Estimation of Residence Information of Twitter Users based on their Posted Messages: Data for Tourism Development

Suguru Tsujioka, Akio Kondo, and Kojiro Watanabe

***Abstract*—In this study, we describe a method for estimating the residence location of a Twitter user based on his/her posted messages. Tourism development requires the determination of the attractiveness of target areas. Polling and prospective tourists are an important means to determine the attractiveness of target areas. Previously, we used questionnaires to obtain data regarding these factors. However, administering questionnaires and performing analyses are expensive. Therefore, we use Twitter post data to obtain information regarding the attractiveness of an area. Twitter is a simple-to-use platform for users to post their messages. Thus, we are able to capture a more large number of messages than questionnaires. However, obtaining a Twitter user's attributes (age, sex, residence, and so on) is impossible in many cases. The opinions of users differ depending on each user's attributes. In particular, opinions regarding the attractiveness of an area differ depending on a user's residence location.**

The primary basis of our estimation method involves machine learning to generate decision trees. Latent semantic analysis was adopted for the dimension reduction of vectors. The target of our estimation was classification between the tweets of Twitter users in Tokushima and those of the users outside Tokushima. As a result, the accuracy of our estimation method is over 60%.

***Keywords*—Twitter, tourism, latent semantic analysis (LSA), machine learning**

I. Introduction

A. Purpose

Tourism is currently expected to improve the social vitality of rural areas in Japan in light of the population decrease in these areas. Tourism is an important industry for rural areas because it increases demand for living in such areas and attracts visitors from other areas. The tourism industry in Japan during 2012 amounted to 21.2 trillion yen, with the added-value-induced effect of 10.9 trillion (2.3% of the nominal GDP) and employment-induced effect of 2.1 million people (3.3% of total workers)[1]. Thus, tourism is expected to become increasingly important for Japanese rural areas in the future.

Manuscript received April 15, 2016. This work was supported by Shikoku University Grant for Innovative Research Project.

Suguru Tsujioka is with Shikoku University, 123-1 Ebisuno, Furukawa, Ojin-cho, Tokushima 771-1192, Japan

Akio Kondo is with Tokushima University 2-24, Shinkura-cho, Tokushima 770-8501, Japan

Kojiro Watanabe is with Tokushima University 2-24, Shinkura-cho, Tokushima 770-8501, Japan

Tourism development requires the determination of the attractiveness of target areas. Polling and prospective tourists are important means to determine the attractiveness of target areas. Previously, we used questionnaires to obtain data regarding these factors [2][3]. However, administering questionnaires and performing analyses require high costs, such as postage. In addition, young people do not tend to answer questionnaires that are conducted via interviews by examiners and mail surveys [4]. Therefore, we use Twitter post data (a.k.a. tweets) to obtain information regarding the attractiveness of an area. Twitter is a user-friendly means for users to post their messages. Twitter is an online social network service used by millions of people worldwide. In particular, Japan had the largest number of Twitter users in the world in 2014 [5]. Thus, we are able to capture a more large number of messages from Twitter post data than questionnaires. However, obtaining the attributes of a Twitter user (age, sex, residence, and so on) is impossible in many cases. The opinions of users differ depending on the attributes of each user. In particular, opinions regarding the attractiveness of an area differ depending on the residence location of a user. However, according to Hecht et al., 34% of Twitter users did not provide real location information, frequently incorporating fake locations such as "Justin Bieber's heart" [6].

In this study, we propose an estimation method for determining the residence location of a Twitter user based on his/her posted messages. Our method enables an analysis of Twitter messages with a distinction between local citizens and tourists.

B. Previous Study

Numerous studies have performed an estimation of a Twitter user's residence information [6]. Hecht et al. [7] performed a simple machine learning experiment to determine whether they can identify a user's location by examining his/her tweets. They successfully distinguished the user's country and state.

Hashimoto and Oka [8] estimated conditions in posted locations based on geo-tagged tweets. They used term frequency–inverse document frequency analysis to obtain the condition data. The estimation method reduced noise factors in the tweet data.

Tsujioka verified the relation between the tweet content and posted locations. The verification method includes correspondence analysis and self-organizing maps. Verification targets were seven towns in Tokyo. In this verification, the characteristics of each town were obtained from tweets [9].

II. Estimation Method and Result

We estimate a Twitter user's residence location based on his/her posted messages. Specifically, among many Twitter users, we extract a list of Twitter users from a certain residence location. We focus on Tokushima Prefecture as the residence. The population density of this prefecture was 187.1 person/km^2 in 2013. The population density was ranked as 33 of 47 prefectures in Japan. The gross annual product of the prefecture was 282 billion yen in 2013, ranking 44th among the 47 prefectures. Based on these rankings, Tokushima prefecture is a typical rural area in Japan.

A. Data to Obtain Corpora of Twitter Users in and outside Tokushima

Our estimation method is based on Tokushima and outside Tokushima user corpora. First, we selected 100 Twitter users who live in Tokushima Prefecture. The residence location of the users was confirmed from the location field (Figure 1). In the same manner, we selected 100 Twitter users who live outside Tokushima Prefecture. We targeted a total of 200 selected Twitter users to obtain the Tokushima and outside Tokushima user corpora.

Figure 1. Location field.

We collected the most recent 200 tweets of each of the selected total 200 Twitter users (2016/4/1). No original tweets as mention or reply are excluded from the tweet collection. Tweets containing only URL are excluded from the collection because they do not comprise any useful information for preparing the corpora. In the collection, we obtained 11,796 tweets of users residing in Tokushima and 9,532 tweets of users outside Tokushima. These tweets were categorized as Tokushima user's collection or outside Tokushima user's collection.

B. Corpora

Japanese language is unsegmented by space. Morphological analysis was performed on the texts of the tweets in our collections to write messages with spaces in clauses (Table 1). As a result of the analysis, we obtained 14,998 nouns, 1,196 adjectives, 2,904 verbs, and 920 adverbs. Therefore, the tweet collections have 18,918 vectors as total words.

We performed latent semantic analysis (LSA) for the dimension reduction of the vectors. LSA [10] is an application of the principal component analysis of language processing. It gives sentence semantic vectors from their co-occurrence.

Table 1. Example of morphological analysis of Japanese text.

Text: 徳島は今日、晴れている。
↓ a result of morphological analysis

Morpheme	***Into English***	***Part of speech***
徳島	Tokushima	noun
は		postpositional particle
今日	today	noun
、	,	sign
晴れ	clear, fine	verb
て		postpositional particle
いる	is	verb

According to the results of LSA, 20 factors were identified for the Tokushima user's corpus from the Tokushima user's collection, and the cumulative contribution ratio of these 20 factors was 0.828. Table 2 shows the Tokushima user's corpus. We adopted factors that have eigenvalues greater than 1.00 as the reference value. Table 2 shows examples of the representative words of the factors with factor loadings greater than 0.4. In the same manner, 22 factors were identified for the corpus of users outside Tokushima from the collection of users outside Tokushima, and the cumulative contribution ratio of these 22 factors was 0.989 (Table 3). Tweets are expressed as the principal component scores of the total 42 factors using these corpora (Table 4).

Table 2. Corpus of Tokushima users obtained using LSA.

Factor No.	***eigenvalue***	***contribution ratio***	***representative words (into English)***
Tokushima1	6.290	0.081	fine, imprtant, smile
Tokushima2	6.236	0.080	this week, tired
Tokushima3	6.156	0.079	today
Tokushima4	5.991	0.077	morning
Tokushima5	5.771	0.074	congratulations
Tokushima6	5.493	0.071	today, good morning, laugh
Tokushima7	5.210	0.067	voltis(*1
Tokushima8	4.879	0.063	
Tokushima9	4.529	0.058	Tokushima
Tokushima10	4.144	0.053	Oosaka, Tokyo
Tokushima11	3.783	0.049	This year
Tokushima12	3.424	0.044	news
Tokushima13	3.047	0.039	
Tokushima14	2.697	0.035	life, time
Tokushima15	2.348	0.030	
Tokushima16	2.029	0.026	sleep
Tokushima17	1.733	0.022	eat, buy
Tokushima18	1.467	0.019	enter
Tokushima19	1.231	0.016	work
Tokushima20	1.031	0.013	a, the

(*1 "voltis" is the name of a soccer team in Tokushima Prefecture.)

Table 3. Corpus of users outside Tokushima obtained using LSA.

Factor No.	*eigenvalue*	*contribution ratio*	*representative words (into English)*
outside1	5.910	0.072	good morning, good night
outside2	5.873	0.071	today, now
outside3	5.829	0.071	like, free
outside4	5.713	0.069	delicious
outside5	5.535	0.067	problem
outside6	5.342	0.065	
outside7	5.144	0.062	good, really, pretty
outside8	4.899	0.059	hot
outside9	4.601	0.056	beautiful
outside10	4.296	0.052	
outside11	3.971	0.048	
outside12	3.659	0.044	work, participation
outside13	3.339	0.040	
outside14	3.025	0.037	present
outside15	2.730	0.033	myself, everyone
outside16	2.446	0.030	
outside17	2.159	0.026	Japan
outside18	1.899	0.023	
outside19	1.646	0.020	
outside20	1.430	0.017	
outside21	1.232	0.015	Mt.Fuji, Haneda
outside22	1.049	0.013	net, blog

Table 4. Example of analysis of tweets.

	Tokushima 1	…	*Tokushima 20*	*Outside 1*	…	*outside 22*
Tweet1	0.012		6.989	1.254		3.256
Tweet2	-0.160		7.265	2.389		-1.560
・ ・ ・						

C. Estimation of Residence Location using Machine Learning

Our proposed method generates a decision tree using machine learning. A decision tree is a decision-making device that assigns a probability to each possible choice based on the context of a decision [11] (Figure 2). In this study, a decision tree is generated based on the corpora of users in and outside Tokushima. Four machine learning algorithms are used as the production method of the decision tree: random forest algorithm, C4.5 algorithm, NBTree, and REPTree. Short introductions of these decision tree algorithms are as follows.

Random forest algorithm: Random forest algorithm is an ensemble classifier. Each classifier uses random feature selection to generate a decision tree. Prediction data is obtained based on the majority result for classification or the average result for regression [12].

C4.5 algorithm: C4.5 algorithm generates a decision tree using recursive partitioning data. The production process is based on a depth-first strategy. The algorithm attempts to perform all possible combinations and then selects a combination that provides the best benefit [13].

NBTree: This algorithm uses naive Bayesian classification. In the production process of a decision tree, a naive Bayes model is constructed for each leaf node using the data associated with that leaf node [14].

REPTree: REPTree is a fast decision tree learner. This algorithm constructs a decision/regression tree using information of the gain/variance and prunes the tree using reduced-error pruning. Values of numeric attributes are only sorted once. Missing values are addressed by splitting corresponding instances into pieces [15].

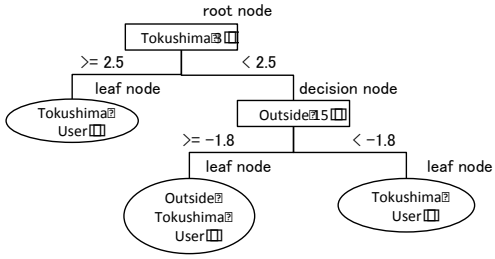

Figure 2. Example of decision tree.

We investigated the performance of the algorithms. The result of the investigation is shown in Table 5. From Table 5, the random forest algorithm is found to have the best performance regarding the classification of tweets. Therefore, the random forest algorithm is adopted for estimating the residence location of Twitter users.

Table 5. Result of comparison of different algorithms.

	Correctly Classified Tweets (%)
Random Forest	82.58
C4.5	80.95
NBTree	78.89
REPTree	66.35

As reported in subsection A of section II, 100 Twitter users who live in Tokushima Prefecture and 100 Twitter users who live outside the prefecture were selected anew for the target data of the estimation. Next, tweets posted by our selected Twitter users were collected (2016/4/2). Original tweets were extracted from the collection. We obtained 11,949 tweets from users in Tokushima and 9,046 tweets from users outside Tokushima. We combined all tweets posted by these 200 users as the test collection. As reported in subsection B of section II, morphological analysis and LSA were performed on the text of each tweet in the test collection.

We estimated the tweets of Tokushima users from the test collection via the decision tree using the random forest algorithm. The result of the estimation is shown in Table 6.

Table 6. Results of estimation of tweets of Tokushima users.

		Real number of tweets	
		Users in Tokushima	***Users outside Tokushima***
Estimated number of tweets	Users in Tokushima	8,424 (70.50%)	3,385
	Users outside Tokushima	3,525	5,661 (62.58%)
	Total	11,949	9,046

Accuracies of the estimation are given in parentheses.

From Table 6, the accuracies of both the tweets of users in and outside Tokushima are worse than the result of the investigation for the performance of random forest algorithm, as shown in Table 5. The tweets of users in Tokushima show better accuracy than those of users outside Tokushima. The tweet corpus of users in Tokushima has a greater number of eigenvalue factors (Table 2) compared with users outside Tokushima. Therefore, it is assumed that the Tokushima user's tweet corpus has a strong influence on the production process of a decision tree. When obtaining opinions of tourists who visit Tokushima, we require a collection of tweets of users outside Tokushima. In this case, using all the Tweets collection excluded that estimated Tokushima user's Tweet is better than using estimated outside Tokushima user's Tweet.

III. CONCLUSION

In this study, we estimated the residence location of Twitter users based on their posted messages. Our estimation method is primarily based on the use of machine learning to generate decision trees. LSA was adopted for the dimension reduction of vectors. The target of our estimation was the classification between the tweets of Twitter users in and outside Tokushima.

First, we collected the Twitter data and performed morphological analysis on the collection. Next, LSA was performed on tweets in the collection to obtain a corpus. A decision tree was generated using the corpus. The random forest algorithm was selected as the best machine learning algorithm for generating a decision tree. Finally, we estimated the tweets of Twitter users in Tokushima using a decision tree. As reported in the result, the accuracy of our estimation method is over 60%.

This study is an important contribution because it demonstrated that tweets have significant differences depending on the residence location of each user. We confirmed the possibility that a tweet collection can replace a questionnaire because determining the residence location is important when obtaining opinions of tourists.

REFERENCES

[1] Japan Association of Travel Agents, quantitative data for tourism 2014, 2014, [https://www.jata-net.or.jp/data/stats/2014/pdf/2014_sujryoko.pdf].

[2] Ryan, Chris, Researching tourist satisfaction: issues, concepts, problems, Routledge, 1995.

[3] McCool, Stephen F., and Steven R. Martin. "Community attachment and attitudes toward tourism development." Journal of Travel research 32.3 pp.29-34, 1994.

[4] Tsujioka, S., A. Kondo, and K. Watanabe. "Relative analysis of factors of place attachment–Case study in the Tokushima urban area." Advances in Civil Engineering and Building Materials IV: Selected papers from the 2014 4th International Conference on Civil Engineering and Building Materials (CEBM 2014), Hong Kong. CRC Press, 2015. http://dx.doi.org/10.1201/b18415-102

[5] 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].

[6] CHENG, Zhiyuan; CAVERLEE, James; LEE, Kyumin. You are where you tweet: a content-based approach to geo-locating twitter users. In: Proceedings of the 19th ACM international conference on Information and knowledge management. ACM, 2010. pp.759-768, 2010.

[7] HECHT, Brent, et al. Tweets from Justin Bieber's heart: the dynamics of the location field in user profiles. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM, 2011. pp. 237-246, 2011.

[8] Yasuhiro Hashimoto and Mizuki Oka, "Statistics of Geo-Tagged Tweets in Urban Areas." Journal of Japanese Society for Artificial Intelligence 27(4), pp.424-431, 2012.

[9] Suguru Tsujioka,"Town Characteristics Estimation using Geotagged Twitter Data -A Case Study in the Tokyo Area-," Proceedings of International conference on "Civil , Architectual and Environmental Engineering, pp.143-147, 2016.

[10] Dumais, Susan T. "Latent semantic analysis." Annual review of information science and technology 38.1, pp.188-230, 2004. http://dx.doi.org/10.1002/aris.1440380105

[11] Magerman, David M. "Statistical decision-tree models for parsing." Proceedings of the 33rd annual meeting on Association for Computational Linguistics. Association for Computational Linguistics, 1995. http://dx.doi.org/10.3115/981658.981695

[12] Breiman, Leo. "Random forests." Machine learning 45.1,pp.5-32, 2001. http://dx.doi.org/10.1023/A:1010933404324

[13] Mitchell, Thomas M. "Machine learning." 1997.

[14] Kohavi, Ron. "Scaling Up the Accuracy of Naive-Bayes Classifiers: A Decision-Tree Hybrid." KDD. Vol. 96. 1996.

[15] Weka project, "Weka- Machine learning software to solve data mining problems," 2016, [http://weka.sourceforge.net/doc.dev/weka/classifiers/trees/REPTree.html].

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 current research interests are town/city characteristics estimation, analysis of decision- making process.

Prof. Akio Kondo was born in Anan city, Japan at 1953. His educational background is urban planning. He received his Ph.D. degree from Kyoto University in 1991. He is working as a professor at the Department of Advanced Technology and Science, Tokushima University (Graduate school), Tokushima, Japan. His current research interests are regional planning, urban planning and transport planning.

Dr. Kojiro Watanabe was born in Fukuoka city, Japan at 1974. His educational background is architecture and urban planning. He was graduated from Toyohashi University of Technology in 2001, and earned PhD (urban planning) in 2001. His major research field is urban planning using GIS. After Post-doctoral researcher in TUT, he is working in Tokushima University as an assistant professor from 2003. He studied following topics: a planning support system by GIS in Asian developing countries, a planning support system for disaster mitigation urban planning and design in Japanese historical built-up area, urban growth modeling by cellular automata and so on. His current study topics is Tsunami disaster mitigation urban planning by GIS.