Massive Online Geo-social Networking Platforms and Urban Human Mobility Patterns: A Process Map for Data Collection

Q. Wang[1] and J. E. Taylor[2]

[1] PhD Student, Department of Civil and Environmental Engineering, Virginia Tech, 121 Patton Hall, Blacksburg, VA 24061, United States. E-mail: wangqi@vt.edu.
[2] Associate Professor, Department of Civil and Environmental Engineering, Virginia Tech, 113B Patton Hall, Blacksburg, VA 24061, United States. E-mail: jet@vt.edu.

ABSTRACT

Human mobility is central to our understanding of design, planning and development of civil infrastructure, particularly in urban areas where large scale mobility flow problems can critically depend on the interface between human mobility and infrastructure. Therefore, researchers have spent considerable effort to understand and predict human mobility patterns. Several recent studies have used geo-social networking platforms to examine human mobility, but the focus of these studies has been on small scale social networking media. In this study, we examined the possibility of using Twitter, a massive online social networking platform with over 400 million users, to collect human mobility data. We developed a process map to collect data from Twitter, and designed two Python modules for its implementation. A case study was conducted and its results confirmed that Twitter can provide a larger quantity of useful human mobility data. In future research, we plan to analyze the data and validate that it can accurately capture mobility patterns. This will provide insight into whether Twitter is a viable resource to study city-scale human mobility. It can also potentially deepen our understanding about the interaction between urban dwellers and civil infrastructure.

INTRODUCTION

Understanding human mobility and its patterns has drawn research interest from multiple fields now for decades. Researchers have found that human mobility plays vital roles in human migration, urban development, epidemic breakout, etc. (Gonzalez et al. 2008; Song et al. 2010a; Song et al. 2010b). A great deal of this research has focused on populous cities like New York City, London, and Beijing, owing to the fact that human mobility dynamics are extremely active and complex in densely populated areas (Noulas et al. 2012). Understanding and predicting the patterns of human mobility can inform urban issues and drive policy-making in these large cities, including urban planning, disaster evacuation and response, public health provisioning, etc. (Noulas et al. 2012).

The current trend of research on human mobility focuses on understanding the movement trajectories of individuals. Due to the difficulty of data collection, human

mobility was historically assumed as random movements (Viswanathan and Afanasyevt 1996; Ramos-Fernandez et al. 2004). The advent of mobile and wireless technologies has made it possible to collect a large quantity of empirical data. Based on this data, studies have proposed several human mobility models (Gonzalez et al. 2008; Song et al. 2010a; Song et al. 2010b). More recently, researchers have begun using geo-social networking platforms to study human mobility in urban areas (Cho et al. 2011; Noulas et al. 2012). These platforms, e.g. FourSquare, Gowalla, Brightkite, allow users to share their locations by checking-in. They not only provide more precise geographical information of a user's location, but also provide opportunities to understand how geography and social networks are interrelated.

Research to date has been limited to the aforementioned platforms, missing the opportunity to track mobility using more massive social networking platforms with larger numbers of users. Some of these more popular social media websites have recently integrated geo-social networking functions into their platforms, such as Facebook, Twitter, and Google+. Until September 2013, one of the most popular geo-social networking platforms, FourSquare, had over 40 million users (Smith 2013). Meanwhile, Facebook had attracted 1.16 billion. Twitter and Google+ had attracted 400 million and 500 million, respectively, at this time (Smith 2013). Unfortunately, we do not yet know if these more massive online social networking platforms are good venues through which to study human mobility.

In this paper, we propose a process map to collect human mobility data from Twitter. The reason Twitter was chosen is because Twitter was designed to share public information and has a much more open API for data collection. Two python modules were developed to implement the process map. In the paper we suggest promising future research directions enabled by using Twitter to study human mobility before concluding with a discussion of limitations and validation plans.

BACKGROUND

Collecting empirical data of human mobility is a difficult task. Some early studies assumed human movements have the same patterns as animal movements and characterized them as random walks (Viswanathan and Afanasyevt 1996; Ramos-Fernandez et al. 2004). This assumed human mobility followed a heavy-tailed probability distribution. Although it was an important discovery, the lack of adequate and precise data of human movement greatly limited the applicability of these studies.

During the past several years, researchers have begun using data collected on mobile phone users to understand human mobility. Gonzalez et al. (2008) studied human mobility by using two datasets. The larger one included 100,000 mobile phone users over a 6-month period. They found that the distribution of incremental displacements of all individuals followed a truncated power-law distribution. They also showed that individual trajectories were largely indistinguishable after rescaling using population density. This study discovered some fundamental laws of human mobility and provided important implications and algorithms for large-scale agent-based models of human mobility.

Song et al. (2010a) developed a human mobility model using a similar method. They used a 3-month-long record of 50,000 mobile phone users. The model adopted

entropy as the fundamental quantity to capture the movement of individuals and predict their locations. The study concluded that a person's daily movements exhibited inherent regularity. Predictability of movement could peak at 93% accuracy, although determining the exact locations of individuals was beyond the model's capability.

Later, Song et al. (2010b) investigated a larger dataset which contained 1 million mobile phone users for a year-long period. They observed three unique characteristics of human mobility which both the Lévy flights model (Brockmann et al. 2006) and the continuous-time random-walk model (Montroll and Weiss 1965; Metzler and Klafter 2000) failed to explain. These characteristics were: (1) a decreasing tendency of a person to visit new locations; (2) an uneven visitation frequency to different locations; and (3) an ultraslow diffusion, which meant people tended to return to the same locations (e.g., home, office, etc.). Based on these observations, Song et al. (2010b) developed a new individual-mobility model by adding two unique generic mechanisms: exploration and preferential return. While this new model was more representative of human mobility patterns compared to other models, its strength was in capturing long-term spatial and temporal scaling patterns.

Recently, researchers have begun to use geo-social networking platforms to study human mobility. Other than data of people's movements, these platforms also provide social relation information of the users. This extra information may enhance models' predictive capability. Cho et al. (2011) used geo-social networking platforms to attempt to answer the question of how social relations influence mobility patterns. Instead of mobile phone users, they collected empirical movement data from Gowalla and Brightkite, two online location-based social networking websites. These platforms not only provide more precise location information because they combined GPS signals and wireless networks, but also disclose the social relationships among users. Using geographical movement, temporal dynamics, and social networks, the researchers developed a periodic mobility model and a periodic and social mobility model. The latter showed up to 30% better predictive performance when compared to the former. Also, the models could predict the exact location of individuals with up to 40% accuracy. In another study, Noulas et al. (2012) used FourSquare to study human mobility in urban areas. Thirty-one large cities around the world were selected, and human movement data from these cities were analyzed. They found that the global movements followed a power-law distribution. Also, human mobility in all the cities studied followed almost the same pattern.

While these studies took initial steps to employ geo-social networking websites to study human mobility, there remain some important research gaps. First, additional research is necessary to resolve contradictory findings. For example, Noulas et al. (2012) found that global movements of humans followed a power-law distribution while no such pattern was observed from the movement data in urban areas. Also, research has not extended to some of the more massive and popular social networking platforms to study urban mobility patterns, including Facebook, Google+, and Twitter, which also allow geo-social networking. Their massive social network structures, enormous user base, and rich communication data can undoubtedly enhance our understanding of human mobility.

These gaps highlight the need for research into human mobility by using massive online geo-social networking platforms. To overcome these gaps, we propose a methodology to collect and analyze data from online geo-social networking platforms. We focus our study on Twitter, and develop a process map to collect human mobility data. We also compare the quantity of data collected from Twitter to the quantities from other geo-social media used in previous studies. The data will be analyzed and the results will be compared to the findings from research based on both mobile data and data from relatively small-scale social networking platforms. The effort will help us develop human mobility models to simulate and predict human mobility in large populous cities in the future.

METHODOLOGY

Twitter is an online social networking media that allows people to post text messages that are limited to 140 characters, called tweets. Users can also select to let Twitter add location information, called a geotag, to each tweet they post. Each geotag contains the exact coordinate at which the tweet was posted. Using the Twitter public API, we developed a process map to collect public tweets. The map contains two steps, as shown in Figure 1. These steps are described below:

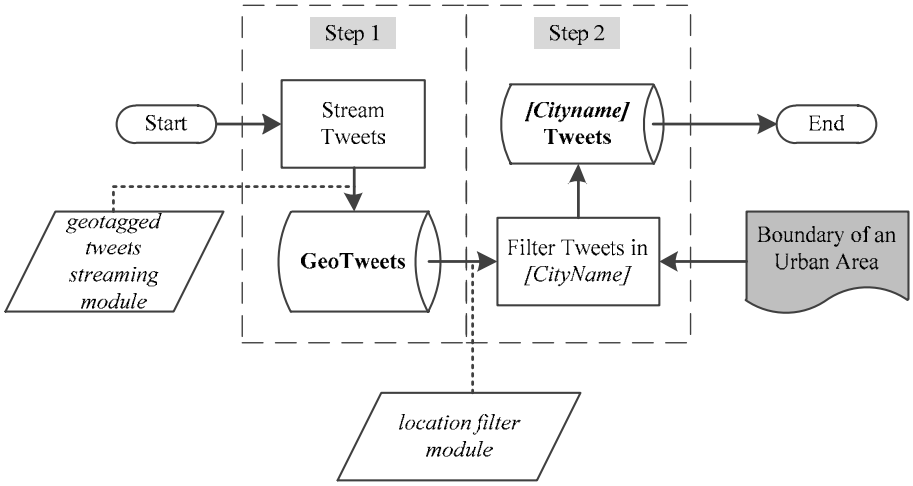

Figure 1. Process Map for Twitter Data Collection

Step 1: Collect Tweets with Geotags

We first collected all tweets that contain geotags. A *geotagged tweets streaming module* was developed. This module establishes a continuous connection between a computer in our research lab and the Twitter server. The connection streams every tweet that contains a geotag. Tweepy, a Python package for implementing the Twitter API, was used to develop the module. In addition to the exact geographical coordinate, each streamed tweet contains a list of extra information, including the tweet's ID, place name, the name and ID of the user who posted the tweet, and the time stamp of when it was posted. The collected tweets were stored in a database called **GeoTweets**. A reconnecting mechanism was coded

so that if the streaming was lost for 30 seconds, a restart message was displayed and a new streaming connection established.

Step 2: Filter Tweets in a Specific Urban Area

In this step, we first identify the geographical boundary from an urban area. While the boundary can be simple boxes, they can also be detailed borderline of a city. Then we test if the geographical coordinate of each tweet is within the boundary. As mentioned before, the geospatial information was embedded in each tweet. Therefore, we designed a *location filter module* for this step. The module checks each tweet in the **GeoTweet** database and retrieves its geographical coordinate. If the coordinate lies in the city, the tweet was selected and imported into a second database, ***[CityName]*Tweets**.

CASE STUDY AND DISCUSSION

To test the process map, we conducted a case study in the urban area of New York City. We tested the two modules in our process map and to ascertain whether human mobility data can be collected. First, *geotagged tweets streaming module* was set up and it ran for 15 days, from June 1, 2013 to June 15, 2013. As mentioned before, a reconnecting mechanism was built into the module to prevent losing the stream. During the 15-day period, we observed no such reconnection. About 69.93 million tweets were streamed into our **GeoTweets** database.

Then, we tested the second module, *location filter module*. We constrained the coordinates to 74˚15'W to 73˚40'W longitude and 40˚30'N to 40˚57'N latitude which covers the entire area of New York City. The module filters every tweet by its embedded coordinate. If the coordinate is within the area, the tweet will be selected, and imported into the **NYCTweets** database. After filtering, the **NYCTweets** contained 1,146,316 tweets from 79,022 users. The distribution of the tweets is shown in Figure 2. Each black dot represented a tweet.

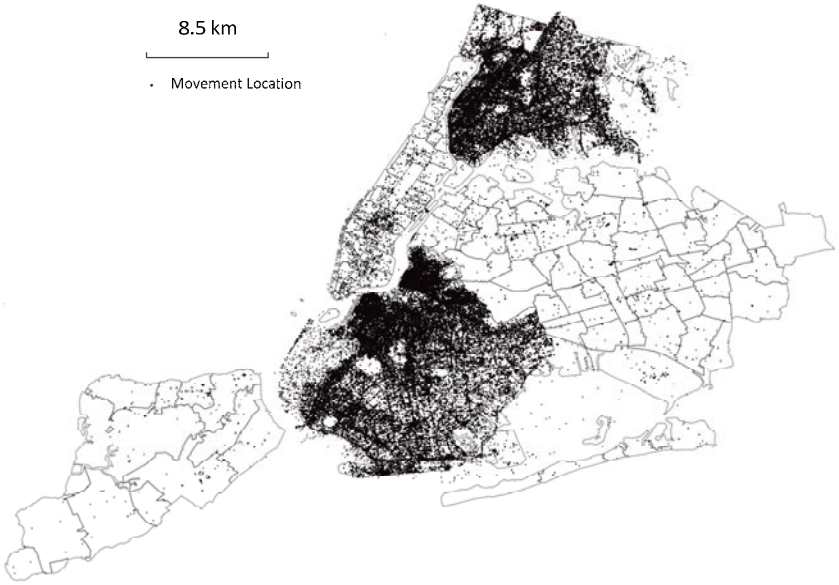

Figure 2. Geotagged Tweets (GeoTweets) Distribution in New York City

Then we retrieved the displacement data from the coordinate information embedded in each tweet. A displacement is the distance of two consecutive locations from the same individual. A total of 1,122,496 displacements were found. For users with at least two displacements, we also connected each individual's visited locations and extracted his/her trajectory. A total of 63,207 individual movement trajectories were identified. Figure 3 shows the movement trajectories on June 1, 2013. Our data showed that most trajectories happened between two boroughs in New York City, The Bronx and Brooklyn. It was also clear that while some trajectories covered long distances, others spanned relatively short ones as shown in the enlarged inset in Figure 3.

Figure 3. Movement Trajectories on June 1, 2013

Compared to other studies, we found Twitter can potentially provide more human mobility data for research. In Cho et al.'s study (2011), the average monthly location data was about 305,000 location data points around the world collected from Gowalla, and the number was about 145,000 from Brightkite. In Noulas et al.'s study (2012), 371,502 displacements were collected from FourSquare during a six-month period in New York City which includes 43,681 individuals. Comparably, Twitter provided a much larger quantity of human mobility data. The movement data is about 37 times greater to the volume of displacement data from the study using FourSquare. Additionally, the number of users was 1.8 times greater using Twitter compared to FourSquare if we assume no new individuals posted geotagged tweets or about 21.6 times otherwise. This substantially larger quantity of data from Twitter was also

collected over a much shorter period, further emphasizing the potential for such data to be used to understand, model and predict human mobility.

LIMITATIONS AND FUTURE STUDY

The study is limited by its data sample. We only used 15 days of Twitter data to analyze the potential of Twitter data to examine human mobility patterns. Also, our analysis was limited to one urban area, New York City. To further develop the study, we plan to collect data for longer time periods and from more cities. Our next step in the research will be to validate, both internally and externally, the ability of massive online geo-social networking platforms to study human mobility. For internal validation, data from different cities will be compared to examine if daily urban movements show consistencies. For external validation, data analysis results will be compared to findings from other studies.

We plan to use the empirical data we have collected and will collect in the future to build a human mobility model. Such a model should integrate agent-based modeling and GIS development. Our candidate platform is Quantum GIS which can also execute Python language. We will compare the results from the model to empirical data, find out the accuracy of model predictions, and thus improve the predictive capability of the model. Both the empirical data and simulation model will be helpful to study critical issues in urban areas, including epidemic spread, disaster response and evacuation, etc.

CONCLUSION

Human mobility has important impacts on policy-making in large cities. Researchers have developed models to understand and predict the patterns of movement trajectories. However, their civil applications are limited. Several reasons cause such limitation, and chief among them is data collection. Although researchers have used geo-social networking platforms to study human mobility, more research is still necessary to explore the viability of the more massive and popular platforms.

In this paper, we introduced a method to collect human mobility data from Twitter. A process map was developed and two Python modules were designed. Our preliminary analysis and case study found that Twitter can provide a much larger quantity of human mobility data compared to other social networking platforms (Cho et al., 2011; Noulas et al., 2012). This is owing to its larger user base and open design.

In the future, we plan to; (1) collect additional mobility data from multiple cities and over longer time scales to validate the methodology internally and externally, (2) develop algorithms that emulate the mobility patterns identified, and (3) use the data to build and validate an agent-based human mobility simulation model. We hope that the model will provide key insights and powerful predictive ability for policymakers in large, global cities, and potentially help inform key urban issues and policies, such as those relating to disaster evacuation, response and relief, and the spread of epidemics.

REFERENCES

Brockmann, D., Hufnagel, L., & Geisel, T. (2006). The scaling laws of human travel. Nature, 439(7075), 462-465.

Cho, E., Myers, S. A. and Leskovec, J. (2011). "Friendship and mobility: User movement in location-based social networks". Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM.

Gonzalez, M. C., Hidalgo, C. A. and Barabasi, A.-L. (2008). "Understanding individual human mobility patterns." Nature, 453(7196): 779-782.

Metzler, R., and Klafter, J. (2000). The random walk's guide to anomalous diffusion: a fractional dynamics approach. Physics reports, 339(1), 1-77.

Montroll, E. W., and Weiss, G. H. (1965). Random walks on lattices. II. Journal of Mathematical Physics, 6, 167.

Noulas, A., Scellato, S., Lambiotte, R., Pontil, M. and Mascolo, C. (2012). "A tale of many cities: Universal patterns in human urban mobility." PloS one, 7(5): e37027.

Ramos-Fernandez, G., Mateos, J. L., Miramontes, O., Cocho, G., Larralde, H. and Ayala-Orozco, B. (2004). "Lévy walk patterns in the foraging movements of spider monkeys (ateles geoffroyi)." Behavioral Ecology and Sociobiology, 55(3): 223-230.

Smith, C. "How many people use the top social media, apps & services?" Digial Marketing Ramblings. http://expandedramblings.com/index.php/resource-how-many-people-use-the-top-social-media/. Last accesed: Semptember 19, 2013.

Song, C., Koren, T., Wang, P. and Barabasi, A.-L. (2010a). "Modelling the scaling properties of human mobility." Nat Phys, 6(10): 818-823.

Song, C., Qu, Z., Blumm, N. and Barabási, A.-L. (2010b). "Limits of predictability in human mobility." Science, 327(5968): 1018-1021.

Viswanathan, G. and Afanasyevt, V. (1996). "Lévy flight search patterns of." Nature, 381: 30.