Breadcrumbs: A Rich Mobility Dataset with Point-of-Interest Annotations

Arielle Moro
University of Lausanne, Switzerland
Arielle.Moro@unil.ch

Vaibhav Kulkarni
University of Lausanne, Switzerland
Vaibhav.Kulkarni@unil.ch

Pierre-Adrien Ghiringhelli
University of Lausanne, Switzerland
Pierre-Adrien.Ghiringhelli@unil.ch

Bertil Chapuis
University of Lausanne, Switzerland
Bertil.Chapuis@unil.ch

Kévin Huguenin
University of Lausanne, Switzerland
Kevin.Huguenin@unil.ch

Benoît Garbinato
University of Lausanne, Switzerland
Benoit.Garbinato@unil.ch

ABSTRACT

Rich human mobility datasets are fundamental for evaluating algorithms pertaining to geographic information systems. Unfortunately, existing mobility datasets–that are available to the research community–are restricted to location data captured through a single sensor (typically GPS) and have a low spatiotemporal granularity. They also lack ground-truth data regarding points of interest and the associated semantic labels (e.g., "home", "work", etc.). In this paper, we present *Breadcrumbs*, a rich mobility dataset collected from multiple sensors (incl. GPS, GSM, WiFi, Bluetooth) on the smartphones of 81 individuals. In addition to sensor data, Breadcrumbs contains ground-truth data regarding people points of interest (incl. semantic labels) as well as demographic attributes, contact records, calendar events, lifestyle information, and social relationship labels between the participants of the study. We describe the data collection methodology and present a preliminary quantitative analysis of the dataset. A sanitized version of the dataset as well as the source code will be made available to the research community.

CCS CONCEPTS

• **Information systems** Spatial-temporal systems.

KEYWORDS

mobility dataset; point of interest annotations

ACM Reference Format:
Arielle Moro, Vaibhav Kulkarni, Pierre-Adrien Ghiringhelli, Bertil Chapuis, Kévin Huguenin, and Benoît Garbinato. 2019. Breadcrumbs: A Rich Mobility Dataset with Point-of-Interest Annotations. In *27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (SIGSPATIAL '19), November 5–8, 2019, Chicago, IL, USA*. ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3347146.3359341

1 INTRODUCTION

Modeling human mobility is gaining importance as cities are experiencing growth and rapid transformations; this modeling demands a good understanding of individual mobility behaviors. Therefore, rich mobility datasets are fundamental for designing and evaluating algorithms pertaining to human-related geographic information systems (GIS) and for facilitating experimental reproducibility. Their availability have spurred different complex problems around the mobility domain, such as predictive queries [9], object tracking [21], trajectory indexing [4], mobility modeling [1], and location privacy [19].

As detailed in Table 1, many mobility datasets have already been made available to the research community (e.g., [13, 16, 17, 22, 23, 25]). Unfortunately, these datasets have several limitations, which include: (1) the lack of location data and related information captured from multiple sensors; (2) the unavailability of location data at a high spatiotemporal granularity throughout the data collection; (3) the lack of ground-truth information regarding participant points of interests (POI); (4) the unavailability of semantic information regarding POIs. For example, despite the proliferation of smartphones equipped with multiple sensors, datasets such as [17, 23, 24] are restricted to location data derived from either GPS, GSM, WiFi or Bluetooth. Gaining access to high granularity multi-sensor location data can lead to richer comparative and compositional studies [16]. Another example relates to the lack of ground-truth and semantic information in existing datasets. This information is crucial for research domains such as social network pattern mining [7, 10], as it is the only credible way to validate certain semantical results.

In this paper, we introduce Breadcrumbs, a rich mobility dataset that contains high-granularity data from GPS, WiFi, Bluetooth and accelerometer sensors from 81 individuals in Lausanne (Switzerland) for a period of 12 weeks that spanned between March and June 2018. This novel dataset addresses the limitations of the aforementioned datasets: it is enriched with POIs ground-truth annotations (incl. semantic labels), demographic attributes, social relationships, health information, mobility information, calendar events and contact records. This information is especially important given that, in the last decade, there has been an increasing demand to understand the behavior of individuals in multiple domains [15]. In the following sections, we describe the data collection methodology and present a preliminary quantitative analysis of the dataset. A sanitized version of the dataset and the source code will be made available to the research community at https://bread-crumb.github.io.

2 DATA COLLECTION METHODOLOGY

In order to build the Breadcrumbs dataset, we organized a data collection campaign in Lausanne in the spring of 2018. We recruited participants through a specialized unit called Labex at the University of Lausanne, which manages a pool of around 8,000 individuals

SIGSPATIAL '19, November 5–8, 2019, Chicago, IL, USA

ACM ISBN 978-1-4503-6909-1/19/11.
https://doi.org/10.1145/3347146.3359341

Dataset	Collection / Publication	#Participants	Duration	#Events	Sampling	Location	GPS	Check-ins	GSM	Wifi	Bluetooth	Annotation
GeoLife (Zheng et al. [25])	2007-2012 / 2012	182	5.5 years	25M	5 sec	Beijing, CN	✓	✗	✗	✗	✗	✗
MDC (Kiukkonen et al. [13])	2009-2011 / 2012	185	3 years	11M	-	Lausanne, CH	✓	✗	✓	✓	✓	relationships
Privamov (Mokhtar et al. [16])	2014-2016 / 2017	100	15 months	15M	-	Lyon, FR	✓	✗	✓	✓	✗	✗
Reality Mining (Pentland [17])	2004 / 2009	100	9 months	5M	-	Boston, US	✗	✗	✗	✗	✓	relationships
FourSquare (Yang et al. [23])	2011-2012 / 2013	3112	10 months	9M	-	New York, US	✗	✓	✗	✗	✗	relationships
blebeacon (Sikeridis et al. [20])	2016 / 2018	46	1 month	5M	-	California, US	✗	✗	✗	✗	✓	✗
hyccups (Ciobanu and Dobre [8])	2012 / 2016	72	63 days	-	-	Bucharest, RO	✗	✗	✗	✓	✗	relationships
sigcomm2009 (Pietilainen and Diot [18])	2009 / 2012	76	2 days	-	120 sec	Barcelona, ES	✗	✗	✗	✓	✓	✗
telefonica (Bogomolov et al. [3])	2013 / 2014	342	4 weeks	-	-	ES	✗	✗	✓	✗	✗	✗
ParticipAct (Chessa et al. [6])	2013-2015 / 2017	300	1 year	-	-	Bologna, IT	✓	✗	✗	✓	✓	✗
Nodobo (Bell et al. [2])	? / 2011	27	4 months	5M	-	Glasgow, GB	✗	✗	✓	✓	✗	✗
d4d challenge (Furletti et al. [11])	2016 / 2016	9M	1 year	-	-	SN	✗	✗	✓	✗	✗	✗
Gowalla (Cho et al. [7])	2008-2010 / 2011	196,591	1.5 years	6M	-	Worldwide	✗	✓	✗	✗	✗	relationships
Brightkite (Chessa et al. [6])	2008-2010 / 2010	58,228	1.5 years	4M	-	Worldwide	✗	✓	✗	✗	✗	relationships
Breadcrumbs	2018 / 2019	81	12 weeks	14M	50 sec	Lausanne, CH	✓	✗	✗	✓	✓	ground-truth semantic labels relationships

Table 1: Comparative summary of popular mobility datasets available to the community (: GPS / : Check-ins / : GSM / : Wifi / : Bluetooth).

(mostly students) who registered for behavioral experiments. We contacted them by e-mail; those who were interested had to fill a short questionnaire (i.e., a screener) in order to verify their eligibility for the experiment. The main criterion was to have an iPhone with a recent version of iOS (≥11.2.6) and to use it as their main phone. Eligible participants had to sign a consent form. Then, they had to install a mobile application (developed by us) on their smartphones and to keep it installed and running during the whole experiment.

The system architecture for collecting the data is presented in Figure 1. The sampling (periodic vs. motion-based) and upload (e.g., GSM vs. WiFi) strategies were carefully calibrated so that the impact on the battery life was acceptable, i.e., the battery life of the phone should be at least one day for a normal usage in the best case scenario with a recent model of iPhone. We put in place a number of mechanisms (e.g., backup, replication, notifications) to ensure a reliable and steady collection of data. The mobile application collected data from various sensors: GPS location, WiFi scans (i.e., neighboring SSIDs) and Bluetooth scans (i.e., neighboring UUIDs), and acceleration. The collected data was pre-processed directly on smartphones, for privacy reasons, and then uploaded to our backend where it was stored in a persistent database (see Figure 2 for the complete schema).

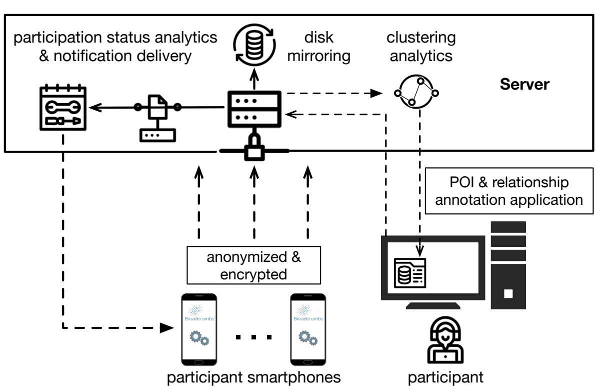


Figure 1: System architecture of the Breadcrumbs data collection.

In the middle of the experiment, we sent a questionnaire to each participant of the study in order to collect demographic (gender, age, etc.) and lifestyle (sport activities, smocking habits, transportation mode preferences, etc.) information. At the end of the experiment, participants had to fill an exit questionnaire in order for us to collect ground-truth data regarding their POIs (incl. semantic labels) and relationship information (e.g., friendship with other participants). To collect the ground-truth, we first extracted points of interest from their full mobility traces (i.e., over the whole experiment). We tested and compared four different clustering algorithms based on the MDC [13] dataset (same region as Breadcrumbs) and on the Geolife [25] dataset: (1) DJ Cluster [26], (2) DT Cluster [5], (3) TD Cluster [12] and (4) Capstone [14], which operates without parameters. Our selection criteria included the number of returned POIs, the minimum distance between distinct POIs, and the number of parameters. We selected DT Cluster [5] and further processed the returned POIs by merging overlapping POIs (a POI consists of a point on the map and a radius) and removing those that the participants visited less than 3 times over the course of the whole experiment. Each participant was shown the POIs resulting from the analysis of her/his mobility trace, then had to validate/invalidate each of them and to annotate each valid one with a semantic label. The set of possible labels was predefined; it contained the following nine categories: transport, study, residency, work, sustenance, shopping, sports, leisure and other (free-text).

The participants were compensated for their participation with CHF 100 (∼USD 100) in cash, which they received at the very end of the experiment. The experiment was approved by the ethical committee of our institution.

3 QUANTITATIVE ANALYSIS

In this section, we report on our preliminary quantitative analysis of the Breadcrumbs dataset and present the different feature sets, alongside with the associated descriptive statistics. The Breadcrumbs dataset contains 34,080,964 records of GPS, WiFi and Bluetooth data points. The aggregate distance travelled by the participants amounts to 548,210 km, and the average distance travelled per participant is 6768 ± 4336 km. We collected the geospatial coordinates at an average of 79 ± 36 points per hour for each participant. The WiFi scans amount to 105 ± 49 SSIDs per hour per participant and the Bluetooth scans result in 7 ± 12 device UUIDs per hour for each participant. Additionally, each participant had an average of 280 ± 183 unique contacts in their contact list.

Table 2 shows the total number of records collected by the different sensors as well as the minimum, the median, the average, the standard deviation and the maximum of records per user. The

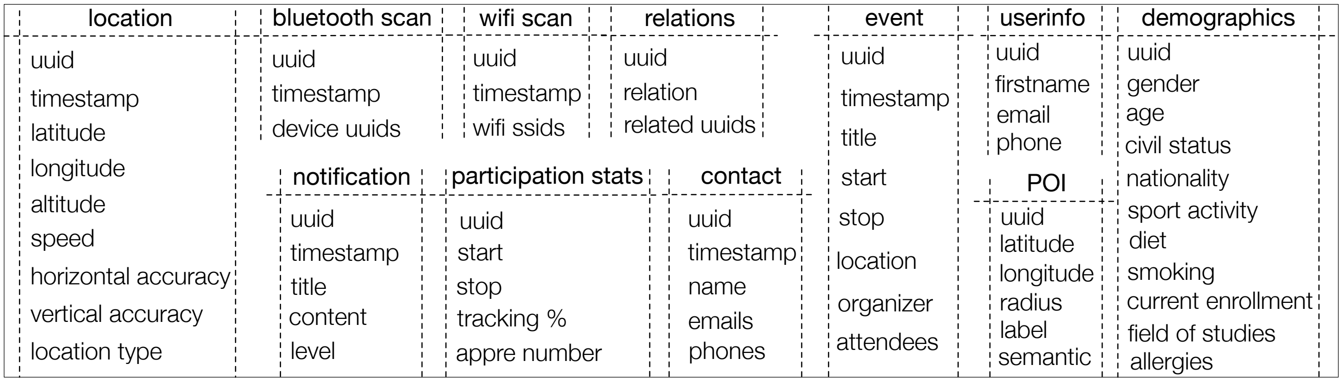
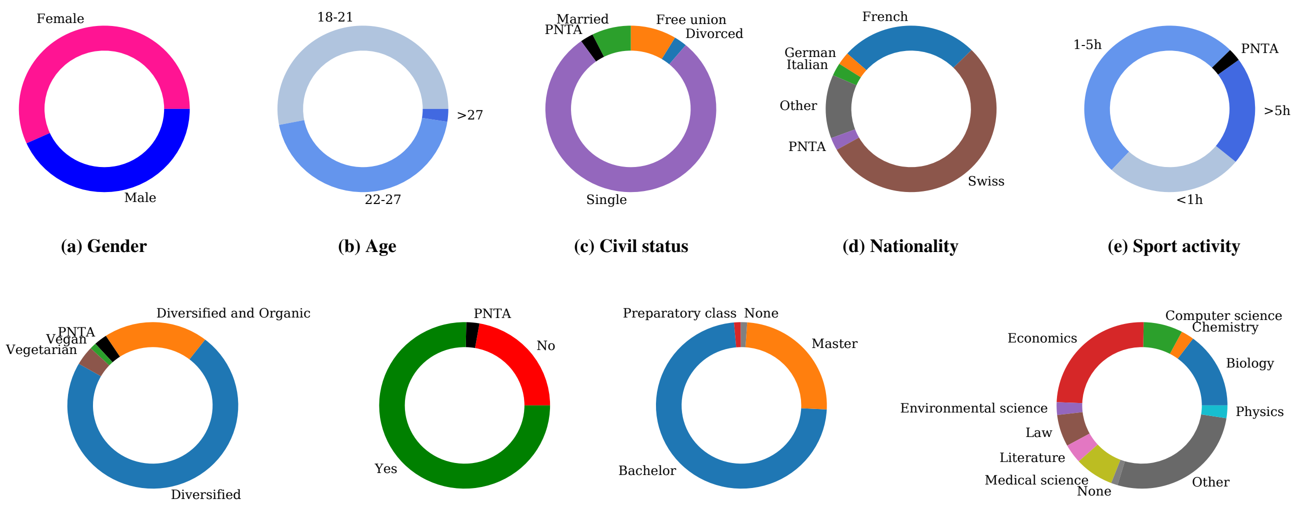


Figure 2: Database schema of the Breadcrumbs dataset.



(a) Gender **(b) Age** **(c) Civil status** **(d) Nationality** **(e) Sport activity**

(f) Diet **(g) Non-smoking** **(h) Current enrollment** **(i) Field of studies**

Figure 3: Demographics of the Breadcrumbs dataset (PNTA means "prefer not to answer").

summary of the GPS location data is presented in Table 3. The horizontal and the vertical accuracy is reported by the *Core Location API* provided by Apple.

Regarding the demographics, 56.79% of the participants identified as females, as shown in Figure 3a. The largest age groups present in the campaign are 18-21 and 22-27, with 53.09% and 44.44% respectively, as depicted in Figure 3b. In Figure 3c, the most represented civil status group is the "Single" category, i.e., 79.01%. The two most important nationality groups are "Swiss" and "French", 54.32% and 25.93% respectively, as indicated in Figure 3d. In terms of sport activities, 25.93% of the participants do sport exercises less than one hour per week, 50.62% between one and five hours per week and 20.99% more than 5 hours (see Figure 3e). Figure 3f and Figure 3g show that 72.84% of the participants have a diversified diet and 75.31% are not smoking. Figure 3h indicates that 72.84% participants were enrolled in a bachelor's degree program and 24.69% in a master's degree program. Finally, we observe that most of the participants are studying economics and biology, 24.69% and 14.81% respectively, as seen in Figure 3i.

Type	#Records	Min/usr	Median/usr	Avg./usr	STD/usr	Max/usr
GPS	13,903,934	22,418	168,050	171,654	7820	469,298
WiFi	18,669,063	15,888	234,550	230,482	107,482	426,885
Bluetooth	51,424	0	93	704	1063	5803
Accelerometer	11,661,738	17,759	131,177	143,972	71,364	415,666

Table 2: Number of data points and ratio per user.

Variable	Q05	Median	Avg.	STD	Q95
Longitude	3.962	6.589	6.618	4.509	8.465
Latitude	44.040	46.520	46.238	1.997	47.407
Altitude	64.583	415.500	465.858	557.575	753.903
Speed	0.001	9.690	13.455	16.965	35.390
Horizontal accuracy	5.000	12.000	70.792	1210.320	200.000
Vertical accuracy	3.000	6.000	14.842	111.470	29.714

Table 3: Descriptive statistics of the GPS data points.

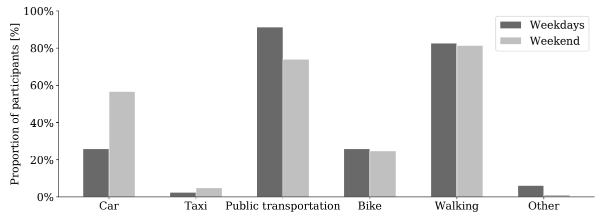


Figure 4: Transportation mode preferences for weekdays and weekend.

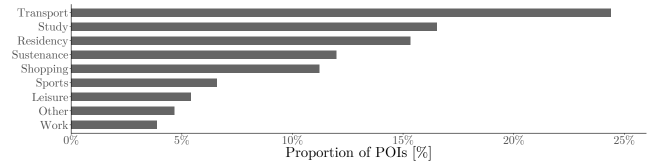


Figure 5: Distribution of POIs according to their semantic labels.

Figure 4 shows the transportation modes utilized during weekdays and weekend by the participants. We observe an increase in the usage of private transportation modes (cars) during the weekend as compared to the weekdays. However, walking and biking habits look similar during the weekdays and the weekend. As shown in Figure 5, the majority of the POIs correspond to the transport, study and residency semantic labels (top-level categories).

4 CONCLUSION

In this paper, we have introduced Breadcrumbs, a rich mobility dataset. In addition to demographic attributes, contacts, calendar records and social relationships, we have provided the semantic labels and the ground-truth for the points of interest. We have described the complete data-collection process and our methodology to collect ground-truth information. Our qualitative analysis sheds light on several aspects of this dataset, including the POI distribution. A sanitized version of the dataset as well as the source code will be made available to the research community at https://bread-crumb.github.io to facilitate and advance GIS research. This new dataset opens plenty of promising research avenues, such as the combination of sensor data (GPS, Wifi, Bluetooth, etc.) with demographic data, and the possibility to validate research results with a ground-truth.

ACKNOWLEDGMENTS

We thank the HEC-Labex team for their help during all the steps of the data-collection campaign. This research work was partially supported by the Business Information Systems and Architecture (BISA) research laboratory and the Faculty of Business and Economics (HEC Lausanne) at the University of Lausanne and by the Swiss National Science Foundation with grant #157160.

REFERENCES

[1] Albert-Laszlo Barabasi. 2005. The origin of bursts and heavy tails in human dynamics. *Nature* 435, 7039 (2005), 207.

[2] Stephen Bell, Alisdair McDiarmid, and James Irvine. 2011. Nodobo: Mobile phone as a software sensor for social network research. In *Proc. of VTC*.

[3] Andrey Bogomolov, Bruno Lepri, Jacopo Staiano, Nuria Oliver, Fabio Pianesi, and Alex Pentland. 2014. Once upon a crime: towards crime prediction from demographics and mobile data. In *Proc. of ICMI*.

[4] V. Prasad Chakka, Adam Everspaugh, and Jignesh M. Patel. 2003. Indexing Large Trajectory Data Sets With SETI. In *Proc. of CIDR*.

[5] Yixin Chen and Li Tu. 2007. Density-based clustering for real-time stream data. In *Proc. of KDD*.

[6] Stefano Chessa, Michele Girolami, Luca Foschini, Raffaele Ianniello, Antonio Corradi, and Paolo Bellavista. 2017. Mobile crowd sensing management with the ParticipAct living lab. *Pervasive and Mobile Computing* 38 (2017).

[7] Eunjoon Cho, Seth A. Myers, and Jure Leskovec. 2011. Friendship and mobility: user movement in location-based social networks. In *Proc. of KDD*.

[8] Radu I. Ciobanu and Ciprian Dobre. 2016. CRAWDAD dataset upb/hyccups (v. 2016-10-17). Downloaded from https://crawdad.org/upb/hyccups/20161017. https://doi.org/10.15783/C7TG7K

[9] Stephen Cronen-Townsend, Yun Zhou, and W. Bruce Croft. 2002. Predicting query performance. In *Proc. of SIGIR*.

[10] Nathan Eagle and Alex Pentland. 2005. Reality mining: sensing complex social systems. *Personal and Ubiquitous Computing* 10 (2005), 255–268.

[11] Barbara Furletti, Roberto Trasarti, Paolo Cintia, and Lorenzo Gabrielli. 2017. Discovering and understanding city events with big data: the case of rome. *Information* 8, 3 (2017), 74.

[12] Sébastien Gambs, Marc-Olivier Killijian, and Miguel Núñez del Prado Cortez. 2010. Show me how you move and I will tell you who you are. In *Proc. of SIGSPATIAL Workshop SPRINGL*.

[13] Niko Kiukkonen, Jan Blom, Olivier Dousse, Daniel Gatica-Perez, and Juha K. Laurila. 2010. Towards rich mobile phone datasets: Lausanne data collection campaign.

[14] Vaibhav Kulkarni, Arielle Moro, Bertil Chapuis, and Benoît Garbinato. 2017. Extracting Hotspots Without A-priori by Enabling Signal Processing over Geospatial Data. In *Proc. of SIGSPATIAL*.

[15] Zhenhui Li, Bolin Ding, Jiawei Han, Roland Kays, and Peter Nye. 2010. Mining periodic behaviors for moving objects. In *Proc. of KDD*.

[16] Sonia Ben Mokhtar, Antoine Boutet, Louafi Bouzouina, Patrick Bonnel, Olivier Brette, Lionel Brunie, Mathieu Cunche, Stephane D 'Alu, Vincent Primault, Patrice Raveneau, Hervé Rivano, and Razvan Stanica. 2017. PRIVA'MOV: Analysing Human Mobility Through Multi-Sensor Datasets.

[17] Alex Pentland. 2009. Reality mining of mobile communications: Toward a new deal on data. *The Global Information Technology Report 2008–2009* 1981 (2009).

[18] Anna-Kaisa Pietilainen and Christophe Diot. 2012. CRAWDAD dataset thlab/sigcomm2009 (v. 2012-07-15). Downloaded from https://crawdad.org/thlab/sigcomm2009/20120715. https://doi.org/10.15783/C70P42

[19] Vincent Primault, Antoine Boutet, Sonia Ben Mokhtar, and Lionel Brunie. 23. The Long Road to Computational Location Privacy: A Survey. *IEEE Communications Surveys & Tutorials* 21, 3 (23), 2772–2793. https://doi.org/10.1109/COMST.2018.2873950

[20] Dimitrios Sikeridis, Ioannis Papapanagiotou, and Michael Devetsikiotis. 2019. CRAWDAD dataset unm/blebeacon (v. 2019-03-12). Downloaded from https://crawdad.org/unm/blebeacon/20190312.

[21] Chieh-Chih Wang, Charles E. Thorpe, Sebastian Thrun, Martial Hebert, and Hugh F. Durrant-Whyte. 2007. Simultaneous Localization, Mapping and Moving Object Tracking. *I. J. Robotics Res.* 26 (2007).

[22] Xiao-Yong Yan, Xiao-Pu Han, Bing-Hong Wang, and Tao Zhou. 2013. Diversity of individual mobility patterns and emergence of aggregated scaling laws. *Nature Scientific reports* 3 (2013).

[23] Dingqi Yang, Daqing Zhang, Zhiyong Yu, and Zhiwen Yu. 2013. Fine-grained preference-aware location search leveraging crowdsourced digital footprints from LBSNs. In *Proc. of UbiComp*.

[24] Dingqi Yang, Daqing Zhang, Vincent W Zheng, and Zhiyong Yu. 2015. Modeling user activity preference by leveraging user spatial temporal characteristics in LBSNs. *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 45, 1 (2015).

[25] Yu Zheng, Xing Xie, and Wei-Ying Ma. 2010. GeoLife: A Collaborative Social Networking Service among User, Location and Trajectory. *IEEE Data Eng. Bull.* 33 (2010).

[26] Changqing Zhou, Dan Frankowski, Pamela Ludford, Shashi Shekhar, and Loren Terveen. 2004. Discovering personal gazetteers: an interactive clustering approach. In *Proc. of GIS Workshops*.