Extracting Land-Use Patterns using Location Data from Smartphones

Kentaro Nishi *
The University of Tokyo
7-3-1, Hongo, Bunkyo-ku,
Tokyo, Japan
nishi@ics.t.u-tokyo.ac.jp

Kota Tsubouchi
Yahoo Japan Corporation
9-7-1, Akasaka, Minato-ku,
Tokyo, Japan
ktsubouc@yahoo-corp.jp

Masamichi Shimosaka
the University of Tokyo
7-3-1, Hongo, Bunkyo-ku,
Tokyo, Japan
simosaka@ics.t.u-tokyo.ac.jp

ABSTRACT

This paper proposes an approach to extract area-by-area and daily land-use patterns using location data obtained from users of Yahoo! Japan's smartphone applications. Information used for extracting patterns is extracted from only location data. In this research, a land-use pattern is defined as how the area is used throughout a day. We extract the land-use patterns based on temporal transition in the number of people in the area. In petterns extraction, a clustering technique with an infinite Gaussian mixture model with Dirichlet process mixtures is used, which can be used to discover the appropriate number of patterns. Experiments were conducted in 34 areas over 56 consecutive days. This means 1,904 conditions were studied. The results of our experiments show that our approach successfully extracts land-use patterns using the temporal transition in a population.

The results also reveal that additional features which are estimated only from spatio-temporal data helps us control the extracted patterns.

Keywords

Data mining; Land-use patterns; Spatio-temporal data

Categories and Subject Descriptors

H.2.8 [**Database Applications**]: Spatial databases and GIS.; H.3.3 [**Information Search and Retrieval**]: Clustering.

1. INTRODUCTION

As smartphones become more widespread, more location data are acquired with application use and uploaded to servers every day. Therefore, mining spatial and spatio-temporal data is becoming popular [12]. Location data enables us to estimate some types of urban properties. For example, Sekimoto et al. presented a people-flow reconstruction using location data extracted from person-trip surveys [16, 15]. Inferring a user's mode of transportation using location data was demonstrated by Stenneth et al [19]. Ratti et al. used location data for urban analysis by geographical mapping of cell phone usage at different times of the day [14]. Look et al. presented a technique for modeling the functional purpose of a given space for generating descriptive walking directions [9].

As a service using location data to visualize urban properties, ZENRIN Datacom [1] provides a web service for visualizing area-by-area congestion estimated using location data from cellular phones by overlaying colors on a map. The population density is estimated using a mesh and can be categorized by population density in the cells.

In this research it is assumed that the daily population trends, in other words, the temporal transition in the number of people in thetarget area, is a good representation of "how" the area is used. For example, let us think the daily population trends near suburb stations. It may have two peaks in the morning and evening, which are rush hours. Daily population trends near main urban stations that have many offices and shops around them may also have two peaks in the morning and evening, but the morning peak may come a little later and the evening peak may come earlier because they are "destinations" for most people.

As mentioned above, we can imagine a correspondence between land-use patterns and daily population trends. We argue that we can reconstruct this correspondence automatically using daily population trends estimated from location data from smart-phones.

Discovering "how" an area is used is important for urban design. Local governments in most developed countries use "zoning" for land-use planning [8]. Zoning is usually assigned manually, which is difficult in terms of how to determine borders. Furthermore, it is uncertain whether land-use patterns fit to actual land-use.

If we can obtain land-use patterns from actual data, we can use patterns to determine the location of new stores, where to place outdoor advertisements, and so on. For example, if the population trends between two areas resemble each other, we can predict that commerce that is going well in one area goes well in the other. Also, we can use the land-use patterns and population trends for allocation of taxis and estimation of epidemics, for example.

In our approach, we extract daily land-use patterns besed on the temporal transition in the number of people in the area. A simple approach is used to estimate the number of people in certain areas, which creates hourly histograms by counting the number of location data records in the area and multiplies that number by a proper factor. With this approach, it is implicitly assumed that the amount of location data records is proportional to the actual number of people in the area. In a similar work, Pulselli et al. analyzed the distribution of people using cell phone chatting data [13].

*This work was conducted during an internship at Yahoo! Japan Research, Tokyo, Japan.

Urb-IoT '14, October 27 - 28 2014, Rome, Italy

http://dx.doi.org/10.1145/2666681.2666688.

Latitude	Longitude	Accuracy	Time stamp	ID
34.871648	135.661309	5.00	20130611000000	UID1
38.721329	139.849629	30.00	20130611000000	UID2
43.092888	141.371409	1414.00	20130611000001	UID3
35.557460	139.445981	1144.88	20130611000000	UID4
35.732912	139.670771	65.00	20130611000012	UID5
35.784689	139.899813	1060.53	20130611000001	UID6
35.652135	140.026954	1414.00	20130611000002	UID7
35.699879	139.841407	165.00	20130611000002	UID8

Table 1: Examples of location data

Moreover, we can also use some additional features to control the extracted patterns. In the experiment, three types of additional features are introduced. They are, of course, extracted from only location data without any previous knowledge about the property of the area.

One difficulty in discovering land-use patterns is that the proper number of patterns is unobvious. Number of patterns should be changed according to the selection of target areas and purposes. Some areas may have unique land-use patterns, and the patterns may be different by days of the week.

Common clustering approaches, like k-means [10] and hierarchical algorithms [7, 18], do not deal with this difficulty. To solve this problem, the infinite Gaussian mixture model (GMM) with Dirichlet process (DP) mixtures [5] is used to discover different land-use patterns for this study. The use of DP mixtures enables us to automatically select an appropriate number of typical land-use patterns.

2. APPROACH

2.1 Dataset

We use location data that smartphone users provided through Yahoo! Japan applications with their agreements. The data contain the following information: latitude, longitude, horizontal accuracy, time stamp (in seconds), anonymized user ID (changes every 24 hours).

Table 1 lists several examples of the location data. Latitudes and longitudes are recorded in degrees, horizontal accuracies are recorded in meters, time stamps are recorded at a second rate in format "YYYYMMDDhhmmss". the user IDs are anonymized and changed every 24 hours to protect the privacy of users.

The number of records is approximately 15 million for a single day. In the experiment, eight weeks of data were analyzed. The total data size was approximately 34.9 GB in compressed form with gzip. All data in a certain day are plotted in Figure 1. Note that no map is overlaid in this figure. The figure implies that data are acquired from all over Japan.

The number of records is temporally biased due to the method of data acquisition. The data are acquired mainly when users used certain applications. In some applications, the data are acquired at regular intervals or when the connected base stations change.

2.2 Method

2.2.1 *Feature extraction and pre-processing*

We introduce $\boldsymbol{x}_j = [x_{1,j}, x_{2,j}, ...]^{\mathsf{T}}$ as the feature vector for extracting land-use patterns, where j denotes the index of target conditions. The main feature we used for extracting patterns is the hourly number of people in areas. We estimated this by creating histograms of which the bar width is one hour.

As mentioned above, we extracte day-by-day land-use patterns

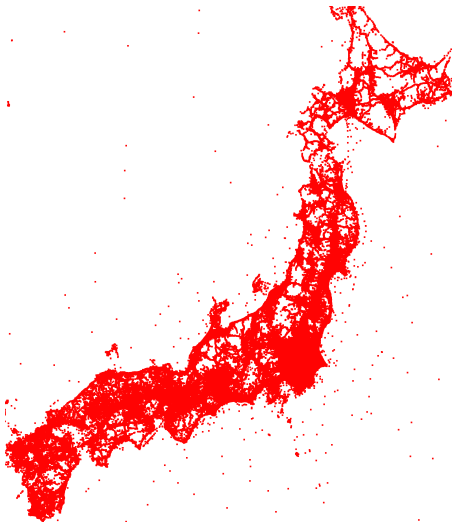

Figure 1: Location data points collected on May 20th, 2013. The shape of the Japanese archipelago is stood out on white canvas.

	Area 1			...	Area n		
	Day 1	...	Day m	...	Day 1	...	Day m
0 - 1 a.m	$a_{1,1}$	...	$a_{1,m}$	...	...	...	$a_{1,nm}$
1 - 2 a.m	$a_{2,1}$	...	$a_{2,m}$	...	...	...	$a_{2,nm}$
...	...	...	...	$a_{i,j}$	...	...	...
11 - 12 p.m	$a_{24,1}$	...	$a_{24,m}$	...	...	...	$a_{24,nm}$
Σ	A_1	...	A_m	A_j	...	...	A_{nm}

Table 2: Matrix of count data with n areas and m days

in target areas. That is, if the number of areas is n and the number of analyzed days is m, we extracte some patterns from total nm conditions. First, we count the number of different IDs by hour for each area and each day. We then create a matrix like Table 2.

Here, $a_{i,j}$ denotes the number of different IDs in the area, day and hour. $j = 1, ..., nm$ is the index of conditions (observations), and $i = 1, ..., 24$ is the index of hours. The count is normalized by dividing by the total number of records (Eqs 1 and 2) to attach the weight to the distribution shape.

$$A_j = \sum_{i=1}^{24} a_{i,j} \tag{1}$$

$$x_{i,j} = \frac{a_{i,j}}{A_j} \tag{2}$$

In $i = 1, 2, ...24$, $x_{i,j}$ denotes the i-th feature for the j-th observation (here, combination of day and area).

2.2.2 *Additional Features*

We can use not only the values described above but also additional features to control the extracted patterns according to applications. This means we can introduce $x_{i,j} (i \geq 25)$ as necessary.

Three additional features are introduced. The details of these additional features are described below.

Additional feature 1.

The first additional feature is the total number of people in one day in an area. In the feature described above, $x_{i,j} (i \leq 24)$ do not

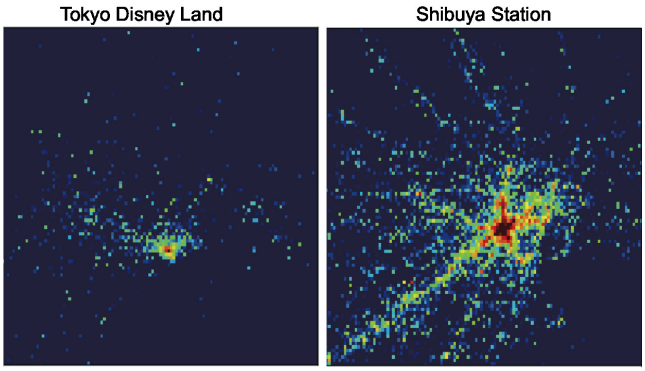

Figure 2: Density map of points after one hour from when people arrived at Tokyo Disney Land gate (left-side) and Shibuya station (right-side).

explain the total population volume because they are normalized. The total number of records is A_j in Eq. 1. A_j is appended to $\boldsymbol{x}_j$.

Additional feature 2.

The second additional feature is the ratio of the number of people in the target area to that in the surrounding area. This means congestion relative to the surrounding area. In our experiment, we determine the surrounding areas as within 300 m from selected points. The threshold of 300 meters is decided empirically.

Let $b_{i,j}$ be the number of people in the surrounding area corresponding to $a_{i,j}$. We can calculate the total number in the area and day B_j in the same manner as calculating A_j (Eq. 1). That is, $B_j = \sum_{i=1}^{24} b_{i,j}$. A_j/B_j is appended to $\boldsymbol{x}_j$.

Additional feature 3.

The third additional feature is the median of moving distances after one hour from when people arrive at a target area. This potentially represents the mean sojourn time in the area. This can be calculated using IDs of records.

We calculate the distance between the target area and the position after one hour from the time the user arrived at the target area. We estimate the position after one hour by linear interpolation using user ID. We calculate this distance for all users who arrived at the target area in the target day and took median. This median is introduced as part of $\boldsymbol{x}_j$.

For example, Figures 2 shows density maps of points after one hour from when people arrived at Tokyo Disney Land gate and Shibuya Station, respectively. They show that people who go to Shibuya Station tend to move farther away in one hour than who go to Tokyo Disney Land gate.

2.2.3 Whitening

Before clustering process is applied, each feature si divided by its standard deviation across all observations to give it unit variance as follows:

$$\sigma_i = \sqrt{\frac{1}{nm}\sum_{j=1}^{nm}(x_{i,j} - \mu_i)}, \quad \text{where} \quad \mu_i = \frac{1}{nm}\sum_{j=1}^{nm} x_{i,j} \tag{3}$$

$$x'_{i,j} = \frac{x_{i,j}}{\sigma_i} \tag{4}$$

This eliminates the negative effect of sampling bias. We use $x'_{i,j}$ to extract patterns for each j. Each feature vector is rewritten as $\boldsymbol{x}_j = [x'_{1,j}, x'_{2,j}, x'_{3,j}...]^\mathsf{T}$

2.2.4 Land-Use clustering using Dirichlet process Mixture

Now that we have made land-use feature vectors $\boldsymbol{x}_j$, we describe the method for extracting patterns. We use a clustering technique with an infinite GMM. A GMM is a probabilistic model with which all the data points $\boldsymbol{x}_j$ are assumed generated from a mixture of a finite number of Gaussian distributions with unknown parameters.

The GMM with K components may be written as:

$$p(\boldsymbol{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{5}$$

where π_k denotes mixture weight, and $\boldsymbol{\mu}_{1...K}$ and $\boldsymbol{\Sigma}_{1...K}$ denote the parameters of each component.

By introducing K-dimentional binary cluster assignment $\mathbf{z} = z_k, k = 1, ..., N$, we can obtain:

$$p(\boldsymbol{x}|\mathbf{z}) = \prod_{k=1}^{K} \mathcal{N}(x|\mu_k, \boldsymbol{\Sigma}_k)^{z_k} \tag{6}$$

The limitation of a GMM is that we have to specify the number of components K. To solve this limitation, we introduce the Dirichlet Process (DP). Clustering using DP mixtures enables the automatic selection of the number of clusters. Using the DP implements a GMM with a variable number of components. The DP is a prior probability distribution on clustering with an infinite number of components.

The method of using an infinite number of components is represented by the Chinese restaurant process [2] and stick-breaking process [17].

The DP has a hyperparameter α. When α is large the DP tends to result in many clusters.

The inference part is implemented variational Bayes techniques [4], which enable us to incorporate this prior structure on GMMs with relatively fast execution time.

2.2.5 Evaluation of extracted patterns

Generally speaking, clustering results are difficult to evaluate. So we evaluate the result from qualitative aspect and quantitative aspect. We manually assigned metadata of area categories to all target areas to evaluate the extracted patterns by comparing with the metadata. Note that we do not argue that the extract patterns should fit to the metadata because the land-use pattern, how the area is used, is not determined only by the area category. The comparison of the extracted patterns with the manually asigned metadata is used just for demonstrating the effects of the additional features.

The target areas are divided into following eight categories: suburb station, main station. Shinkansen station (Shinkansen is a network of high-speed railway lines in Japan), downtown area, amusement park, business district, leisure venue, big market.

As quantitative evaluation, purity, inverse purity and F measure, which is the harmonic average of the other two [11] [6] [3] as evaluation metrics, are used in this research. These are standard evaluation metrics in the Web People Search Task.

Let $n_{i,j}$ be the number of conditions which are assigned to cluster i and the metadata j and N denote the total number of conditions.

Purity represents how the generated cluster is occupied by major metadata. It is explained as:

$$\text{Purity} = \frac{1}{N}\sum_{i} \max_{j}(n_{i,j}) \tag{7}$$

In cluster i, $\max_j(n_{i,j})$ represents the number of conditions with the most major metadata. Purity is obtained by summing this for all i and dividing it by N.

The range of purity is 0 to 1, where "1" means that all clusters consist of data with a single metadatum. In terms of the purity metric, "1" is the best clustering result. Though purity can represent how "pure" the cluster is, a problem is remained. If the number of clusters is the same as the number of conditions, in other words, all clusters have only one datum, purity becomes "1", which is not a fair evaluation.

To solve this problem, inverse purity, which is represented as:

$$\text{InversePurity} = \frac{1}{N}\sum_j \max_i(n_{i,j}), \tag{8}$$

is used. This is the same as purity but i and j are switched.

Purity and inverse purity are similar to the precision and recall measures. The range of inverse purity is also 0 to 1, where "1" means all data with the same metadata is in a single cluster. A larger value means the better clustering. The problem of inverse purity is that the maximum inverse purity can be similarly achieved by putting all data into one cluster.

There is a tradeoff between purity and inverse purity; therefore we use the F measure as an evaluation metric. The F measure is the harmonic mean of purity and inverse purity. i.e.

$$F = \frac{2}{1/\text{Purity} + 1/\text{InversePurity}} \tag{9}$$

$$= 2 \cdot \frac{\text{Purity} \cdot \text{InversePurity}}{\text{Purity} + \text{InversePurity}} \tag{10}$$

We now focus on the F measure to assess the clustering quality and discuss the patterns through histograms of population trends.

3. RESULTS AND DISCUSSION

We conducted land-use clustering for 34 areas over 56 days from May 22 to July 14 in 2013. This means 1,960 conditions were studied. The target regions were within a 200-meter radius from selected points in these areas. The target regions were selected from the areas which have remarkable characteristics in the meaning of land-use patterns so that metadata are added clearly.

3.1 Qualitative evaluation of extracted patterns

The clustering results with the additional features are shown in Figure 3, and the mean histograms of the temporal transition in the number of people for the extracted petterns are shown in Figure 4. For example, Pattern 5 has a peak around 9 o'clock and, it is shown in business district in weekday or amusement park. Commuters in the morning around 9 o'clock are commen scene in Japan and almost tourists of famous amusement parks enter the parks around 9 o'clock.

First, each different kind of station was well-discriminated by manually added metadata. Pattern IDs are shown in Figure 3. Patterns 3 and 6 indicate suburb stations and main stations on weekdays, respectively and Pattern 2 indicates Shinkansen (high-speed bullet trains) stations. These differences were not extracted well with none of the additional features.

As discussed above, we can obtain natural clustering results from qualitative evaluations, which can be selected on demand. This fact that we can manipulate the results at some level is helpful because obtaining patterns fitting for purposes is important in location aware services.

3.2 Quantitative evaluation by purity

Features				Alpha		
f_{hist}	f_{add1}	f_{add2}	f_{add3}	0.01	0.1	1
✓				0.428	0.524	0.523
✓	✓			0.381	0.531	0.520
✓		✓		0.455	0.551	0.497
✓			✓	0.381	0.537	0.428
✓	✓	✓		0.456	0.596	0.569
✓	✓		✓	0.430	0.571	0.540
✓		✓	✓	0.431	0.560	0.509
✓	✓	✓	✓	0.431	**0.612**	0.559
	✓			0.430	0.430	0.457
		✓		0.381	0.381	0.381
			✓	0.424	0.414	0.414
	✓	✓		0.486	0.486	0.487
	✓		✓	0.436	0.399	0.498
		✓	✓	0.426	0.415	0.415
	✓	✓	✓	0.484	0.513	0.519

Table 3: F measure according to feature combination.

In this section, we consider which features contribute to controlling the patterns. We conducted land-use patterning with combinations of the features and assessed the results by using the F measure of purity and inverse purity. We conducted clustering three times and take one with the maximum F measure because the results were different for every trial.

Table 3 lists the results. Our main feature is daily population trend (let the feature be f_{hist}), where f_{add1}, f_{add2}, f_{add3} denote each additional features described above, respectively. The left of Table 3 denotes the selected features. Features with a ✓are selected. We conducted clustering with three patterns of alpha (0.01, 0.1 and 1).

Note that the F measures without f_{hist} are small. This shows that daily population trend is a good representation of land-use patterns. The results with alpha as 0.1 tend to be good. The F measures are higher with any single additional feature than without additional features. When we used two features, the F measures became higher than with a single one. Moreover, the F measures reached maximum when using all additional features besides daily population trends. Thus, the daily population trend is important to represent land-use patterns, and all additional features are helpful for controlling the extracted patterns.

4. CONCLUSION

An approach to extract land-use patterns, how the area is used, using location data from users of Yahoo! Japan's applications are proposed. Our approach is innovative because this does not require any specified knowledge and skill for the clustering of areas. All users have to do is select the hyper parameter.

The hourly transition in the density of people was used as the main feature. It is estimated by making a histogram of the number of points in a target area with a bar width of one hour. Three additional features extracted from location data were introduced to obtain natural clustering results. An infinite GMM is used for land-use clustering, which uses the DP to select the appropriate number of patterns.

The experiments were conducted for 34 areas over 56 days, that is, a total of 1,904 conditions. The experimental results showed that our approach successfully discovers typical land-use patterns and

area names	metadata	Mon	Tue	Wed	Thu	Fri	Sat	Sun	Mon	Tue	Wed	Thu	Fri	Sat	Sun	Mon	Tue	Wed	Thu	Fri	Sat	Sun
Kokubunji Sta.	suburb station	3	3	3	3	3	7	1	3	3	3	3	3	7	7	3	3	3	3	3	7	1
Tachikawa Sta.	suburb station	3	3	3	3	3	7	1	3	3	3	3	3	7	1	3	3	3	3	3	7	1
Mitaka Sta.	suburb station	3	3	3	3	3	7	1	3	3	3	3	3	7	1	3	3	3	3	3	7	1
Kashiwa Sta	suburb station	3	3	3	3	3	7	1	3	3	3	3	3	7	1	3	3	3	3	3	7	1
Minamiosawa Sta.	suburb station	3	3	3	3	3	7	1	3	3	3	3	3	7	1	3	3	3	3	3	7	1
Tama-center Sta.	suburb station	3	3	3	3	3	7	1	3	3	3	3	3	7	1	3	3	3	3	3	1	1
Hachioji Sta.	suburb station	3	3	3	3	3	7	1	3	3	3	3	3	7	1	3	3	3	3	3	7	1
Takao Sta.	suburb station	3	3	3	3	3	7	1	3	3	3	3	3	7	1	3	3	3	3	3	7	7
Shinagawa Sta.	main station	6	6	6	6	6	7	7	6	6	6	6	6	7	7	6	6	6	6	6	7	7
Shinjuku Sta.	main station	6	6	6	6	6	7	7	6	6	6	6	6	7	7	6	6	6	6	6	7	7
Shibuya Sta.	main station	6	6	6	6	6	7	7	6	6	6	6	6	7	7	6	6	6	6	6	7	7
Ikebukuro Sta.	main station	6	6	6	6	6	7	7	6	6	6	6	6	7	7	6	6	6	6	6	7	7
Shinkobe Sta.	Shinkansen station	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
Okayama Sta.	Shinkansen station	2	2	2	2	2	2	2	2	2	2	2	2	2	1	2	2	2	2	2	1	1
Hiroshima Sta.	Shinkansen station	2	2	2	2	2	2	1	2	2	2	2	6	2	1	6	2	2	2	2	7	1
Kokura Sta.	Shinkansen station	2	6	2	2	2	1	1	2	2	2	2	2	1	1	2	2	2	2	2	1	1
Shin-Yokohama Sta.	Shinkansen station	3	3	3	2	3	2	1	3	2	2	3	3	2	2	3	3	3	3	3	1	1
Dogenzaka	downtown	1	7	6	6	6	7	7	1	7	7	1	6	7	7	7	7	7	7	6	7	7
Kabuki-cho	downtown	6	6	6	6	6	7	7	6	6	7	6	6	7	7	6	6	6	6	6	7	7
Nakasu	downtown	6	2	6	6	6	7	7	2	6	2	6	6	7	7	2	7	6	6	7	7	7
Gate of Universal Studios Japan	amusement park	5	5	1	1	5	5	5	5	5	1	1	5	5	5	5	5	5	5	5	5	5
Gate of Tokyo Disney Land	amusement park	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
Gate of Tokyo Disney Sea	amusement park	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
Tokyo Metropolitan Government	business district	5	5	5	5	5	1	7	5	5	5	5	5	1	1	5	5	5	5	5	1	1
Marunouchi	business district	5	5	5	5	5	1	1	5	5	5	5	5	1	1	5	5	5	5	5	1	1
Tokyo Midtown	business district	5	5	5	5	5	1	1	5	5	5	5	5	7	1	5	5	5	5	5	7	7
Kasumigaseki	business district	5	5	5	5	5	7	1	5	5	5	5	5	1	1	5	5	5	5	5	7	7
Tokyo Dome	leisure venue	1	6	6	6	6	1	1	5	6	6	2	6	1	1	2	6	6	6	6	1	1
Hibiya Park	leisure venue	1	1	6	1	1	1	1	5	1	1	5	1	1	1	5	1	1	5	1	1	1
Tokyo Skytree Town	leisure venue	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Sensoji temple	leisure venue	1	1	1	1	1	1	1	1	1	7	1	1	1	1	7	1	1	1	1	7	1
Nishiki Market	Market	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Temma Market	Market	6	6	6	6	6	7	7	6	6	6	6	6	7	7	6	6	6	6	6	1	7
Tsukiji Market	Market	4	4	1	4	4	4	1	4	4	4	4	4	4	7	4	4	5	4	4	4	1

Figure 3: Clustering results with additional features. Areas could be successfully divided into manually added metadata groups. The number in the table shows the id of the temporal transition pattern acquired from clustering. Metadata are added manually from ordinal land use pattern. Day of the week column shows day of week from May 22 to July 14 in 2013.

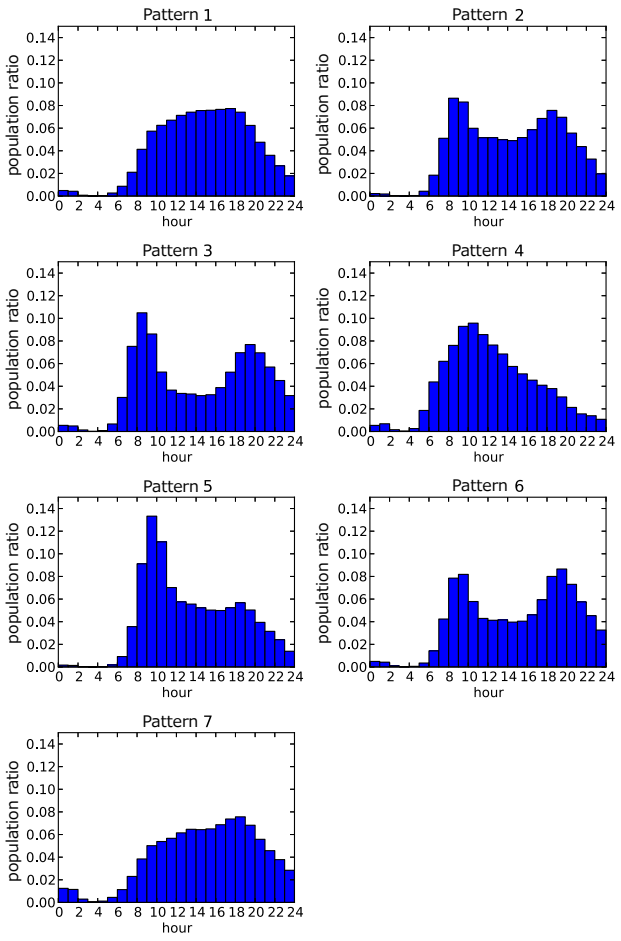

Figure 4: Histogram of mean temporal transition in the number of people of each pattern in Figure 3.

different patterns between weekdays and holidays.The results also showed that using additional features can control the patterns.Though whether the values of F measure are enough or not for clustering is not discussed, it can be confirmed that the additional features make clustering result more natural relatively.

Directions for further work in this area include conducting experiments for more various areas and exploring more efficient features appropriate to specific applications. We will also obtain more useful patterning results by using personal data that comes with location data, for example, genders and ages. Introducing more features is also one of future works.

5. REFERENCES

[1] ZENRIN Datacom. `http://www.zenrin-datacom.net/en/`.

[2] Antoniak, C. E. Mixtures of dirichlet processes with applications to bayesian nonparametric problems. *The Annals of Statistics 2*, 6 (1974), 1152–1174.

[3] Artiles, J., et al. The semeval-2007 WePS evaluation: Establishing a benchmark for the web people search task. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, Association for Computational Linguistics (2007), 64–69.

[4] Blei, D. M., et al. Variational inference for dirichlet process mixtures. *Bayesian Analysis 1* (2006), 121–144.

[5] Escobar, M. D., et al. Bayesian density estimation and inference using mixtures. *Journal of the american statistical association 90*, 430 (1995), 577–588.

[6] Hotho, A., Nürnberger, A., and Paaß, G. A brief survey of text mining. In *Ldv Forum*, vol. 20 (2005), 19–62.

[7] Johnson, S. C. Hierarchical clustering schemes. *Psychometrika 32*, 3 (1967), 241–254.

[8] Lefcoe, G. The regulation of superstores: the legality of zoning ordinances emerging from the skirmishes between wal-mart and the united food and commercial workers union.

[9] Look, G., et al. A location representation for generating descriptive walking directions. In *Proceedings of the 10th International Conference on Intelligent User Interfaces*, ACM Press (2005), 122–129.

[10] MacQueen, J., et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, vol. 1, California, USA (1967), 14.

[11] Manning, C. D., et al. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[12] Parthasarathy, S. Mining Spatial and Spatio-Temporal Patterns in Scientific Data. *22nd International Conference on Data Engineering Workshops (ICDEW'06)* (2006), x146–x146.

[13] Pulselli, R., et al. Computing urban mobile landscapes through monitoring population density based on cellphone chatting. *International Journal of Design & Nature and Ecodynamics 3*, 2 (2008), 121–134.

[14] Ratti, C., et al. Mobile Landscapes: using location data from cell phones for urban analysis. *Environment and Planning B: Planning and Design 33*, 5 (2006), 727–748.

[15] Sekimoto, Y., et al. Digital archiving of people flow using person trip data of developing cities. *The First Workshop on Pervasive Urban Applications (PURBA) in conjunction with the Ninth International Conference on Pervasive Computing*, 1 (2011), 1–8.

[16] Sekimoto, Y., et al. PFlow: Reconstructing People Flow Recycling Large-Scale Social Survey Data. *IEEE Pervasive Computing 10*, 4 (2011), 27–35.

[17] Sethuraman, J. A constructive definition of dirichlet priors. *Statistica Sinica 4* (1994), 639–650.

[18] Steinbach, M., et al. A comparison of document clustering techniques. In *KDD workshop on text mining*, vol. 400 (2000), 525–526.

[19] Stenneth, L., et al. Transportation mode detection using mobile phones and gis information. In *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, ACM (2011), 54–63.