Town Characteristics Estimation using Geotagged Twitter Data - A Case Study in the Tokyo Area -

Suguru Tsujioka

***Abstract*—Here, we explain a method for analyzing Twitter post data (i.e., tweets) to identify specific characteristics of various towns. Our analysis targets seven towns in Tokyo. Tweets were collected from August 25 to September 14, 2015. As a result of this three-week data collection period, we obtain 64,174 geotagged tweets. Our proposed method includes correspondence analysis, self-organizing maps, and cross-tabulation. From our analysis results, we show that tweets have certain fixed trends depending on their corresponding posted towns. Moreover, we describe the characteristics of the tweets originating in each target town. Our proposal methods are useful for urban planning, marketing, and regional design. Because they provide a means in place of a questionnaire survey.**

***Keywords*—Twitter, town characteristics, correspondence analysis, self-organizing maps.**

I. Introduction

A. Purpose

We propose a method for analyzing the characteristics of various towns from Twitter data (i.e., tweets). Individual towns have unique characteristics based on their geographic and historical background [1]. Town characteristics are broadly classified into physical and emotional factors (Table 1). Emotional factors are generally more important than physical factors for town characteristics. Clarifying these factors is useful in various fields such as marketing, city planning, and civic activities. Thus, we observe data from these factors to estimate town characteristics. Previously, we used questionnaires to obtain data regarding these factors; however, administering questionnaires and performing analyses are expensive.

TABLE I: Examples Of Factors Of Town Characteristics

Category	*Factor*
Physical factors	Transport (e.g., public transportation, traffic jam), Shopping, Medical, Education
Emotional Factors	Quiet, Energetic, Refined, Luxury

Therefore, in this study, we use tweets to obtain factors regarding town characteristics. Twitter is an online social network used by millions of people around the world. In particular, as of 2014, Japan had the largest Twitter user base in the world [2]. An important characteristic of Twitter is its simplicity for users to post tweets. Although blog users typically update their blogs once every few days, Twitter users post tweets several times in a single day. Thus, we are able to capture a large number of tweets in a relatively short period of time and use such data to analyze town characteristics. We estimate such town characteristics as weather reports or commerce surveys on the basis of these tweets.

B. Previous Study

Numerous studies have considered Twitter users as social sensors [3][4]. The purpose behind some of these studies is event detection via geotagged tweets; here, the tweets immediately respond to various events. Therefore, tweets are useful to track earthquakes as they occur [3] and sports events [4]. For our purpose, though, our focus is not on event detection, but rather on estimating town characteristics.

Java et al. focused on examining the geographical properties of Twitter's social network [5]. Likewise, Kinsella et al. estimated the positions of nongeotagged tweets using a supervised learning approach with small amounts of geotagged tweets as labeled data [6]. Our method uses geotagged tweets.

Green [7] studied the notion of "town characteristics" within the context of an Australian coastal town from a community perspective. He showed that a positive character image was strongly supported by natural landscape features associated with their naturalness, beauty, pleasantness, distinctiveness, and interest.

II. Target Area and Analysis Data

Tweets were collected from the Twitter streaming application program interface to estimate town characteristics. We targeted seven towns in the Tokyo area. The target towns are defined as 1 km $\times$ 1 km squares centered around Tokyo, Akihabara, Ueno, Ikebukuro, Shinjuku, Shibuya, and Shinagawa (Table 2 and Figure 1). Tweets were obtained from August 25 to September 14, 2015. From this time period, we collected 64,174 geotagged tweets. These tweets were stored in JSON format and included screen names, post times, post positions (i.e., latitude and longitude information), and posted content.

Manuscript received Dec. 5, 2015. This work was supported by Shikoku University Grant for Innovative Research Project. S. Tsujioka is with Shikoku University, 123-1 Ebisuno, Furukawa, Ojin-cho, Tokushima-shi, Tokushima Prefecture 771-1192, Japan.

TABLE II: TARGET AREAS

	Target area (latitude, longitude)	***Number of geotagged tweets***
Tokyo	(35.676895, 139.760417) - (35.685854, 139.771532)	7,318
Akihabara	(35.694053, 139.767794) - (35.702818, 139.778755)	16,862
Ueno	(35.707637, 139.770713) - (35.716505, 139.781836)	3,448
Ikebukuro	(35.725639, 139.706454) - (35.73438, 139.717552)	7,212
Shinjuku	(35.68507, 139.695415) - (35.694243, 139.706432)	12,274
Shibuya	(35.654179, 139.696602) - (35.662609, 139.707041)	14,182
Shinagawa	(35.624508, 139.73329) - (35.633282, 139.744279)	2,878

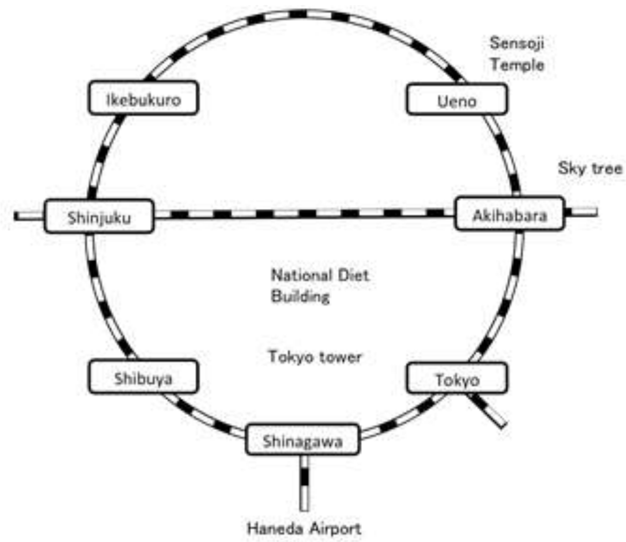

Fig. 1. Target areas

III. ANALYSIS

A. Correspondence Analysis

To obtain a rough outline of the target towns, the collected tweets were analyzed using correspondence analysis based on their constituent words. More specifically, KH Coder was used for the analysis [8]; note that KH Coder is a quantitative content analysis tool based on R [9]. Our analysis considered words for which frequency of occurrence was over 500; accordingly, 31 words were adopted as factors of the analysis.

Results of this correspondence analysis are shown in Figure 2. In the figure, squares represent towns, while circles represent words identified as factors of tweets. The size of each square expresses the number of tweets for that town, and the size of each circle denotes the frequency of that word. In our analysis, the sum of the contribution rate was 78.46%.

Figure 2 shows that the target towns were classified into the following three groups:

- Group 1: This group consisted of Tokyo and Shinagawa. These towns have arrival and departure stations of Shinkansen. Therefore, tweets in this group consisted of words related to trips, such as Shinkansen, station platform, platform number, and bus.
- Group 2: This group consisted of Shibuya, Shinjuku, and Ikebukuro. These towns are relatively closer to one another. Many verbs were included in this group. Therefore, it seems that active people posted tweets in the towns in this group.
- Group 3: This group consisted of Ikebukuro and Akihabara. Tweets in this group consisted of words related to food, such as curry, noodles, and café.

Note that Ueno belonged to no group. Our correspondence analysis showed that tweets had a certain fixed trend depending on their posted towns.

B. Clustering with a Self-Organizing Map

Next, tweets were clustered using a self-organization map (SOM) to clarify the factors of town characteristics. Clustering was performed for each target town using KH Coder. Tweets were clustered according to their constituent words. During clustering, place names were ignored. Parameters of the SOM in our study are summarized in Table 3. An example result of this clustering is shown in Figure 3. From our clustering results, five clusters were obtained as common clusters in the target towns. These clusters and examples of constituent words in the clusters are shown in Table 4.

TABLE III: PARAMETERS OF THE SOM

Number of nodes	400 (20 × 20)
Number of clusters	8
Learning frequency	2000

C. Characteristics of Each Town

Next, we performed cross-tabulation based on the themes, i.e., the clusters shown in Table 4, to quantitatively clarify the characteristics of each town. The resulting cross-table is shown in Table 5. In the cross-table, a p value less than 0.01 was considered statistically significant, and chi-square values were large for the degree of freedom. Therefore, each town has significant differences, as shown in Table 5.

In the table, Position Report accounted for the highest percentage among all target towns. In particular, there were 27,105 "I'm at" phrases in the Twitter data. This phrase is posted when Twitter works in conjunction with Foursquare, a popular location-based social networking service [9]. More specifically, the phrase is posted when the Foursquare post data is passed to Twitter. Many Foursquare users add their location to their Tweets using this mechanism.

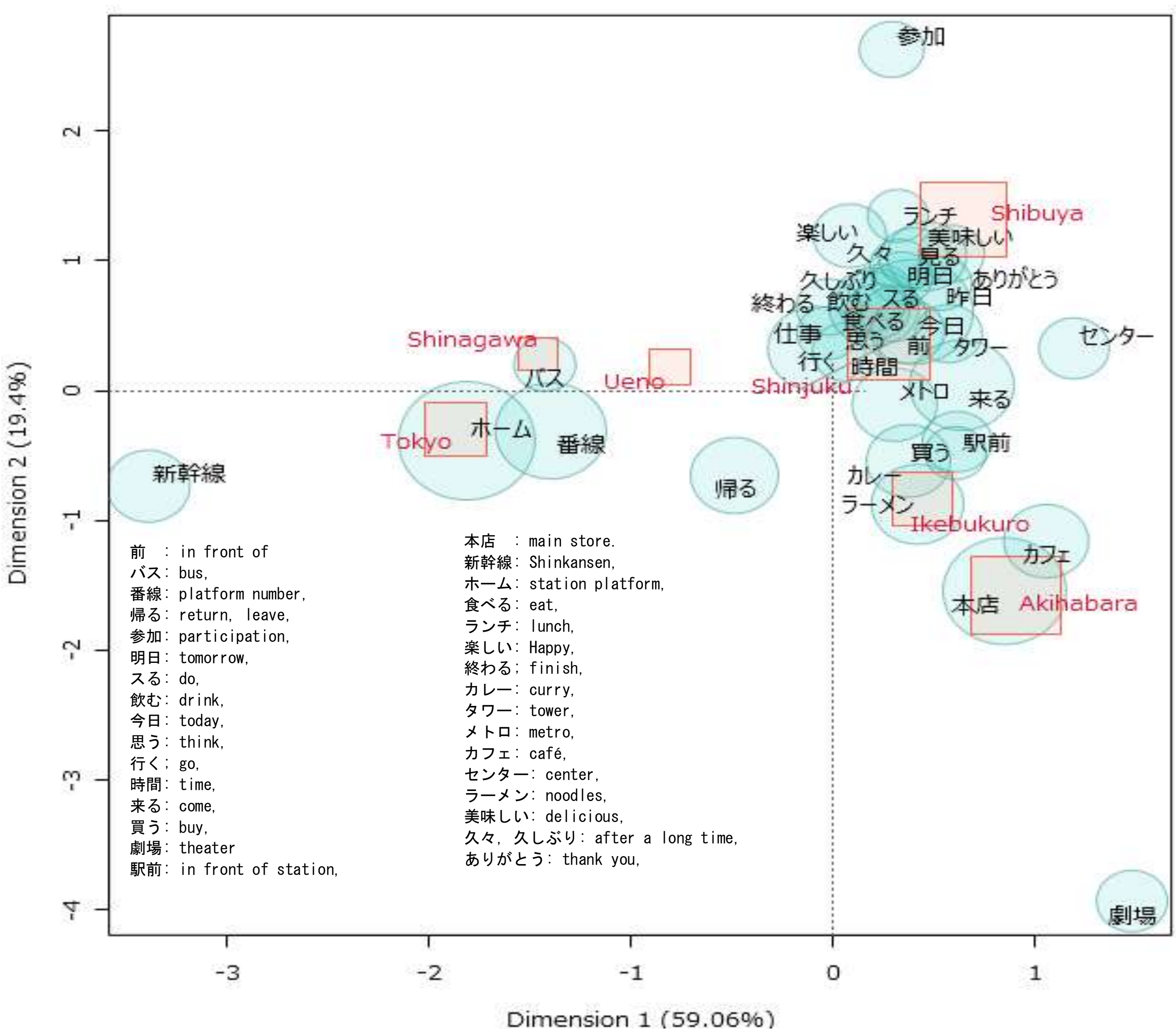

Fig. 2. Results of our correspondence analysis

Table 5 shows the same results as our correspondence analysis, i.e., (1) in Tokyo and Shinagawa, the trip characteristics accounted for a higher percentage than the other towns; (2) Shibuya, Shinjuku, and Ikebukuro showed a similar trend, i.e., Happy and Shopping had large values in all three towns; (3) in Akihabara and Ikebukuro, Food and Drink accounted for a higher percentage than the other towns; and (4) Ueno had no particular identifiable theme. The four points are quantitatively shown in Table 5.

IV. CONCLUSIONS

In this study, we analyzed town characteristics using geotagged tweets. Our analysis was based on correspondence analysis, a self-organizing map, and cross-tabulation. The targets of our analysis were seven towns in the Tokyo area. First, the tweets were analyzed using correspondence analysis based on their constituent words to obtain a rough outline of the target towns. Analysis results showed that tweets had clear distinctions among target towns. Next, tweets were clustered using a SOM to clarify the factors of each town's characteristics. Finally, we performed cross-tabulation based on the themes (the clusters shown in Table 4)to quantitatively clarify the characteristics of each town. The cross-table quantitatively showed these characteristics.

This study is an important contribution because it shows how tweets have significant trends per posted town or region. When tweets are obtained for analysis, they should be considered in part based on their posted location.

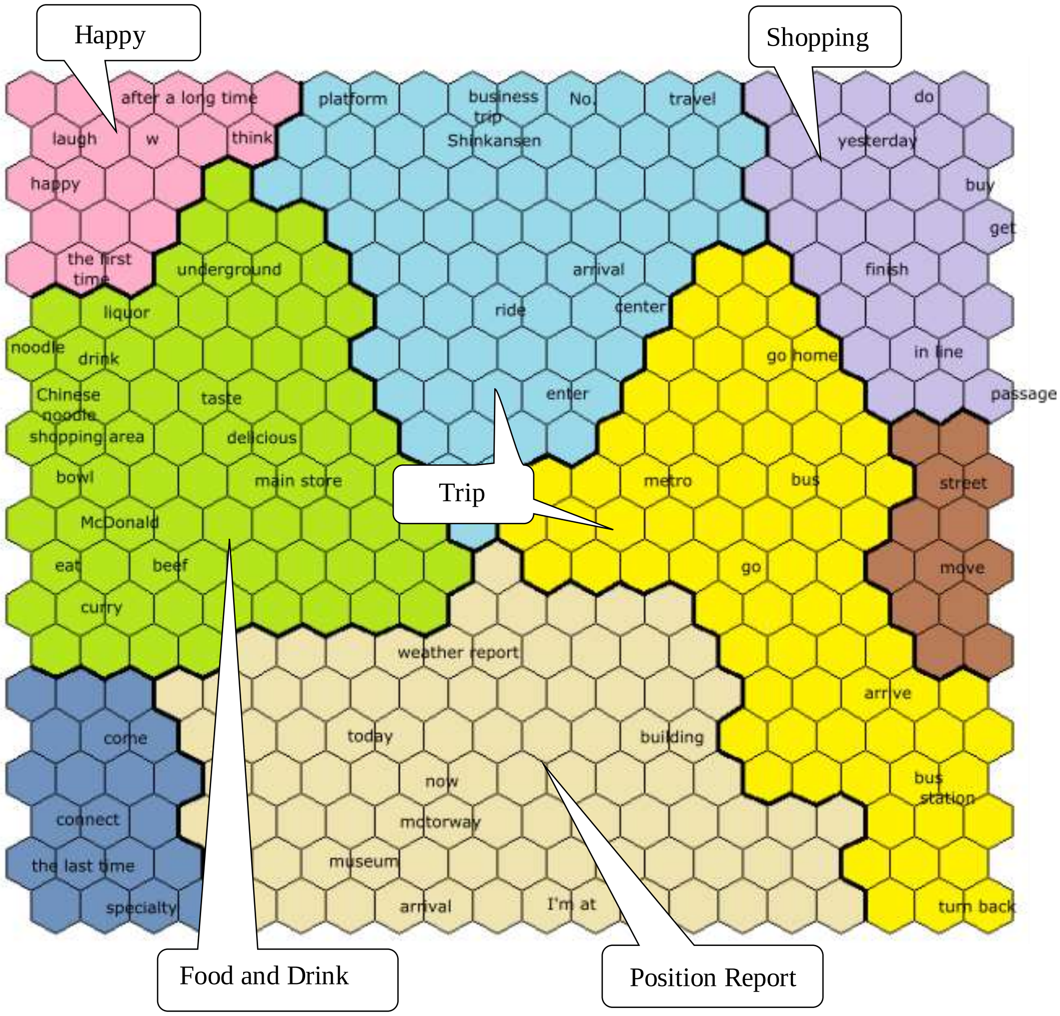

Fig. 3. Results of clustering using a SOM for the case of the Tokyo town.

TABLE IV: CLUSTERS AND EXAMPLES OF THEIR CONSTITUENT WORDS

Cluster name	***Constitutive words or phrase***
Happy	happy, enjoy, !!, laugh, KAWAII, and w[*1]
Shopping	buy, get, purchase, shop, and store
Trip	go, come, ride, leave, move, and bus
Food and Drink	eat, drink, lunch, dinner, beef, and taste
Position Report	I'm at, @, arrive, now, and reach

(*1 "w" expresses laughing in Japan.)

TABLE V: THE CROSS TABLE BASED ON THE THEME

	Position Report		***Trip***		***Food & Drink***		***Happy***		***Shopping***		***N of Tweet***
Tokyo	3,469	47.40%	1,782	24.35%	492	6.72%	273	3.73%	126	1.72%	7,318
Akihabara	8,805	52.22%	379	2.25%	2,436	14.45%	222	1.32%	446	2.64%	16,862
Ueno	1,507	43.71%	206	5.97%	295	8.56%	122	3.52%	104	3.00%	3,448
Ikebukuro	3,541	49.10%	424	5.88%	1,173	16.26%	288	3.99%	237	3.29%	7,212
Shinjyuku	5,784	47.12%	674	5.49%	1,046	8.52%	672	5.47%	407	3.31%	12,274
Shibuya	4,450	31.37%	614	4.33%	1,533	10.81%	1,032	7.28%	500	3.52%	14,182
Shinagawa	1,357	47.14%	302	10.49%	136	4.73%	101	3.49%	50	1.72%	2,878
total	28,912	45.05%	4,381	6.83%	7,111	11.08%	2,709	4.22%	1,868	2.91%	64,174
chi-square	305.54**		3069.71**		672.05**		1047.57**		171.67**		

REFERENCES

[1] K. Lynch, The Image of the City, The M.I.T. Press. 1960.

[2] eMarketer.com, Asia-Pacific grabs largest twitter user share worldwide -Japan, Indonesia and India are the region's top Twitter markets-, http://www.emarketer.com/Article/Asia-Pacific-Grabs-Largest-Twitter-User-Share-Worldwide/1010905.

[3] Y. Takeichi, K. Sasahara, R. Suzuki, and T. Arita, Twitter as social sensor: Dynamics and structure in major sporting events. Artificial Life 14: In Proc. 14th International Conference on the Synthesis and Simulation of Living Systems, pp.778-784. The MIT Press, 2014.

[4] T. Sakaki, M. Okazaki, Y. Matsuo, Earthquake shakes Twitter users: real-time event detection by social sensors, In Proc. 19th International Conference on World Wide Web, pp. 851-860, ACM New York, 2010.

http://dx.doi.org/10.1145/1772690.1772777

[5] A. Java, X. Song, T. Finin, and B. Tseng. Why we twitter: Understanding microblogging usage and communities. In Proc. Joint 9th WEBKDD and 1st SNA-KDD Workshop 2007, pp.56-65, 2007.

http://dx.doi.org/10.1145/1348549.1348556

[6] S. Kinsella, V. Murdock, and N. O'Hare, "I'm eating a sandwich in Glasgow: Modeling locations with tweets". In Proc. SMUC, pp.61-68, 2011.

http://dx.doi.org/10.1145/2065023.2065039

[7] Green, R. Meaning and form in community perception of town character. Journal of Environmental Psychology, 19(4), pp.311-329, 1999.

http://dx.doi.org/10.1006/jevp.1999.0143

[8] K. Higuchi, KH Coder - Quantitative content analysis or text mining, http://sourceforge.net/projects/khc/, 2001.

[9] The R foundation, The R project for statistical computing, https://www.r-project.org.

[10] FOURSQUARE, FOURSQUARE, https://foursquare.com.

Dr. Suguru Tsujioka was born in Tokushima city, Japan at 1974. His educational background is information science and urban planning. He received his master's degree in engineering from Tokushima University in 2000. And he received his Ph.D. degree in management and information science from Shikoku University in 2004.

He is working as an associate professor at the Department of Information and Management Science, Shikoku University, Tokushima, Japan. His latest work is: ("Relative analysis of factors of place attachment–Case study in the Tokushima urban area-" Hong Kong, *Advances in Civil Engineering and Building Materials IV*, 2015). His current research interests are town/city characteristics estimation, analysis of decision- making process.