Spatial-Temporal Distribution of Mobile Traffic and Base Station Clustering Based on Urban Function in Cellular Networks

Tong Wang(✉), Xing Zhang, and Wenbo Wang

Wireless Signal Processing and Network Laboratory,
Beijing University of Posts and Telecommunications,
Beijing 100876, People's Republic of China
zqwt199439@bupt.edu.cn

Abstract. With the rapid development of mobile internet, it's essential to understand the spatial-temporal distribution of mobile traffic. Based on the mobile traffic data collected from a large 4G cellular network in northwestern China, this paper presents detailed analyses of the traffic data on base stations in two aspects: (1) spatial-temporal distribution, (2) clustering based on physical context, i.e., urban function. We introduce the concept of traffic density to measure the traffic level, according to the Voronoi diagram to partition the covering area of BSs. Both spatial and temporal dimensions show distinct inhomogeneity property of mobile traffic. Furthermore, we cluster BSs utilizing urban function information, which enables us to identify and label base stations. The diverse application usage patterns of each cluster of BSs are obtained, which could be applied in resource cache policy and BS loading allocation.

Keywords: Spatial-temporal distribution · Mobile traffic
BS clustering · Urban function · Application usage pattern

1 Introduction

Global mobile data traffic has reached 7.2 exabytes per month at the end of 2016. And with the increasing scale of the cellular network, the fourth-generation (4G) traffic has increased up to 69% of mobile traffic [1]. Hence, from the network operators' point of view, to implement high-efficiency network planning and intelligent base station sleeping mechanisms are now in urgent need, which are key elements of establishing a green network.

For each single base station (BS), it has a specific covering area determined by its configuration and surrounding BSs, making it quite difficult for us to find a specific traffic usage pattern. As a result, through learning the spatial-temporal distribution of traffic, we understand the network traffic distribution, resource consumption and so on, thus making it easier for the resource allocation and efficiency promotion.

C. Li and S. Mao (Eds.): WiCON 2017, LNICST 230, pp. 300–312, 2018.
https://doi.org/10.1007/978-3-319-90802-1_26

Models of spatial-temporal patterns and traffic load of BSs have both been studied in various existing works. For the spatial distribution of traffic load, it has been studied respectively for 2G [2] and 3G cellular networks [3]. Among those studies, applying log-normal distribution is the most common method to characterize the spatial traffic variations. But few papers study the spatial-temporal distribution of 4G cellular network traffic, and they often neglect the factor of covering area of the BSs. Besides, existing literatures focus on behavior modeling and prediction mainly from the perspective of users. However, those works don't focus on the BSs from application usage perspective, which is more important for network planning and optimizing. Cranshaw *et al.* [4] proposed an approach to discover urban functional regions based on the check-in data in FourSquare. However, the penetration rate of FourSquare is very low in China.

Based on the above-mentioned limitations of the related works, we utilize the 4G network traffic data and analyze the integral BSs of a metropolis from both temporal and spatial dimension. Furthermore, we cluster BSs utilizing the data of Points of Interest (PoIs) [5] so as to get the application usage patterns. The main contributions of our work are summarized as follows:

- **Both temporal and spatial distribution of traffic are analyzed in 4G cellular networks:** We utilize the 4G HTTP data and analyze the BSs of a metropolitan in temporal dimension by comparing the traffic data between weekday and weekend. And in spatial dimension, based on the Voronoi diagram to partition the covering area of BSs, we introduce the concept of traffic density to measure the traffic level.
- **The PoIs data are used to depict the physical context of BSs:** By crawling PoIs data through the Amap's application programming interface (API), We depict the BSs using a PoIs vector and cluster BSs by applying K-means algorithm, which is much easier to be generalized to other cities.
- **Diverse application usage patterns of each cl uster of BSs are obtained through horizontal and vertical comparison:** With a view to the heterogeneous usage of BSs on different applications, we employ the horizontal and vertical comparison to mine the diverse application usage patterns and verify our clustering results.

The rest of the paper is organized as follows. Section 2 introduces our dataset, details of application catalogues and data preprocessing. In Sect. 3, we analyze the distribution of traffic in temporal and spatial dimension. In Sect. 4, we cluster BSs using the PoIs and analyze the application usage patterns of each cluster. Finally, we discuss the conclusions in Sect. 5.

2 Dataset and Preprocessing

2.1 Dataset

In this section, we will introduce the dataset in detail, also including the application catalogues. The collected traffic data come from a large Chinese 4G-LTE

service provider, which covers over 25,000 cells in a northwest province in China. The billions of HTTP records last for 2 days (April 3–4, 2015), spanning from weekday to weekend. The fields 'etac' and 'eci' identify a specific cell.

Moreover, we crawl the PoIs via the Amap's API. As we choose one city (provincial capital) to analyze, the number of PoIs is about 9,600. PoIs are divided totally into 10 class: business, recreation, hospital, hotel, scenic spot, residence, government, education, transportation and company.

2.2 Preprocessing

Application Catalogue Update. Due to the inaccuracy of the application (APP) label, we update it by analyzing the following three fields: 'URL', 'host', 'user agent'. Enough key words are extracted to match more than 85 percent traces. Finally, we categorize the applications to 15 service types: instant messaging (IM), reading, blog, navigation, pic & video, music, app store, game, e-commerce, e-mail, social network, news, download, search and others. Table 1 shows the detailed statistic information about the new application catalogues.

Table 1. The percentage of metrics for each application category.

Application category	Traffic			Packets number			User number	Flow number	Duration
	Uplink	Downlink	Up & down	Uplink	Downlink	Up & down			
Instant messaging	13.50%	9.50%	9.64%	13.84%	11.34%	12.40%	90.92%	17.11%	8.27%
Reading	1.06%	0.51%	0.53%	0.70%	0.57%	0.62%	23.20%	0.95%	1.21%
Blog	1.17%	1.77%	1.75%	1.84%	1.64%	1.72%	19.43%	2.09%	1.63%
Navigation	2.59%	0.42%	0.49%	1.46%	0.96%	1.17%	77.72%	3.07%	3.71%
Pic & video	12.23%	46.06%	44.91%	30.95%	39.52%	35.91%	86.30%	12.92%	16.06%
Music	1.34%	2.59%	2.55%	1.92%	2.38%	2.19%	42.63%	1.06%	1.66%
App store	4.07%	9.08%	8.91%	7.59%	8.39%	8.05%	91.69%	5.26%	4.57%
Game	3.17%	1.12%	1.19%	1.79%	1.43%	1.59%	62.28%	2.89%	2.01%
E-commerce	10.91%	4.45%	4.67%	6.64%	5.32%	5.88%	63.96%	9.04%	12.56%
E-mail	0.24%	0.07%	0.08%	0.10%	0.09%	0.09%	5.95%	0.13%	0.13%
Social network	10.64%	5.31%	5.49%	8.46%	6.34%	7.23%	82.15%	11.23%	12.14%
News	7.95%	3.83%	3.97%	5.67%	4.45%	4.97%	83.92%	8.14%	7.73%
Download	7.90%	9.21%	9.17%	7.25%	8.72%	8.10%	85.19%	5.45%	5.49%
Search	8.95%	2.52%	2.74%	4.30%	3.29%	3.71%	87.26%	6.41%	8.96%
Others	14.28%	3.56%	3.92%	7.48%	5.57%	6.38%	90.92%	17.11%	8.27%

Data Aggregation. To analyze the traffic consumption from the view of BSs, we aggregate the data records to 1-h granularity. Each BS's traffic is characterized by a 48*14 matrix (2 days correspond to 48 h, and application numbers are 14 without regard to the unmatched traces).

3 Spatial-Temporal Traffic Analysis

In this section, we will describe our analysis result for mobile traffic on BSs from the temporal and spatial dimension using the 4G traffic data of a metropolis including downtown, suburb and subordinate rural area.

3.1 Temporal Dimension

As for the temporal dimension, we compare the traffic of different time slots in a day, along with the same time slot between weekday and weekend. To reflect the objective law, we choose five typical time slots, i.e., midnight (2:00–3:00), morning (7:00–8:00), noon (11:00–12:00), afternoon (17:00–18:00), evening (20:00–21:00), and aggregate the traffic for each BS and each time slot.

Figure 1 shows empirical CDFs of the traffic on all BSs. The 5 solid lines represent traffic on weekday, and the 5 dotted lines represent traffic on weekend. Blue lines on the top show that most of the BSs consume less traffic during midnight. However, dotted blue line is a little low, which means that BSs consume more traffic during midnight on weekend than that on weekday and we conclude that users tend to sleep late. But in the morning, BSs become more active during weekday. And among those 5 time slots, BSs consume more traffic during noon and afternoon than other time slots, which may be led by the behaviors of users to eat lunch and get off work. In summary, those results conform with people's daily activities and can be used to optimize the traffic resource allocation between BSs on temporal dimension.

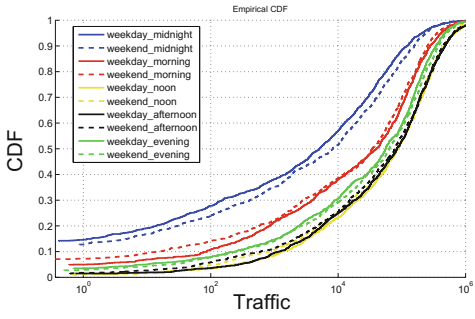

Fig. 1. The distribution of traffic on BSs. (Color figure online)

3.2 Spatial Dimension

Voronoi. While the BSs have a specific geographic location which could be identified by longitude and latitude, we use a software to transform them into Cartesian coordinate system and use x and y value to represent the BSs' position. Moreover, due to the complex practical environment such as the topography and

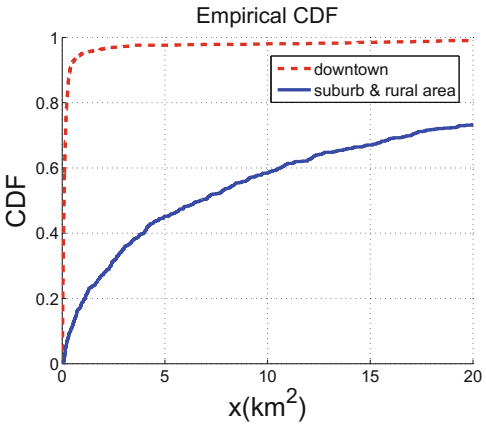

Fig. 2. The distribution of covering area.

buildings' distribution, the covering area of BSs are quite anomalous. To simplify the schematic diagram, we use Voronoi [6] to divide the covering area.

Utilizing the API of Amap, we divide the BSs of one city into two groups: (1) downtown, (2) suburb & rural area. Figure 2 evidently shows that the covering area of BSs in downtown are remarkable smaller than that in suburb & rural area. That means, the developed area has a higher BS density than developing area.

In order to combine the covering area and mobile traffic, we define the traffic density as the ratio of traffic and covering acreage to reflect the traffic consumption level [7]. Figure 3 shows a scatter diagram of traffic density versus BS covering area (including BS in both downtown and suburb & rural area). Traffic per BS and BS covering area exhibit some degree of negative correlations. The Spearman's correlation coefficient and the Pearson's correlation are -0.5208 and -0.0947, respectively. We check correlations between them by using a linear regression function to fit the dots and get a rough relation:

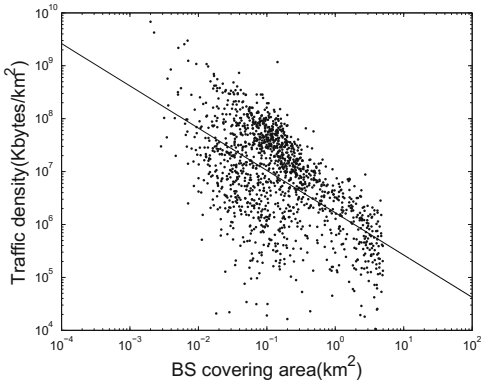

Fig. 3. A scatter diagram of traffic density versus BS covering area.

$$\log_d \rho = -0.7989 * \log_d s + 6.2231 \tag{1}$$

where ρ represents the BS traffic density and s represents the BS covering acreage. We conclude that the number of deployed BSs and BS locations are closely affected by the traffic density of a specific area. So in reality, when deploying a new BS, the network operators will take the traffic density of this area into account.

Traffic Density Distribution. Figure 4 shows the empirical CDFs of the traffic density and its fitting with log-normal, Weibull, gamma and exponential distributions. For both downtown and suburb & rural area, the log-normal and Weibull distribution show a better fitting result. We introduce K-S test [8] to measure the fitting performance and the test results are shown in Table 2. The last column 'cv' represents the critical value of the test. The Weibull distribution is accepted at the 5% significance level for downtown ($0.0406 < 0.0407$), while the log-normal distribution is accepted for suburb & rural area ($0.0316 < 0.0773$).

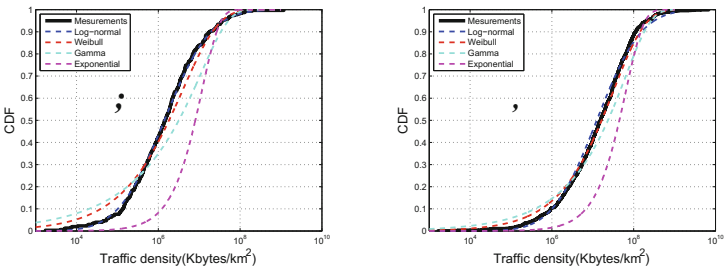

Fig. 4. The distribution of traffic density in (a) downtown, (b) suburb & rural area.

Table 2. Fitting result in k-s test.

	Log-normal	Weibull	Gamma	Exponential	cv
Downtown	0.0488	0.0406	0.1141	0.2939	0.0407
Suburb & rural area	0.0316	0.0847	0.1878	0.4365	0.0773

4 BS Clustering and Application Usage Patterns

In this section, we combine the cellular network traffic data with the PoIs data crawled from Amap's API to cluster BSs. Besides, in view of the advantage that traffic data contain APP types, horizontal and vertical comparison are applied to help us understand the APP patterns existing in different BS clusters.

4.1 Methodology

Normalization Utilizing TF-IDF. The traffic pattern of BSs largely depends on the area they cover, so we utilize the PoIs data to cluster BSs and allocate labels. The PoIs are classified into 10 class: business, recreation, hospital, hotel, scenic spot, residence, government, education, transportation and company. By counting the PoIs in the area the BS covers, we get a 1*10 PoI vector to characterize each BS: $X_i = \left[p_i^1, p_i^2, ..., p_i^{10}\right]$. In consideration that the numbers of different PoI types vary considerably, term frequency inverse document frequency (TF-IDF) [9] is used as a method to balance the weights between those PoI categories. TF-IDF is a non-linear transformation, which is widely used in document classification to reflect how important a word is to a document. It's the product of term frequency (TF) and inverse document frequency (IDF). In our research, it's used to measure the importance of a specific PoI type to the BS. Specifically, for a BS $i \in \{1, 2, \ldots, N\}$ and a PoI type $j \in \{1, 2, \ldots, 10\}$, the TF-IDF value (F_i^j) can be calculated as follows:

$$F_i^j = \mathrm{TF}_i^j \cdot \mathrm{IDF}_i^j \tag{2}$$

$$\mathrm{TF}_i^j = p_i^j / \sum_1^{10} p_i \tag{3}$$

$$\mathrm{IDF}_i^j = \log_d(N/n) \tag{4}$$

where N is the total number of BSs and n is the number of BSs that have the PoI type i in their area ($p_n^i \neq 0$). Multiply the TF and IDF factor, the final TF-IDF vector can be described as: $F_i = \left[F_i^1, F_i^2, ..., F_i^{10}\right]$.

K-Means++ Clustering. As for the clustering algorithm, we cluster the BSs based on the classic K-means algorithm. Taking into consideration of relevance of 10 kinds of PoIs, Pearson correlation is adopted as the distance depiction. Due to the randomness of initial seeds selection in standard K-Means algorithm, we use K-Means++ [10] instead and set 10 linearly independent vectors as initial cluster centers. Table 3 shows the BS clustering result when k = 10. Through observing the mean TF-IDF value, we conclude that each cluster has a unique dominant PoI type. In Fig. 5, each color represents a specific BS cluster, which reveals that the BS clusters don't have aggregation effect but truly contain some adjacent BSs. This clustering method based on the urban function could be applied to many researches. Take traffic prediction for example, when building the prediction model, the label of BSs should be taken into consideration, thus models with various parameters are built for different clusters to enhance the accuracy.

Table 3. The BS clustering result.

BS label	BS number	PoI type (TF-IDF value)									
		Business	Recreation	Hospital	Hotel	Scenic spot	Residence	Government	Education	Transportation	Company
Business	130	**0.3137**	0.0007	0.0233	0.0126	0	0.0546	0.0269	0.0362	0.0019	0.0396
Recreation	19	0.0586	**0.5277**	0.0313	0.0130	0.0146	0.0389	0.0273	0.0637	0	0.0437
Hospital	109	0.0146	0.0017	**0.2812**	0.0136	0.0014	0.0456	0.0270	0.0535	0	0.0255
Hotel	101	0.0349	0.0035	0.0261	**0.3081**	0.0066	0.0540	0.0361	0.0481	0.0014	0.0290
Scenic spot	44	0.0279	0.0109	0.0439	0.0145	**0.4442**	0.0483	0.0521	0.0378	0	0.0414
Residence	137	0.0094	0	0.0221	0.0078	0.0028	**0.2210**	0.0249	0.0450	0.0013	0.0181
Government	138	0.0133	0	0.0191	0.0160	0.0060	0.0382	**0.2744**	0.0521	0.0014	0.0391
Education	179	0.0069	0	0.0148	0.0064	0.0030	0.0275	0.0153	**0.2135**	0.0008	0.0220
Transportation	27	0.0195	0	0.0260	0.0319	0.0056	0.0327	0.0601	0.0350	**0.6018**	0.0634
Company	116	0.0207	0	0.0099	0.0094	0.0033	0.0335	0.0149	0.0306	0.0009	**0.3316**

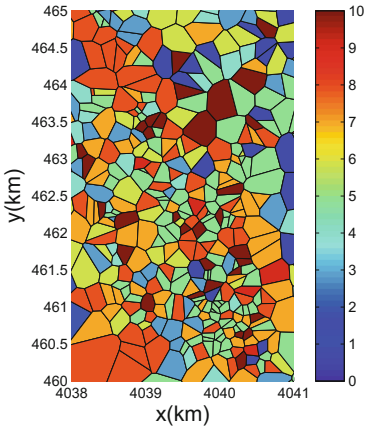

Fig. 5. Voronoi cells with BS clusters padding.

4.2 Traffic Patterns

Entropy of APPs. Just as the Table 1 shows, the proportion of 14 kinds of APPs differ significantly. To characterize the specific distribution of those APPs, we import entropy to measure their diversity, which is defined as below:

$$H(x) = -\sum_{1}^{N} p(x_i) \log_b p(x_i) \tag{5}$$

where N is the number of BSs, each different i represents a specific BS, and x represents an APP category. $p(x_i)$ is the proportion of the APP traffic consumption on one BS. The entropy reaches maximum when the distribution is uniform. Here, we use the normalized entropy as below to compare the different patterns between APP categories.

$$H_{norm}(x) = H(x)/H_{max}(x) \tag{6}$$

$$H_{max}(x) = \log_b N \tag{7}$$

Figure 6 shows the normalized entropy of each APP for every hour in two days. The trend of the entropy for 14 kinds of APP catalogues are approximately

similar, which is high in the day and low during the night, but their average levels exhibit considerable divergence. The entropy of IM is the highest, with a maximum of 0.9042, which means IM is most popularized. But e-mail exhibits a quite low entropy, with a maximum of 0.6992 and a minimum of 0.1926. It reveals that e-mail is mainly used on a small proportion of BSs, more likely in workplaces.

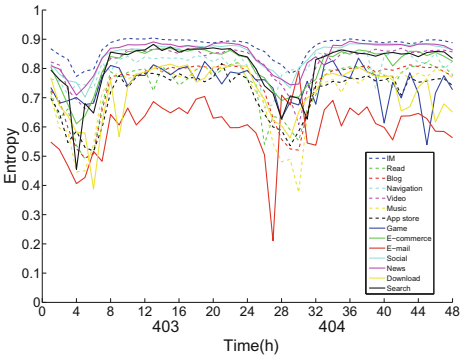

Fig. 6. The entropy of APPs with time varying.

Horizontal Comparison Between Clusters. According to the above result using entropy, the 14 kinds of applications exhibit high difference in mobile traffic and flow numbers due to their natural properties. To break through this limitation, we use horizontal comparison method to observe the traffic patterns of each BS cluster without normalization. Specifically, the overall traffic consumption ratio of each kind of application is calculated as a baseline, then we compare the traffic of a specific application in each cluster based on the baseline. Figure 7 describes the traffic pattern of each cluster, where red pillars are applications exceeding the baseline and blue ones on the contrary. To illustrate them more clear, we set a threshold to filter the histogram. That means, we only take the red ones with value bigger than 0.1 as dominant services and blue ones with value smaller than -0.1 as weak services. Table 4 gathers all the filtered information, some objective law and interesting phenomena are revealed:

- In hotel and scenic spot areas, navigation APPs are dominant. Tourists and the people on a business trip are main users in those areas, and they tend to be unfamiliar with the local traffic information. In those condition, Navigation APPs such as Amap and Didi Taxi are frequently used.
- Cluster shopping doesn't show a specific preference on APPs. It may be because that BSs in shopping areas have a large number of floating population. No regular application usage pattern is acquired.
- Recreation area is mainly constituted by theater, bar and so on. Users are more likely to use entertainment APPs to relax such as games and videos.

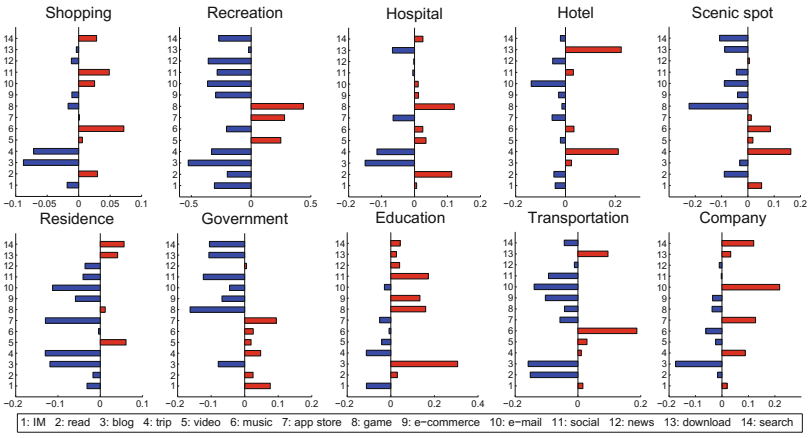

Fig. 7. APP pattern – horizontal comparison result.

- BSs in education area cover many middle school students and undergraduates. They show a preference on blog, game and social network APPs, while social networking and navigation APPs are in a weak position.
- In transportation area such as railway station, bus station and airport, people listen to music more than other places. It shows a popular user behavior about music service.
- Games are played less in the government, while companies use more e-mail service. Both of them show their characteristic of workplaces.

For BS clusters under different urban function, they show various preference on APP usage, which conform to the distribution of urban PoIs. Those above APP patterns answer to popular user behaviors, and can be utilized to help network operators understand the traffic distribution, such as applying some targeted resource cache policy. For example, because of the preference that people tend to listen to music in transportation junctions, we could buffer more music information on BSs in those areas and let the people nearby access to those BSs preferentially when they request music information. In this way, the requests to core network will decrease, which will both save the network energy and promote the network efficiency.

Vertical Comparison Between Weekday and Weekend. In order to understand the APP patterns comprehensively, we also apply the vertical comparison thought. Specifically, the data records on April 3 and 4 are used to represent the traffic on weekday and weekend, and we compare the APP patterns using the ratio of the weekend service traffic to weekday service traffic. In Fig. 8, the APP patterns acquired through vertical comparison are showed by aggregating the red pillars and blue pillars separately, which is more intuitive. The whole traffic of cluster shopping, recreation, residence, government, transportation grow

Table 4. Filtration result in horizontal comparison.

BS clusters	Dominant	Weak
Shopping	/	/
Recreation	Game, app store, pic & video	Blog, e-mail, navigation
Hospital	Read, game	Blog, navigation
Hotel	Navigation, download	e-mail
Scenic spot	Navigation	Game, search
Residence	/	App store, navigation, blog
Government	IM, navigation	Game, social networking, download
Education	Blog, game, social networking	IM, navigation
Transportation	Music	Read, blog, e-mail
Company	Email, app store, search	Blog

rapidly on weekend, with the traffic of cluster hotel, scenic spot and education on the contrary. As for cluster company and transportation, plot the detailed comparison histogram and some interesting phenomena emerge. In company area, the traffic of blog, music and e-commerce increase 21.36%, 8.29% and 9.32% respectively, while the traffic of e-mail reduces more than 22%. The traffic of transportation cluster BSs increases clearly on all APP types, as the traffic on weekend is almost 1.5 to 3 times of that on weekday for every application. It reveals that the number of traveling people has an alarming rise on weekend.

Through vertical comparison between weekday and weekend, we obtain the APP patterns on each BS cluster. Depend on the activities of users in each BS's covering area, the app usage patterns show a certain degree of periodicity. Those would give us an essential cognition of their APP patterns, which could be used to optimize the BS loading with day-varying allocation. Based on this, for a dataset that lasts more than one week, the periodicity of traffic pattern for those clusters could be extracted. And due to the differences of urban planning and economic development degree among cities, more studies are further needed to compare the traffic patterns between cities of varying degrees of development.

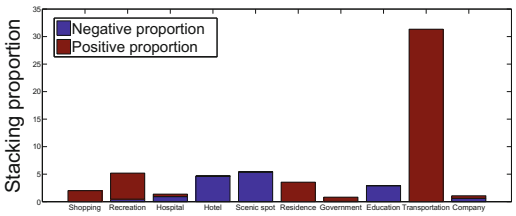

Fig. 8. APP pattern – vertical comparison result.

5 Conclusion

In this paper, through utilizing the mobile traffic data collected from 4G cellular networks in northwestern China, we comprehensively study the temporal and spatial distribution of mobile traffic data on BSs. The Voronoi diagram and traffic density are imported in order to make the results more clearly. The result shows quite an obvious heterogeneous distribution of traffic data on both temporal and spatial dimension. Based on the Voronoi diagram, we combine the traffic data with the PoIs data crawled from Amap to cluster BSs into 10 groups, which give each BS a specific label. The methodology of combining datasets could be applied to many research. And the APP patterns obtained by horizontal and vertical comparison are very practical and meaningful in resource cache policy and BS loading allocation.

We are continuing our research focusing on various interesting and practical dimensions, including predicting the traffic trend of BSs and introducing various complex network theories. We hope our research result inspire you for further research in this area.

Acknowledgements. This work is supported by the National Science Foundation of China (NSFC) under grant 61571054, 61771065 and 61631005, by the New Star in Science and Technology of Beijing Municipal Science and Technology Commission (Beijing Nova Program: Z151100000315077).

References

1. Cisco visual networking index: Global Mobile Data Traffic Forecast Update, 2016–2021. https://www.cisco.com
2. Gotzner, U., Rathgeber, R.: Spatial trac distribution in cellular networks. In: Vehicular Technology Conference, VTC 1998, Ottawa, vol. 3, pp. 1994–1998 (1998)
3. Paul, U., Subramanian, A.P., Buddhikot, M.M., Das, S.R.: Understanding traffic dynamics in cellular data networks. In: 2011 Proceedings IEEE INFOCOM, Shanghai, pp. 882–890 (2011)
4. Cranshaw, J., Schwartz, R., Hong, J., Sadeh, N.: The livehoods project: utilizing social media to understand the dynamics of a city. Social Science Electronic Publishing (2012)
5. Xu, F., Zhang, P., Li, Y.: Context-aware real-time population estimation for metropolis. In: ACM International Joint Conference on Pervasive and Ubiquitous Computing ACM, pp. 1064–1075 (2016)
6. Guruprasad, K.R.: Generalized Voronoi partition: a new tool for optimal placement of base stations. In: 2011 Fifth IEEE International Conference on Advanced Telecommunication Systems and Networks (ANTS), Bangalore, pp. 1–3 (2011)
7. Zhou, S., Lee, D., Leng, B., Zhou, X., Zhang, H., Niu, Z.: On the spatial distribution of base stations and its relation to the traffic density in cellular networks. IEEE Access **3**, 998–1010 (2015)
8. Woodruff, B.W., Moore, A.H., Dunne, E.J., Cortes, R.: A modified Kolmogorov-Smirnov test for Weibull distributions with unknown location and scale parameters. IEEE Trans. Reliab. **R–32**(2), 209–213 (1983)

9. Leng, B., Liu, J., Pan, H., Zhou, S., Niu, Z.: Topic model based behaviour modeling and clustering analysis for wireless network users. In: 2015 21st Asia-Pacific Conference on Communications (APCC), Kyoto, pp. 410–415 (2015)
10. Agarwal, S., Yadav, S., Singh, K.: Notice of violation of IEEE publication principles K-means versus k-means ++ clustering technique. In: 2012 Students Conference on Engineering and Systems, Allahabad, Uttar Pradesh, pp. 1–6 (2012)