts the VIVO architecture. At the top level, experimenters and researchers are provided with a **Web User Interface** for access. They can define new experiments, upload the corresponding source code and parameterise them; e.g. define the number of volunteers to be engaged or the environment in which the experiment will be conducted (indoor, outdoor, in a smart building, etc). At this layer the front-end management of users' authentication takes place and corresponding security preferences are defined. The main back-end platform noted as the **VIVO Server** lies below the Web user interface. It consists of the following elements:

1. The *VIVOManager* handles requests from the testbed users and, based on their preferences, forwards experiments to be run either on real devices or on an emulation environment provided by the *EmulVIVO* component. Once an experiment has been completed, it sends the results to the testbed user conducting the experiment in a secured and anonymised way. This component also performs the back-end management of the actual identification keys as well as the authentication and security preferences.
2. The central database of the system constitutes the anchor point via which the other components are able to exchange data. Here, for each defined experiment the corresponding data structures are maintained. Collected data are then provided to the experimenter and are also available for "a posteriori" analysis; e.g. to be stripped from potentially sensitive information and to be stored in a repository for future reference. The database will also be equipped with mechanisms enforcing privacy and handling encrypted data.
3. The *Experiment Broker* provisions and orchestrates the experiments to be conducted by using devices provided by *VIVO Volunteers*. This component also takes care of aspects such as the time scheduling of the experiments as well as load balancing issues among the available volunteers. While the experiments are running, data collected from *VIVO Volunteers* is stored in the corresponding micro-DB of each experiment. At this layer authentication and security issues related to the *VIVO Volunteers* are addressed. Also, mechanisms regarding truthfulness, encryption and privacy are implemented, thus mitigating such issues from the side of the volunteers.
4. *EmulVIVO* offers an environment to run experiments on existing data, available in the *EmulVIVO-DB*. The reasoning component of this module is the *EmulVIVO Manager*, which is in charge of receiving requests from *VIVO-Manager*, retrieving the corresponding data from the *EmulVIVO-DB* and allocating the emulation-running environment for the requested experiments.

The final layer includes the **VIVO Volunteers**, who are people equipped with smartphones or other personal devices able to run experiments and who accept to run VIVO experiments. Volunteers provide their characteristics (e.g. socioeconomic profile) and also define their availability. The experiments proposed by the VIVO platform must first be checked and validated (during an alpha testing phase) in terms of respecting privacy and trust issues. Also, authentication and security mechanisms are incorporated in the experiments' source code.

The Crowd-augmented Experimenting Facility. Syndesi 2.0 [1] based on [7] is an IoT testbed architecture for smart buildings, which enables the seamless and scalable integration of crowdsourced resources, provided by the end-users of the facility. The end-users of the smart building are equivalent to VIVO *volunteers*, and different from *testbed users*. End-users are not necessarily VIVO volunteers and vice versa. This integration increases the awareness of the facility both in terms of sensory capabilities as well as in terms of end-user preferences and experienced comfort. The purpose of integrating crowdsourced resources, such as smartphones and tablets, is two-fold. First, the sensory capabilities of the resources provided by the crowd are combined with those of the building for smart actuations. Second, the system is able to interact directly with end-users, both to incentivize them to provide sensory data from their devices and to receive feedback. functionalities will be abstracted to the experimenters as services via RESTful APIs, thus enabling their usage in the context of webservices. Given the testbed APIs, an experimenter canto use them while being agnostic of the technical details. Such architectures, in which testbed functionalities are exposed as services, have led to the notion of Testbed as a Service (TBaaS). Thanks to its modular architecture, Syndesi 2.0 can be integrated into the VIVO testbed presented above. All testbed resources of Syndesi 2.0, along with the accompanying mechanisms (e.g. defining the incentivizing strategies towards the end-users) are exposed as services via RESTful APIs. These services can be consumed by the VIVO testbed, thus leading to the integration of VIVO and Syndesi.

VIVO Privacy and Security. The security issues facing the VIVO testbed (including the IoT unit) can be defined by specifying different trust models. First, we can assume that the user trusts the application, but may not trust the central VIVO Server. The user definitely does not trust the intervening network. The server, on the other hand, cannot be sure that the application (or users) are providing truthful information. Many security components are available to make sure that the system is functioning properly. Mechanism design can be used to give incentives to users to provide truthful information. Differentially private statistical models can be used to optimally trade off user privacy requirements with utility of the barometric service, in a task-dependent manner. Finally, cryptographic methods can be used for secure communication between the server and the users. The particular provisioning of the VIVO testbed for trust and privacy preserving issues along with the capability of supporting a heterogeneous set of information will enable the facility to be used in more diverse experiments, by a higher number of end-users, e.g., for monitoring and collecting data on environmental conditions in out-door settings (via sensors for ambient noise and luminance levels, pressure, etc.) and their correlation to user preferences. The extracted data can then be utilized in order to emulate and study more populous crowds in the EmulVIVO running environment.

4.2 Tackling the Heterogeneous Data Challenge

In addition to the challenges of collecting and unifying the data, our proposed platform also needs an appropriate data model that allows easy and efficient querying, processing and analytics. Efficiently storing, processing and analyzing continuous streams of heterogeneous and dynamic data is a complex task [4,13,14]. The goal of analytics is to identify and exploit relationships in data. A graph-based model is the natural data model choice, as widely recognized (e.g.

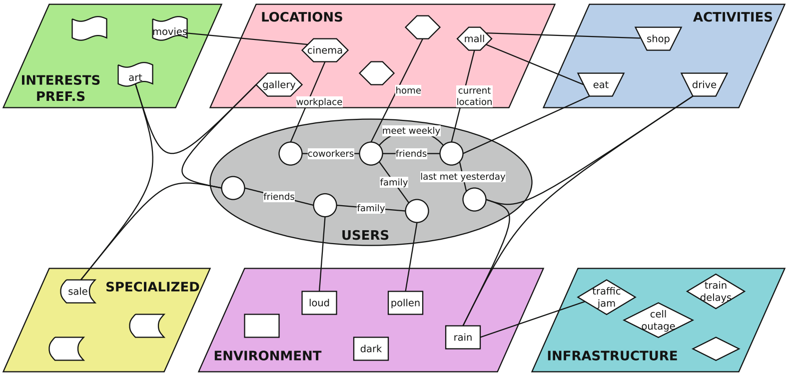

Fig. 3. Graph model for synergistic user ↔ context analytics

Google's knowledge graph, Facebook's social graph and Twitter's interest graph) Other growing commercial uses include cloud management, bioinformatics, content management, and security and access control.

In the case of Synrgistic Analytics, we are dealing with multiple node types (users, locations, activities etc.) and multiple link types ("knows", "is interested in", "is currently at" etc.). In addition, both nodes and links may have attributes, such as demographic information for users, usage for locations or statistical information for links. Finally, while graphs normally only support edges between two nodes, it would be clearly beneficial to be able to represent links among several nodes, forming hypergraphs. For example, as shown in Fig. 3, an *interest* in art is connected both to the interested *user* and to a *gallery*. Storing this type of information in an efficient, but easy to handle manner is challenging. The two main options are: (i) the new generation (hyper)graph databases and the RDF (Resource Description Framework) databases. Choosing between the two (or additional options) will highly depend on the type of processing to be done on the graph, which we discuss in the next section.

4.3 Prediction Tasks

The prediction engine of our synergistic analytics platform enables different types of predictions, such a user mobility, behavior or service use predictions, as shown in Fig. 1. This engine uses social and physical data, environmental and infrastructure information, and application-specific data to predict the users' next place and behavior, the users' service usage, as well as any required application-specific predictions. Social data mainly consists of the user's social profile (e.g., habits, interests, etc.) and social contacts/activity information. Physical data contains the user's mobility history, activity, sensed data from different embedded sensors in mobile phones and physical contacts with other people. Environmental and infrastructure information may include ambient noise levels, ambient luminance, barometric pressure, public transportation schedules, road traffic data etc.

Finally, the application-specific data (e.g., shopping behavior) should be provided by the contracting entity (e.g. retailer, hotel owner).

The heterogeneity of the collected data gives high potential to the prediction engine, which is then able to perform a deeper analysis of the user and context related data. In terms of mobility, it predicts the user's next-visited physical location together with its semantic meaning (i.e., where the user is willing to go), and it predicts the user's next physical contact. The behavioral prediction includes user activity (i.e., what the user is willing to do), mood (i.e., how the user feels), social contacts, and activity (i.e., who the user is willing to meet).

The prediction methodology is based on both historical and current data. The historical data is analyzed to create a user-dynamic mobility and behavioral model. This allows a user characterization in terms of mobility aspects (i.e., more active or sedentary persons) and the identification of the locations that are relevant for both the user itself and the social community he belongs to (according to users' social profile similarities). The model dynamically adequate to changes in the user mobility and behavior. The current data allows adaptivity to the current user's context, providing so more accurate predictions.

The potentials of including social aspects to location prediction was confirmed in some preliminary study: we showed that with the analysis of the user's mobility history we can classify the visited locations according to their relevance to the user. This classification is then used to retrieve the user's mobility and behavioral *characteristics*. Even this simple information about the user profile already improves the next-visited location prediction [18,17]. The synergistic platform will further combine our initial results with personality and social behavior information to improve the location prediction.

5 Conclusions

We introduced synergistic user $\leftrightarrow$ context analytics, a concept extending recent proposals for (socio-)physical or personal analytics by including more comprehensive data sources. We argued that, in addition to smartphone sensors and (online) social interactions, the environment and application-specific information is valuable for gaining insights into interactions between users and their context. We presented a testbed, based on mobile crowdsensing and the IoT, a data model for representing the different sources of data and their connections, and a prediction engine for analyzing the data and producing the insights.

Acknowledgments. We thank Steven Mudda and Alan Ferrari for their contributions. This work is supported by the Swiss National Science Foundation via the SwissSenseSynergy project and by the COST Action IC1303.

References

1. Angelopoulos, C.M., Evangelatos, O., Nikoletseas, S.E., Raptis, T.P., Rolim, J.D.P., Veroutis, K.: A user-enabled testbed architecture with mobile crowdsensing support for smart, green buildings. In: IEEE ICC 2015, London, U.K., 2015.

2. Angelopoulos, C.M., Nikoletseas, S.E., Raptis, T.P., Rolim, J.D.P.: Characteristic utilities, join policies and efficient incentives in mobile crowdsensing systems. In: IFIP Wireless Days, Rio de Janeiro, Brazil, 2014. pp. 1–6
3. Balan, R.K., Misra, A., Lee, Y.: Livelabs: Building an in-situ real-time mobile experimentation testbed. In: Proceedings of the 15th Workshop on Mobile Computing Systems and Applications. p. 14. ACM (2014)
4. Beach, A., Gartrell, M., Xing, X., Han, R., Lv, Q., Mishra, S., Seada, K.: Fusing mobile, sensor, and social data to fully enable context-aware computing. In: Proceedings of the Eleventh Workshop on Mobile Computing Systems & Applications. pp. 60–65. ACM (2010)
5. Dimitrakakis, C., Nelson, B., Mitrokotsa, A., Rubinstein, B.: Robust and private Bayesian inference. In: Algorithmic Learning Theory (2014)
6. Dwork, C.: Differential privacy. In: Encyclopedia of Cryptography and Security, pp. 338–340. Springer (2011)
7. Evangelatos, O., Samarasinghe, K., Rolim, J.: Syndesi: A framework for creating personalized smart environments using wireless sensor networks. In: DCOSS. pp. 325–330 (2013)
8. Ganti, R.K., Ye, F., Lei, H.: Mobile crowdsensing: current state and future challenges. IEEE Communications Magazine 49, 32–39 (2011)
9. Garg, K., Giordano, S.: Towards developing a generalized modeling framework for data dissemination. EWSN 2015 p. 9 (2015)
10. Gentry, C.: A fully homomorphic encryption scheme. Ph.D. thesis, Stanford University (2009)
11. Goldreich, O.: Secure multi-party computation (1998)
12. Larkou, G., Mintzis, M., Taranto, S., Konstantinidis, A., Andreou, P.G., Zeinalipour-Yazti, D.: Demonstration abstract: Sensor mockup experiments with smartlab. In: IPSN-14. pp. 339–340. IEEE (2014)
13. Lee, Y., Balan, R.K.: The case for human-centric personal analytics. In: ACM MobiSys WPA 2014
14. Misra, A., Jayarajah, K., Nayak, S., Prasetyo, P.K., Lim, E.p.: Socio-physical analytics: Challenges & opportunities. In: Proceedings of the 2014 workshop on physical analytics. pp. 19–24. ACM (2014)
15. Nandakumar, R., Rallapalli, S., Chintalapudi, K., Padmanabhan, V.N., Qiu, L., Ganesan, A., Guha, S., Aggarwal, D., Goenka, A.: Physical analytics: A new frontier for (indoor) location research. Tech. Rep. MSR-TR-2013-107 (2013)
16. Nandugudi, A., Maiti, A., Ki, T., Bulut, F., Demirbas, M., Kosar, T., Qiao, C., Ko, S.Y., Challen, G.: Phonelab: A large programmable smartphone testbed. In: First International Workshop on Sensing and Big Data Mining. pp. 1–6. ACM (2013)
17. Papandrea, M., Giordano, S.: Location prediction and mobility modelling for enhanced localization solution. Journal of Ambient Intelligence and Humanized Computing 5(3), 279–295 (2014)
18. Papandrea, M., Zignani, M., Gaito, S., Giordano, S., Rossi, G.P.: How many places do you visit a day? In: PERCOM Workshops. pp. 218–223. IEEE (2013)
19. Spiekermann, S., Rothensee, M., Klafft, M.: Street marketing: how proximity and context drive coupon redemption. Journal of Consumer Marketing 28(4), 280–289 (2011)