Spatio-temporal knowledge discovery from georeferenced mobile phone data

Yihong Yuan, Martin Raubal

Department of Geography, University of California, Santa Barbara, USA, 93106

yuan@geog.ucsb.edu, raubal@geog.ucsb.edu

1. Introduction

Information and communication technologies (ICTs), such as mobile phones and the Internet, are increasingly pervasive in modern society. These technologies provide greater flexibility regarding when, where, and how to travel. Understanding the influence of ICTs in our current mobile information society will be essential for updating environmental policies, and maintaining sustainable mobility and transportation (De Souza e Silva 2007). Moreover, ICTs have provided a wide range of spatio-temporal data sources, which can be used for geographic knowledge discovery and data mining in studies on geographic dynamics, such as human travel behavior and mobility patterns (Song et al. 2010; Yuan 2009; Miller 2009). There have been several studies focusing on extracting spatio-temporal data from georeferenced mobile phone data. For example, Ahas' social positioning method (SPM) combines both location data and social attributes of mobile phone users to study the dynamics of urban systems (Ahas and Mark 2005, Ahas et al. 2007). Gonzalez et al. (2008) studied the individual trajectory of 100,000 mobile phone users based on tracked location data, providing new insights to understanding the basic law of human motion.

As a generalized research frame, Miller (2009) discussed five major tasks in geographic data mining and knowledge discovery: spatial classification and capturing spatial dependency, spatial segmentation and clustering, spatial trends, spatial generalization, and spatial association. Traditional geographic knowledge discovery mainly focuses on obtaining new knowledge from a relatively comprehensive dataset, such as extracting movement patterns based on high resolution trajectories. However, several spatio-temporal datasets (e.g., georeferenced mobile phone data), only provide incomplete data with relatively low resolution and few individual attributes. Therefore, it is important to determine how much and to what extent we can extract knowledge from sparse data sources, as well as dealing with uncertainty in incomplete datasets. In this paper, we will provide a framework of extracting spatio-temporal knowledge in a typical georeferenced mobile phone dataset. This will be helpful in updating the research tasks of geographic knowledge discovery in the age of instant access.

2. Dataset

This research utilizes a dataset from Harbin City, China. Harbin city is a major commercial, industrial, and transportation center situated in northeast China. It was ranked as one of the top ten populated cities in China in the year 2009. The dataset covers over one million people from Harbin city, including mobile phone connection records for a time span of 9 days (07/21/07-07/29/07). The data include the time, duration, and location[1] of mobile phone connections, as well as the age and gender attributes of the users. Moreover, it provides the phone number and city code[2] of the other end of each phone call. Note that the location records in the dataset cannot represent the accurate moving trajectory of each user, since the locations are recorded only when there is a phone call connection. However, based on a summary of 9 days' records, the data are still useful for depicting the general characteristics of individual travel mobility.

3. Extracting spatio-temporal knowledge in georeferenced mobile phone data

Generally, the dataset provides us with three types of directly recorded information for each mobile phone user:

1) Cell phone usage
2) Social attributes (age, gender)
3) Spatio-temporal points within a given time span

All the data mining and knowledge discovery tasks are based on the combination and interaction of the above three information categories. Moreover, since urban systems are considered organized aggregations of human settlements, we can also obtain inferential knowledge for the city based on the behavior of associated citizens, such as indentifying spatial cluttering of traffic in different urban areas. Therefore, the research questions are divided into two categories: *individual-oriented* research and *urban-oriented* research.

[1] For each user, the location of the nearest cell phone tower is recorded both when the user makes and receives a phone call. Since the towers are located every 300m-500m in the city, the location accuracy is about 300m-500m.

[2] The city code indicates in which city the other side of a particular phone call is located.

3.1 Individual-oriented research

3.1.1 Analysis of movement patterns

Although human activity is potentially random and irregular, there still exist identifiable patterns in every person's life. It is easier to predict the behavior of people who travel little than those who travel more frequently (Gonzalez et al. 2008).

(1) **Trajectory patterns:** Based on the scattered spatio-temporal points provided in the dataset, it is possible to identify spatio-temporal paths through interpolation methods. However, this method may not be applicable for every user, since we need a certain number of well-distributed points to create a user trajectory. Based on trajectories, it is feasible to study the patterns (e.g., clustering, periodicity, predictivity) of individual trajectories. Figure 1 depicts the activity path for a specific individual during a week. Furthermore, we can identify particular patterns for population groups divided by social attributes.

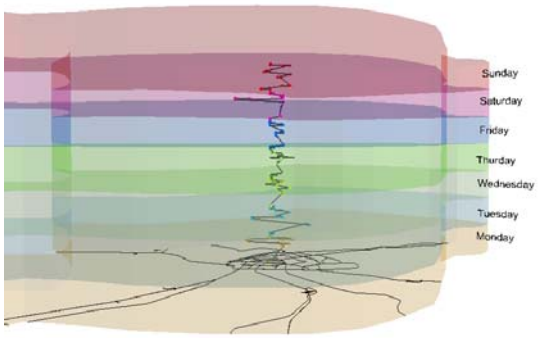

Figure 1: A week-long individual travel-activity path.

(2) **Point patterns (points of interest - POIs):** Based on the analysis of trajectory patterns, the POIs associated with each individual can be extracted by applying predefined discovery rules. POI discovery provides a method to obtain more sufficient user attributes in an incomplete dataset. For example, it is feasible to extract the work and home locations, regular entertainment places, and other POIs for each mobile phone user.

(3) **Correlation between mobile phone usage and movement patterns:** Previous studies have focused on the interaction between ICT and human activity-travel behavior. However, due to the lack of sufficient data and the complicated nature of the interaction, there is still a continuing debate on how it works in everyday life. Based on the phone usage information and social attributes, we examined how population heterogeneity impacts the relationship between mobile phone usage and individual movement radius. We argue that the heterogeneity of the population should be taken into account when analyzing the correlation between ICT and human activity-travel behavior. Therefore it is important to specify

individual attributes (e.g., cultural, social, institutional, physical aspects) when investigating this problem. The mobile phone usage and travel behavior correlate differently among various social groups. Therefore, a general conclusion for the population is insufficient to represent the complicated nature of this problem. Future research should focus on studying the correlation between phone usage and trajectory patterns. This would provide more specific results on how mobile phone usage impacts an individual's daily life, as well as offering references to policy makers.

3.1.2 Analysis of social networks

Janelle (1995) introduced four types of communication modes based on different spatio-temporal constraints: Synchronous Presence (SP), Asynchronous Presence (AP), Synchronous Tele-presence (ST), and Asynchronous Tele-presence (AT). Communication often occurs within members of a particular social network. Therefore, geographic mobility and cell phone usage could both be considered as connections in the same social network. Traditional research questions include the prediction and inference of network topology, the flow of information, and the interaction among networks. For instance, Pultar and Raubal (2009) studied the integration of social, transportation, and data networks. Particularly, in the case of georeferenced mobile phone data, a potential research question would be the combination, differentiation, and interaction of social networks associated with different communication modes.

3.2 Urban-oriented research

Cities are complex systems constituted by a myriad of processes and elements (Batty 2005). Since individuals are atoms in an urban system, the spatio-temporal characteristics of an urban system could be viewed as a conceptual generalization of individual behavior.

(1) **Spatial division**: Mobile communications might alter the traditional spatial division of urban spaces (Kwan et al. 2007), resulting in a change in urban planning and transportation systems. Potential research questions include the involvement of city structure under the impact of ICT (Torrens 2008), the comparison of "phone usage flow maps" and "travel flow maps", etc.

(2) **Spatial clustering**: Clustering refers to different types of hotspots in the city, for example, the hotspots of mobile phone usage, traffic jams, or night life (Figure 2). These patterns can be found on a range of timescales: from daily scale to yearly scale. The study of hotspot clustering patterns would be helpful for constructing a more efficient urban system.

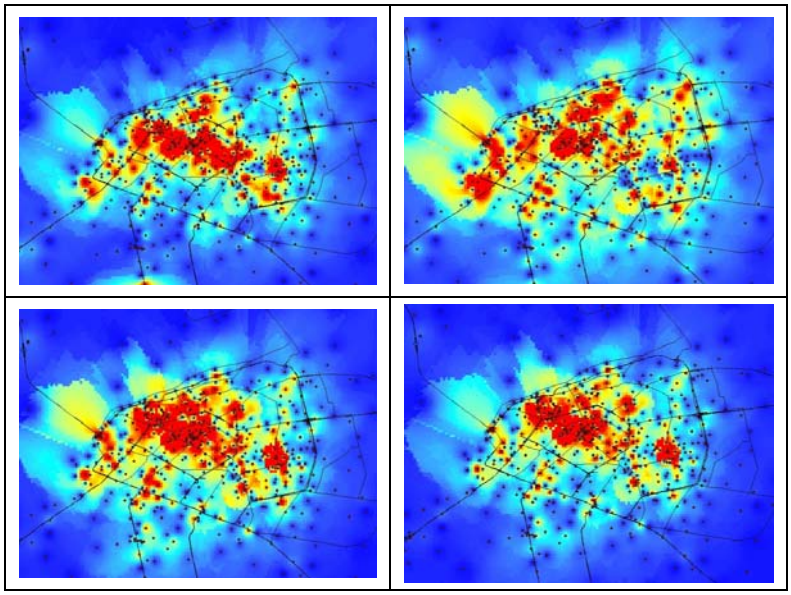

Figure 2: The changing clustering of mobile phone usage in Harbin city at different time points.

(3) Spatial central tendency and spread: Central tendency refers to a "middle" value or a typical value of the distribution. For a given city, the central tendency of spatio-temporal behavior reflects various characteristics of the urban system. For example, Figure 3 shows the distribution of individual travel radius in Harbin city. As can be seen, the movement radius of most people is around 3km. However, if we switch to another target city, will the distribution be similar to this one? Therefore, it would be interesting to find the correlation between travel distance distribution and the attributes (area, structure, population, average phone usage) of a given city.

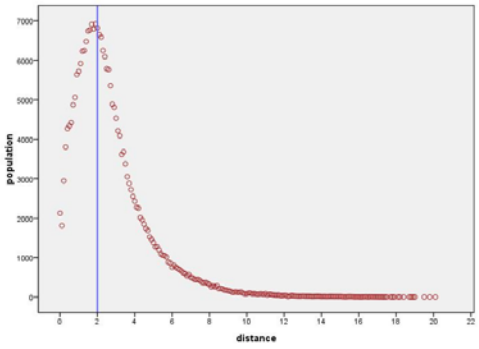

Figure 3: The distribution of individual movement radius.

4. Conclusions

Spatio-temporal knowledge discovery has gained wide attention due to the increasing availability of georeferenced data. Much progress has been made regarding the theories, methodologies, and applications in this field. In this research, we focus on

extracting spatio-temporal knowledge in a restricted and sparse dataset (mobile phone data). A framework of potential research questions is provided as two parts: individual-oriented research and urban-oriented research. This also provides a guideline for our future research on extracting spatio-temporal knowledge in incomplete datasets.

References

Ahas R and Mark Ü, 2005, Location based services - new challenges for planning and public administration. *Futures*, 37(6): 547-561.

Ahas R, Aasa A, Slim S, Aunap R, Kalle H and Mark Ü, 2007, Mobile Positioning in Space – Time Behaviour Studies: Social Positioning Method Experiments in Estonia. *Cartography and Geographic Information Science*, 34(4): 259-273.

Batty M, 2007, *Cities and complexity*. MIT Press, Cambridge, MA, USA.

De Souza e Silva A, 2007, Mobile phones and places: The use of mobile technologies in Brazil. In: Miller HJ (ed), *Societies and cities in the age of instant access*, Dortdrecht, The Netherlands, Springer, 295-310.

Gonzalez MC and Hidalgo CA, et al., 2008, Understanding individual human mobility patterns. *Nature*, 453(7196):779-782.

Janelle D, 1995, Metropolitan expansion, telecommuting, and transportation. In: Hanson S (ed), *The Geography of Urban Transportation*, New York, The Guilford Press, 407-434.

Kwan MP, Dijst M and Schwanen T, 2007, The interaction between ICT and human activity-travel behavior. *Transportation Research Part A-Policy and Practice*, 41(2):121-124.

Miller H, 2009, Geographic data mining and knowledge discovery: An overview. In: Miller HJ and Han J (eds), *Geographic Data Mining and Knowledge Discovery (Second Edition)*, London, CRC Press, 3-32.

Pultar E and Raubal M, 2009, Progressive Tourism: Integrating Social, Transportation, and Data Networks. In: Sharda N (ed), *Tourism Informatics: Visual Travel Recommender Systems, Social Communities, and User Interface Design*, Hershey, PA, USA, IGI Global, 145 -159.

Song CM and Qu ZH, et al., 2010, Limits of Predictability in Human Mobility. *Science*, 327(5968):1018-1021.

Torrens PM, 2008, Wi-Fi geographies. *Annals of the Association of American Geographers*, 98(1): 59-84.

Yuan M, 2009, Toward Knowledge Discovery about Geographic Dynamics in Spatiotemporal Databases. In: Miller HJ and Han J (eds), *Geographic Data Mining and Knowledge Discovery (Second Edition)*, London, CRC Press, 347-365.