Available online at www.sciencedirect.com

ScienceDirect

Transportation Research Procedia 00 (2018) 000–000

Transportation Research Procedia

www.elsevier.com/locate/procedia

World Conference on Transport Research - WCTR 2019 Mumbai 26-31 May 2019

Topic Modeling Approach to Analyze Students' Activities on University Campus using Smartphone-based Survey Data

Rintaro Kizaki[a], Tomoki Kawano[a], and Takuya Maruyama[a,*]

[a] *Kumamoto University, Kumamoto 860-8555, Japan*

Abstract

GPS-based tracking data of students in university campuses can be used to examine various campus issues, including illegal bicycle parking and in deciding where to position parking lots. Existing studies have attempted to extract behavioral patterns from GPS-based data; some studies have applied topic modeling to this problem. However, the application of this method to small-scale areas (e.g., a university campus) has been limited. In this study, we apply topic modeling to the data collected via a smartphone-based GPS survey at Kumamoto University Campus, Japan in January 2016. We found several travel patterns among students in the campus using topic modeling. These patterns can be explained using individual attributes, demonstrating the usefulness of this method. It is worth noting that only tracking data are needed for such an analysis, which indicates that in the future, we may infer the individual attributes of tracking data by using the topic modeling approach.

© 2018 The Authors. Published by Elsevier B.V.
Peer-review under responsibility of WORLD CONFERENCE ON TRANSPORT RESEARCH SOCIETY.

Keywords: travel behavior survey, GPS, smartphone, topic model, pattern mining

1. Introduction

A typical university campus has several transport problems. For example, many bicycles located in a cluttered way can impede the safety of the campus and worsen the campus landscape. In addition, congestion is often observed at the cafeteria during lunch time. The behavioral data of students, faculty, and staff within the university would be useful when considering solutions to these problems.

A traditional method of collecting these data has been paper-based travel surveys; however, the method is not sufficiently accurate for a detailed planning of the campus. Furthermore, the survey cost and burden are high. Thus,

* Corresponding author. Tel.: +81-96-342-3489.
E-mail address: takumaru@kumamoto-u.ac.jp

2352-1465 © 2018 The Authors. Published by Elsevier B.V.
Peer-review under responsibility of WORLD CONFERENCE ON TRANSPORT RESEARCH SOCIETY

the use of smartphones and Wifi trace data for surveys has attracted considerable attention in recent time (Danalet et al., 2014; Oosterlinck et al., 2017; Petre et al., 2017; Versichele et al., 2012; Yoshimura et al., 2017, 2014). A large amount of data can be obtained through such surveys, and an efficient procedure is needed to analyze the collected data. This research focuses on topic modeling among the various methods used to analyze tracking data.

The topic model was originally developed as a machine learning method for analyzing text data. An increasing number of studies using topic model have been published since the pioneering work by Blei et al.(2003). Some studies used topic models to examine the trends in transportation research (Das et al., 2017, 2016; Sun and Yin, 2017). Kuhn (Kuhn, 2018) used structural topic modeling to identify latent topics and trends in aviation incident reports. Hasan and Ukkusuri (2014) classified activity patterns using topic models from online geo-location data. Furuya et al.(2018) proposed a classification method of GPS log data of foreign visitors to Japan by using a topic model. Among the existing studies, this is the most relevant research for the present study. They applied their method to the whole of Japan, while we applied their model to a small area, i.e., a university campus.

The objectives of this study are as follows.

- To apply the topic model to classify the behavioral pattern from the data obtained through a smartphone-based survey conducted in Kumamoto University campus.
- To examine the individual attributes of the classified data.

The smartphone-based survey was conducted in January 2016 by using an application developed by our research group. The application has been used in household travel surveys and visitor surveys in downtown Kumamoto, Japan (Maruyama et al., 2015, 2014). A university campus usually has many Wifi access points and examination of Wifi trace data is possible (Danalet et al., 2016). Although we made use of data obtained from the smartphone-app-based survey, the method can be applied to analyze Wifi trace data more easily. Wifi and Bluetooth scanner data usually lack demographic information or individual attributes. Thus, we intend to infer such information from tracking data using the proposed method in the future.

2. Method and Data

2.1. Topic model: Overview

Latent Dirichlet Allocation (LDA), a type of topic model, was used to extract behavioral patterns. Here, we briefly describe the outline of LDA.

LDA is a probabilistic model that can be used to generate a document set represented by Bag of Words (BoW). BoW expressions are vector expressions of words appearing in sentences. In addition, BoW expressions do not consider the structure of sentences but represent the number of occurrences and co-occurrence of words. LDA is a method of clustering words and documents using co-occurrences of words obtained from BoW.

LDA assumes latent variables (topics) that cannot be directly obtained from BoW for each word in the document. As a feature of LDA, a document is composed of multiple topics and has a probability distribution as a composition ratio of the topics. Specifically, we define i th word of document d as $w_{d,i}$ and the corresponding latent variable as $z_{d,i}$. Here, the number of topics is K, $\theta_{d,k}$ (k = 1, 2, …, K) is the probability that the topic k appears in document d. The topic distribution is given by $\theta_d = (\theta_{d,1}\ldots, \theta_{d,K})$. Each topic has an appearance distribution ϕ_k (k = 1, 2, …, K) corresponding to each topic. We denote the number of documents as D, and the sentence length (total number of words) of document d is N_d. $\phi_{d,v}$ is the occurrence probability of word v in topic k, and the appearance distribution of the word is $\phi_k = (\phi_{k,1}\ldots, \phi_{k,V})$. θ_d and ϕ_k are assumed to be generated by the Dirichlet distribution and are arranged as follows.

$$\theta_d \sim \mathrm{Dir}(\alpha),\ d = 1,\ldots, M \quad (1)$$

$$\phi_k \sim \mathrm{Dir}(\beta),\ k = 1,\ldots, K \quad (2)$$

$$z_{d,i} \sim \mathrm{Multi}(\theta_d),\ i = 1,\ldots, N_d \quad (3)$$

$$w_{d,i} \sim \mathrm{Multi}(\phi_{z_{d,i}}),\ i = 1,\ldots, N_d \quad (4)$$

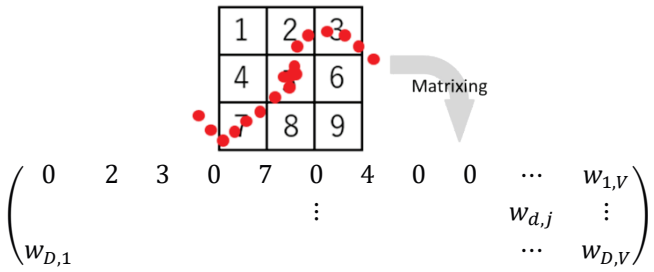

Fig. 1. BoW expression of GPS log data

Table 1. Survey Overview.

Name	Kumamoto University Kurokami campus behavioral survey
Dates	January 19 (Tue) - January 22 (Fri), 2016 (21 and 22 are supplemental days)
Targets	Students in Kumamoto University having an Android or iOS smartphone
Target area	Kumamoto University Kurokami campus and its surroundings
Used devices	Participants' smartphone
Used application	Travel survey application "Suma Kuma"

2.2. Application of Topic model in tracking data

To apply the topic model to the location data, this study divided the target area into meshes. We regard the document d as a trip chain, the word w as a mesh, and the appearance frequency of the word w in the document d, i.e., $n_{w,d}$ as the number of points observed in the mesh. Fig. 1 illustrates the BoW expression of GPS tracking data where w is a mesh, d is GPS log data, D is the total number of trip chains, and V is the total number of meshes in the target area. Note that LDA assumes independence among topics (i.e. behavioral patterns), but a spatial correlation will exist in actual behavioral patterns. This study, however, assumes no correlation in behavioral patterns.

2.3. Data

Table 1 outlines the survey. The survey was conducted at the Kurokami campus of Kumamoto University. We requested the students to participate at least two days during the survey. Standard survey date was January 19 and 20, 2016, but for students that did not go to the campus on these days, they could participate on January 21 and 22 instead. The target sample was undergraduate and graduate students of Kumamoto University owning Android or iOS smartphone. We asked the participants to track their record while they had moved away from their homes until they went back to their homes. However, the data outside campus was deleted for privacy protection. The information obtained in the survey includes location data, the name of faculty, school grade, gender, and OS in a smartphone. Figure 2 shows the map of target campus. In this study, we divided the area into 10-m meshes. Figure 2 illustrates the main facilities in the campus that attracted many visitors. The summary of data is illustrated in Table 2 and Table 3.

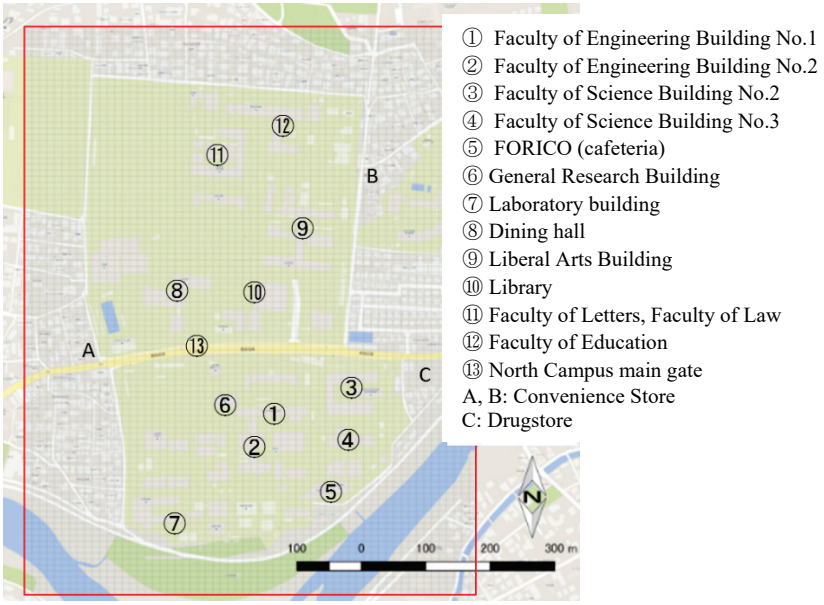

Fig. 2. Kumamoto University Campus

Table 2. Summary of observed data.

		Male	Female	total
Android	(device)	51	10	61
iOS	(device)	61	27	88
Participants (2 days or more)	(person)	112	37	149
Participants 1 day or more)	(person)	174	50	224
Registered ID	(person)	387	88	475
Participation rate (1 day or more)	(%)	45.0	56.8	47.2
Valid participation rate (2 days or more)	(%)	28.9	42.0	31.4

Table 3. Summary of student participation for faculty and school grade. (unit: person)

	literature	education	law	engineering	science	N.A.	Total
1st grader		1		26	3		30
2nd grader		1		42	1		44
3rd grader	4	2	1	35	10		52
4th grader				17	19		36
MC 1				8			8
MC 2				6	8		14
DC					2		2
N.A.						38	38
Total	4	4	1	134	43	38	224

3. Results and Discussion

3.1. Classification of tracking data by topic model

Collapsed Gibbs sampling was used for sampling method. The hyperparameter was set uniformly as $\alpha = (0.1, ..., 0.1)$, $\beta = (0.1, ..., 0.1)$, the number of samplings was 1,000, and the number of topics was 10. Fig. 3 shows the classification of tracking data by the topic model. Each mesh is colored based on $\phi_k = (\phi_{k,1}\ldots, \phi_{k,V})$, where ϕ_k indicates the appearance probability of the word v in the topic k. For the behavioral data analysis, ϕ_k can be interpreted as the contribution ratio of the mesh to the behavioral pattern k.

Fig. 3(a) demonstrates the behavioral patterns observed mainly around the Faculty of Engineering Building No. 1. The behavioral patterns of students who attended lectures in this building is shown in the figure; we speculated that most of them would be students in the school of engineering. Fig. 3(b) demonstrates the behavioral patterns found mainly around Faculty of Science Building No. 2. Fig. 3(c) demonstrates the behavioral patterns around the university cafeteria (FORICO), Faculty of Engineering Building No.1 and No. 2. Fig. 3(d) shows the pattern around the Faculty of Science Buildings, indicating the students in the faculty of science. Fig. 3(d) shows the behavioral pattern around the liberal arts building and the general research building. These facilities are used mainly by freshman and sophomore. Therefore, it will indicate the behavioral pattern of students of lower school grades.

Fig. 3(f) demonstrates the pattern around the library and cafeteria, indicating the students who use such facilities. Fig. 3(g) shows a pattern around the faculty of education, literature, and law. Fig. 3(h) shows a pattern around cafeteria (FORICO) and Faculty of Engineering Building No. 2. The Faculty of Engineering Building No. 2 is mainly used for lectures in the school of engineering for 1 to 3 graders. These patterns are generated by students taking lecture and going for lunch. Fig. 3(i) reveals the behavioral patterns around the general research building and the experiment building, indicating that it will be given by the students in the school of engineering fourth grader and graduate student. Fig. 3(j) demonstrates patterns along the western side of the campus, indicating students commuting to school or returning home.

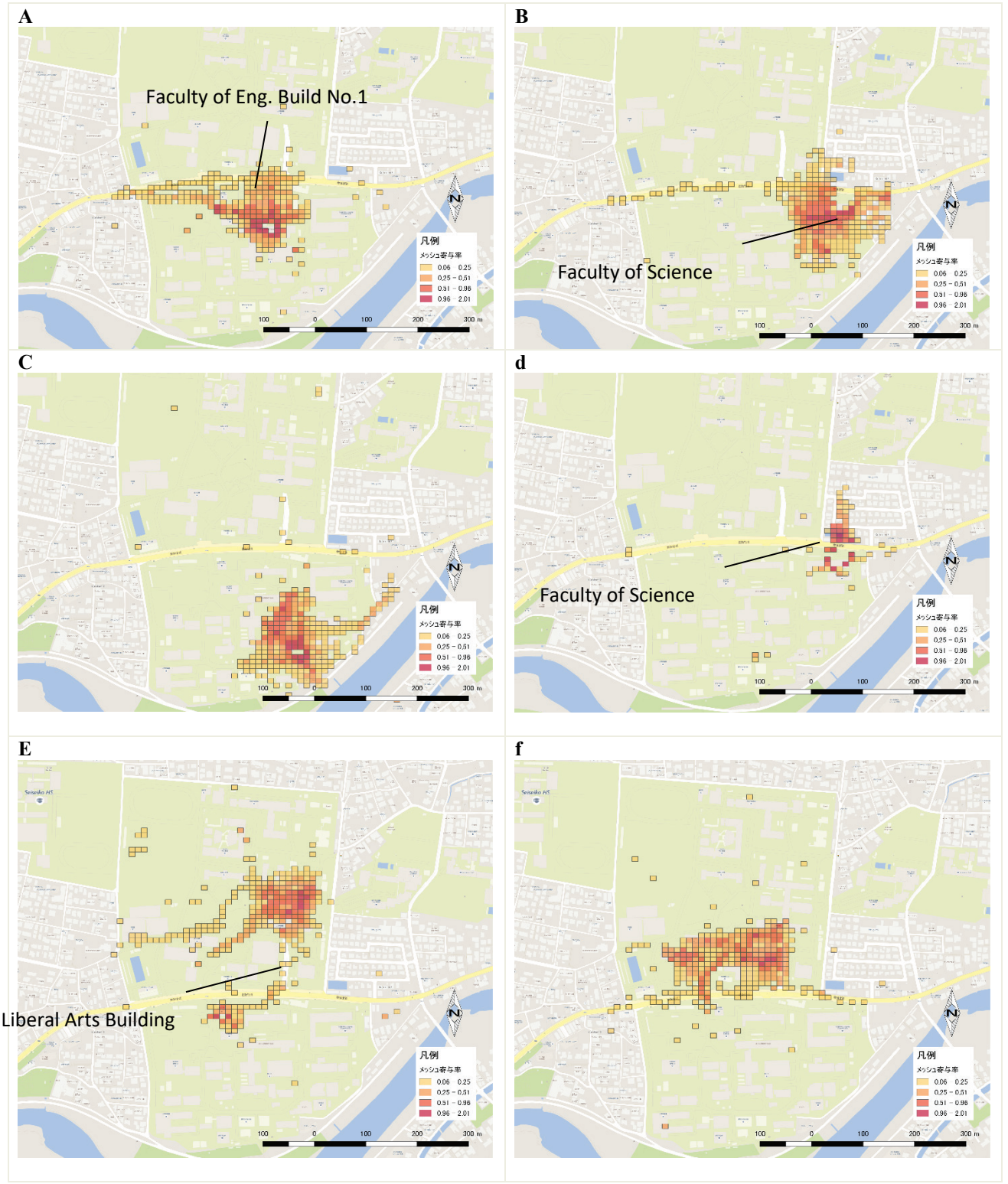

Fig. 3. Classified tracking data by topic model

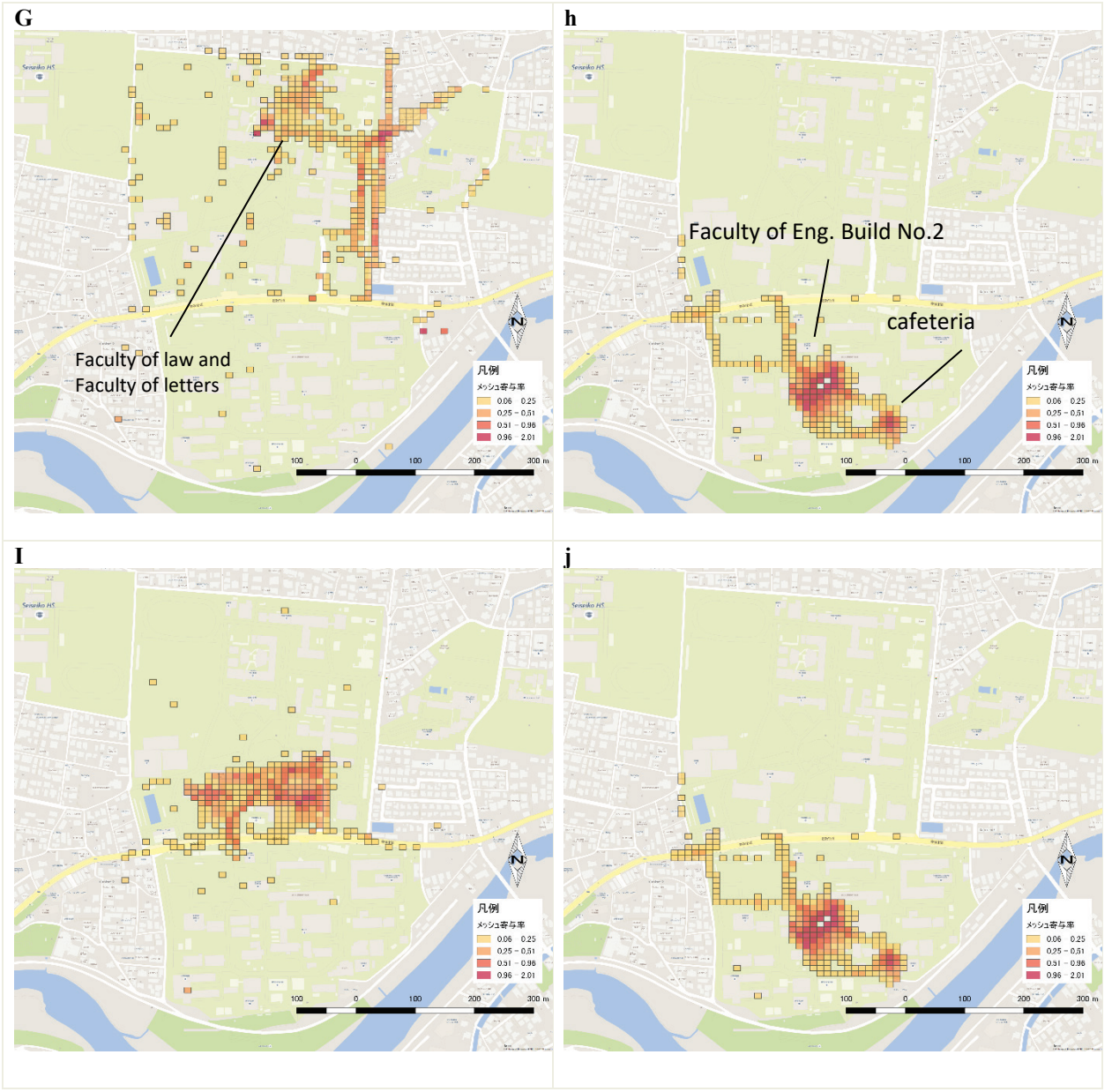

Fig. 3. Classified tracking data by topic model (continued)

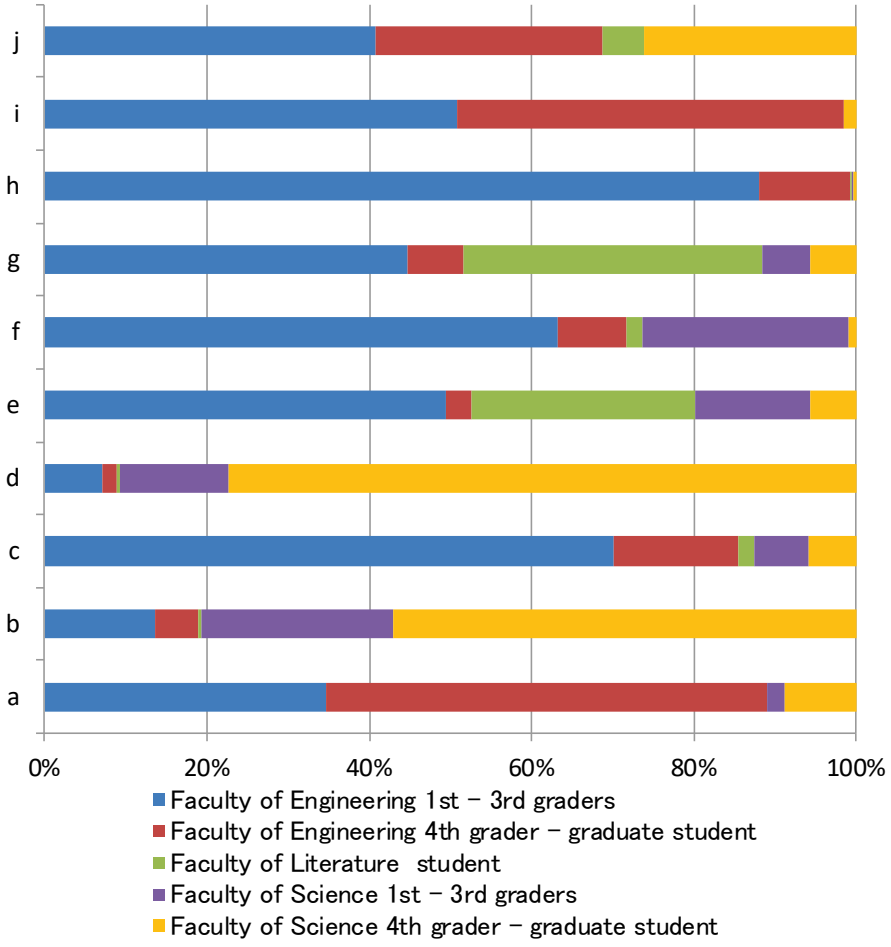

Fig. 4. Classified tracking data and its individual attributes by faculty and school grade

3.2. Classified patterns and individual attribute

Here, we analyze the relationship between the classified patterns by topic models and individual attributes collected by the smartphone-based survey. The probability $\theta_{d,k}$ that the topic k appears in the sentence d in the original topic model corresponds to the probability that sample d is classified as behavioral pattern k.

Fig. 4 demonstrates this analysis based on the probability $\theta_{d,k}$. Here, we classified student category into five groups by faculty and school grade, which are assumed to be homogeneous. Faculties in human and social sciences are classified as one group because the samples are small. Patterns (a), (c), (h), and (i) reveals a high proportion of engineering students, and patterns (b) and (d) reveals a high proportion of students in the school of science. This is consistent with their daily usage of facilities. Pattern (e) has a high proportion of 1st−3rd grade students in the school of engineering, science, and engineering which is consistent with the discussion in the previous subsection. The pattern (f) reveals a high proportion of 1 - 3 graders in the school of engineering and science. This is because fourth graders and graduate students spend more time in the laboratory, and thus, their use of libraries decreases, while the first to third graders would likely go to the library to seek out a learning space.

4. Conclusion

In this research, we analyzed smartphone-based tracking data collected at the Kumamoto University campus using topic models. Topic models were originally used to analyze text data, and this study considered the document and words in the original topic model as trip chain and mesh, respectively. We successfully retrieved 10 behavioral patterns by topic models, and these patterns can be explained using individual attributes, demonstrating the usefulness of this approach. Only tracking data are needed for such an analysis, indicating that in the future, we may infer the individual attributes of tracking data by using the topic modeling approach.

References

Blei, D.M., Ng, A.Y., Edu, J.B., 2003. Latent Dirichlet Allocation. J. Mach. Learn. Res. 3, 993–1022.

Danalet, A., Farooq, B., Bierlaire, M., 2014. A Bayesian approach to detect pedestrian destination-sequences from WiFi signatures. Transp. Res. Part C Emerg. Technol. 44, 146–170. https://doi.org/10.1016/j.trc.2014.03.015

Danalet, A., Tinguely, L., Lapparent, M. de, Bierlaire, M., 2016. Location choice with longitudinal WiFi data. J. Choice Model. 18, 1–17. https://doi.org/10.1016/j.jocm.2016.04.003

Das, S., Dixon, K., Sun, X., Dutta, A., Zupancich, M., 2017. Trends in Transportation Research. Transp. Res. Rec. J. Transp. Res. Board 2614, 27–38. https://doi.org/10.3141/2614-04

Das, S., Sun, X., Dutta, A., 2016. Text Mining and Topic Modeling of Compendiums of Papers from Transportation Research Board Annual Meetings. Transp. Res. Rec. J. Transp. Res. Board 2552, 48–56. https://doi.org/10.3141/2552-07

Furuya, H., Okamoto, N., Nozu, N., 2018. Development of the Combination Analysis Method of Visited Places of Foreign Visitors to Japan by GPS Log Data. Transp. Policy Stud. Rev. 20, 20–29. https://doi.org/10.24639/tpsr.20.0_020

Hasan, S., Ukkusuri, S. V., 2014. Urban activity pattern classification using topic models from online geo-location data. Transp. Res. Part C Emerg. Technol. 44, 363–381. https://doi.org/10.1016/j.trc.2014.04.003

Kuhn, K.D., 2018. Using structural topic modeling to identify latent topics and trends in aviation incident reports. Transp. Res. Part C Emerg. Technol. 87, 105–122. https://doi.org/10.1016/j.trc.2017.12.018

Maruyama, T., Mizokami, S., Hato, E., 2014. A smartphone-based travel survey trial conducted in Kumamoto, Japan: An examination of voluntary participants' attributes, in: Transportation Research Board 93rd Annual Meeting Compendium of Papers. p. #14-0997.

Maruyama, T., Sato, Y., Nohara, K., Imura, S., 2015. Increasing smartphone-based travel survey participants, in: Transportation Research Procedia. pp. 280–288. https://doi.org/10.1016/j.trpro.2015.12.024

Oosterlinck, D., Benoit, D.F., Baecke, P., Van de Weghe, N., 2017. Bluetooth tracking of humans in an indoor environment: An application to shopping mall visits. Appl. Geogr. 78, 55–65. https://doi.org/10.1016/j.apgeog.2016.11.005

Petre, A.C., Chilipirea, C., Baratchi, M., Dobre, C., van Steen, M., 2017. WiFi Tracking of Pedestrian Behavior. Smart Sensors Networks Commun. Technol. Intell. Appl. 309–337. https://doi.org/10.1016/B978-0-12-809859-2.00018-8

Sun, L., Yin, Y., 2017. Discovering themes and trends in transportation research using topic modeling. Transp. Res. Part C Emerg. Technol. 77, 49–66. https://doi.org/10.1016/j.trc.2017.01.013

Versichele, M., Neutens, T., Delafontaine, M., Van de Weghe, N., 2012. The use of Bluetooth for analysing spatiotemporal dynamics of human movement at mass events: A case study of the Ghent Festivities. Appl. Geogr. 32, 208–220. https://doi.org/10.1016/j.apgeog.2011.05.011

Yoshimura, Y., Amini, A., Sobolevsky, S., Blat, J., Ratti, C., 2017. Analysis of pedestrian behaviors through non-invasive Bluetooth monitoring. Appl. Geogr. 81, 43–51. https://doi.org/10.1016/j.apgeog.2017.02.002

Yoshimura, Y., Sobolevsky, S., Ratti, C., Girardin, F., Carrascal, J.P., Blat, J., Sinatra, R., 2014. An analysis of visitors' behavior in the louvre museum: A study using bluetooth data. Environ. Plan. B Plan. Des. 41, 1113–1131. https://doi.org/10.1068/b130047p