Computers, Materials & Continua CMC, vol.61, no.1, pp.227-241, 2019

Semantics Analytics of Origin-Destination Flows from Crowd Sensed Big Data

Ning Cao[1, 2], Shengfang Li[1], Keyong Shen[1], Sheng Bin[3], Gengxin Sun[3, *], Dongjie Zhu[4], Xiuli Han[5], Guangsheng Cao[5] and Abraham Campbell[6]

Abstract: Monitoring, understanding and predicting Origin-destination (OD) flows in a city is an important problem for city planning and human activity. Taxi-GPS traces, acted as one kind of typical crowd sensed data, it can be used to mine the semantics of OD flows. In this paper, we firstly construct and analyze a complex network of OD flows based on large-scale GPS taxi traces of a city in China. The spatiotemporal analysis for the OD flows complex network showed that there were distinctive patterns in OD flows. Then based on a novel complex network model, a semantics mining method of OD flows is proposed through compounding Points of Interests (POI) network and public transport network to the OD flows network. The propose method would offer a novel way to predict the location characteristic and future traffic conditions accurately.

Keywords: Origin-destination (OD) flows, semantics analytics, complex network, big data analysis.

1 Introduction

In recent years, with the development of up-to-date technology in wireless network communication, such as 5G and Global Position System, a dramatic rise of crowd sensed data collecting and processing had been seen. Analytics of sensing data has been widely used to enable a broad spectrum of applications, ranging from city planning [Horner and O"Kelly (2001)] or traffic [Kitamura, Chen, Pendyala et al. (2000); Lakhina, Mark, Christophe et al. (2005)] to epidemic disease monitoring [Colizza, Barrat, Barthelemy et al. (2007); Hufnagel, Brockmann and Geisel (2004)] or real-time reporting from disaster situations [Li, Li, Chen et al. (2018)].

In the field of mobile crowd sensing, for example, cellphones, vehicular sensors, or people themselves collected information. Hence, the obtained data through using crowd sensing methods is a new trend for big data acquisition [Sun and Bin (2017)]. Position information

[1] College of Computer Information and Engineering, Nanchang Institute of Technology, Nanchang, China.

[2] College of Information Engineering, Sanming University, Sanming, China.

[3] School of Data Science and Software Engineering, Qingdao University, Qingdao, China.

[4] School of Computer Science and Technology, Harbin Institute of Technology, Weihai, China.

[5] Public Teaching Department, Qingdao Technical College, Qingdao, China.

[6] School of Computer Science, University College Dublin, Dublin, Ireland.

[*] Corresponding Author: Gengxin Sun. Email: sungengxin@qdu.edu.cn.

would become a type of core data for constructing smart vehicles [Pan, Xu, Wu et al. (2011); Wu, Wu, Cheng et al. (2007)]. These core data can form position-based social networks [Song, Hu, Leng et al. (2015)].

The most important position-based social networks which stand for behavior of crowds in a town are origin-destination flows. It describes a journey by its departure point (Origin) and arrival point (Destination) [Sun and Bin (2018)]. OD flows not only can reflect people' behavior but also traffic jam. However, a major challenge for broader adoption of these patterns under OD flows is that the sensed data is not always reliable [Han, Dai, Paritosh et al. (2016)]. Taxi is acted as the most frequently used means of transportation, its tracks can be accurately recorded with the help of GPS. So it is a very appropriate data for gathering and evaluating OD flows.

We firstly build a taxi flow complex network by GPS tracks and detect some distinctive and implicit patterns through detecting community structure [Bin and Sun (2011)]. Then we use a novel complex network model to build a complex network [Shao and Sui (2014)] through compounding POI network and public transport network to OD flows network. Based on the composited complex network, spatiotemporal analysis is done to those patterns and discovers that there are close relationships between the semantics of OD flows and those patterns. At last, we design a new method to analyze semantics of OD flows through multiple relationships, and the new method is verified on actual dataset.

Our contribution lies on the following two aspects: Firstly, a novel method to evaluate the OD flows between geographical positions is proposed. We use multi-subnet composited complex network model to express multiple kinds of actual impact factors for OD flows in a city. Secondly, through topological analytics of the composited complex network, we discover that there are distinctive patterns which have tight relations with semantics of OD flows. Through spatiotemporal analysis, geographical location of boarding and disembarking can be discovered. Combined with POIs and public transport lines, we can get more accurate semantics of OD flows.

2 Related work

Research on taxi trajectory for understanding people behavior in location-based social networks is a very active research field at present. There had been many related research results.

Yuan et al. [Yuan, Zheng, Xie et al. (2012)] presented a decision model for statistical analysis of the dataset of taxi trajectory, the model can predict the passenger flow of taxis. Ying et al. [Ying, Kuo, Tseng et al. (2014)] proposed a new algorithm that depends on historical data to compute the shortest path for a given departure position and arrival position. Zhang et al. [Zhang, Sun, Li et al. (2015)] proposed a data mining algorithm to find abnormal driving behavior based on taxi's tracks, it can be used to automatically detect dangerous driving behavior or traffic jam. Chang et al. [Chang, Tai and Hsu (2009)] proposed a taxi passenger flow forecasting model based on multiple demand factors. Based on historical data, the model can successfully predict passenger demand in different time periods.

Human travel behavior had tight relationship with social data. Li et al. [Li, Wu, Xu et al. (2014)] studied taxi users' social network information, and they found the intrinsic relationship between taxi trajectory and users' sharing of social network information. The most major function of taxi tracks research is detecting urban areas of different roles in a

town. Zhong et al. [Zhong, Huang, Stefan et al. (2014)] investigated the relation between the location of users getting on and getting off and the function of urban areas. Zheng et al. [Zheng, Capra, Wolfson et al. (2014)] designed a method which maybe detect various functional areas of a town through using points of interests.

3 Preliminaries

This section introduces compounding mapping operation and subnet compounding operation of multi-subnet composited complex network model.

Definitions 1 (Compounding mapping): Given subnet network $G_a = (V_a, E_a, R_a, F_a)$, $G_b = (V_b, E_b, R_b, F_b)$, R' is called as set of compounding interrelations, $r' \in R'$, $\Psi : V_1 \times V_2 \to r'$ is called as compounding mapping between G_1 and G_2 according to r', which is called as compounding relation. R' is called as set of compounding relations.

Definitions 2 (Subnet compounding): Given subnet network$G_a = (V_a, E_a, R_a, F_a)$, $G_b = (V_b, E_b, R_b, F_b$, and $R_1 = R_{11} \times ... \times R_{1i} \times ... \times R_{1n_1} = \{(r_{11},...,r_{1i},...,r_{1n_1}) \mid r_{1i} \in R_{1i}, 1 \le i \le n_1\}$ $R_2 = R_{21} \times ... \times R_{2j} \times ... \times R_{2n_2} = \{(r_{21},...,r_{2j},...,r_{2n_2}) \mid r_{2j} \in R_{2j}, 1 \le j \le n_2\}$, compounding mapping $\Psi : V_1 \times V_2 \to r'$, $r' \in R'$, compounding subnet G_1 to G_2 would generate a new composited one network $G = (V, E, R, F)$,

(1) $V = V_1 \cup V_2$;

(2) $E \subseteq E_1 \cup E_2 \cup (V_1 \times V_2)$;

(3) $R = \{(r_1,...,r_{k'},...,r_k,...,r_{k''},...,r_n) \mid r_{k'} \in R_{1k'}, r_{k''} \in R_{2k''},\ r_{k'} \ne r_{k''}, 1 \le k' \le n_1, 1 \le k'' \le n_2, 1 \le n \le n_1 + n_2 + 1\}$;

(4) $F : E \to R$, when $< v_h, v_l > \in E_1$, $F(< v_h, v_l >) = (F_1(< v_h, v_l >), \varnothing, ..., \varnothing)$. when $< v_h, v_l > \in E_2$, $F(< v_h, v_l >) = (\varnothing, ..., \varnothing, F_2(< v_h, v_l >))$. when $< v_h, v_l > \in V_1 \times V_2$, $F(< v_h, v_l >) = (\varnothing, ..., \varnothing, \Psi(< v_h, v_l >), \varnothing, ..., \varnothing)$, where $v_h, v_l \in V, 1 \le h, l \le |V|$. Thereinto, $< v_h, v_l > \in V_1 \times V_2$ is called as outside edge and v_h, v_l as border nodes.

An example of subnet compounding is illustrated in Fig. 1.

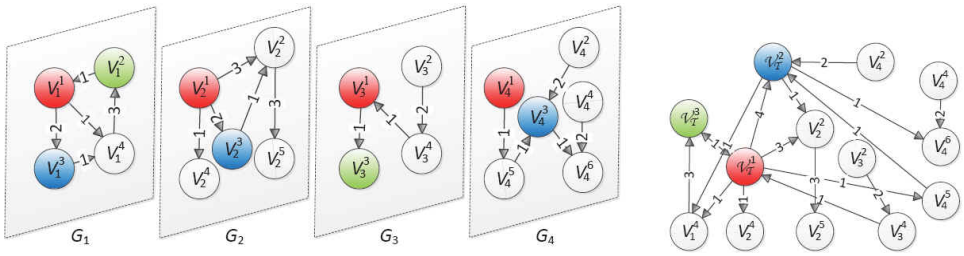

Figure 1: Subnet compounding of multiple network (G1, G2, G3, G4)

4 Dataset description

Firstly, the taxi trace dataset provided by Transportation Committee of Qingdao city is introduced. The dataset with about 20 million taxi-GPS records consists of 5872 taxi and covers 371 days. State of taxi is defined in a predetermined time interval of one minute,

the state includes some fields as follows:

- ID: the identification of data record;
- GPS LONGITUDE: longitude of a record;
- GPS LATITUDE: latitude of a record;
- LADEN/UNLADEN STATE: whether a taxi is laden at sampled time, 1 represents it is laden and 0 represents it is unladen;
- TIME: the sampled time.

An example of state explanation is show in Tab. 1.

Table 1: An example of state explanation

ID	LONGITUDE	LATITUDE	LADEN/UNLADEN STATE	TIME
2601	120.399201	36.087059	1	2014-5-2 09:03:31
5112	120.382474	36.085295	0	2014-5-2 09:03:31
716	120.379564	36.111745	1	2014-5-2 09:03:31
1998	120.426168	36.107182	1	2014-5-2 09:03:31
2119	120.437271	36.070068	1	2014-5-2 09:05:31
4571	120.397763	36.060571	1	2014-5-2 09:05:31
3309	120.334145	36.072825	0	2014-5-2 09:05:31

Abnormal data cleaning process is a necessary step in big data analysis. We remove taxi traces whose length is less than 500 m and more than 30 km or travel time less than 2 mins.

5 Spatiotemporal study and pattern analysis

For the purpose of analysis, Qingdao urban map is divided into cells of 0.5×0.5 km^2. To estimate the OD flows, we count the quantity of taxi traces from position L_i to position L_j. The quantity of taxi traces c_{ij} can be approximated as OD flow between position L_i and position L_j. Through statistical analysis, we found that c_{ij} is rather uneven. Statistical analysis indicates that most of human behavioral activities by taxi can be reflected by OD flows. The quantity of OD flows whose c_{ij} value is more than1000 per month is 237, and the quantity of grids bound up with those 237 OD flows is 75. We think that they can represent typical human behavior by taxi.

We use the 75 location grids as nodes and those 237 OD flows as edges to build a complex network, which is shown as Fig. 2.

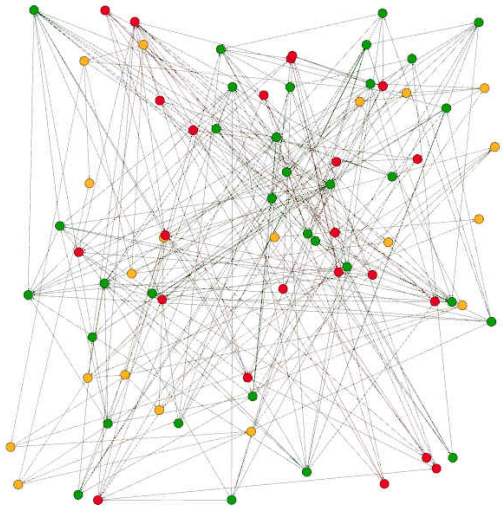

Figure 2: The complex network of OD flows

For the complex network, we use Mapping Vertex into Vector algorithm to detect community structure. Nodes of the complex network are divided into three communities (green grids, red grids, orange grids) as shown in Fig. 3.

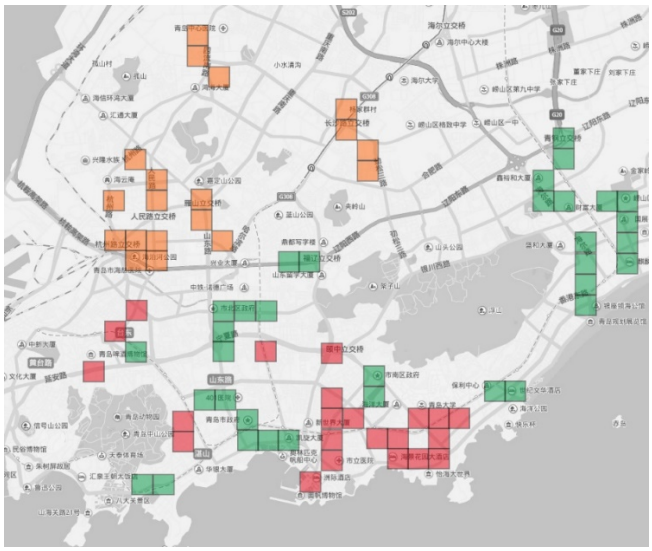

Figure 3: Distribution of grids belonged to three communities in Qingdao urban map

For better understanding OD flows and identifying emerging patterns, then we explore spatial and temporal distribution of OD flows.

According to the LADEN/UNLADEN STATE and TIME in source dataset, we can get taxi demands variation trend varying time. The taxi demands with hours in a day is shown in Fig. 4.

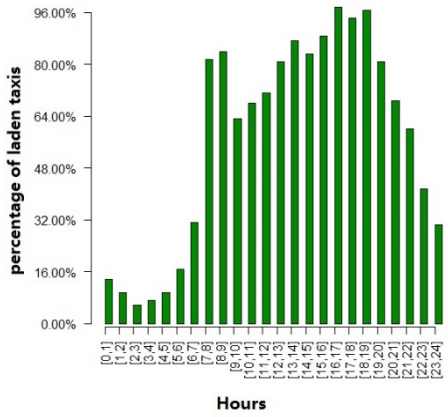

Figure 4: Percentage of laden taxis according to the hours of day

As expected, the percentage of laden taxis varies with working hours. It begins to increase sharply from 7:00, it will gradually reach peak value between17:00 and 19:00, then it will slowly fall back at night.

Percentage of taxi traces over time of the day and over weekday and weekend are individually shown in Fig. 5 and Fig. 6.

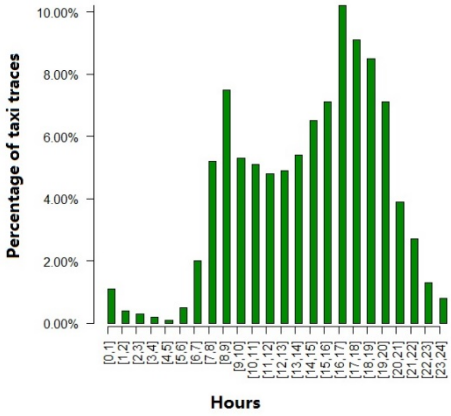

Figure 5: Percentage of taxi traces over time of the day

From Fig. 5 we can see that the percentage of taxi traces over time of the day also follows the business hours, time interval from 7 a.m. to 8 a.m., and from 4 p.m. to 5 p.m. form two peaks. The result is basically consistent with laden taxis variation.

From Fig. 6 we can see that there are more taxis carrying passengers on weekdays than on weekends.

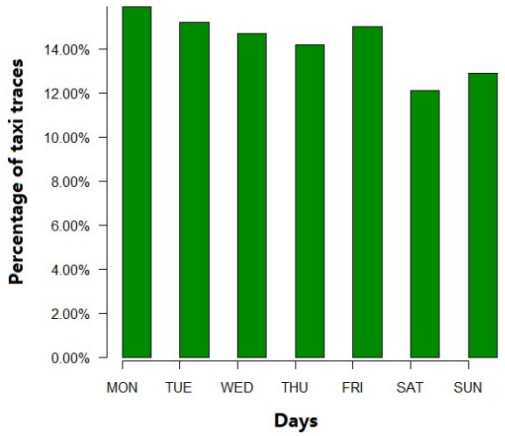

Figure 6: Percentage of taxi traces over time of weekday and weekend

We use vertex in-degree and out-degree of complex networks [Barabási and Albert (1999)] to identify some major locations. The top-10 largest in-degree and out-degree of grid locations is shown in Fig. 7 and Fig. 8.

Figure 7: The top-10 largest in-degree of grid locations

Fig. 7 presents the major locations of taxi drop-offs distribution in Qingdao, these locations mainly includes downtown (C, G, H), hospitals (A, E, J), governments (B, F, I) and university (D).

Figure 8: The top-10 largest out-degree of grid locations

Fig. 8 presents the major locations of taxi pick-ups distribution in Qingdao, these locations mainly include Central Business Districts and large residential districts.

The related stopping grid positions are shown as Fig. 9, where the thickness of links stands for intensity between two grid positions.

Figure 9: The related taxi stopping positions

6 OD flows semantics mining method

POIs are grouped into seven categories including downtown, education, health facilities, public transport hub, central business districts, governments and residential district.

Percentage of POIs Categories is shown as Fig. 10.

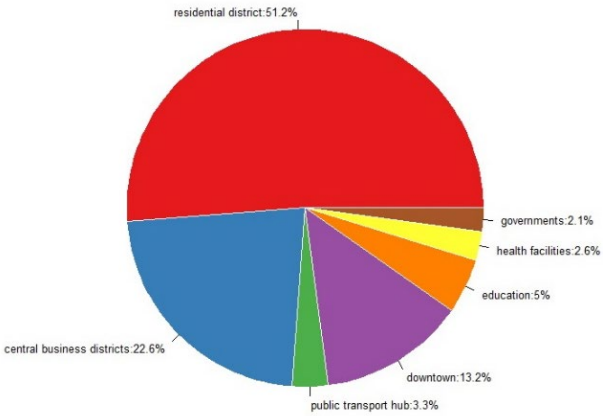

Figure 10: Percentage of POIs Categories

Fig. 11 shows the POI distribution for distinguishing the main POI on each position grid.

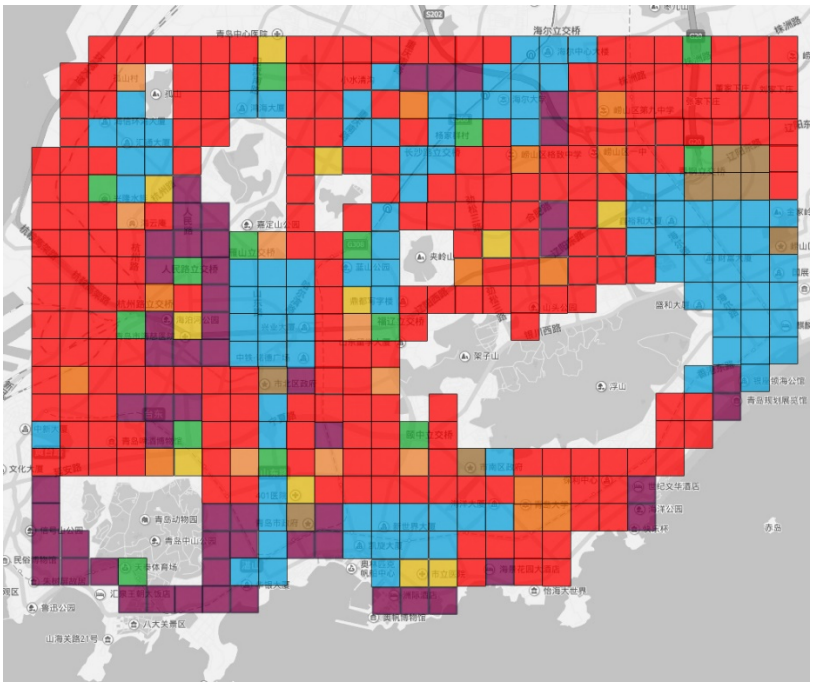

Figure 11: Predominant POI category on each location grid

We use multi-subnet composited complex network model to compound OD flows network and POI network. Then the semantics of OD flow is defined by the semantics of its starting position grid and ending position grid, such as residential district to public transport hub or central business districts to governments. Through topological analytics of the composited complex network, the quantity of OD flows with each kind of semantics is shown in Tab. 2.

Table 2: Quantity of each semantics

Semantics Type	Number
residential district to health facilities	67
residential district to education	43
downtown to residential district	32
health facilities to residential district	27
central business districts to downtown	17
residential district to central business districts	12
others	39
total	237

We select 3 representative semantics to explore their relations with behavioral patterns.

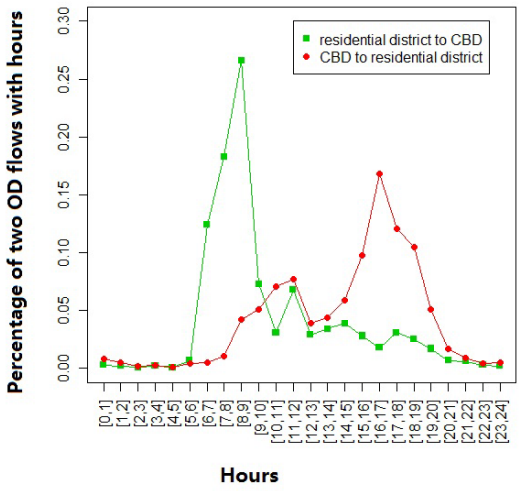

Figure 12: Percentage of OD flow from residential district to central business districts and OD flow from central business districts to residential district

From Fig. 12 we can see that the OD flow from residential district to central business districts has a peak from 8:00 a.m. to 9:00 a.m. and the OD flow from central business districts to residential district has a peak value from 16:00 to 17:00. The two patterns are in accordance with daily behavior experience which people go to work in the morning and return home in the evening.

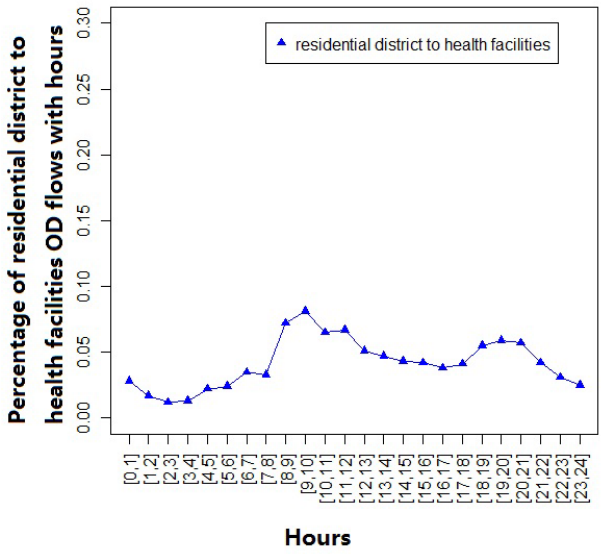

Figure 13: Percentage of residential district to health facilities OD flows

From Fig. 13 we can see that residential district to health facilities OD flow is flat distributed in day-time. It means that there are no peaks for some OD flows. Based on the analysis mentioned above, the percentage of OD flows with hours distribution can be acted as their feature to identify these OD flows. So, we could explain each OD flow by using a feature vector S_d.

$S_d = (c_{ij}^1, c_{ij}^2, ..., c_{ij}^{24}) / c_{ij}$ represents the percentage of OD flows with hour in a day. c_{ij}^k represents the quantity of taxi traces in *k*-th hour, c_{ij} represents total quantity of taxi traces.

So, in like manner, we could explain each OD flow with another feature vector W_d.

$W_d = (S_d^{Mon}, S_d^{Tue}, S_d^{Wed}, S_d^{Thu}, S_d^{Fri}, S_d^{Sat}, S_d^{Sun})$ represents the percentage of OD flows over time of week. S_d^{Mon} represents S_d on Monday.

We have divided the primary 75 location grids into three communities, there are dense OD flows in the same community, and there are sparse OD flows between two communities. To analyze the empirical observation, we use multi-subnet composited complex network again to compound public transportation network to the former composited network. The public transport network consists of 873 bus station nodes and 1522 lines between bus stations, its topology is shown as Fig. 14.

Figure 14: The complex network of Qingdao public transport

We found that the more there are public transport lines between two grid locations, the less there are OD flows between them. Distance is not the most important factor of OD flows. An example is shown in Fig. 15.

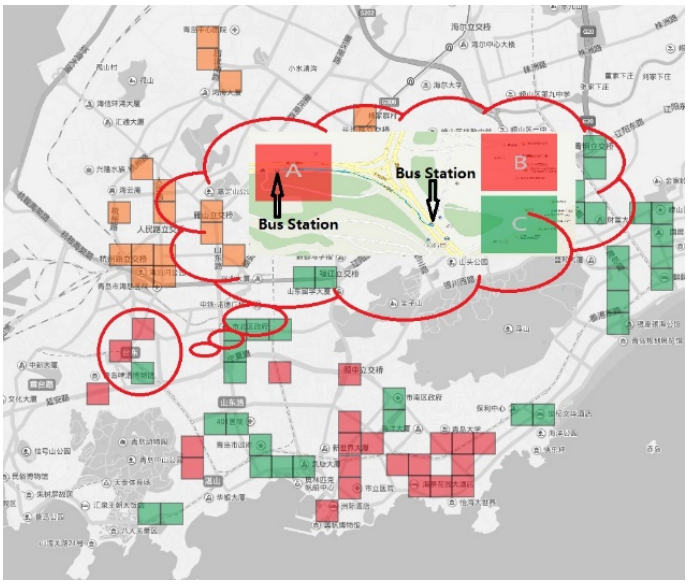

Figure 15: Relationship between OD flow and public transport line

From Fig. 15 we can see that the distance of grid A-grid B and he distance of grid A-grid C are almost the same, but there are much more OD flows between grid A and grid B than them between grid A and grid C. It is because that there are public transport stations nearby

grid A and grid C. So taking public transport into consideration, it will mine better the semantics of OD flows.

We use an improved Support Vector Machine [Fung and Mangasarian (2005)] to classify above defined feature vectors. Our experimental dataset is actual taxi trajectory data of Qingdao. The actual dataset is stochastically divided into three subsets, train set accounts for 70%, validation accounts for 20% and test set accounts for 10%. The results are limited to several semantic types shown in Tab. 2. The classification process is run 100 times and the accurate rate is shown in Tab. 3.

Table 3: The predictive accuracy for each type of feature vectors

Feature Vector Types	Predictive Accuracy
S_d	87.7%
W_d	84.6%

7 Conclusion

In this paper, our research pays close attention to the OD flows from taxi-GPS traces and understands crowd movement. Through data gathered in Qingdao, China, the distinctive human behavioral patterns which closely related with OD flows are found. Then, a semantics mining method of OD flows is proposed through compounding Points Of Interests (POI) network and public transport network to OD flows network. Experimental results show that we can mine more accurate unknown rules based on the method.

Future work includes being able to accurately predict taxi flow, comparing pattern of OD flow under different conditions, and suggesting for urban traffic planning.

Acknowledgement: This work is supported by Shandong Provincial Natural Science Foundation, China under Grant No. ZR2017MG011. This work is also supported by Key Research and Development Program in Shandong Provincial (2017GGX90103).

References

Bin, S.; Sun G. X. (2011): An algorithm for detecting community structure of complex networks based on clustering. *International Journal of Digital Content Technology and Its Applications*, vol. 5, no. 7, pp. 326-334.

Barabási, A. L.; Albert, R. (1999): Emergence of scaling in random networks. *Science*, vol. 286, no. 5439, pp. 509-512.

Chang, H.; Tai, Y.; Hsu, Y. (2009): Context-aware taxi demand hotspots prediction. *International Journal of Business Intelligence and Data Mining*, vol. 5, no. 1, pp. 3-18.

Colizza, V.; Barrat, A.; Barthelemy, M.; Valleron, A.; Vespignani, A. (2007): Modeling the worldwide spread of pandemic influenza: baseline case and containment interventions. *PLoS Medicine*, vol. 4, no. 1, pp. 95-110.

Fung, G. M.; Mangasarian, O. (2005): Multicategory proximal support vector machine classifiers. *Machine learning*, vol. 59, no. 2, pp. 77-97.

Han, S.; Dai, P.; Paritosh, P.; Huynh, D. (2016): Crowdsourcing human annotation on web page structure. *ACM Transactions on Intelligent Systems and Technology*, vol. 7 no. 4, pp. 1-25.

Hufnagel, L.; Brockmann, D.; Geisel, T. (2004): Forecast and control of epidemics in a globalized world. *Proceedings of the National Academy of Sciences*, vol. 101, no. 42, pp. 15124-15129.

Horner, M. W.; O"Kelly, M. E. (2001): Embedding economies of scale concepts for hub network design. *Journal of Transport Geography*, vol. 9, no. 4, pp. 255-265.

Kitamura, R.; Chen, C.; Pendyala, R. M.; arayanan, R. (2000): Micro-simulation of daily activity-travel patterns for travel demand forecasting. *Transportation*, vol. 27, no. 1, pp. 25-51.

Lakhina, A.; Mark, C.; Christophe, D. (2005): Mining anomalies using traffic feature distributions. *ACM SIGCOMM Computer Communication Review*, vol. 35, no. 4, pp. 217-228.

Li, Y.; Li, J.; Chen, J.; Lu, M. (2018): Seed selection for data offloading based on social and interest graphs. *Computers, Materials & Continua*, vol. 57, no. 3, pp. 571-587.

Li, Y.; Wu, D.; Xu, J.; Choi, B.; Su, W. (2014): Spatial-aware interest group queries in location-based social networks. *Data & Knowledge Engineering*, vol. 92, no. 1, pp. 20-38.

Pan, G.; Xu, Y.; Wu, Z.; Li, S.; Yang, L. et al. (2011): Takshadow: toward seamless task migration across smart environments. *IEEE Intelligent Systems*, vol. 26, no. 3, pp. 50-57.

Shao, F. J.; Sui, Y. (2014): Reorganizations of complex networks: compounding and reducing. *International Journal of Modern Physics C*, vol. 25, no. 5, pp. 112-124.

Song, Y.; Hu, Z.; Leng, X.; Tian, H.; Yang K. (2015): Friendship influence on mobile behavior of location based social network users. *Journal of Communications and Networks*, vol. 17, no. 2, pp. 126-132.

Sun, G. X.; Bin, S. (2018): A new opinion leaders detecting algorithm in multi-relationship online social networks. *Multimedia Tools and Applications*, vol. 77, no. 4, pp. 4295-4307.

Sun, G. X.; Bin, S. (2017): Router-level internet topology evolution model based on multi-subnet composited complex network model. *Journal of Internet Technology*, vol. 18, no. 6, pp. 1275-1283.

Wu, Z.; Wu, Q.; Cheng, H.; Pan, G.; Zhao, M. et al. (2007): Scudware: a semantic and adaptive middleware platform for smart vehicle space. *IEEE Transactions on Intelligent Transportation Systems*, vol. 8, no. 1, pp. 121-132.

Ying, J. C.; Kuo, W. N.; Tseng, V. S.; Lu, H. C. (2014): Mining user check-in behavior with a random walk for urban point-of-interest recommendations. *ACM Transactions on Intelligent Systems and Technology*, vol. 5, no. 3, pp. 1-26.

Yuan, J.; Zheng, Y.; Xie, X.; Sun, G. (2012): T-drive: enhancing driving directions with taxi drivers. *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, no. 1, pp. 220-232.

Zhang, D.; Sun, L.; Li, B.; Chen, C.; Pan, G.; Li, S. (2015): Understanding taxi service strategies from taxi gps traces. *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 1, pp. 123-135.

Zheng, Y.; Capra, L.; Wolfson, O.; Yang, H. (2014): Urban computing: concepts, methodologies, and applications. *ACM Transactions on Intelligent Systems and Technology*, vol. 5, no. 3, pp. 38-55.

Zhong, C.; Huang, X.; Stefan, M. A.; Schmitt, G.; Batty, M. (2014): Inferring building functions from a probabilistic model using public transportation data. *Computers Environment & Urban Systems*, vol. 48, no. 6, pp. 124-137.