Discovering Social Relationship between City Regions using Human Mobility

Ya-Jing Xu[1], Chao Xue[1],Gong-Fu Li[1],An-Gen Luo[1],Yi-Zhe Song[2]

1.School of ICE
Beijing University of Posts and Telecommunications
Beijing, China
memoryse7en@gmail.com

2.School of EECS
Queen Mary University of London
London, British
yizhe.song@qmul.ac.uk

***Abstract*— The development of a city gradually fosters different functional regions, and between these regions there exists different social information due to human activities. In this paper, a Region Activation Entropy Model (RAEM) is proposed to discover the social relations hidden between the regions. Specifically we segment a city into coherent regions according the base station (BS) position and detect the stay and passing regions in trajectories of mobile phone users. We regard one user's trajectory as a short document and take the stay regions in the trajectory as words, so that we can use Natural Language Processing (NLP) method to discover the relations between regions. Furthermore, the Region Activation Force (RAF) is defined to measure the intensity of relationship between regions. By measuring the Region Activation Entropy (RAE) based on RAF, we find an 88% potential predictability in regional mobility. The result generated by RAEM can benefit a variety of applications, including city planning, location choosing for a business and predicting the spread of human. We evaluated our method using a one-month-long record collected by mobile phone carriers. We believe our findings offer a new perspective on research of human mobility.**

Keywords-region relationship; activation force; entropy; human mobility

I. INTRODUCTION

The advance in location acquisition technologies has generated a myriad of spatial trajectories representing the mobility of various moving objects, such as people and vehicles. Such trajectories offer us unprecedented information to understand moving objects and locations, foster a broad range of applications in location-based social networks [1], intelligent transportation systems and urban computing [2].

Mining human location history has attracted intensive attention in the past years [3][4].Here, we briefly reviews related works on the exploration of correlation between different regions based on human historical trajectories. Yu Zheng et al [5] proposed several efficient metrics to mine the correlation between locations with a large-scale real-world GPS dataset. Giannotti et al. [6] mined similar sequences from users' moving trajectories. They introduced trajectory patterns as concise descriptions of frequent behaviors, in terms of both space and time. Mamoulis et al. [7] proposed a top-down framework with an indexing scheme for retrieving maximum periodic patterns in spatio-temporal data. Jing Yuan et al [8] build a LDA model using the regions' pick up/drop off data to discover the functions of city regions.

Comparing with the co-location pattern mining [9] [10] [11] which aims to find classes of spatial objects that are frequently located together. The major differences between these work and ours lie in two aspects: 1) We address the challenge to infer the correlation between each pair of locations rather than the co-located patterns of location categories. 2) We use human behaviors to estimate the correlation between two locations rather than the geospatial distance between them.

Our key contributions are summarized as follows:

(1) Region Activation Force (RAF) is proposed to quantify the strength of relationship between city regions using human trajectories, which can help us understand the regional mobility rules.

(2) Region Activation Entropy (RAE) based on RAF is defined to calculate the upper bound of the accuracy rate of the prediction using the Fano's inequality [14]. Therefore, we find a 88% potential predictability in regional mobility.

The rest of this paper is organized as follows. In Section 2, some definitions are given. In addition, the activation entropy model is proposed to mining the relationship between regions. In section 3, some experiment results are studied to evaluate the performance of our model. Finally, we briefly conclude the paper and point out future works.

II. SYSTEM MODEL

A. Preliminary

Here, a city is divided into individual regions by the position of each Base Station (BS) using Voronoi diagram (refer to Fig. 1). Compared with the dividing method of road network, a voronoi region is the covers of a cell-tower and there is just

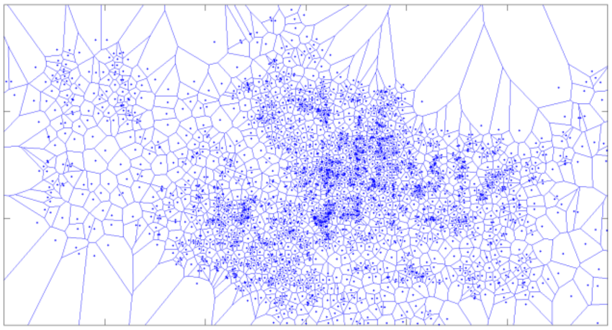

Figure 1. Map segmentation by Voronoi diagram(130km*70km)

QSHINE 2015, August 19-20, Taipei, Taiwan

DOI 10.4108/eai.19-8-2015.2260775

one tower in each region. Thus, a region R_i is defined as the voronoi cell divided by cell-tower i.

In this paper, we aim to discover the relationships between regions by Human mobility. Human mobility is represented by user movement's trajectories, which are the sequences of cell-tower traces given in definition 1.

Definition 1. (Trajectory) Trajectory of a user is a series of regions, $Tr_k^d = \{R_1,\ R_2 \dots R_n\}$, where $R_i = \langle c, t_a, t_l \rangle$; Tr_k^d is the trajectory of user k in the period d, c is the coordinates of the region, that is, the longitude and latitude of the cell tower which user k connected between arriving time t_a and leaving time t_l. Thus, there would be n trajectories of each user if we have a dataset of n periods.

Definition 2. (Stay and Passing Region) R_i is a stay region only if it is a region where a user moved less than a distance threshold D_{th} over a time threshold T_{th}[16], otherwise it is a passing region. Then we label $LR_i = \langle c, t_a, t_l, \alpha \rangle$, where α is the property of region i , $\alpha = 's'$ stands for stay while $\alpha = 'p'$ stands for passing.

For a set of consecutive $\{LR_0, \mathrm{LR}_1,\ LR_2 \dots, \mathrm{LR}_n\}$ in trajectory Tr_k^d , the following criteria are given.

a) $\forall p \le k \le q,\ distance\left(LR_p.c\,, \mathrm{LR}_q.c\right) \le D_{th}$

b) $distance\left(LR_p.c\,, \mathrm{LR}_{q+1}.c\right) > D_{th}$

c) $LR_q.t_l - LR_p.t_a \ge T_{th}$

Then if $\{LR_p, LR_{p+1},\ \dots, LR_q\}$ satisfy the above criteria, the region LR_x which is nearest to the central point would be labeled $'s'$, others labeled $'p'$. For example, as shown in Fig. 2, $\{\ l_1, l_2 \dots l_7\ \}$ is a trajectory, where l_i represents $LR_i.\mathrm{c}$, if $\mathrm{LR}_6.t_l - LR_3.t_a \ge T_{th}$, then LR_5 will be labeled $'s'$ while the others will be labeled $'p'$.

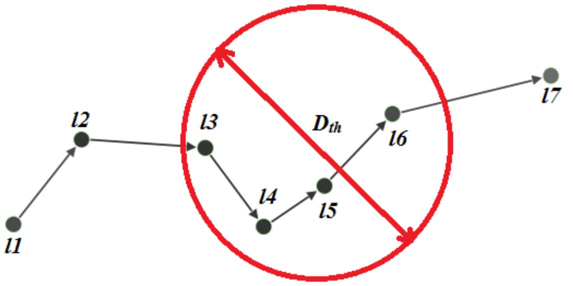

Figure 2. A stay region in a trajectory

Firstly, we analyze the distribution of distances between neighboring regions for all of 2440 towers in Fig. 3(A). Most neighboring regions have a distance less than 1 kilometer and 0.83 kilometer is the average. So 0.8 kilometer is selected as the distance threshold D_{th} in our stay region detection. In addition, it peaked around at d=0.3km, which indicates that the coverage of many cell towers is about 300 meters. Because the cell towers are intensive in the city center.

Fig. 3(B) shows the distribution of stay time by all users in each region: almost all of the stay time are below half an hour. It means the time intervals in which most users remain in one region are very short. Therefore, we choose 30min as the time threshold T_{th}.

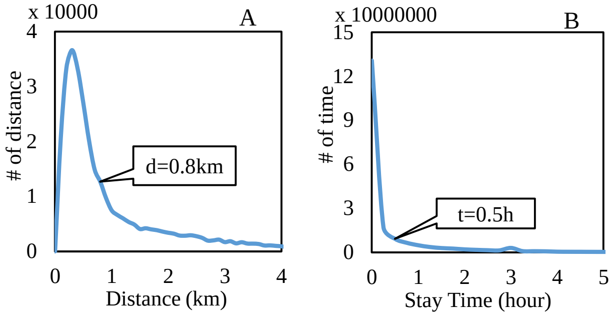

Figure 3. Distribution of (A) distance between neighboring regions (B) stay time of all users

Definition 3. (Labeled Trajectory) A labeled trajectory of user k is $LT_k^d = \{LR_1, LR_2 \dots LR_n\}$, which is a time series of his/her stay and passing regions in the period d.

B. Region Activation Force

The labeled trajectories have extracted in the previous work, we will propose a new method to measure the regions relation in this part.

The activation force [12] is an effective approach to measure the strength of the link between two nodes in complex network, which is commonly used in Natural Language Processing (NLP) to find the relationship between words. In this paper, we use the activation force to calculate the relationship of regions as the definition 4.

Definition 4. (Region Activation Force, RAF) If we regard a trajectory as a short document, then a region could be treated as a word. The RAF from R_i to R_j is defined as

$$raf_{ij} = \frac{f_{ij}}{f_i} \cdot \frac{f_{ij}}{f_j} \cdot \frac{1}{d_{ij}^2}$$

f_i is the frequency of R_i in all trajectories, f_{ij} is the co-occurrences of R_i to R_j in the trajectories where i precedes j by up to L regions (L is called window); RAF from region *i* to itself is useless in our research so we set $f_{ii} = 0$); d_{ij} is the average distance between a pair of regions in their co-occurrences.

Before we start the RAF calculation, one question should be considered: which kind of region in each trajectory should be chosen?

There are two kinds of regions, passing regions and stay regions, in the trajectories. As we know the stay regions represent the sources and the destinations of a user which contains the truly mobility features of regions. That is to say: the purpose of RAF model is to mine where the people in the target region come from, not where they pass by. So only the stay regions in trajectories are selected to calculate RAF.

Another question is what is the proper L to be used in RAF calculation. In the field of Natural Language Processing, word windows L is generally set around 5 since researchers believe that the relation between the words far away is very weak. In RAF model, the amount of stay regions in one user's trajectory is commonly no more than 5, therefore we set L to infinity, that is to say, calculate the RAF with all stay regions.

For example, suppose we have a total of 4 regions (R_1, R_2,

R_3, R_4), and 3 labeled trajectories are captured as follows.

$$LT_1 = \{R_1, R_2, R_3, R_4\}$$

$$LT_2 = \{R_1, R_2, R_1,\}$$

$$LT_3 = \{R_4, R_1, R_3, R_2, R_4\}$$

In order to simplify the calculation process, there are only stay regions in labeled trajectories. Firstly, we can count the frequencies f_i of the 4 regions.

$$f_1 = 4,\ f_2 = 3,\ f_3 = 2,\ f_4 = 3$$

Then co-occurrences matrix $F = \{f_{ij}\}$ and average distance matrix $D = \{d_{ij}\}$ of each pair of regions are calculated.

$$F = \{f_{ij}\} = \begin{bmatrix} 0 & 3 & 2 & 2 \\ 1 & 0 & 1 & 2 \\ 0 & 1 & 0 & 2 \\ 1 & 1 & 1 & 0 \end{bmatrix}$$

$$D = \{d_{ij}\} = \begin{bmatrix} 0 & 4/3 & 3/2 & 3 \\ 1 & 0 & 1 & 3/2 \\ 0 & 1 & 0 & 3/2 \\ 1 & 3 & 2 & 0 \end{bmatrix}$$

Finally, we can get the RAF between these regions:

$$A = \{raf_{ij}\} = \begin{bmatrix} 0 & 0.422 & 0.222 & 0.037 \\ 0.083 & 0 & 0.167 & 0.198 \\ 0 & 0.167 & 0 & 0.296 \\ 0.083 & 0.012 & 0.042 & 0 \end{bmatrix}$$

By this step, we get the RAF matrix $\mathbf{A} = \{raf_{ij}\}$, which record the activation force between each pair of regions. The nonzero elements in the *i*-th row provide the out-links of region *i*, while nonzero elements in the *i*-th column provide its in-links. The asymmetry of matrix **A** can be inferred from the definition 4.

We also try to use Association Rules (AR) to mine the relationship between these regions. AR can capture the support of these pairs of regions with the same sample. Obviously $Support(R_i, R_j) = Support(R_j, R_i)$. Table 1 shows the region relationship captured by the two different methods.

RAF from R_1 to R_2 is strongest since they frequently occur together; $raf_{31} = 0$ because of R_1 never appears after R_3 in all of the labeled trajectories. Naturally, region *i* would have a greater activation force to region *j* if they co-occur more or they are nearer in labeled trajectories.

Table 1.Regions Relationship by AR and RAF

	RAF	*AR*
R_1R_2	0.422	1
R_1R_3	0.222	0.667
R_1R_4	0.037	0.667
R_2R_1	0.083	1
R_2R_3	0.167	0.667
R_2R_4	0.198	0.667
R_3R_1	0	0.667
R_3R_2	0.167	0.667
R_3R_4	0.296	0.667
R_4R_1	0.083	0.667
R_4R_2	0.012	0.667
R_4R_3	0.042	0.667

The relation captured by AR is either 1 or 0.667, because the frequency of occurrence and distance between regions are not considered in this method.

We can infer from Table 1 that the degree of differentiation by RAF is much higher than AR. Consequently, we choose RAF to measure the regions relation in the following research.

C. *Region Activation Entropy Model*

We have obtained the RAF between each pair of regions, but how could we separate the different activate patterns of every region? In this part, we will define three kinds of entropy [15] to measure the uncertainty of the regions that activate the target region.

Definition 5. (Random Entropy) The random entropy of region *i* is $E_r^i = \log n_i$, where n_i is the number of regions with a nonzero activation force to region *i*, characterizing the degree of diversity of where the users in region *i* come from. A region would be more private if it has a lower random entropy.

Definition 6. (Distribution Entropy) The distribution entropy of region *i* is $E_d^i = -\sum_{j=1}^{n} p(j) \log p(j)$, where $p(j) = af_{ji} / \left(\sum_{j=1}^{n} af_{ji}\right)$. It captures the heterogeneity of activation patterns.

Definition 7. (Region Activation Entropy, RAE) The region activation entropy is defined to measure the randomness, which depends on the diverse distributions of RAF in different time intervals. A period can be divided into *m* intervals, then the RAE of region *i* is:

$$RAE_i = \frac{1}{m} \sum_{h=1}^{m} \left[-\sum_{j=1}^{n} \frac{raf_{ji}^h}{\sum_{j=1}^{n} raf_{ji}^h} log \frac{raf_{ji}^h}{\sum_{j=1}^{n} raf_{ji}^h} \right]$$

where raf_{ji}^h represents the RAF form region *j* to region *i* in time interval *h*.

In general, $RAE_i \leq E_d^i \leq E_r^i$.

Fig. 4 shows the distribution of the three kinds of entropy. Here we take the hourly interval in RAE calculation. $\overline{E_r} \approx 6.4$, which indicates that, the number of regions which activate the target region averagely provide 6.4 bit information; in other words, a region possesses around $2^{6.4} \approx 84$ regions which have a nonzero RAF to it. In contrast, $\overline{\text{RAE}} \approx 1.4$, indicates that, on average for per hour, the real uncertainty of a region's activation patterns is only $2^{1.4} \approx 2.64$, less than 3 regions.

Considering the mobility of human, people always return to starting point *S* after visiting his/her destination *D*, thus *D* is not only the in-link but also the out-link of *S*. Therefore, we measure the three kinds of entropy from one region to other regions. As shown in Fig. 5, it is quite similar with the distribution in Fig. 4. Although the RAF matrix is asymmetry, the difference between in-links and out-links is slight, so we only use in-links RAF in the following research.

Entropy indicates the degree of uncertainty in a probability distribution. We cannot predict where people in region *i* come from with an accuracy more than p_i no matter how excellent our predictive algorithm since there is a randomness in the relationship between region i and other regions. The predictability p_i can be evaluated by Fano's inequality [14]. If a region has an entropy E with N regions, p_i is given by $E = -p_i \log p_i - (1 - p_i) \log(1 - p_i) + (1 - p_i) \log(N - 1)$.

We determine p_i separately for each region using the three kinds of entropy. As shown in Fig. 6, random entropy gives the worst predictability, that is, we could hardly predict the stream of people if the amount of related regions is the only information provided to us. By comparison, predictability given by RAE is 0.88 on average, indicates that, a historical record of the daily activation patterns of a region implies a high degree of predictability. In addition, we have also obtained the predictability of distribution entropy. It is widely distributed and peaked at 0.5, which means that the predictability is extremely different between regions.

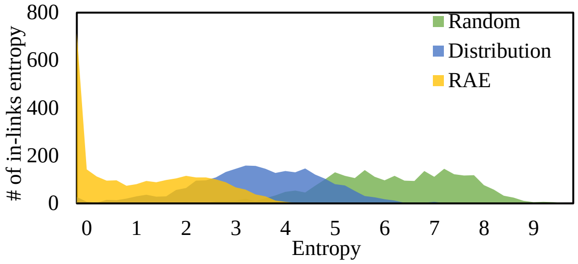

Figure 4. Distribution of in-links entropy.

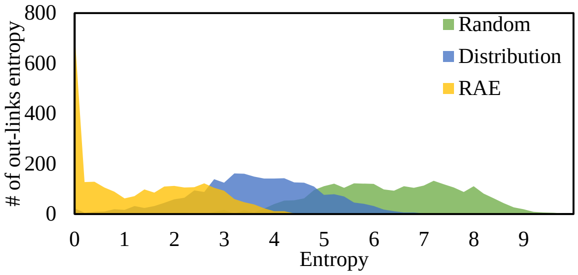

Figure 5. Distribution of out-links entropy.

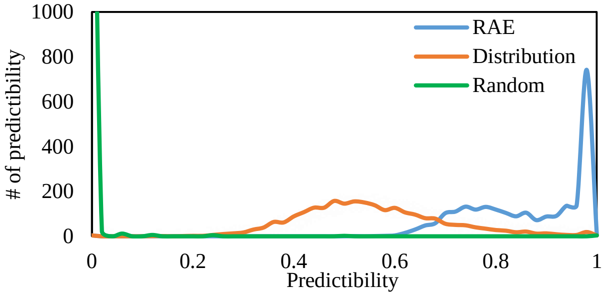

Figure 6. Distribution of predictability across all regions

III. Experiment and Result

In order to evaluate the model mentioned in this article, we conduct a series of experiments. The dataset contains one-month-long records of approximately 4 million anonymized mobile phone users from a medium-sized city with 2440 cell towers. In this experiment, we choose one day as a period d, $D_{th} = 0.8km$ and $T_{th} = 30min$ in stay region detection algorithm. Finally, 65 million stay trajectories are captured into our entropy model.

In our previous work [13], we have classified the city regions into 9 different social functions. Here we will measure the various activation patterns in every function.

Fig.7 plots the various activation distance for 4 kinds of functions, education (A), residential area (B), commercial/ entertainment (C) and governmental agencies/public organizations (D). We can see from the figure that the peak values of all the curves are at about 2 kilometers. The average distance of different function regions are $d_{education} = 3.63$km , $d_{residential} = 3.66$km , $d_{commercial} = 2.95$km and $d_{govenment} = 3.22$km. That is, people tend to visit the places near to them, especially in entertainment areas. The difference of distribution among them is that for education and government regions the curve are multi-peaks and others are single-peak. This is because the locations of government agencies and school are in accordance with the division of administrative areas, but those of commercial districts follows the distribution of population.

Then we choose a sample region (WANDA PLAZA, a commercial place) to display the hourly activation patterns on map in Fig.8. The green cell in the centre of each picture represents the target region, while the other cells in different color stand for the regions possess a nonzero RAF to the target region. The deeper the color is, the greater the RAF is. Before 8 a.m., there is no region activate it. The number of activate regions climbs slowly in the morning, then keeps stable in the afternoon, peaks in the evening and declines rapidly at night.

Fig.9 shows the activation patterns of the other three

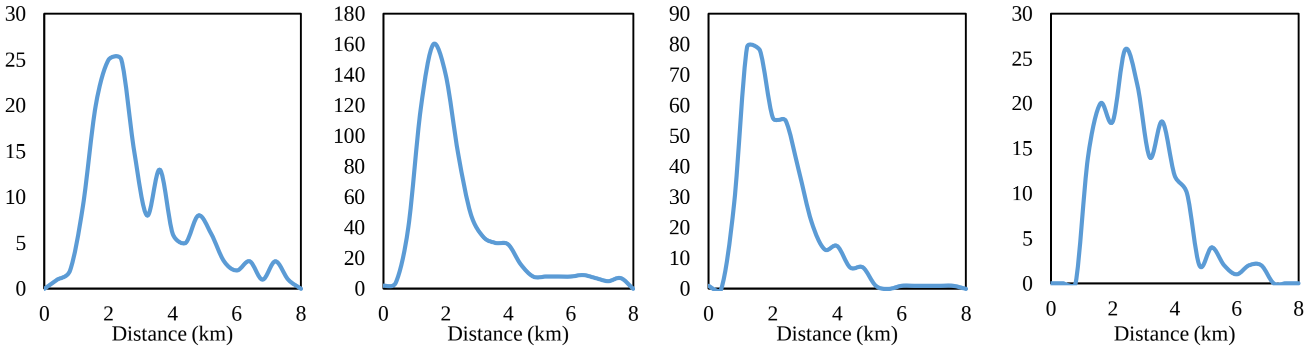

Figure 7. Activation distance of 4 different functions: (A) education, (B) residential area, (C) commercial/entertainment and (D) governmental agencies/ public organizations. The x-axes are the average distance from activation regions to the target region and y-axes are the frequentness of different distance.

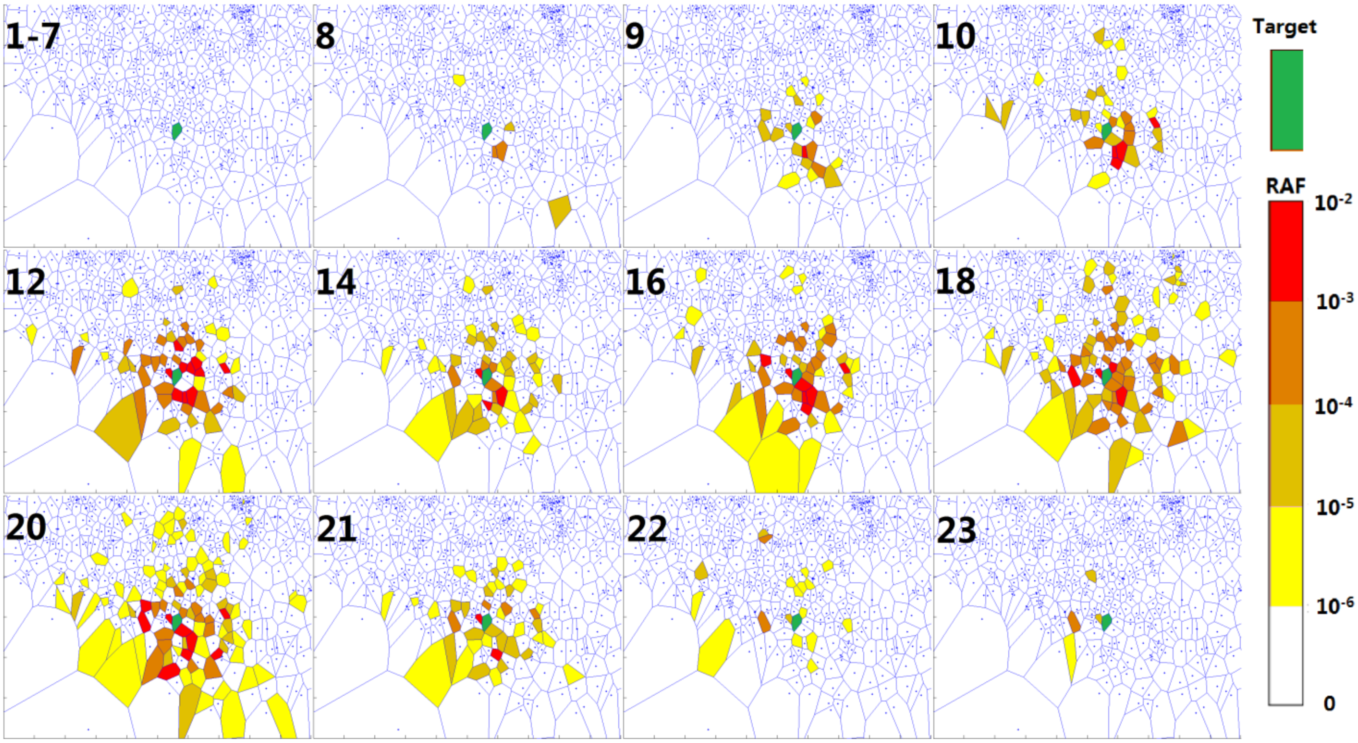

Figure 8.Distribution of RAF to a sample region (WANDA PLAZA, a commercial place) in different hours. The area of each picture is 9km*6km.

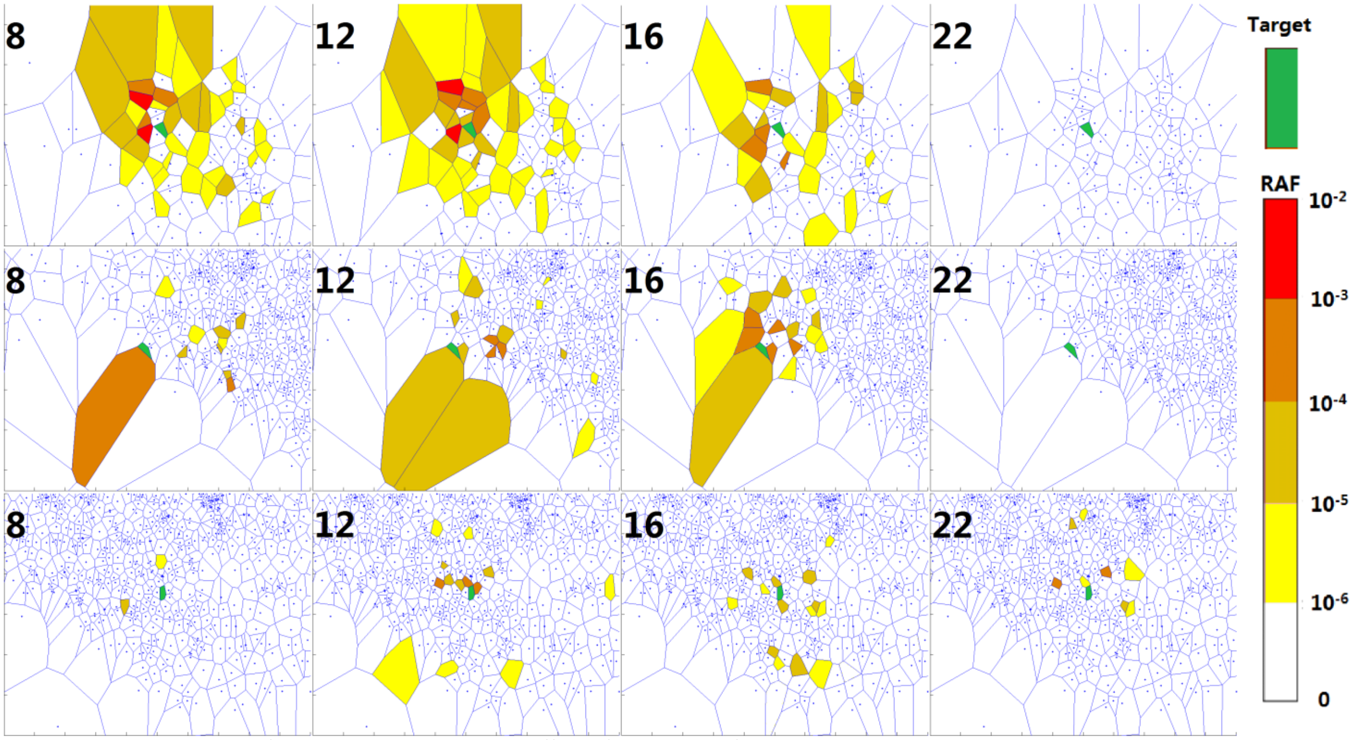

Figure 9. Distribution of RAF to a sample region in 3 different functions. The first row is education, the second is residential area and the last is governmental agencies/public organizations. The area of each picture is 9km*6km.

functions. What the different color represent is same with Fig.8. Residential area have the least activation regions since it possesses the strongest privacy, while the commercial area (Fig.8) is just on the opposite.

The activation patterns in Fig.8 and Figure.9 accord closely to the real world. It indicates that RAF can measure the relations between city regions very well.

IV. Conclusion and Future Work

In this article, we propose a novel method to explore the relevance between city regions by using human mobility.

Furthermore, we use an Activation Entropy Model (AEM) to measure the relevance, give the predictability for each region. At last, we import the social function of the regions to discover the various activation patterns. Researchers might find our result useful in smart city, such as people stream prediction and urban planning.

There are several types of relationship between words in traditional textual syntax, such as subject-predicate, verb-object and modifier-core structure. We only measure the intensity of the relevance between regions without identifying the types like words. In addition, we ignore the difference between in-links and out-links entropy in our work, but there are still part of the regions that have a great gap between the two patterns. It should catch our great attention in the following work.

Acknowledgment

This work is supported by the National Nature Science Foundation of China (NSFC) Grant No. 61511130081, NSFC 61471048 and NSFC 61273217.

References

[1] Y. Zheng. 2011. Location-Based Social Networks: Users. Computing with Spatial Trajectories, Y. Zheng and X. Zhou, eds. Springer, 243-276.

[2] Y. Zheng, L. Capra, O. Wolfson, H. Yang. 2014. Urban Computing: Concepts, Methodologies, and Applications. ACM

[3] González, M. C., Hidalgo, C. A., Understanding individual human mobility patterns. Nature 453 (June 2008), 779-782.

[4] Y. Zheng, Y. Chen, X. Xie and W. Y. Ma, GeoLife2.0: A Location Based Social Networking Service, In Proc. of MDM 2009, pp.357-358.

[5] Yu Zheng, Lizhu Zhang, Xing Xie, Wei-Ying Ma Mining Correlation between Locations Using Human Location History

[6] Giannotti F, Nanni M, Pinelli F, et al. Trajectory pattern mining. In Proc. of KDD'07, pp.

[7] Mamoulis N, Cao H, Kollios G, et al. Mining, Indexing and Querying Historical Spatiotemporal Data. In Proc. of KDD'04, pp. 236-245.

[8] Jing Yuan,Yu Zheng,Xing Xie. Discovering Regions of Different Functions in a City Using Human Mobility and POIs. ACM KDD '12 Beijing, China,2012 ACM 978-1-4503-1462-6 /12/08

[9] Huang, Y., Shekhar, S., and Xiong, H.. Discovering co-location patterns from spatial datasets: A general approach. TKDE, 16(12):1472-1485. 330-339.

[10] Morimoto, Y.. Mining frequent neighboring class sets in spatial databases. In Proc. of SIGKDD, 2001, pp. 353-358.

[11] Vladimir, E. and Lee, I.. Data mining techniques for autonomous exploration of large volumes of geo-referenced crime data. In Proceedings of Geocomputation, 24-26, 2001.

[12] Guo J, Guo H, Wang Z. An activation force-based affinity measure for analyzing complex networks. Scientific reports, 2011, 1.

[13] Y. Xu, G. Li, Integrating Context with Human Mobility Pattern for Improved Region Function Discovering, unpublished.

[14] R. M. Fano, Transmission of Information .the MIT Press and Wiley, New York and London, 1961.

[15] C. Song, Z. Qu, N. Blumm, and A. Barabási. Limits of predictability in human mobility. Science, 2010.

[16] Hariharan R, Toyama K. Project Lachesis: Parsing and Modeling Location Histories, In Proceedings of GIScience, (Park Utah, October 2004), ACM Press: 106-124.