Computers, Environment and Urban Systems 67 (2018) 55–67

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

journal homepage: www.elsevier.com/locate/ceus

Activity patterns mining in Wi-Fi access point logs

Guilhem Poucin[a], Bilal Farooq[b,*], Zachary Patterson[c]

[a] Laboratory of Innovations in Transportation (LITrans), Département de Génies Civil, Géologique et des Mines, École Polytechnique Montréal, 2500 Ch. Polytechnique Montréal, H3T 1J4 Montréal, Canada
[b] Laboratory of Innovations in Transportation (LITrans), Department of Civil Engineering, Ryerson University, 350 Victoria Street, M5B 2K3 Toronto, Canada
[c] Transportation Research for Integrated Planning (TRIP) Laboratory, Geography, Planning and Environment Department, Concordia University, Montreal, Canada

ARTICLE INFO

Article history:
Received 21 June 2016
Received in revised form 12 September 2017
Accepted 13 September 2017
Available online xxxx

Keywords:
Information mining
Activity pattern recognition
Wi-Fi networks
Clustering
Facility usage

ABSTRACT

This article proposes a methodology to mine valuable information about the usage of a facility (e.g. building, open public spaces, etc.), based *only* on Wi-Fi network connection history. Data are collected at Concordia University in Montréal, Canada. Using the Wi-Fi access log data, we characterize activities taking place within a building without any additional knowledge of the building itself. The methodology is based on identification and generation of pertinent variables derived by Principal Component Analysis (PCA) for clustering (i.e. PCA-guided clustering) and time-space activity identification. K-means clustering algorithm is then used to identify 7 activity types associated with buildings in the context of a campus. Based on the activity clusters' centroids, a search algorithm is proposed to associate activities of the same types over multiple days. The spatial distribution of the computed activities and building plans are then compared, which shows a more than 85% match for the weekdays.

1. Introduction

Most traditional efforts to collect data on mobility and human behavior involve surveys (e.g. regional origin-destination surveys) are based on sampling a small proportion of the known population (see e.g. Ortúzar and Willumsen (2011)). Such data collection methods are expensive in terms of direct costs and time required. They may not represent actual behavior of users due to sampling biases (Richardson, Ampt, & Meyburg, 1995; Zmud, Lee-Gosselin, Munizaga, & Carrasco, 2013) and the reliability of data reported by respondents may suffer due to difficulty in recalling the activities (Frick & Grabka, 2005). Moreover, due to their predominantly cross-sectional nature, traditional mobility data are not able to observe the evolution of an individual's behavior over time. The development of free Wi-Fi networks in cities (e.g. Smart Sidewalks in the UK) and the spread of smartphones represent an opportunity to capture a larger sample of the population continuously at very low cost. While such passive collection technologies and longitudinal behavioral data represent tremendous opportunities, methodologies and tools to exploit these new sources are in their infancy.

Using data from pervasive and ubiquitous networks for mobility studies is an emerging area of research (Cao et al., 2015). Munizaga and Palma (2012), Kusakabe and Asakura (2014), Long and Thill (2015) and other recent studies have used smart-card transaction data from Automated Fare Systems (AFS) to study urban mobility patterns. Iqbal, Choudhury, Wang, and González (2014) used cellular phone network data to develop origin-destination matrices in an urban area. Meneses and Moreira (2012) used Wi-Fi network data for localization and routing on a university campus. The main challenge faced while using these datasets is the lack of information concerning users, which is caused by the common necessity of anonymizing the data. Recent studies have tried to incorporate other data sources (e.g. land use data, travel diary surveys, time tables etc.) to overcome such issues (Calabrese, Lorenzo, & Ratti, 2010; Danalet, Farooq, & Bierlaire, 2014; Grapperon, Farooq, & Trépanier, 2016). However in many cases, it is very difficult to access such data, especially at a very disaggregate level.

The Media Access Control (MAC) address is a unique identifier associated with each network interface and is used as a unique address in a Wi-Fi network. This address is fixed for a Wi-Fi enabled device and remains the same throughout the life of a device. Wi-Fi networks are composed of sets of Access Points (AP) to which a device can connect using its MAC address. APs provide Wi-Fi services, i.e. a connection to the internet. APs are spatially distributed covering large areas (e.g. campus, shopping center, etc.) and collectively comprise a Local Area Network (LAN). Our methodology only uses

* Corresponding author.
E-mail addresses: guilhem.poucin@polymtl.ca (G. Poucin), bilal.farooq@ryerson.ca (B. Farooq), zachary.patterson@concordia.ca (Z. Patterson).

http://dx.doi.org/10.1016/j.compenvurbsys.2017.09.004
0198-9715/© 2017 Published by Elsevier Ltd.

communication between devices and APs over the LAN to develop the traces of people over time and space. We advance the current state of research by proposing a PCA-guided K-means clustering to associate to each Wi-Fi AP the dominant activity performed at its location over time. The methodology is applied at a large scale in terms of space (i.e. a high-rise building) and on a very disaggregate time scale (i.e. day level), without any previous spatial knowledge of the infrastructure. To confirm the consistency of mined activities, we extend our analysis to a period of an entire week. Furthermore, to indirectly test the accuracy of our methodology, we compare our results to the designated usage of spaces in the building.

The rest of the article is structured as follows: a review of current literature on the use of ubiquitous networks, especially Wi-Fi networks is followed by a description of the case study location, and the dataset used in the analysis. This leads us to a section describing the methodology adopted and results related to the classification of access points in terms of their surrounding activities. The next section compares the activity inference results with designated usage of the space on building plans. In the end we discuss our conclusions, limitation, and possible applications.

2. Literature review

The study of pervasive systems such as cellular networks, Global Positioning Systems (GPS) or Wi-Fi networks, has received growing attention from researchers during the last decade. Indeed, the development and growth of these ubiquitous networks, combined with the improvement of data collection processes, the spread of smartphones, and the emergence of data science have opened promising perspectives towards the understanding and characterization of human behavior. This interest has resulted in applications related to network optimization, urban modeling and even transportation policy. In the following section, we inventory few of the studies analyzing the relationship between human activities and infrastructure space. We then explore the potential of network trace data (especially Wi-Fi) to study human behavior, proposing an overview of the literature from the different data collection processes, to the challenges encountered and the different methodologies proposed to analyze the data.

2.1. Relationship between human activity and space

Recent studies on the relationship between activities/behavior and spatial data have benefited from the large improvement of the datasets available. These studies show a high regularity in human trajectories (Gonzalez, Hidalgo, & Barabasi, 2008) and daily routine (Song, Qu, Blumm, & Barabási, 2010). Wang, Pedreschi, Song, Giannotti, and Barabasi (2011) study the relationship between human mobility and their social network connections.

However, privacy issues surrounding such data reduce the accuracy of the analysis–especially with respect to the socio-demographic variables driving human behavior. Some studies are based on geographic data, for instance Eagle and Pentland (2009). In parallel, other studies, such as (Jiang, Ferreira, & González, 2012), base themselves on more conventional survey data (travel diary survey), benefiting from the richness of these datasets, but limiting their large scale applicability due to cost.

2.2. Network traces

As suggested in Aschenbruck, Munjal, and Camp (2011), user traces can be acquired through three different methods: monitoring location, communications or contacts.

The monitoring of location involves collecting successive positions of a user's device and is mostly done with GPS. Using a network of 72 satellites, this technology can furnish a user's position within a few meters. In the past, Liu, Andris, and Ratti (2010) used GPS traces to study the mobility behavior of taxi drivers. Patterson and Fitzsimmons (2016) analyzed trace data collected through the smartphone travel survey application DataMobile. However, GPS data show limited application in indoor and dense urban environments, where obstacles can create a shadowing effect. This problem can be overcome by coupling the information from GSM and Wi-Fi networks (Aschenbruck et al., 2011), or in some cases with additional data sources like GTFS, as Zahabi, Ajzachi, and Patterson (in press) did to infer transit itineraries from smartphone data. However, as mentioned in J.~Su, Chin, Popivanova, Goel, and De~Lara (2004), some users can be reluctant to share the history of their positions. Since this method is device-centric, it needs a user's cooperation by accepting the burden of an additional device (e.g. GPS unit) or an energy consuming application on their own device (e.g. smartphone). We refer to this kind of location monitoring (i.e. location monitoring by a user's device) as device-centered monitoring. This is distinguished from network-centered monitoring when device information is collected passively and automatically by Wi-Fi or GSM networks (Nguyen-Vuong, Agoulmine, & Ghamri-Doudane, 2007), which we divide into two broad categories.

The first type of network-centered monitoring relies on the monitoring of communication and it uses interactions between devices and a communication system (cellular or Wi-Fi) to recreate a user's mobility history. This information is regularly collected by network operators and represents a low-cost (and low-burden on the user) source of locational information. The pervasive nature of these networks allows the capture of a large sample of the population; even if the characterization of this sample is still a challenge (Calabrese, Diao, Di~Lorenzo, Ferreira, & Ratti, 2013). The improvement of the relatively low accuracy of the locational data obtained is considered in Mao, Fidan, and Anderson (2007) and Wymeersch, Lien, and Win (2009). Usually this leads to work on symbolic spaces rather than geographic spaces, as described in Meneses and Moreira (2012). The use of signal strength can improve location accuracy and can be further improved with other information fusion processes (Aschenbruck et al., 2011). Cellular (GSM) data, which can provide locational accuracy at the size of neighborhoods in cities, are discussed in Calabrese et al. (2013). Wi-Fi data have been used at finer scales in locations such as campuses, offices or festivals. (More detail on this literature is discussed in the next section.)

The second type of network-centered monitoring of locational data is done by monitoring contact between users through technologies such as Bluetooth or Wi-Fi. Monitoring of contacts can allow the characterization of social networks existing in closed environments. This topic has received growing attention thanks to developments in multi-hop networks (Conti & Giordano, 2007). Monitoring of contacts involves unloading a part of the Wi-Fi network and making information transit through a chain of user devices. In J.~Su et al. (2004) and J.~Su, Goel, and De~Lara (2006), a sample of students are monitoring their contacts with other users within a campus to explore the feasibility of such a technology, and such an approach has been used to highlight the social behavior of the students. Another use of this process is proposed in Naini, Dousse, Thiran, and Vetterli (2011), where phone Bluetooth activity at a festival allowed the estimation of the size of the entire population of festival goers using a statistical algorithm derived from biology.

2.3. Challenges to using network-centric Wi-Fi data

While the information from a phone's connection history gives us the advantage of collecting data on a large sample of individuals at low cost, important challenges to using this data have been highlighted in the literature. The first challenge results from the need for anonymization of user data, essential to guaranteeing user