A Study of Users' Movements Based on Check-In Data in Location-Based Social Networks

Jinzhou Cao[1], Qingwu Hu[1,*], and Qingquan Li[2,3]

[1] School of Remote Sensing and Information Engineering,
Wuhan University, Wuhan 430079, P.R. China
{caojinzhou,huqw}@whu.edu.cn
[2] Shenzhen Key Laboratory of Spatial Smart Sensing and Services,
Shenzhen University, Shenzhen 518060, P.R. China
[3] State Key Laboratory of Information Engineering in Surveying,
Mapping and Remote Sensing, Wuhan University, Wuhan 430079, P.R. China
liqq@szu.edu.cn

Abstract. With the development of GPS technology and the increasing popularity of mobile device, Location-based Social Networks (LBSN) has become a platform that promote the understanding of user behavior, which offers unique conditions for the study of users' movement patterns.

Characteristics of users' movements can be expressed by places they've visited. This paper presents a method to analyze characteristics of users' movements in spatial and temporal domain based on data collected from a Chinese LBSN Sina Weibo. This paper analyzes spatial characteristics of users' movement by clustering geographic areas through their check-in popularity. Meanwhile, temporal characteristics and variation of users' movements on the timeline is analyzed by applying statistical method.

Keywords: Check-In, Location-based Social Networks, Users' movements.

1 Introduction

The improvement of means of geographic data acquisition and the thriving rise of mobile Internet technology make it possible to create location data in social networks anytime and anywhere. This social networks driven by geographic location are called Location-based Social Networks (LBSN). This kind of network not only adds a location to existing social network, but also generates a knowledge database inferred from an individual's location (history) and location tagged data, e.g., common interests, behavior, and activities [1]. For instance, a user's trajectory movement often appearing in the stadium indicates that the user might like sports; the trajectory of the user frequently crossing the wild shows his preferences for outdoor activities.

LBSN has become a platform to promote the understanding of user behavior, which offers unique conditions for the study of users' movement patterns. Hence, how to take full advantage of huge geographic data generated in LBSN to mine knowledge becomes particularly important.

D. Pfoser and K.-J. Li (Eds.): W2GIS 2014, LNCS 8470, pp. 54–66, 2014.

Mobile social networking services has been the concern by many scholars at home and abroad in recent years. In early years, most of the studies were based on non-geospatial networks and the impact of geographical space was ignored. However, follow-up studies suggest that geographical space play a restrained role on social networks and many complex networks are embedded in it [2]. Zheng et al. mined recommendatory locations and representative activities to provide a roadmap for travelers using a large amount of GPS trajectories [3]. Liang et al. raised a way through the study of check-in data to help urban public space managers to make improvements in the spatial arrangement and operation of urban space at a lower cost and higher efficiency [4].

Unlike the traditional GPS data that were collected passively, data generated by LBSN is characterized by large amount, high efficiency, and high socialization. As a result, the subjective desire of users like interests, habits can be well reflected. Hence, if location check-in data could be fully mined we argue that a higher level of knowledge and information can be obtained, e.g., understanding the similarity between users based on their location histories [5]. Commercial social media itself analyze users' check-in records actively to recommend and push advertisement in order to create new profits [6].

Characteristics of users' movements can be expressed by places they've visited. In this paper, we present an approach to analyze of user's daily movement patterns from spatial and temporal perspective using check-in data in Sina Weibo, which is one of the most popular social network in China. First, we provide a general overview of the dataset collected from Sina Weibo and briefly analyze the spatial and the frequency distribution of the data. Then, we introduce the principles and methods to process spatial modeling analysis and temporal statistical analysis on users' movement patterns. After that, we collect data in specific regions and users through Sina API interface, and conduct experiments. The results are analyzed and discussed. Finally, we conclude with a discussion and highlight directions for future work.

2 Location Check-In Dataset

Social behavior is directly related to the location in users' daily life. When a user arrives at a place (e.g., restaurants or gymnasium), he will usually be associated with the activities of this place (e.g., eating or fitness). Nevertheless, we need lots of data sources for further research on the law of statistical characteristics in order to confirm this correlation is not accidental.

Sina Weibo is a Chinese microblogging website, a hybrid of Twitter and Facebook with a market penetration similar to what Twitter has established in the USA. Users check-in at places through a dedicated mobile device using GPS and other sensing technologies to automatically detect their location and post on the Sina Weibo platform. It has more than 0.5 billion registered users as of 2013, 57% of total number of microblogging users in China, and the number of daily active users has reached more than 60 million, with frequent information update, which provides powerful data

guarantee[7]. Moreover, there has accumulated more than 600 million check-in records in Sina Weibo. The fact that most of the records are in three major cities in China: Beijing, Shanghai and Guangzhou and about 60% of them are restaurant spot, 20% scenic spot among the records confirms the relationship between users' check-in activities and their movements.

Fig. 1. Sina Weibo mobile client check-in interface

Previous research may only use the two attributes (e.g. geographic coordinates and timestamp) of check-in data, with no more detailed information to further support, to analyze. Sina Weibo API provides location service interfaces freely, however, and we can acquire the following various attributes about a place： name, category, geographic coordinates, total number of check-ins, number of visitors checked-in, etc. Thus, it can meet the needs of the multi-level and multi-angle analysis and processing.

We have crawled data in Shanghai, China, between January 1st 2013 and March 31th 2013.Due to the data generated by users voluntarily, data quality issues, such as low accuracy, data redundancy, incorrect formatting, should be taken into account [8]. Thus it's necessary to data preprocess to get standard data. We have selected 1514470 check-ins after data preprocessing. Each record corresponds to a check-in at one of the 34963 POIs. A spatial distribution of collected dataset is depicted in Fig.2. A circle represents a geographic venue and its radius the popularity of it in units of number of check-ins. Each color corresponds to one of 10 categories shown in Table 1. The distribution of spatial dataset highlights the diversity of users' movements.

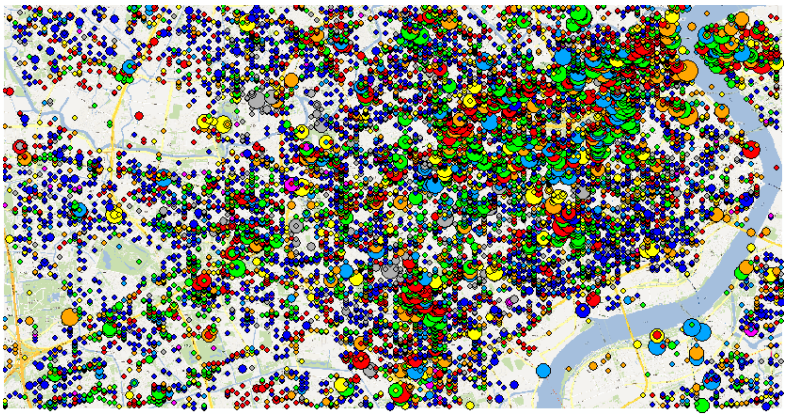


Fig. 2. Spatial distribution of collected dataset in Shanghai

The number of check-ins is an indicator of popularity for places among users [9].The complementary cumulative distribution function (CCDF) of the number of check-ins at different places is shown in Fig.3: there is a significant heavy tail in the distribution and the data approximately exhibit log-normal distribution. Only a few places have a large number of check-ins, while a higher number of places have only few check-ins; about 20% of places have just one check-in, with 30% above 10, whereas there is around 50% of places that have more than 100 check-ins. It well reflects the heterogeneity in users' movements, and the reasons behind it could be many, ranging from subjective reasons (e.g., forgetting check-in at a place), to social ones (e.g., sharing location with others). Users checking-in has always been voluntary rather than mandatory, anyhow, for which reason characteristics of users' check-ins can be a good sign to characterize the users' movements.

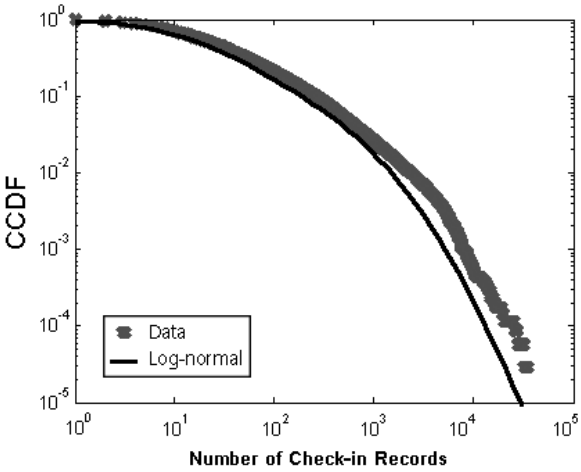


Fig. 3. Complementary Cumulative Distribution Function (CCDF) of the number of check-ins at different places. The data approximately exhibit log-normal distribution.

3 Users' Movement Pattern Analysis

People can be profiled according to the categories of places they visit, whereas geographic areas can be modelled according to their constituent venues. In this section we model the users' movement patterns by clustering geographic areas through their check-in popularity. In particular, we propose the use of place categories to create the squared area feature vector, define the similarity measurement and then apply the spectral clustering algorithm [10].In the meantime, we analyze temporal patterns of users' movements by applying statistical method in order to demonstrate the characteristics and variation on the timeline. Flow chart is shown in Fig.4.

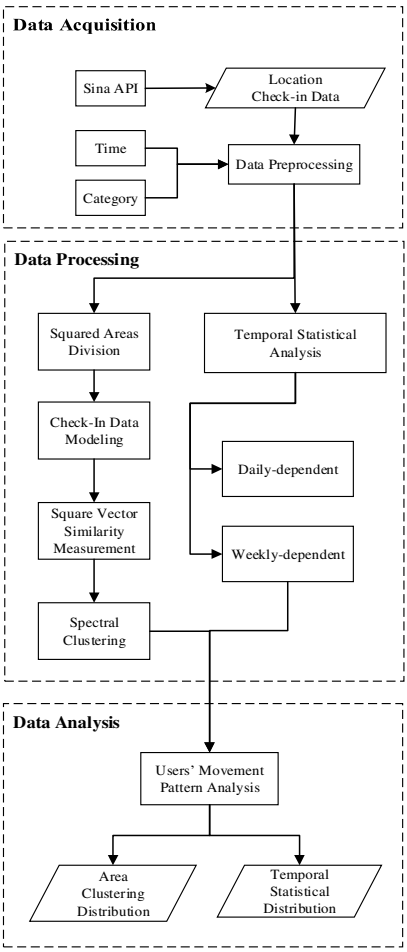


Fig. 4. The flow chart of users' movement pattern analysis

3.1 Spatial Modeling of Users' Movements

Squared Areas Division. Squared areas division effectively is a basis for subsequent operations. The square size of each squared area is an important factor to consider. If the size is too large, check-in records may contain multiple categories, thus the characterization of area is hard to determine. On the contrary, the amount of data inside the area can be too small to generate reasonable statistical representation. We set a threshold of the number of check-ins per area and finally calculate a reasonable square size and the number of area.

158 square kilometers in the central area of Shanghai was chosen to be the dataset in the experiment. Imposing the threshold of at least 30 check-in records per area has generated 559 areas. Spatial distribution of squared areas is shown in Fig.5.

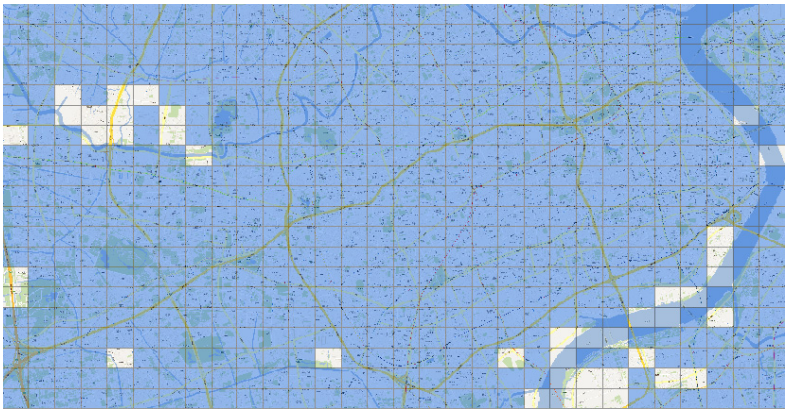


Fig. 5. Spatial distribution of Squared Areas. The squared areas not covered by color blue mean that there are less than 30 check-in records within them.

Location Check-In Data Modeling. There is a need to merge and split location category according to the characteristics of users' movements due to location category provided by Sina Weibo differing from what we need. Finally we classified into 10 categories, as shown in Table 1 and manually modified the category attributes of acquired data.

Table 1. The location category classification

1	Home	2	Work	3	Education
4	Shopping	5	Travel	6	Outdoors
7	Food	8	Life services	9	Leisure
10	Fitness				

Detailed description of location check-in data modeling is the following: Considering a squared area A within a city, we divide A into a certain number of equally sized

squares, each one representing a smaller local area *a*. The representation of *a* is defined according to the categories of nearby places and the number of check-ins took place at those. In this way not only we know what types of places are in an area, but we also have a measure of their importance from the perspective of users' movements. We define $TC_{c,a}$ of a category *c* to a geographic area *a*, for all places *p* that belong to category *c* within *a*, as follows:

$$TC_{c,a} = \sum_{p \in c} Allcheckins(p) \ , \ \forall p \in a \tag{1}$$

Hence, any area *a* can be represented using a vector $\boldsymbol{TC_a}$, the dimensionality of which is the number of the classified categories and each feature value is equal to $TC_{c,a}$ corresponding to a particular category. Particularly, $TC_{c,a}$ can be normalized in order to facilitate the research.

Square Vector Similarity Measurement. Supposing feature samples constituted by all the values of $TC_{c,a}$ as *X*, number of squared areas *a*, the dimensionality (number of categories) *c*, the matrix form is shown in Equation (2).

$$X = \begin{bmatrix} TC_{1,1} & \cdots & TC_{c,1} \\ \vdots & \ddots & \vdots \\ TC_{1,a} & \cdots & TC_{c,a} \end{bmatrix} \tag{2}$$

Where $TC_{i,j}$ represents the number of check-ins that belong to category *i* within area *j*.

We now define the similarity $Sim(i,j)$ between two square vectors *i* and *j*. Distance calculation (e.g., Euclidean Distance, Ming Distance and Mahalanobis Distance) and the similarity function (e.g., SMC, Cosine, Correlation Coefficient) are the common similarity measurement methods [11, 12]. Nevertheless, the similarity matrix calculated by different formulae will be very different and also different matrices will have different clustering results. For instance, Euclidean Distance is commonly used in image segmentation, and Cosine is often used in text data clustering. Because Cosine similarity has the property that it can be used in any dimension vector comparison, especially in high-dimensional space, we adopt the Cosine similarity measure as similarity measurement. See Equation (3), (4).

$$Sim(i,j) = \frac{1-d_{ij}}{2} \tag{3}$$

$$d_{ij} = \sum_{k=1}^{c} \frac{TC_{ki} TC_{kj}}{\left(\sum_{k=1}^{c} {TC_{ki}}^2 \sum_{k=1}^{c} {TC_{kj}}^2\right)^{1/2}} \tag{4}$$

Similarities between all vectors constitute the similarity matrix *W*, as shown in Equation (5).

$$W = \begin{bmatrix} w_{1,1} & \cdots & w_{n,1} \\ \vdots & \ddots & \vdots \\ w_{1,n} & \cdots & w_{n,n} \end{bmatrix} \tag{5}$$

Where $w_{i,j}$represents the similarity between sample *i* and *j*, equaling to $Sim(i,j)$.

Spectral Clustering. The impact of the similarity matrix for clustering results doesn't been taken into consideration in traditional clustering algorithms. The direct analysis of similarity matrix itself can avoid the limitations of the introduction of distribution hypothesis of sample space to a great degree in spectral clustering algorithm, however. Spectral clustering algorithm is capable of clustering on all sample space that is arbitrary shape theoretically and has been applied to speech recognition, text mining and other fields widely [13].

Spectral clustering method views samples as vertex, and similarity between two samples is considered as edge with weight. From this point of view, clustering problem is converted into graph division problem: find a method to divide a graph into groups so that weight of edges inside groups is as low as possible (namely similarity between groups as low as possible) and weight of edges among groups is as high as possible (namely similarity within group as high as possible) [14].

In this paper, we treat each squared area as a vertex in graph. The graph is generated by connecting the vertexes according to similarities between squared areas. Then divide the graph into groups and each group is a cluster. Detailed steps are listed as follows:

1. Create similarity graph from squared areas, and generate weight matrix W.
2. Compute Laplacian matrix L by Equation (6), in which D is degree matrix:

$$L = D - W \tag{6}$$

3. Compute k smallest eigenvector of L.
4. Combine the k eigenvectors together and generate an $N * k$ matrix, in which every row is a k-dimension vector. Finally conduct *k-means* algorithm to cluster the data and get result [15].

3.2 Temporal Statistical Analysis of Users' Movements

Characteristics of users' movement is largely associated with time. Temporal patterns of check-in data can be acquired by conducting statistical analysis on check-in data's time attribute, and it is presented as temporal characteristics and variation of users' movements on the timeline. Generally statistical analysis on time can be conducted in two different temporal bands, day and week [16, 17].

Generally speaking, users' dining and sleeping behavior are daily-dependent. This kind of activities take place each day and are closely related with time of the day. Thus we can conduct statistical analysis on daily-dependent behavior based on categories of the location separately. Meanwhile, users' working and entertaining behavior are weekly-dependent: users show different behavior in weekends and weekdays. Because of this, users' weekly-dependent behaviors are analyzed weekly.

4 Experimental Results and Analysis

4.1 Area Clustering Results

We now demonstrate the results yielded by clustering the 559 areas. Eight clusters are displayed in different colors, as seen in Fig.6.Each cluster is represented in Table 2 with top 5 categories ranked according to their popularity amongst the cluster members.

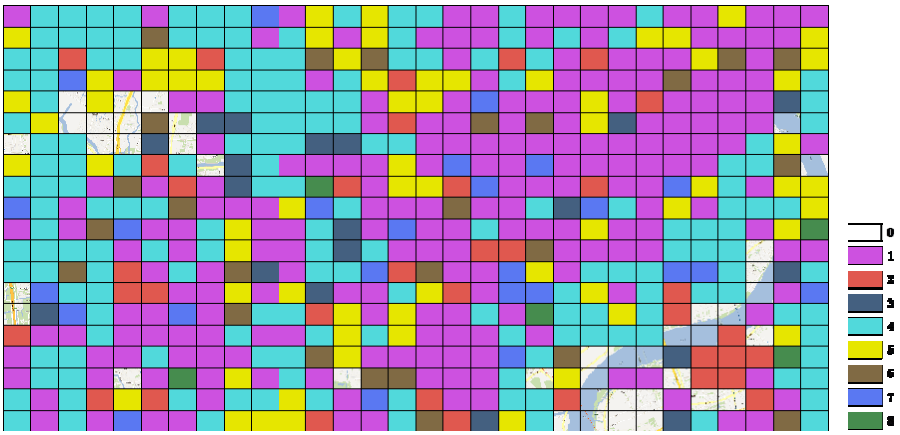


Fig. 6. Spectral Clustering Results. Correspondence between the color and the cluster number is shown in the right.

A common observation from Table 2 is the fact that each area has a dominant category, usually much higher than the second. The proportion of category ranking first are more than 50% in addition to Cluster 1.Cluster 1 suggests the coexistence of Food and Travel, covering the most central area of Shanghai with lots of famous scenic spots，which is the highest membership score amongst all clusters. Cluster 4 may signify residential areas, ranking second amongst all clusters. These two clusters share close to 60% of all squared areas, which is not only in line with the characteristics of urban POI category, mainly in restaurants and residential areas, but also the characteristics of users' movements in urban areas. It is notable to observe that categories Food and Home being the top five categories in all clusters also more confirms this conclusion.

Table 2. Squared Area Clustering. The category of Life Services is abbreviated as Life.

Cluster1 (211)		**Cluster2 (36)**		**Cluster3 (19)**		**Cluster4 (172)**	
Food	0.379	Leisure	0.644	Outdoors	0.649	Home	0.564
Travel	0.253	Home	0.088	Work	0.116	Education	0.124
Leisure	0.084	Travel	0.079	Home	0.076	Travel	0.087
Shopping	0.081	Food	0.077	Food	0.05	Food	0.072
Home	0.068	Outdoors	0.043	Travel	0.03	Work	0.056
Cluster5 (66)		**Cluster6 (25)**		**Cluster7 (25)**		**Cluster8 (5)**	
Work	0.507	Shopping	0.579	Life	0.549	Fitness	0.785
Food	0.126	Food	0.098	Home	0.113	Education	0.088
Home	0.106	Work	0.082	Travel	0.109	Home	0.063
Travel	0.087	Home	0.069	Food	0.08	Food	0.02
Leisure	0.052	Travel	0.064	Work	0.036	Leisure	0.018

4.2 Temporal Distribution Results

We will find very meaningful patterns closely related to users' movements from a temporal point of view by applying statistical measures to check-ins over hours and days.Fig.7 provides a general overview of temporal distribution of check-ins.

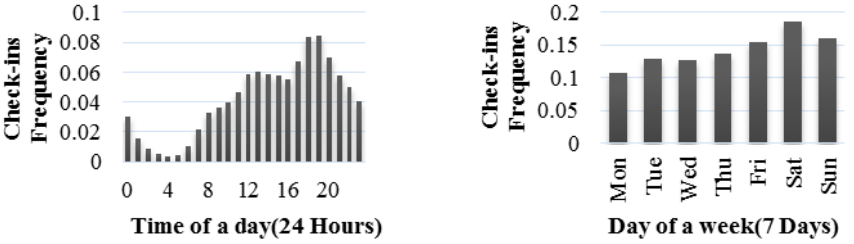


Fig. 7. (a) Daily and (b) Weekly Temporal Distributions of check-ins

As depicted in Fig.7 (a), users typically check-in frequently at noon and in the evening, most occurring at 9:00 to 23:00, with two peaks at around 13:00 and 7:00.This is due to the fact that most POIs are related to restaurants and food, and check-in activities are mostly concentrated in dinner time. A related observation can be made for Fig.7 (b).As users' movements related to dining, shopping, and leisure are over-represented in this figure, and we find the highest volume of check-ins on Saturdays and Sundays. Overall, we can see that data has been reasonably well reflected, and no evidence for contrary to common sense can be found in our data, e.g., higher number on check-ins in the middle of the night or lower during weekends. In this way it ensures that the characteristics extracted from them would be meaningful.

For better analyzing characteristics and variation on the timeline, we can apply statistical measures to those categories which are daily-dependent and weekly-dependent. Fig.8 plots the daily check-ins patterns to three different categories: Home, Food, and Work.

As can be seen in Fig.8 (a), home related check-ins increase from 6am, reaching a long lasting plateau between 10am and 3pm yet. This may be related with the fact that people go out for work or other things at this time. But when they return home linearly – increasing distribution is observed between 3pm and 11pm, which rather indicates that more and more people commute to home for rest.

Places related with food patterns is shown in Fig.8 (b) significantly, with two peaks: at 12pm, at 6pm, demonstrating that users check-in at restaurants at the peak dining time, while almost no check-ins can be observed from 12pm to 6am.Those findings are in line with what may be expected by a human observer and daily living habits. A specific point to note, however, is that check-ins don't show a continuous rise at breakfast time and between 6am and 9am in the morning. The reason behind this pattern may be that mostly breakfast restaurants are not fixed and people would not stay too long in the purchase of breakfast. This also demonstrates that most office-goers are used to solve his breakfast in his way to work rather than at breakfast restaurants.

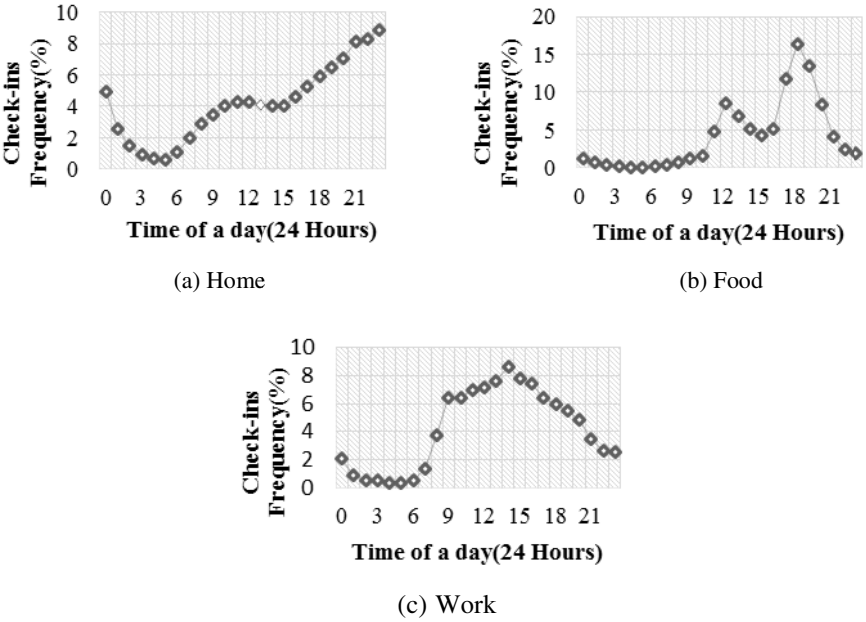


Fig. 8. Daily temporal distributions of check-ins to different daily-dependent categories

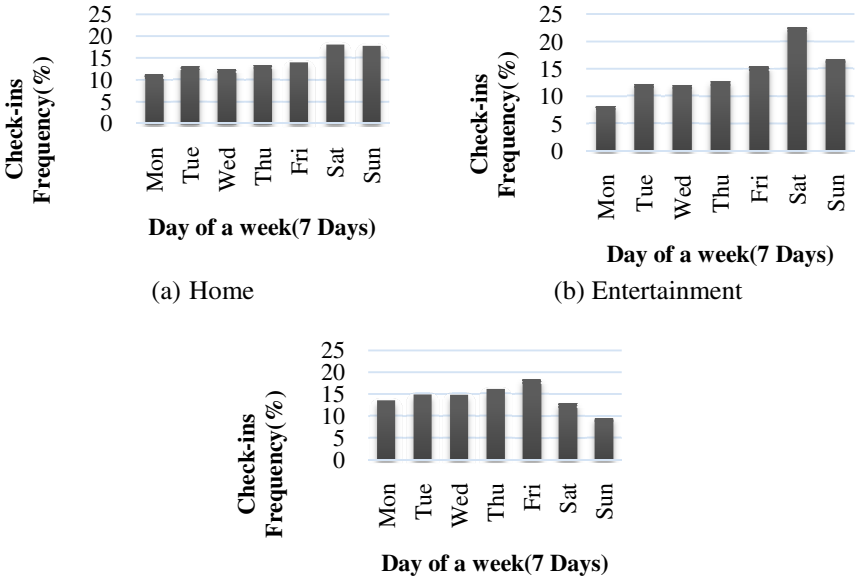
Check-ins show a steeper 2% increase at 9am with regard to 7am, indicating the rush hour at this time. Although a drop on growth rate from 9am can be observed the frequency maintains at a high position. Check-ins decreased from 2pm.

Figure 9 adds the weekly check-ins patterns for three different categories: Home, Entertainment, and Work. Check-ins related with home, as shown in Fig.9 (a), stay relatively rich throughout every day in a week with frequency at above 10%, and the higher number of check-ins takes place at weekends with above 15%.In contrast to the characteristics depicted in Fig.9(c), places tagged as work show a significant check-in decay during the weekend, which is in line with common sense.Fig.9 (b) plots the variation of check-ins related with entertainment. This distribution do not show such significant patterns on weekdays but rises straight up on weekends, especially Saturday.

Discussed above, we can draw the following conclusions:

The frequency statistics of users' movements is concordant with users' daily schedule and behavior. Daily-dependent behaviors is closely tied to eating, work, commute and other daily periodic activities, and shows cyclic effect to some degree. Weekly-dependent behaviors exhibit weekend effect, referred to a significant difference check-in frequency between weekdays and weekends, which is related with the time in working or non-working day.

Finally, while a single temporal band may not be sufficient to identify unique patterns for users' movements, we argue that multiple temporal bands can be combined to provide an accurate and meaningful descriptions of different users' movement patterns [18].



(a) Home (b) Entertainment

(c) Work

Fig. 8. Weekly temporal distributions of check-ins to different weekly-dependent categories

5 Discussion and Future Work

As discussed in the previous section, we can get a general consensus that LBSN offers opportunities of easily relating users with specific locations in reality and users' movement patterns can be extracted quickly by analyzing the attributes of check-in data (e.g., category, the number of check-ins). We argue that users' movements and preferences have been deeply embedded in the digital geographic space, and shared and access to public. It benefits sociologists to understand users' movement patterns by data generated from LBSN and urban scientists could plan layout of the city better.

In terms of future work we intend to improve clustering algorithm, evaluate the accuracy of clustering and improve it, thereby improving the accuracy of users' movements' analysis. Moreover, additional semantic information such as comments, tags could be discussed and mined deeply. Hence, extraction and modeling of semantic information can allow a deeper study of motivation of users' movement and experience degree of movement etc.

Acknowledgment. The authors would like to thank National Natural Science Foundation of China to support the project (Grand No.41371377).

References

1. Zheng, Y., Zhou, X.: Computing with spatial trajectories. Springer Science+Business Media (2011)
2. Garlaschelli, D., Loffredo, M.I.: Structure and evolution of the world trade network. Physica A: Statistical Mechanics and its Applications 355, 138–144 (2005)
3. Zheng, Y., Zhang, L., Xie, X., Ma, W.: Mining interesting locations and travel sequences from GPS trajectories, pp. 791–800 (2009)
4. Liang, L.Y., Ren, L.L., Wan, Y.H.: "LBS-based Social Network" of the Management and Operations in Urban public Space. Information Security and Technology 7, 56–63 (2011)
5. Li, Q., Zheng, Y., Xie, X., Chen, Y., Liu, W., Ma, W.: Mining user similarity based on location history, p. 34 (2008)
6. Zheng, Y., Zhang, L., Ma, Z., Xie, X., Ma, W.: Recommending friends and locations based on individual location history. ACM Transactions on the Web (TWEB) 5, 5 (2011)
7. Wikipedia, http://en.wikipedia.org/wiki/Sina_Weibo
8. Goodchild, M.F., Glennon, J.A.: Crowdsourcing geographic information for disaster response: a research frontier. International Journal of Digital Earth 3, 231–241 (2010)
9. Scellato, S., Mascolo, C.: Measuring user activity on an online location-based social network. In: 2011 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), pp. 918–923 (2011)
10. Noulas, A., Scellato, S., Mascolo, C., Pontil, M.: Exploiting semantic annotations for clustering geographic areas and users in location-based social networks (2011)
11. Bishop, C.M., Nasrabadi, N.M.: Pattern recognition and machine learning, vol. 1. Springer, New York (2006)
12. Ng, A.Y., Jordan, M.I., Weiss, Y., et al.: On spectral clustering: Analysis and an algorithm. In: Advances in Neural Information Processing Systems, vol. 2, pp. 849–856 (2002)
13. Hagen, L., Kahng, A.B.: New spectral methods for ratio cut partitioning and clustering. IEEE Transactions on Computer-aided Design of Integrated Circuits and Systems 11, 1074–1085 (1992)
14. Ng, A.Y., Jordan, M.I., Weiss, Y., et al.: On spectral clustering: Analysis and an algorithm. In: Advances in Neural Information Processing Systems, vol. 2, pp. 849–856 (2002)
15. Mei, Y.C., Wei, Y.K., Yit, K.C., Angeline, L., Teo, K.T.K.: Image segmentation via normalised cuts and clustering algorithm. In: 2012 IEEE International Conference on Control System, Computing and Engineering (ICCSCE), pp. 430–435 (2012)
16. Noulas, A., Scellato, S., Mascolo, C., Pontil, M.: An empirical study of geographic user activity patterns in foursquare. In: ICWSM 2011 (2011)
17. Aubrecht, C., Ungar, J., Freire, S.: Exploring the potential of volunteered geo-graphic information for modeling spatio-temporal characteristics of urban population. In: Proceedings of 7VCT 11, p. 13 (2011)
18. Ye, M., Janowicz, K., Mülligann, C., Lee, W.: What you are is when you are: the temporal dimension of feature types in location-based social networks. In: Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, pp. 102–111. ACM (2011)