# Mining the Semantics of Origin-Destination Flows using Taxi Traces

**Wangsheng Zhang, Shijian Li, Gang Pan**
Department of Computer Science, Zhejiang University
{zws10, shijianli, gpan}@zju.edu.cn

**ABSTRACT**

Origin-destination(OD) flows reflect both human activity and urban dynamic in a city. However, our understanding about their patterns remains limited. In this paper, we study the GPS traces of taxis in a city with several millions people, China and find that there are significant patterns under the OD flows constructed from taxis' random motion. Our spatiotemporal analysis shows that those patterns have close relationship with the semantics of OD flows, hence we can mine the semantics of OD flows from raw GPS trace data. The approach we proposed offers a novel way to explore the human mobility and location characteristic.

**Author Keywords** Urban computing, GPS trace, spatiotemporal analysis, LBSN

**ACM Classification Keywords** I5.2 [Patter Recognition]: Pattern analysis

**General Terms** Algorithms, Experimentation

**INTRODUCTION**

In recent years, as advanced technologies in sensor and communication, such as GPS and 3G, make massive urban data collecting and processing feasible, ubiquitous sensing has been widely applied in various areas(city planning [6], traffic engineering [9], public health [2, 7], and so on) to enable us better understand and coordinate the relationship between human and city. Location is a kind of critical information for building smart environments from smart vehicles to smart cities [13, 15, 18]. It helps bridge the gap between the physical world and cyber social network. People can expand their social structure with the new interdependency derived from their locations. These kinds of location-embedded and location-driven social structures are known as location-based social networks(LBSN) [22]. It can be used for location-based services, and also reveals mobility information of local residents.

One of the most important information source of LBSN that represent the relationship among communities in a city is origin-destination (OD) flows, which count the number of individual movements between locations in city. OD flows reflect not only human activity but also urban dynamic and they are widely used in city planning and traffic engineering. However, our knowledge about their patterns remains limited, partly due to the inefficient and expensive census-based methodologies. With the help of pervasive computing devices(mobile phone, travel card, GPS, and so on), we can improve our ability to gather and analyze raw data about OD flows. As a kind of frequently-used public vehicle which conveys passengers to location of their choice, taxi's trace corresponds precisely to individual movement. Hence it is a good data source for estimating OD flows.

In this paper, we first estimate OD flows from the GPS traces of taxis and find some significant patterns under those OD flows via clustering. Then we do spatiotemporal analysis to those patterns and reveal that they have close relationship with the semantics of OD flows. After that we propose a method to mine the semantics of OD flows via those relationship and execute our method on real data.

Based on the above steps, the main contributions of this paper are:

- We propose a new way to estimate the OD flows among locations in a city. Previous researches on OD flows mainly rely on inefficient and expensive census-based methodologies, limited our knowledge about OD flows. As a kind of frequently-used public vehicle, taxi's trace corresponds precisely to individual movement. It is a cheap and efficient data source for estimating OD flows;
- We find that there are significant patterns under OD flows and those patterns have close relationship with the semantics of OD flows. For example, the commute flow (Station to Market) aggregate in early morning while the transfer flow (Station to Station) is flat distributed in day-time;
- We exploit the relationship observed to mine the semantics of OD flows. According to the relationship, we designed three types of feature vectors extracted from taxi traces data. The best type of feature vector achieves a recognition accuracy of 83.7% using Neural Network.

The remainder of this paper is organized as follows. In the next section we review the related work. In the third section we descript the taxi traces data set we used. In the fourth section we estimate the OD flows from taxis' trace data and analyze the patterns under those OD flows. In the fifth section we mine the semantics of those OD flows and use


*UbiComp'12,* September 5-8, 2012, Pittsburgh, USA.
ACM 978-1-4503-1224-0/12/09.

this knowledge to infer location characteristic. Finally, our concluding remarks are given.

## RELATED WORK

In this section, we briefly review the related works on human mobility, location-based social networks and taxi traces data.

Recent researches have revealed that there are significant patterns under human mobility. Gonzalez *et al.* [5] find that human traces show a high degree of temporal and spatial regularity, each individual being characterized by a time independent characteristic travel distance and a significant probability to return to a few highly frequented locations. Jiang *et al.* [8] find that the human mobility pattern is mainly attributed to the underlying street network. The goal-directed nature of human movement has little effect on the overall traffic distribution. Calabrese *et al.* [1] use an algorithm to analyze opportunistically collected mobile phone location data and estimate weekday and weekend travel patterns of a large metropolitan area with high accuracy.

Human mobility data also have close relation with social networks. Eagle *et al.* [4] show that data collected from mobile phones have the potential to provide information about the relational dynamics of individuals. Cranshaw *et al.* [3] examine the traces of users of a location sharing social network for relationships between the users' mobility patterns and structural properties of their underlying social network.

As a kind of float sensors in city, taxis attract many researchers' attentions. Veloso *et al.* [17] present a spatiotemporal analysis of taxis GPS traces collected in Lisbon, Portugal and discuss the taxi driving strategies and respective income. They also carry out the analysis of predictability of taxi trips for the next pick-up area type given history of taxi flow in time and space [16]. Other researchers propose many useful ideas based on taxi. Zheng *et al.* [23] detect flawed urban planning using the GPS traces of taxis traveling in urban areas and find that pairs of regions with salient traffic problems and the linking structure as well as correlation among them. Zhang *et al.* [21] propose a method to discover anomalous driving patterns from taxi's GPS traces, targeting applications like automatically detecting taxi driving frauds or road network change in modern city. Li *et al.* [11] develop an improved ARIMA-based prediction method to forecast the spatiotemporal distribution of passengers in urban environment. Li *et al.* [10] present a trip analysis system which identifies the travel mode and purpose of the trips sensed by mobile devices and provides trip summaries and insights to mobile subscribers.

One major application of taxi traces is discovering regions of different functions in city. Qi *et al.* [12, 14] establish and confirm the relationship between the pick-up/drop-off characteristics of taxi passengers and the social function of city regions with qualitative and quantitative analysis. Yuan *et al.* [19] propose a framework that discovers regions of different functions in a city using both human mobility among regions and points of interests (POIs) located in a region. They segment an urban road network into regions by an image-processing-based approach [20]. In their work, a region is represented by a distribution of functions, and a function is featured by a distribution of mobility patterns.

## DATASET DESCRIPTION

We use trace dataset provided by the Traffic Bureau of Hangzhou City, which contains 7952 taxis and covers a period of 385 days. Taxis' state is sampled in a fixed time interval of 1 minutes and an extra sampling will be performed when the taximeter turn on or off. The position was obtained by GPS equipped in a taxi, so its precision was not affected by local tower density, which limited the spatial resolution of mobile-phone data. Each state consists of following fields:

- TAXI ID: the unique ID of sampled taxi;
- GPS POSITION: the longitude and latitude of that taxi at the sampling time;
- SPEED: the taxi speed at the sampling time, in kilometer per hour;
- ORIENTATION: the direction of that taxi at the sampling time, from 0° to 360° in clockwise with 0° indicates the north;
- METER STATE: indicates whether the taxi is heavy at the sampling time, 1 means the taxi is heavy(with passenger) and 0 means the taxi is empty(without passenger);
- TIME: the sampling time, with timestamp format 'YYYY-MM-DD HH:MM:SS'.

And a segment of state records in dataset is show in Table 1.

The state records of each taxi are extracted from dataset and sorted by time. Then, we define METER STATE turning from 0 to 1 as a pick-up event and turning from 1 to 0 as a drop-off event. A taxi trace is a series of state records begin with a pick-up event and last until encounter a drop-off event. The METER STATE may be incorrect because it is hard to avoid hardware faults thoroughly and taxi drivers may turn on the taximeter to avoid being interrupted when they have a rest, so a filtering process is necessary to remove these incorrect state records in order to recover taxi's actual traces from raw state records. Here we simply filter out taxi traces with distance less than 300m or travel time less than 2mins.

## PATTERN ANALYSIS

To estimate the OD flows, we divide the urban area into locations with size 0.001 degree in longitude and 0.001 degree in latitude. Then we measure the number of taxis' traces that pick up a passenger in location $L_i$ and drop off him/her in the location $L_j$. The number of taxis' traces $c_{ij}$ is a good approximation of OD flow from the location $L_i$ to the location $L_j$. $c_{ij}$ is rather uneven. The frequency $f(k)$ of the *k*th most visited OD flow follows Zipf's law

$$f(k) \sim k^{-\zeta}$$

with $\zeta = 0.4337 \pm 0.0063$, indicating most of human movements in the city occur on some major OD flows. The number of OD flows with $c_{ij} \geq 1000$ is 633 and the number of locations related to those 633 OD flows is

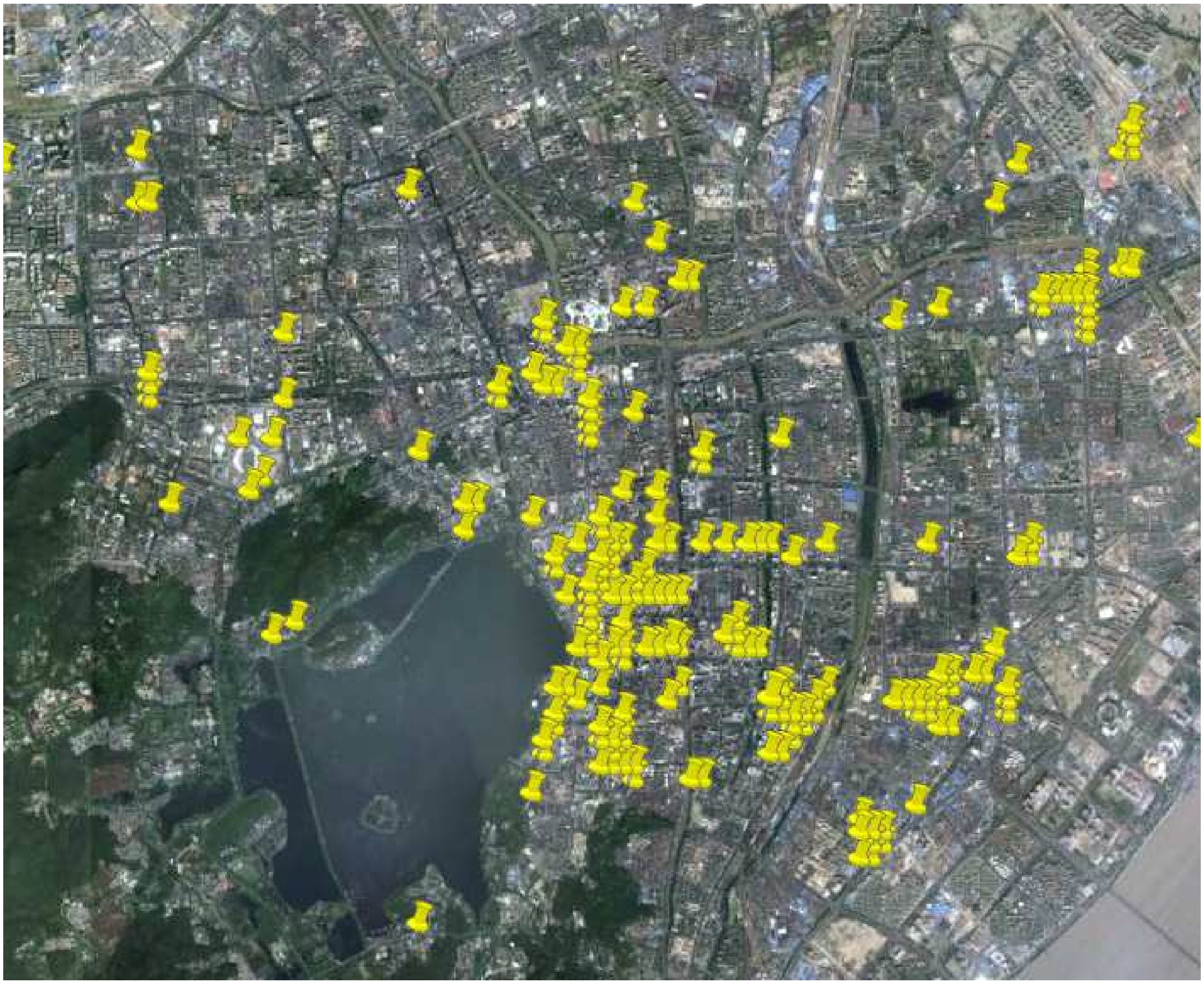

**Figure 1. The map of Hangzhou city. Yellow Pins indicate origin/destination locations of OD flows with the number of taxis' traces $c_{ij} \geq 1000$. Note that locations are rather uneven distributed.**

| TAXI ID | LONGITUDE | LATITUDE | SPEED | ORIENTATION | METER STATE | TIME |
|---|---|---|---|---|---|---|
| 1876 | 120.157295 | 30.241793 | 0.00 | 170.00 | 1 | 2009-4-1 00:00:04 |
| 14273 | 120.161180 | 30.272419 | 29.63 | 90.00 | 1 | 2009-4-1 00:00:04 |
| 2471 | 120.167820 | 30.284243 | 51.86 | 260.00 | 0 | 2009-4-1 00:00:04 |
| 14883 | 120.067444 | 30.090492 | 3.70 | 80.00 | 0 | 2009-4-1 00:00:04 |
| 18336 | 120.154850 | 30.290527 | 0.74 | 0.00 | 1 | 2009-4-1 00:00:07 |
| 10323 | 120.144110 | 30.327316 | 44.45 | 260.00 | 0 | 2009-4-1 00:00:07 |

**Table 1. A segment of state records in dataset.**

233(see Figure 1 and 2). Both number are very small compare with the total number of OD flows and locations but they indeed represent main human movements in the city. So we focus on analyzing those 633 OD flows and 233 locations.

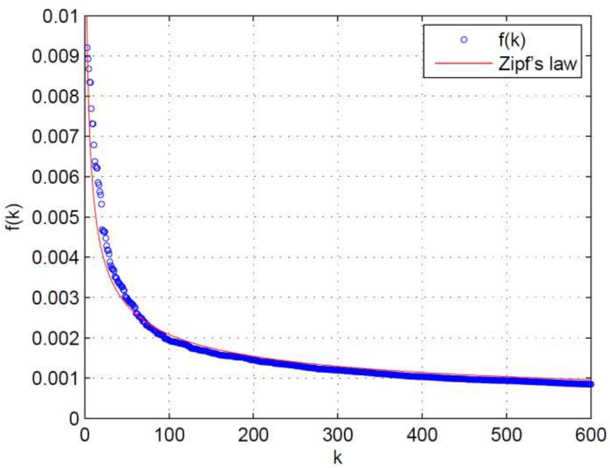


**Figure 2. The frequency $f(k)$ of the kth most visited OD flow follows Zipf's law, $f(k) \sim k^{-0.4337 \pm 0.0063}$**

To analyze the empirical observations, we measure the change of $c_{ij}$ over time. This fine-grained result shows significant periodic pattern, reflecting the short-term dynamic of city. We define the power spectral density of $c_{ij}$ as

$$S(d) = \frac{1}{N}\left|\sum_{n=1}^{N} c_{ij}(n)e^{-j(2\pi dn)}\right|^2$$

where $c_{ij}(n)$ is the the number of taxis' traces from location $L_i$ to location $L_j$ in time interval n. We find that two major components of its spectrum are 1cyc/day and 1cyc/week, this result is consistent with our daily experience.

After getting the period of $c_{ij}$, we can now depict each OD flow with a feature vector. Here we define three feature vectors:

- $V_d = \{c_{ij}^1, c_{ij}^2, \dots, c_{ij}^{24}\}/c_{ij}$: Visit frequency over time of day. $c_{ij}^k$ is the number of traces in the kth hour, $c_{ij}$ is total number of traces.
- $V_w^1 = \{V_d^W, V_d^H\}$: Visit frequency over weekday and weekend. $V_d^W$ is weekday's $V_d$ and $V_d^H$ is weekend's $V_d$.
- $V_w^2 = \{V_d^{Mo}, V_d^{Tu}, V_d^{We}, V_d^{Th}, V_d^{Fr}, V_d^{Sa}, V_d^{Su}\}$ : Visit frequency over time of week. $V_d^{Mo}$ is $V_d$ on Monday, etc.

We find those feature vectors can more or less reflect the characteristic of OD flow. For example, For a OD flow from location $L_{15}$(a scenic spot) to location $L_{32}$(a luxury hotel), its $V_d$ have peaks at 11:00AM and 15:00PM and its $V_w^2$ 's weekend components are larger than weekday components(see Figure 3). So we can assume human activity mainly occur on day-time and weekend for this OD flow.

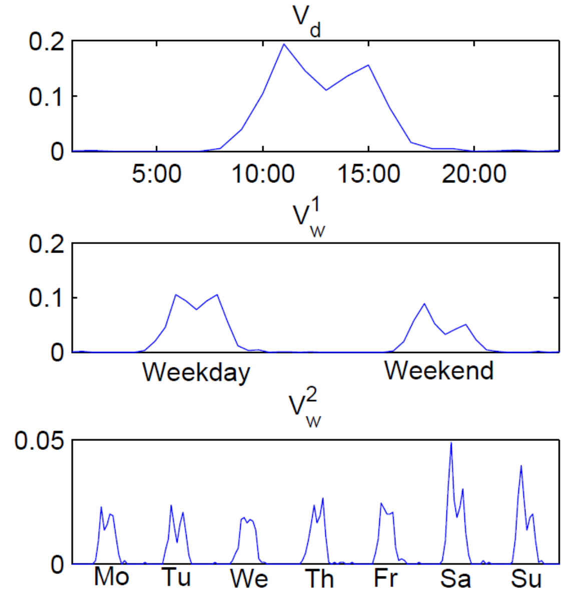


**Figure 3. Feature Vectors of OD flow from location $L_{15}$ (a scenic spot) to location $L_{32}$ (a luxury hotel). $V_d$ is the visit frequency over time of day; $V_w^1$ is the visit frequency over weekday and weekend; $V_w^2$ is the visit frequency over time of week. For this OD flow, human activity mainly occur on day-time and weekend.**

As feature vectors reflect the characteristic the OD flow, we can do cluster to group OD flows with similar character. To compare the performance of those three feature vectors, we do K-means clustering based on them. We define the BSS/TSS factor as

$$BSS/TSS = \frac{\sum_{S_i \neq S_j} \|V_i\text{-}V_j\|^2}{\sum_{i \neq j} \|V_i\text{-}V_j\|^2}$$

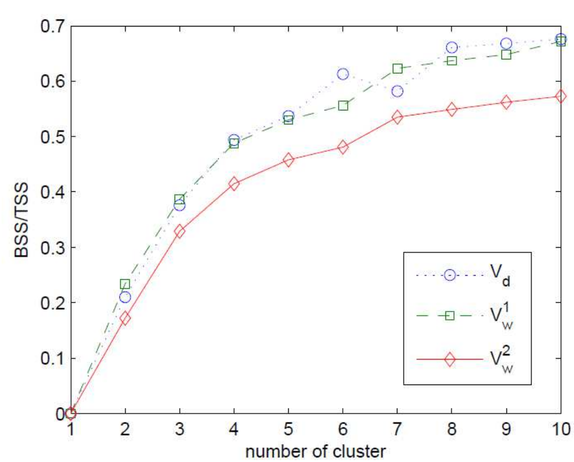


**Figure 4. The factor BSS/TSS versus the number of clusters. Higher BSS/TSS indicates larger difference among clusters.**

where $V_i$ is a type of feature vector of location $L_i$ and $S_i$ is the cluster of location $L_i$. Notice that the factor for all three feature vectors increase quickly, indicating there are significant difference among OD flow clusters(see Figure 4). The cluster center can be viewed as principal pattern of OD flows belong to that cluster.

### SEMANTICS MINING

To explore the semantics of OD flows, first we label location with semantics of main building in it such as Station or Hospital and investigate the cluster result. We find that most of the origin locations of OD flows in same cluster have same semantics, as well as destination locations. For all clusters, there are a few major semantics of locations: Station, Market, Hotel, Hospital, Mall, Dwelling and Bar. Station includes railway station, coach station, airport and large bus station; Market are places where merchants trade with each other while Mall are places people do shopping. Then we can define the semantics of OD flow by the semantics of its origin location and destination location such as Station to Station or Dwelling to Bar. The number of OD flows belong to each semantics type is also uneven (See Table 2).

| Semantics Type | Number |
|---|---|
| Station to Station | 184 |
| Mall to Station | 78 |
| Hospital to Station | 48 |
| Market to Station | 46 |
| Mall to Mall | 43 |
| Station to Hospital | 42 |
| Other 42 Types | 192 |
| Total | 633 |

**Table 2. Number of each semantic kind of OD flows.**

We compare 4 semantics of OD flows to show their relationship with patterns. The cluster center's $V_d$ of OD flow from Dwelling to Bar has a peak at 21:00 and that of OD flow from Bar to Dwelling has a peak at 4:00, Those patterns are consistent with our daily experience that people go to entertainment place before mid-night and return after mid-night(see Figure 5). For semantics of OD flows with same origin Station, commute flow (Station to Market) aggregate in early morning while transfer flow (Station to Station) is flat distributed in day-time (see Figure 6). Based on the relationship mentioned above, we can now mine the semantics of OD flows by their feature vectors. We use a two-layer feed-forward Neural Network with sigmoid hidden and output neurons to classify the semantics of OD flows. The Neural Network is trained with scaled conjugate gradient back propagation.

To verify the performance of our method, we execute our method on taxi traces data of Hangzhou. The input data is randomly divided into three parts: 70% for training, 15% for validation and 15% for testing. The output is limited in six largest semantics types: Station to Station, Mall to Station, Hospital to Station, Market to Station, Mall to Mall and Station to Hospital. We run the classification process for 10 times and the average of their accurate rates is show in Table 3.

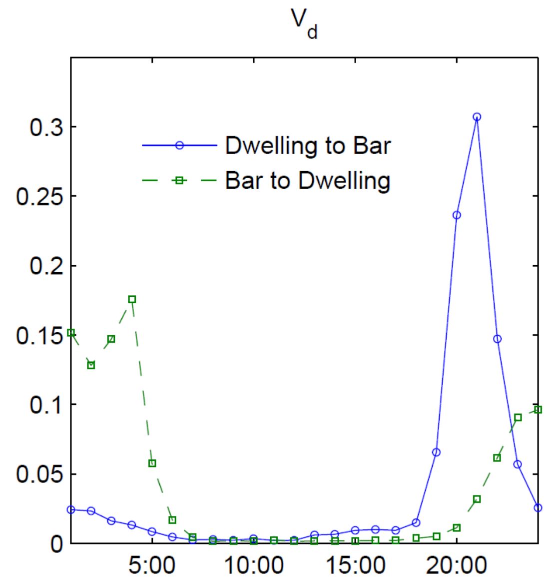


**Figure 5. $V_d$ of OD flow from Dwelling to Bar and OD flow from Bar to Dwelling. Note that people go to entertainment place before mid-night and return after mid-night.**

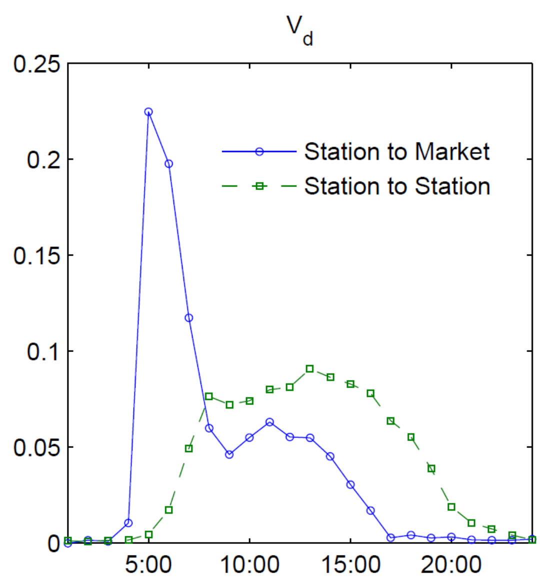


**Figure 6. $V_d$ of OD flow from Station to Market and OD flow from Station to Station. Note that commute flow(Station to Market) aggregate in early morning while transfer flow(Station to Station) is flat distributed in day-time.**

| Feature Vector Type | Average Accuracy |
|---|---|
| $V_d$ | 80.7% |
| $V_w^1$ | 83.4% |
| $V_w^2$ | 83.7% |

**Table 3. The average accuracy for three types of feature vectors.**

Note that the average accurate rates of $V_w^1$ and $V_w^2$ are higher than that of $V_d$, indicating that the information of weekly repeated patterns can help us in mining semantics of OD flows. However, the average accurate rate of $V_w^2$ is nearly the same as that of $V_w^1$ while the length of $V_w^2$ is 3.5 times of that of $V_w^1$, so the weekday/weekend treatment is enough to represent the weekly repeated pattern.

## CONCLUSIONS

In this paper, We estimate the origin-destination(OD) flows from taxis' traces and find that they have significant periodic patterns which closely related with their semantics. We mine the semantics of OD flows based on those patterns and the experiment result achieves a recognition accuracy of 83.7%. Our finding is useful to many LBSN applications and the approach we proposed offers a novel way to explore the human mobility and location characteristic.

Future work includes analyzing the semantics change of OD flow to discover urban events, comparing OD flow's pattern under different conditions such as urban-size or develop-level and detecting communities in city via the semantics of OD flows among them.

## ACKNOWLEDGEMENTS

The authors would like to thank anonymous reviewers for the helpful comments. This work is partly supported by High-Tech Program of China (No.2011AA010104) and Qianjiang Talent Program of Zhejiang (2011R10078). The corresponding author is Dr. Gang Pan.

## REFERENCES

1. F. Calabrese, G. D. Lorenzo, L. Liu, and C. Ratti. Estimating origin-destination flows using mobile phone location data. *IEEE Pervasive Computing*, 10(4):36–44, 2011.
2. V. Colizza, A. Barrat, M. Barthelemy, A. Valleron, and A. Vespignani. Modeling the worldwide spread of pandemic influenza: baseline case and containment interventions. *PLoS Medicine*, 4(1):95-110, 2007.
3. J. Cranshaw, E. Toch, J. Hong, A. Kittur, and N. Sadeh. Bridging the gap between physical location and online social networks. In *Proceedings of the 12th ACM International Conference on Ubiquitous Computing*, pages 119–128. ACM, 2010.
4. N. Eagle, A. Pentland, and D. Lazer. Inferring friendship network structure by using mobile phone data. *Proceedings of the National Academy of Sciences*, 106(36):15274–15278, 2009.
5. M. Gonzalez, C. Hidalgo, and A. Barab´asi. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, 2008.
6. M. Horner and M. O'Kelly. Embedding economies of scale concepts for hub network design. *Journal of Transport Geography*, 9(4):255–265, 2001.
7. L. Hufnagel, D. Brockmann, and T. Geisel. Forecast and control of epidemics in a globalized world. *Proceedings of the National Academy of Sciences*, 101(42):15124-15129, 2004.
8. B. Jiang, J. Yin, and S. Zhao. Characterizing the human mobility pattern in a large street network. *Phys. Rev. E*, 80(2):021136, Aug 2009.
9. R. Kitamura, C. Chen, R. Pendyala, and R. Narayanan. Micro-simulation of daily activity-travel patterns for travel demand forecasting. *Transportation*, 27(1):25–51, 2000.
10. M. Li, J. Dai, S. Sahu, and M. Naphade. Trip analyzer through smartphone apps. In *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 537–540. ACM, 2011.
11. X. Li, G. Pan, Z. Wu, G. Qi, S. Li, D. Zhang, W. Zhang, and Z. Wang. Prediction of urban human mobility using large-scale taxi traces and its applications. *Frontiers of Computer Science*, 6(1):111–121, 2012.
12. G. Pan, G. Qi, Z. Wu, D. Zhang, and S. Li. Land-use classification using taxi GPS traces. *IEEE Transactions on Intelligent Transportation Systems,* , 2012. DOI 10.1109/TITS.2012.2209201
13. G. Pan, Y. Xu, Z. Wu, S. Li, L. Yang, M. Lin, and Z. Liu. Taskshadow: Toward seamless task migration across smart environments. *IEEE Intelligent Systems,* 26(3):50–57, May-June 2011.
14. G. Qi, X. Li, S. Li, G. Pan, Z. Wang, and D. Zhang. Measuring social functions of city regions from large-scale taxi behaviors. In *2011 IEEE International Conference on Pervasive Computing and Communications, Work-in-progress*, pages 384–388. 2011.
15. J. Sun, Z. Wu, and G. Pan. Context-aware smart car: from model to prototype. *Journal of Zhejiang University - Science A*, 10(7):1049–1059, 2009.
16. M. Veloso, S. Phithakkitnukoon, and C. Bento. Urban mobility study using taxi traces. In *Proceedings of the 2011 International Workshop on Trajectory Data Mining and Analysis*, pages 23–30. ACM, 2011.
17. M. Veloso, S. Phithakkitnukoon, C. Bento, N. Fonseca, and P. Olivier. Exploratory study of urban flow using taxi traces. In *The First Workshop on Pervasive Urban*

*Applications (PURBA'11), in conjunction with PERVASIVE'11*. 2011.

18. Z. Wu, Q. Wu, H. Cheng, G. Pan, M. Zhao, and J. Sun. Scudware: A semantic and adaptive middleware platform for smart vehicle space. *IEEE Transactions on Intelligent Transportation Systems*, 8(1):121–132, March 2007.
19. J. Yuan, Y. Zheng, and X. Xie. Discovering regions of different functions in a city using human mobility and POIs. In *Proceedings of the 18th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2012.
20. N. Yuan, Y. Zheng, and X. Xie. Segmentation of urban areas using road networks. *MSR-TR-2012-65*, 2012.
21. D. Zhang, N. Li, Z. Zhou, C. Chen, L. Sun, and S. Li. ibat: detecting anomalous taxi trajectories from gps traces. In *Proceedings of the 13th International Conference on Ubiquitous Computing*, pages 99–108. ACM, 2011.
22. Y. Zheng. Location-based social networks: Users. *Computing with Spatial Trajectories*, Y. Zheng and X. Zhou, Eds. Springer, 2011.
23. Y. Zheng, Y. Liu, J. Yuan, and X. Xie. Urban computing with taxicabs. In *Proceedings of the 13th International Conference on Ubiquitous Computing*, pages 89–98. ACM, 2011