Journal of Big Data

RESEARCH Open Access

A method for extracting travel patterns using data polishing

Mio Hosoe*, Masashi Kuwano and Taku Moriyama

*Correspondence:
d19t4003b@edu.tottori-u.ac.jp
Tottori University, 4-101 Koyama-Minami, Tottori, Japan

Abstract

With recent developments in ICT, the interest in using large amounts of accumulated data for traffic policy planning has increased significantly. In recent years, data polishing has been proposed as a new method of big data analysis. Data polishing is a graphical clustering method, which can be used to extract patterns that are similar or related to each other by identifying the cluster structures present in the data. The purpose of this study is to identify the travel patterns of railway passengers by applying data polishing to smart card data collected in the Kagawa Prefecture, Japan. To this end, we consider 9,008,709 data points collected over a period of 15 months, ranging from December 1st, 2013 to February 28th, 2015. This dataset includes various types of information, including trip histories and types of passengers. This study implements data polishing to cluster 4,667,520 combinations of information regarding individual rides in terms of the day of the week, the time of the day, passenger types, and origin and destination stations. Via the analysis, 127 characteristic travel patterns are identified in aggregate.

Keywords: Public transport, Smart card, High order data

Introduction

With recent developments in ICT, various types of data are being generated and accumulated in real time. The amounts and types of available data have increased considerably, and big data, in particular, has garnered significant attention. Big data comprises multiple kinds of data from various fields, including social media data, multimedia data, sensor data, and log data. Analysis and visualization of big data is expected to enable the recognition of phenomena that are not apparent otherwise, thereby making the creation of new knowledge possible.

In the field of transportation research, big data such as GPS data and probe vehicle data have been analyzed to understand behaviors of travelers [1–3]. In particular, several researchers have analyzed smart card data to draw conclusions regarding the behavior of transit users [4–6]. Smart cards were originally developed for efficient fare payment and/or toll collection. However, smart card data also contain information about particular ticket gates at particular stations that were passed by passengers at particular times of the day as well as at their destination stations. Therefore, they allow analysts to understand the temporal and spatial travel behaviors of smart card users.

Analyzing smart card data is important for understanding such travel patterns of individuals. In addition, it is expected that the results of such analysis may be used as new material to be considered during the development of traffic policy. However, most previous studies have focused on only a small number of data items present in smart card data—the number of uses on each day of the week and different times of the day, the stations of origin and destination at different times of the day, etc. [7, 8]. Alternatively, traditional travel behavior analysis has focused on single data items and their specific elements [9–11]. Thus, the aforementioned studies have failed to simultaneously consider multiple data items in smart card data.

It is understood that the different attributes present in smart card data (such as origin station, destination station, the time of the day, the day of the week, and passenger type) affect each other. However, consideration of specific data items extracted from smart card data impedes the exploration of effects wielded by the excluded data items on travel behaviors. Moreover, the results may vary widely depending on the data items being analyzed. From the perspective of effective big data analysis, this study advocates the simultaneous analysis of as many data items as possible. To this end, it studies patterns in travel behaviors of individuals based on smart card data collected from a provincial city in Japan. Five data items (henceforth referred to as attributes) are used in the analysis—boarding day of the week (henceforth referred to as day), boarding time of day (henceforth, time), passenger type, origin station, and destination station.

Methodologies based on tensor decomposition have been proposed for the simultaneous consideration of multiple attributes [12–14]. Tensor decomposition can be an effective method to analyze data of the 3rd order or higher [15]. Moreover, it enables analysis without disturbing the original data structure [16]. A tensor representation allows the summarization of multivariate data in a multi-dimensional array. The tensors of the lowest order are referred to by specific common names—a 0-order tensor is called a scalar, a 1st order tensor is called a vector, and a 2nd order tensor is called a matrix [14, 17]. Tucker decomposition—a particular model of tensor decomposition—estimates the factor matrices that represent the characteristics of each attribute in high-order data [18–20]. The characteristics of each attribute are called factors. The number of factors are determined arbitrarily [19]. In addition, a core tensor representing the combination of factors corresponding to each attribute is estimated alongside the factor matrices [18–20]. Tucker decomposition reveals the interactions between attributes in the original data based on the estimated factor matrices and core tensors. However, tensor decompositions, such as factor analysis, exhibit greater complexity of results when the number of attributes is increased. In addition, as the number of elements corresponding to each attribute is increased, the unique determination of the number of factors using tensor decomposition and interpretation of the components of the factors becomes progressively more difficult [12]. This complicates the understanding and interpretation of factor matrices and core tensors [12]. Finally, tensor decomposition is incapable of extracting the characteristics of elements based on a small number of samples.

This study attempts to analyze multiple attributes simultaneously by constructing a graph. This approach is different from those used in previous studies and is novel to this study. An increase in the number of combinations of attributes, i.e., the number of vertices and edges, increases the complexity of the graph structure. It is difficult to grasp data characteristics

from a complex graph. To address this problem, this study extracts groups of more relevant vertices from a graph based on the similarity between vertices. Several pattern extraction methods using the similarity index have been proposed in the literature [21–23]. However, they are unsuitable for the extraction of patterns from graphs. Graphs are represented by two-dimensional tables, which are symmetric matrices with zero diagonal elements. In this case, the combinations of column information for each row are different. This should be noted during the selection of the optimal pattern extraction method.

This study implements a data polishing approach to extract travel patterns from the graph. It clarifies group boundaries based on a hypothesis—two vertices have multiple common neighbors in a graph if both are included within a dense sub-graph of a certain size [24]. In data polishing, all vertex pairs possessing at least a certain number of common neighbors (i.e., those whose neighbors have similarity that is not less than the given threshold) are identified, and each pair is connected by an edge [24]. On the other hand, all edges whose endpoints do not satisfy the condition are deleted because they are not considered to lie within the same cluster [24]. The graph is progressively updated by repeating this operation.

Data polishing modifies the input data, enabling the extraction of groups of related vertices without the loss of group structures in the data [24]. In addition, it enables the extraction of vertex groups without depending on the number of samples by focusing on the similarity of vertices instead. This advantage is particularly useful in this study as smart card data of elderly citizens or children may have small sample sizes. Therefore, we adopt data polishing as the preferred method to extract the travel behaviors of such passengers. In addition, as data polishing is a soft clustering method, it enables multiple characteristics of each vertex to be captured. This makes the extraction and analysis of travel patterns via this method particularly flexible.

This study proposes an improved version of the existing data polishing method to analyze smart card data, which considers multiple attributes simultaneously. We extract the travel patterns of smart card users in terms of five attributes. Initially, the “usage vertices” comprising three attributes—day, time, and passenger type—are classified in terms of the strength of connection with “Origin and Destination (henceforth, OD) vertices” composed of two attributes—origin station and destination station. Then, groups of usage vertices with similar connections to OD vertices are extracted. In addition, the origin station and destination station combinations are clarified with the largest number of users for each usage group. Via this process, an understanding of the characteristic travel patterns is gathered in terms of the origin and destination stations of passengers on particular days of the week and at particular times of the day.

The remainder of this paper is organized as follows. The smart card data used in this study is introduced in "Data description and aggregate analysis" section. The proposed data polishing-based method used to extract travel patterns is described in "Method" section. The results of the analysis are presented in "Results" section and discussed in "Discussion" section. Finally, the paper is concluded in "Conclusions" section.

Data description and aggregate analysis

Data description

This study uses smart card data collected in the Kagawa Prefecture in Japan (Fig. 1). Kagawa is located in the northeast part of Shikoku, one of the four main islands of Japan. It has a population of approximately 980,000 people (as of October 1, 2015). Most of them depend on automobiles for transportation. However, people who cannot drive automobiles, such as the elderly, rely on public transportation. The aging rate of the population in Kagawa is approximately 30%, and it is expected to increase. This makes the maintenance of public transportation particularly important. Improvement of transit services requires a thorough understanding of the travel behavior of current users.

The smart card used in Kagawa is called IruCa. IruCa was introduced to be used on Kotoden trains or buses operated by the Takamatsu Kotohira Electric Railroad Company and on buses operated by other bus companies in the prefecture. As of March 2016, 341,706 IruCa cards had been issued. Of these, 75,169 were commuter passes and 266,546 were non-commuter passes. It is understood that most IruCa users are residents of the Kagawa Prefecture because IruCa can be used only within the prefecture. Therefore, it is reasonable to expect that this study sheds light on the travel behaviors of the residents of the Kagawa Prefecture based on IruCa data. Although IruCa can be used on both trains and buses, this study only focuses on smart card data related to Kotoden trains.

The associated train network comprises 52 stations in total, including the Kotohira, Nagao, and Shido lines (as depicted in Fig. 2). Two stations, Takamatsu-Chikko and Kataharamachi, are connected to both the Kotohira and Nagao lines. The Kawaramachi station is connected to all three lines, enabling Kotoden passengers to transfer to any line at this station.

This study focuses on five attributes related to each trip (day, time, passenger type, origin station, and destination station) collected within the smart card data as depicted

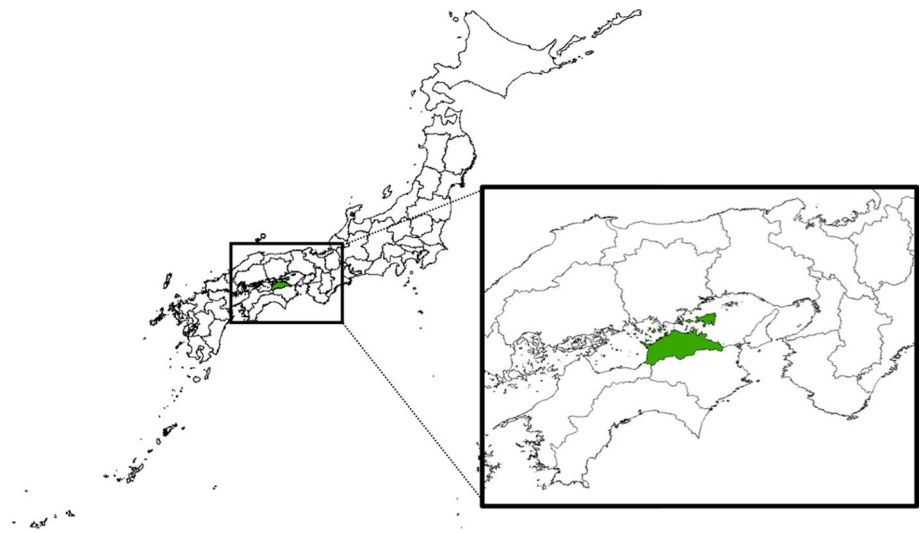

Fig. 1 Kagawa prefecture

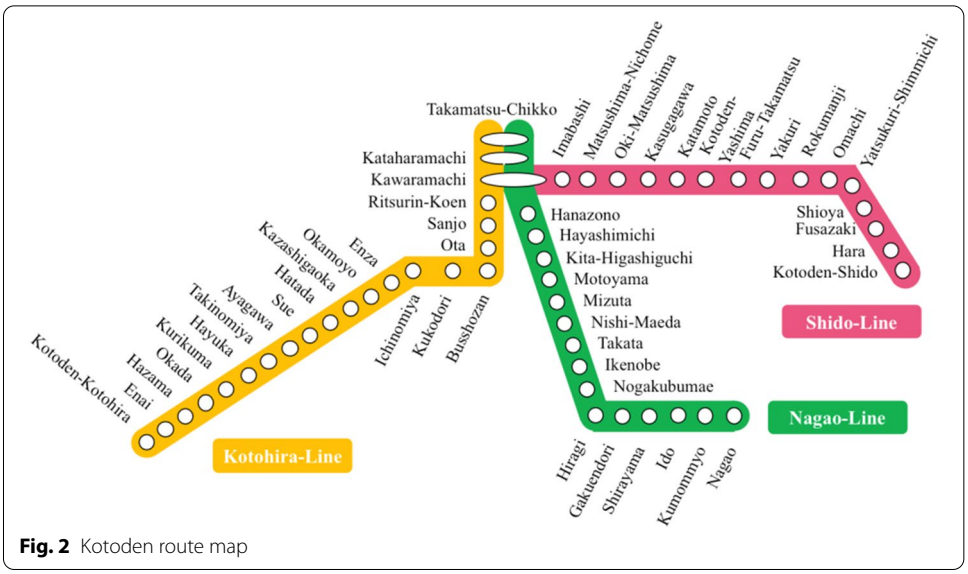

Fig. 2 Kotoden route map

Table 1 Five attributes

Attribute	No. of categories	Description
Day	8	(1) Monday, (2) Tuesday, (3) Wednesday, (4) Thursday, (5) Friday, (6) Saturday, (7) Sunday, (8) Public holiday
Time	20	(1) 5:00–5:59, (2) 6:00–6:59, (3) 7:00–7:59, …, (20) 24:00–24:59
Passenger type	11	(1) Adult, (2) Student, (3) Child, (4) Elderly person, (5) Person with disability, (6) Adult commuter for work, (7) Adult commuter for school, (8) Child commuter not for school, (9) Child commuter for school, (10) Disabled commuter for work, (11) Disabled commuter for school
Origin station	52	52 stations
Destination station	52	52 stations

in Table 1. "Passenger type" is determined based on the indicated categories. Passengers of types (1)–(5) are users of non-commuter passes. On the other hand, passengers of types (6)–(11) are commuter pass users. "Commuter for work" represents a person who uses a commuter pass to travel to work. "Commuter for school" represents a person who uses a commuter pass to travel to school, and "Commuter not for school" represents a person who uses a commuter pass to travel but not to school. Smart card data collected over a period of 15 months, ranging from December 1, 2013 to February 28, 2015, is used. 9,033,748 data points were collected during this period. Of these, this study uses 9,008,709 data points corresponding to "valid" Kotoden trips determined on the basis of the following criteria—(1) the card was used for any train ride during the hours of operation between 5 a.m. and 12 p.m. on any of the three lines, and (2) it took at least 60 s to move between the origin and destination stations.

Aggregate analysis

Figures 3, 4 and 5 depict the average number of daily users, sorted by day, time, and passenger type, respectively. These reveal the usage patterns of the Kotoden train network.

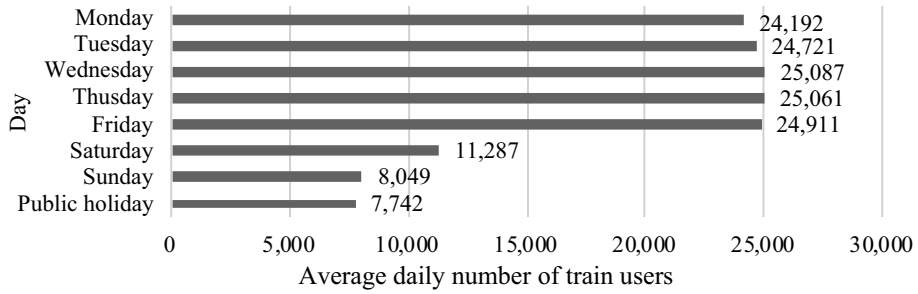

Fig. 3 Average number of train users sorted by the day of the week

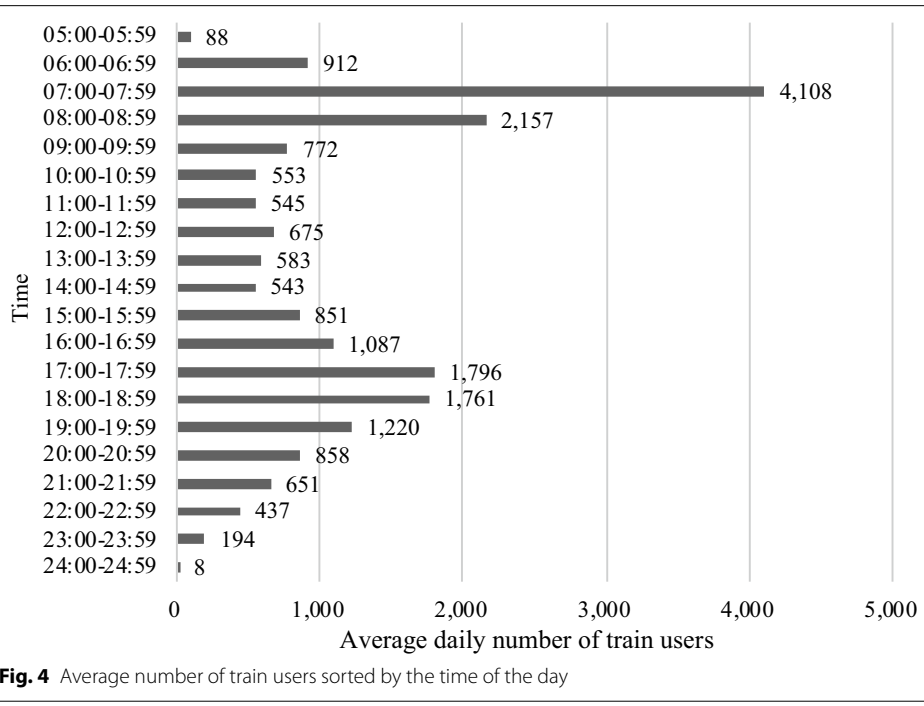

Fig. 4 Average number of train users sorted by the time of the day

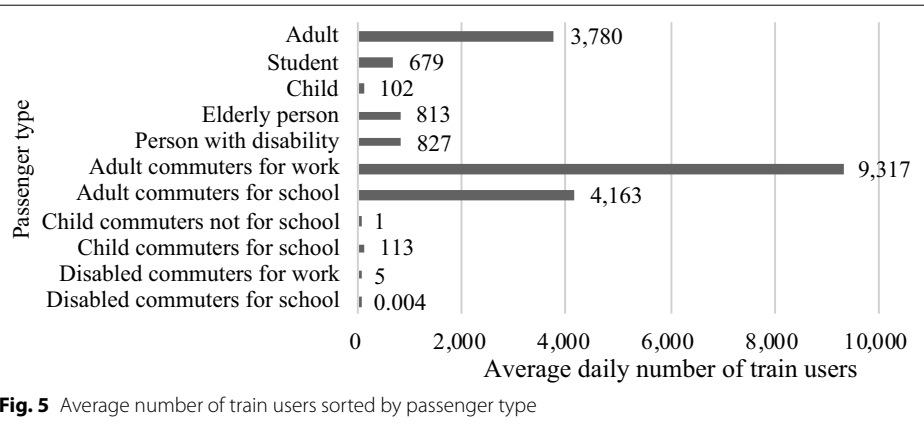

Fig. 5 Average number of train users sorted by passenger type

From the results sorted by the days of the week presented in Fig. 3, it is evident that the average numbers of users on different weekdays are nearly identical. The average number of users on weekdays is 24,794, which is approximately 2.7 times the average number of users on Saturdays, Sundays, and public holidays (approximately 9026). This is consistent with the large number of weekday trips that are taken for commuting to schools and offices.

From the results of usage sorted by the times of the day presented in Fig. 4, it is evident that the average number of users is highest between 7:00 and 7:59 hrs. The average number of users steadily increases between 5:00 and 7:59 hrs, which is understood to be accounted for by commuters to work and school. The average number of users then decreases between 8:00 and 11:59 hrs, and thereafter, the average daily number of users remains approximately constant until 14:59 hrs. After 15:00 hrs, the average daily number of users exhibits another increment, and is particularly high between 17:00 and 18:59 hrs. This is understood to be accounted for by people returning home after 17:00 hrs.

The results of usage sorted by passenger type, as presented in Fig. 5, reveals that the average number of users classified as "Adult" is high, and that the average number of adult commuters to work is as high as 47% (almost half) of the total number of users. By contrast, it is revealed that the average number of child commuters to work and persons with disabilities commuting to work or school are as small as 0.00497%, 0.02696%, and 0.00002%, respectively.

Method

Preliminaries

A graph consists of a vertex set, V, and an edge set, E. All graphs denoted in this paper are undirected graphs.

Given a pair of vertices, u and v, they are termed "adjacent" if they are connected by an edge. A vertex, u, adjacent to v is also called a neighbor of v. The set of neighbors of v is denoted by $N(v)$. A vertex, w, is a common neighbor of the vertices, u and v, if w is adjacent to both u and v. $N[v]$ denotes $N(v) \cup \{v\}$ and is called the closed neighborhood of v. The number of vertices adjacent to the vertex v is denoted by$|N(v)|$.

A vertex set in which every pair of vertices is connected by an edge is called a clique and is usually denoted by C. Although cliques are usually in terms of a subgraph, we adopt the aforementioned definition in his paper, following Uno et al. [24]. A clique that is not completely included in any other distinct clique is called a maximal clique.

Extraction of travel patterns via data polishing

This section explains a new methodology for the extraction of travel patterns from smart card data based on the method of data polishing. We define travel patterns in terms of combinations of five attributes—day (8 categories), time (20 categories), passenger type (11 categories), origin station (52 categories), and destination station (52 categories). These five attributes are considered simultaneously to analyze the types of people who move between particular origin and destination stations at different times of the day on different days of the week.

Uno et al. [24] focused on the extraction of maximal cliques and proposed a clustering method comprising three sub-procedures—(1) construction of a similarity graph, (2) application of data polishing to the similarity graph, and (3) enumeration of maximal cliques. This study introduces additional steps to the aforementioned clustering method proposed by Uno et al. and proposes a corresponding method for the analysis of smart card data. In addition, we do not focus solely on maximal cliques, but on all cliques (Please consult Enumeration of Cliques (Sect. 4 in this section) for further details).

The procedure for extracting travel patterns proposed in this study comprises five steps—(1) construction of the co-occurrence graph, (2) construction of the similarity graph, (3) application of data polishing to the similarity graph, (4) enumeration of cliques, and (5) extraction of combinations of origin and destination stations related to each clique. Each step is explained below in order by using conceptual figures. In this study, an iMac (CPU: Intel Core i7, Memory: 32 GB) was used to execute all the aforementioned steps. The total execution time was approximately 30 minutes (using programs written in the R language).

1) Construction of the co-occurrence graph.

- This study defines the graph constructed based on the co-occurrence relationships between usage vertices and OD vertices to be the co-occurrence graph, G_c. The graph depicted in Fig. 6 is denoted by G for the construction of the co-occurrence graph, G_c. The graph, G, is constructed based on the matrix (OD information × Usage information) generated by smart card data. However, the diagonal component of the matrix is 0. The sum of the row corresponds to an OD vertex, the sum of each column corresponds to a usage vertex, and each element in the matrix corresponds to edge information. The vertices and edges in the graph, G, are defined as follows.
- A vertex representing a particular combination of day, $l = 1, \ldots, 8$; time, $m = 1, \ldots, 20$; and passenger type, $n = 1, \ldots, 11$ is denoted by "usage vertex x_{lmn}," and the vertex set of all usage vertices is denoted by "usage vertex set $\boldsymbol{X}$." The number of elements in the usage vertex set $\boldsymbol{X}$ is 1,760 because it contains one user vertex for each of the total number of day × time × passenger type combinations. Each usage vertex, x_{lmn}, encodes the information about the number of passengers of type n who used the train network at time m on day l. For example, in Fig. 6, the usage vertex ($x_{l=1,m=1,n=1}$) representing the combination of Monday, 5:00–5:59 hrs, and Adult passenger type contains information on the total number

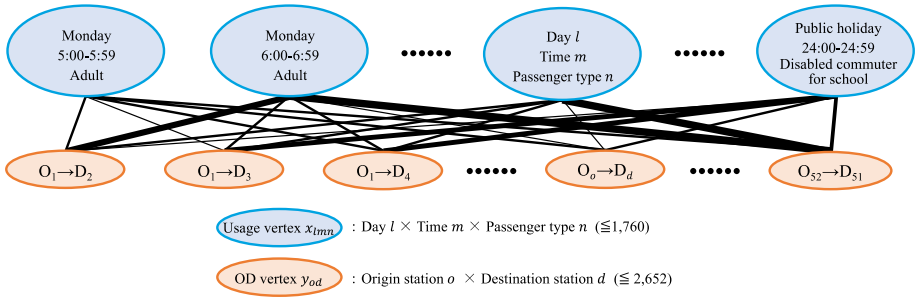

Fig. 6 Graph G representing connections between usage vertices and OD vertices

of adult users on Mondays during 5:00–5:59 hrs. In addition, a vertex representing a particular combination of origin station $o = 1, \ldots, 52$ and destination station $d = 1, \ldots, 52$ is denoted by "OD vertex y_{od}," and the vertex set of all OD vertices is denoted by "OD vertex set $\boldsymbol{Y}$." The number of elements in the OD vertex set $\boldsymbol{Y}$ is $52 \times 52 - 52 = 2{,}652$ because it contains one vertex for each of the total number of origin stations × total number of destination stations combinations, except duplicates. Each OD vertex, y_{od}, encodes information regarding the total number of users travelling from the origin station o to the destination station d. For example, in Fig. 6, the OD vertex ($y_{o=1,d=2}$) representing $O_1 \rightarrow D_2$ captures the information on the total number of users travelling from origin station O_1 to destination station D_2. Further, each edge connecting a usage vertex x_{lmn} with an OD vertex y_{od} encodes the information regarding the number of users of passenger type n who travelled from the origin station o to the destination station d at time m on day l. For example, in Fig. 6, the edge connecting the usage vertex ($x_{l=1,m=1,n=1}$) corresponding to Monday × 5:00–5:59 × Adults and the OD vertex ($y_{o=1,d=2}$) corresponding to $O_1 \rightarrow D_2$ encodes information regarding the number of adult users travelling from origin station O_1 to destination station D_2 on Monday during 5:00–5:59 hrs. The maximum number of edges in the graph, G, is 4,667,520 ($= 1760 \times 2652$).

- We construct the co-occurrence graph, G_c, by extracting combinations that share co-occurrence relationships with respect to all combinations of usage vertices and OD vertices in the graph G. In this case, co-occurrence is expressed by the ratio of common users among the users corresponding to each pair of usage and OD vertices. Further, a statistical test is performed to determine the significance of co-occurrence to rule out the possibility that its manifestation is coincidental instead of causal. In this paper, t-values are used as the criteria for co-occurrence in the natural language processing field, and the statistical significance of co-occurrence is adjudged by a t-test. The t-value used as the test statistic for the t-test is calculated using (1), where W denotes the total number of users (= 9,008,709).

$$t\text{-value} = \frac{\left(\left| x_{lmn} \cap y_{od} \right| - \frac{|x_{lmn}| \times |y_{od}|}{W} \right)}{\sqrt{\left| x_{lmn} \cap y_{od} \right|}} \tag{1}$$

This study considers co-occurrence to be significant if the absolute value of the t-value is equal to or greater than 1.65 (significance level 10%). If the combination of a usage vertex and an OD vertex is determined to exhibit a statistically significant co-occurrence relation, the number of users associated with the edge in the graph, G, is replaced with 1, and it is set to 0 otherwise. Conversion of the numerical information associated to the edges of G into binary variables enables the detection of travel patterns of boarding and alighting, as well as combinations with high-frequency users, even if the combinations contain relatively low frequency users. The co-occurrence graph, G_c, is constructed using only combinations that exhibit significant co-occurrence. In the graph, G_c, the usage vertex set

is denoted by $\boldsymbol{U} = \{u_i | i = 1, \ldots, I(I \leq 1{,}760)\}$ and the OD vertex set is denoted by $\boldsymbol{V} = \{v_j | j = 1, \ldots, J(J \leq 2{,}652)\}$.

2) Construction of the similarity graph.

The similarity graph, G_s, is constructed on the basis of the co-occurrence graph, G_c, and captures the similarity between different usage vertices. In this study, we focus on the determination of similar connections between usage vertices and OD vertices and construct a similarity graph with high-similarity usage vertices. Although Simpson and Dice coefficients can be used as similarity measures, this study uses the Jaccard coefficient following Uno et al. [24]. The similarity between any two usage vertices, u_i and u_i', is defined to be (2).

$$sim(u_i, u_i') = \frac{|N(u_i) \cap N(u_i')|}{|N(u_i) \cup N(u_i')|} s.t. u_i, u_i' \in U, v_j \in V \tag{2}$$

Using this, a similarity graph is constructed comprising usage vertices with similarities exceeding a predefined threshold, θ_s. The construction of the similarity graph from the co-occurrence graph is illustrated using an example presented in Fig. 7 (the example graph comprises a usage vertex set $\boldsymbol{U} = \{u_i | i = 1, \ldots, 6\}$, each vertex of which is highlighted by a blue circle and an OD vertex set $\boldsymbol{V} = \{v_j | j = 1, \ldots, 7\}$, each vertex of which is highlighted by an orange circle). For usage vertices, u_1 and u_2, $|N(u_1) \cap N(u_2)| = 1$ and $|N(u_1) \cup N(u_2)| = 4$ since $N(u_1) = \{v_1, v_3\}$ and $N(u_2) = \{v_3, v_4, v_5\}$. Therefore, by (2), the similarity between u_1 and u_2 is $1/4 = 0.25$. Following the same procedure, the similarities between all pairs of usage vertices are calculated, and a new graph is constructed using the usage vertices as vertices. Edges are only inserted between those pairs of usage vertices whose similarities exceed a predefined threshold. In the example depicted in Fig. 7, the threshold value, θ_s, is set to 0.4, and the similarity graph depicted in Fig. 8 is obtained from it. In this case, each usage vertex of the co-occurrence graph u_1, u_2, u_3, u_4, u_5, and u_6 in Fig. 7 is renamed by s_1, s_2, s_3, s_4, s_5 and, s_6, respectively. The construction of the similarity graph dictates that usage vertices that exhibit similar connections with OD vertices are connected by edges in the similarity graph, i.e., neighbors in the similarity graph correspond to similar user travel behaviors in terms of origin and destination stations. For example, usage vertices s_1, s_4, and s_5 are connected to each other in Fig. 8, and this implies that the users grouped in these vertices share similar travel behaviors. In the similarity graph G_s, the usage vertex set is denoted by $\boldsymbol{S} = \{s_k | s_k \in \boldsymbol{U}, k = 1, \ldots, K(K \leq 1{,}760)\}$.

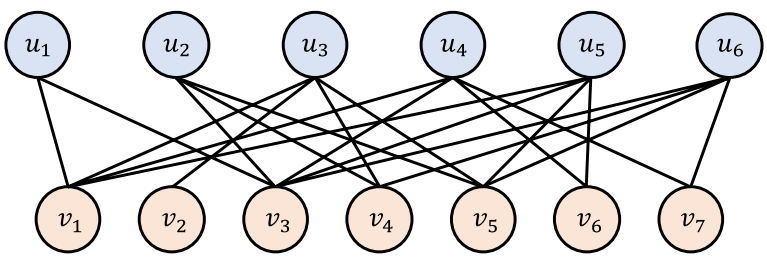

Fig. 7 An illustrative co-occurrence graph

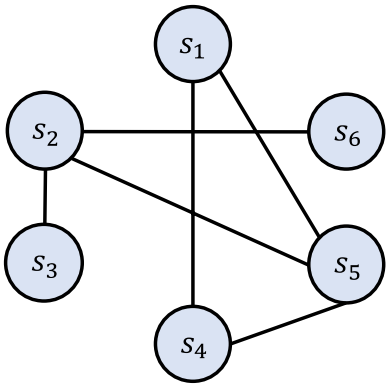

Fig. 8 The corresponding similarity graph

3) Application of data polishing to the similarity graph.

- Next, data polishing is applied to the similarity graph obtained in the previous step to group usage vertices $s_k \in \boldsymbol{U}$ in a fashion that ensures that only pairs of usage vertices with strong connections in the similarity graph remain connected by edges. The similarity measure of sets is used to adjudge whether usage vertex pairs share a strong connection. In this study, the Jaccard coefficient is used as the similarity measure as in the case of the construction of the similarity graph. The similarity between any two usage vertices, s_k and s'_k, is defined by (3).

$$sim(s_k, s'_k) = \frac{\left|N[s_k] \cap N\left[s'_k\right]\right|}{\left|N[s_k] \cup N\left[s'_k\right]\right|} s.t. s_k, s'_k \in \boldsymbol{S} \tag{3}$$

Equation (3) represents the similarity between the closed neighborhoods of s_k and s'_k. We illustrate the application of the data polishing procedure using the similarity graph presented in Fig. 8. First, the similarities between all pairs of usage vertices constituting the similarity graph are calculated using (3). For example, in the case of and $s_2, |N[s_1] \cap N[s_2]| = 1$ and $|N[s_1] \cup N[s_2]| = 7$ since the closed neighborhood of s_1 is $N[s_1] = \{s_1, s_4, s_5\}$ and the closed neighborhood of s_2 is $N[s_2] = \{s_2, s_3, s_5, s_6\}$. Therefore, the similarity between the usage vertices, s_1 and s_2, is $1/7 = 0.14$. Then, we select a threshold value, θ_p, and vertex pairs whose similarities are equal to or greater than θ_p and are connected by edges. By contrast, edges between pairs of vertices whose similarities are less than the threshold are deleted. If θ_p is set, data polishing is repeated using this newly constructed graph as the input graph, and the process is performed until the deformation of the graph converges. Normally, data polishing is required to be applied several times to achieve convergence. However, in this example, the shape of the graph does not change from that of Fig. 9 even if data polishing is repeated. Therefore, data polishing is terminated after only one application in this case. This example requires only one polishing because of the simplicity of the co-occurrence graph. The final graph achieved via data polishing is called the polishing graph, G_p. In G_p, the usage vertex set is denoted by $\boldsymbol{P} = \{p_t | p_t \in \boldsymbol{S}, t = 1, \ldots, T(T \leq 1{,}760)\}$.

4) Enumeration of cliques.

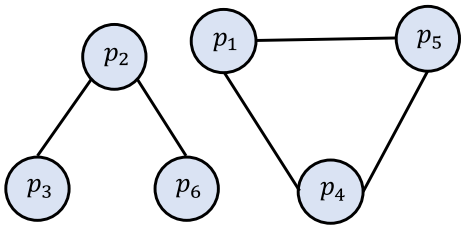

Fig. 9 The corresponding polishing graph

This step involves the enumeration of all cliques in the polishing graph. In the illustrative polishing graph depicted in Fig. 9, there are three cliques—$C_1 = \{p_2, p_3\}$, $C_2 = \{p_2, p_6\}$, and $C_3 = \{p_1, p_4, p_5\}$. Among them, there is one maximal clique—$C_3 = \{p_1, p_4, p_5\}$. In this study, we contend that various travel patterns of IruCa users can be understood by enumerating all cliques, including maximal ones. To this end, we extract all cliques, in contrast to the approach undertaken by Uno et al. [24], in which they extracted only the maximal cliques. Cliques extracted by the proposed method represent groups of usage vertices, $p_t \in \boldsymbol{S}$, corresponding to similar user travel behaviors, as described in subsection 2 of this section. This reveals groups of users who exhibit similar travel behaviors. If a usage vertex belongs to more than one clique (such as p_2 in Fig. 9), it suggests that users grouped within this vertex exhibit more than one travel pattern. By contrast, users in maximal cliques exhibit unique travel patterns.

5) Extraction of the combination of origin station and destination station associated to each clique.

- Next, we attempt to estimate the most frequent origin and destination stations corresponding to users in each extracted clique. As an example, we consider the case of an extracted clique consisting of two usage vertices$x_{l=1,m=1,n=1}$ (Monday × 5:00–5:59 hrs × Adult) and $x_{l=1,m=2,n=1}$ (Monday × 6:00–6:59 hrs × Adult). First, we extract the OD vertices exhibiting co-occurrence with both usage vertices on the basis of the co-occurrence graph, G_c. Then, we identify the most frequent OD vertices in the co-occurring combinations on the basis of the graph, G. Via this process, we identify the types of passengers who travel between different sets of origin and destination stations at different times of the day on different days of the week.

Results

The threshold

In the proposed method, two threshold parameters are used—the threshold, θ_s, during the construction of the similarity graph, G_s, and the threshold, θ_p, during the construction of the polishing graph, G_p. As both θ_s and θ_p are criteria for judging similarity, this study considers them to share the same value. Therefore, only one parameter needs to be defined to extract the cliques. Although the threshold value λ $(= \theta_s = \theta_p)$ influences the extraction of cliques, there is no clear criterion for selecting its optimal value. In

previous studies [24], the threshold was set arbitrarily. This study uses the average of the clustering coefficients to determine λ. This approach is unique to this study.

The clustering coefficient is a measure of the degree to which nodes in a graph tend to cluster together. It is higher in a graph in which vertices adjacent to each other are connected by edges. In our case, as all the vertices in a clique are connected by edges, it can be argued that the clustering coefficients of the vertices in the polishing graph increase on average when the clique is generated. The cluster coefficient of vertex, i,, is defined by (4), where e_i denotes the number of edges connecting the neighborhoods of vertex i and k_i denotes the number of vertices adjacent to i.

$$C_i = \frac{e_i}{k_i(k_i - 1)/2} \tag{4}$$

The average clustering coefficient is the average of the clustering coefficients of all vertices in a graph. It is calculated using (5), where N denotes the total number of vertices.

$$C = \frac{1}{N}\sum_{i=1}^{N} C_i \tag{5}$$

In this study, we focus on the relationship between the average clustering coefficient and similarity. To construct the cliques, we set the threshold λ to the maximum value of the average clustering coefficient. First, the similarities between all pairs of usage vertices in the co-occurrence graph are calculated, and the value of λ is set to 0.01. Next, any edge is deleted if the similarity between the associated pair of vertices is less than λ, and the average clustering coefficient is calculated. Subsequently, usage vertices that are not adjacent to any other usage vertices are removed, λ is increased by 0.01, edges corresponding to similarity less than λ are removed, and the average clustering coefficient is calculated again. These steps are repeated until λ reaches 1.00.

The average clustering coefficient corresponding to each threshold is depicted in Fig. 10. Although the decision was made to set the threshold to be the maximum of the average clustering coefficients, there are multiple maxima for the average clustering coefficients.

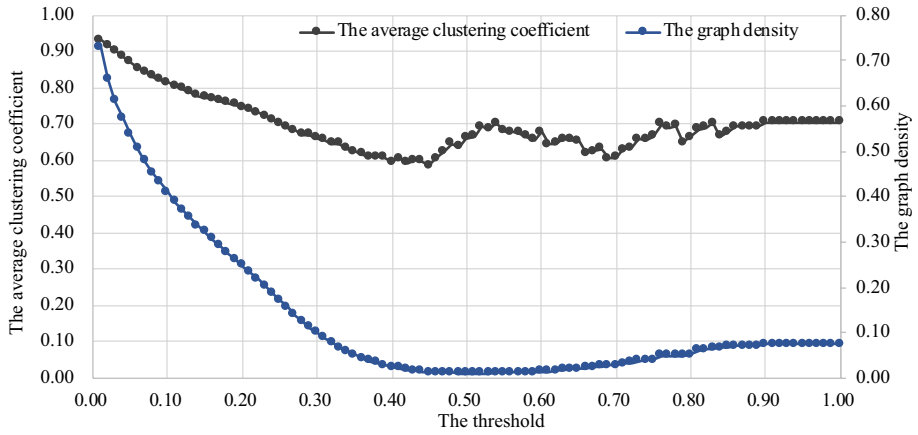

Fig. 10 The average clustering coefficient and the graph density

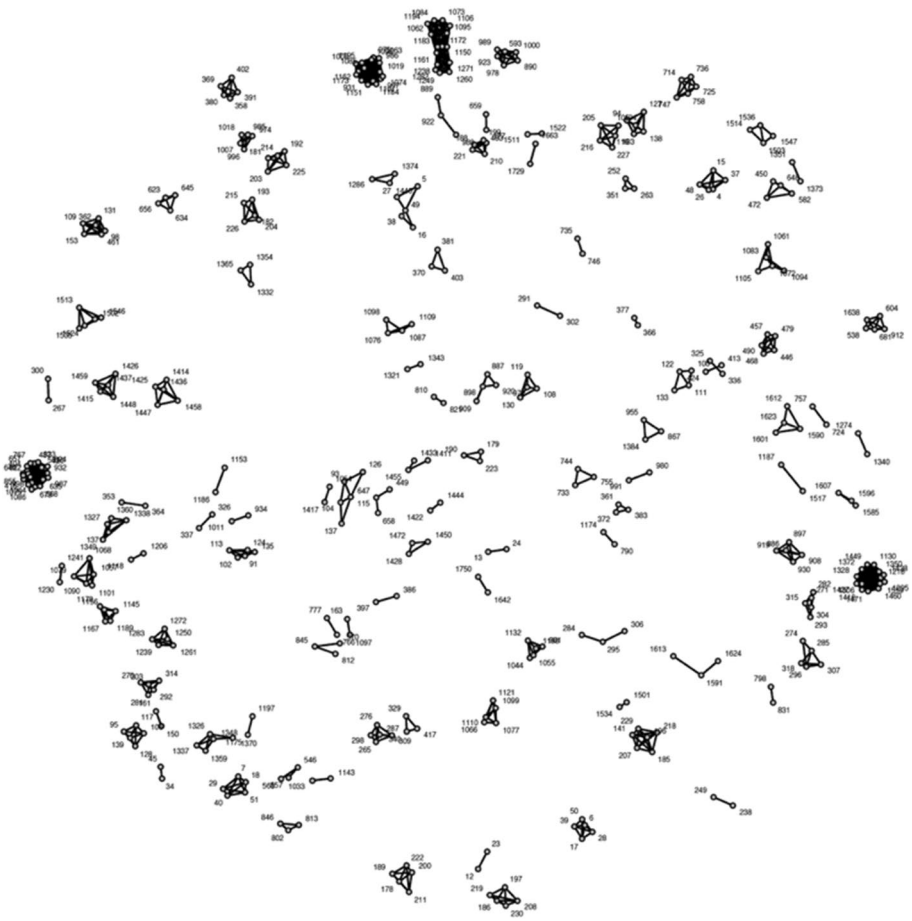

Fig. 11 The polishing graph obtained in this study

Therefore, it is not possible to unambiguously select the threshold solely based on average clustering coefficients. To address this problem, the graph density was also calculated. This is defined using (6), where E denotes the edge set, and V denotes the vertex set.

$$D = \frac{|E|}{|V|(|V|-1)/2} \tag{6}$$

The graph density is observed to increase as the number of edges approaches the maximal number of edges. In other words, a dense graph exhibits high graph density. However, the boundary of the vertex groups (cliques) cannot be determined when the graph is dense. Therefore, this study utilizes the threshold that corresponds to the lowest graph density. By using two indices, it is possible to extract cliques such that the boundaries between them are clear.

The graph density corresponding to each threshold is presented in Fig. 10. In this study, the threshold was set at the point where the average clustering coefficient was at a maximum and the graph density was at a minimum. As is evident from the results presented in Fig. 10, the threshold was set at 0.54.

Table 2 52 cliques consisting of two usage vertices

Combination	No. of cliques
Time and passenger type are the same	37
Day and passenger type are the same	2
Only passenger type is the same	10
Day, time, and passenger type are all distinct	3

Table 3 32 cliques consisting of three usage vertices

Combination	No. of cliques
Time and passenger type are the same	28
Day and passenger type are the same	0
Only passenger type is the same	4
Day, time, and passenger type are all distinct	0

Table 4 5 cliques consisting of four usage vertices

Combination	No. of cliques
Time and passenger type are the same	4
Day and passenger type are the same	0
Only passenger type is the same	1
Day, time, and passenger type are all distinct	0

Table 5 30 cliques consisting of five usage vertices

Combination	No. of cliques
Time and passenger type are the same	28
Day and passenger type are the same	0
Only passenger type is the same	2
Day, time, and passenger type are all distinct	0

Similarity of usage vertices

By applying the proposed method to the smart card data, the polishing graph depicted in Fig. 11 was constructed. To simplify the references to each usage vertex, the numbers assigned to them have been indicated in Fig. 11.

The total number of extracted cliques in the polishing graph is observed to be 127. Of these, 52 cliques consist of two usage vertices, 32 of three usage vertices, 5 of four usage vertices, 30 of five usage vertices, 3 of six usage vertices, and 1 each of nine, ten, fourteen, sixteen, and twenty-one usage vertices.

It is not possible to explicitly depict the compositions of all extracted cliques for the want of space. Instead, we highlight the differences between different combinations of day, time, and passenger type. The total number of cliques in terms of different combinations are depicted in Tables 2, 3, 4, 5, 6, 7, 8, 9, 10 and 11, each of which presents

Table 6 3 cliques consisting of six usage vertices

Combination	No. of cliques
Time and passenger type are the same	0
Day and passenger type are the same	0
Only passenger type is the same	3
Day, time, and passenger type are all distinct	0

Table 7 1 clique consisting of nine usage vertices

Combination	No. of cliques
Time and passenger type are the same	0
Day and passenger type are the same	0
Only passenger type is the same	1
Day, time, and passenger type are all distinct	0

Table 8 1 clique consisting of ten usage vertices

Combination	No. of cliques
Time and passenger type are the same	0
Day and passenger type are the same	0
Only passenger type is the same	1
Day, time, and passenger type are all distinct	0

Table 9 1 clique consisting of fourteen usage vertices

Combination	No. of cliques
Time and passenger type are the same	0
Day and passenger type are the same	0
Only passenger type is the same	1
Day, time, and passenger type are all distinct	0

Table 10 1 clique consisting of sixteen usage vertices

Combination	No. of cliques
Time and passenger type are the same	0
Day and passenger type are the same	0
Only passenger type is the same	1
Day, time, and passenger type are all distinct	0

Table 11 1 clique consisting of twenty-one usage vertices

Combination	No. of cliques
Time and passenger type are the same	0
Day and passenger type are the same	0
Only passenger type is the same	1
Day, time, and passenger type are all distinct	0

the results for the cliques with a given number of usage vertices (as listed in the previous paragraph). For example, Table 3 presents the results for cliques consisting of three usage vertices. Among the 32 extracted cliques, the number of combinations in which time and passenger type are the same is 28; in the other 4 cliques, only the passenger type is identical.

From the data presented in Tables 2, 3, 4, 5, 6, 7, 8, 9, 10 and 11, it is apparent that many of the extracted cliques contain combinations in which time and passenger type are identical. Indeed, cliques with this combination constitute 76% of the total number of extracted cliques. These cliques represent behaviors of users of the same type at the same time but on different days of the week.

The number of cliques in which only the passenger type is identical comprises the second largest proportion of the total (approximately 19%). For the cliques with six, nine, ten, fourteen, sixteen, and twenty-one usage vertices, this is the only combination observed. This can be interpreted to reflect the behaviors of the same type of users at different times of the day on different days of the week.

Combinations in which day and passenger type are identical or in which day, passenger type, and time are all distinct exist only in cliques with two usage vertices. For example, in the clique containing "Thursday × 10:00–10:59 hrs × Disabled commuter for work" and "Thursday × 12:00–12:59 hrs × Disabled commuter for work," the day of travel and passenger type are identical. This suggests that card users identified as "Disabled commuter for work" exhibit similar travel behaviors between "10:00–10:59 hrs" and "12:00–12:59 hrs" on every "Thursday." However, in the clique containing "Saturday × 23:00–23:59 hrs × Child" and "Public holiday × 24:00–24:59 hrs × Adult," day, time, and passenger type are all distinct. Thus, it can be concluded that "Child" passengers and "Adult" passengers exhibit similar travel behaviors at midnight on holidays.

Further, the extracted cliques are compared in terms of passenger type. Cliques with two, three, four, and five usage vertices are observed to be related to various passenger types, as depicted in Fig. 12. Thus, it can be concluded that several travel patterns of different passenger types can be surmised based on the results of cliques comprising small numbers of usage vertices. On the other hand, travel patterns of three specific passenger types can be effectively assumed from cliques comprising large numbers of usage vertices, as all cliques with six or more usage vertices are observed to be related to adult or child commuters.

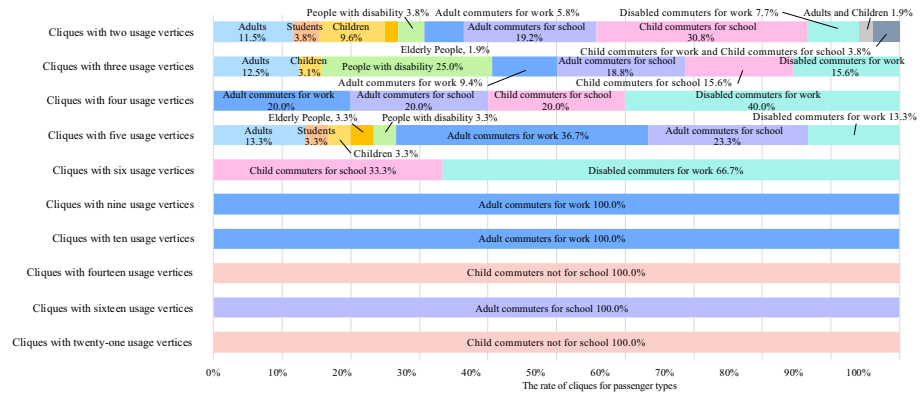

Fig. 12 Cliques compared in terms of passenger types

Table 12 Combinations of origin and destination stations

Cliques	Origin station	Destination station
(C1) "Sunday × 6:00–6:59 × Child commuter for school""Thursday × 17:00–17:59 × Child commuter not for school"	Ota	Shioya
(C2) "Public holiday × 13:00–13:59 × Child commuter for school" "Wednesday × 18:00–18:59 × Child commuter not for school"	Sanjo	Fusazaki
(C3) "Saturday × 23:00–23:59 × Child" "Public holiday × 24:00–24:59 × Adult"	Kawaramachi	Kotoden-Kotohira

Travel patterns of IruCa users

In this section, we present the characteristic combinations of origin and destination stations corresponding to each clique based on the results of the extracted cliques. This corresponds to step (5) of the aforementioned procedure. Due to the large number of cliques, it is impossible to explicitly present the combinations of origin and destination stations corresponding to all cliques. Hence, we focus on cliques related to children ("Child," "Child commuter not for school," and "Child commuter for school"), the elderly, and people with disabilities ("Person with disability," "Disabled commuter for work," and "Disabled commuter for school"), as examples. The number of extracted cliques associated to child commuters is 35, that associated to the elderly is 2, and that associated to people with disabilities is 28. Although most of these cliques consist of identical passenger types, three cliques comprise different passenger types. Henceforth, we focus on these three cliques and clarify the origin and destination station combinations corresponding to the largest number of users in each. The details of the three cliques are as follows:

(C1) "Sunday × 6:00–6:59 hrs × Child commuter for school" and "Thursday × 17:00–17:59 hrs × Child commuter not for school."

(C2) "Public holiday × 13:00–13:59 hrs × Child commuter for school" and "Wednesday × 18:00–18:59 hrs × Child commuter not for school."

(C3) "Saturday × 23:00–23:59 hrs × Child" and "Public holiday × 24:00–24:59 hrs × Adult."

The origin and destination station combinations corresponding to the largest number of users in each of the aforementioned cliques are depicted in Table 12. From the table, it is evident that the travel behaviors of "Child commuter for school" at "6:00–6:59 hrs" on "Sunday" and the travel behavior of "Child commuter not for school" at "17:00–17:59 hrs" on "Thursday" are similar, and they travel from "Ota" to "Shioya." Moreover, it can be concluded that they travel from "Ota" to "Kawaramachi" and then to "Shioya," based on the Kotoden route map in section II. Thus, it is likely that the users in C2 also transfer at Kawaramachi. In C3, we conclude that passengers of type "Child" at "23:00–23:59 hrs" on "Sunday" and those of type "Adult" at "24:00–24:59 hrs" on "Public holiday" travel from "Kawaramachi" to "Kotoden-Kotohira." By extracting the OD vertex corresponding to the largest number of users in each clique, the types of passengers who travel between different stations at different times on different days can be ascertained. In other words, the characteristic travel patterns can be discovered using the proposed method. However, it should be noted that the origin and destination stations thus identified need not correspond to the actual origin and destination of any individual user. Although this study attempts to estimate the actual origins and destinations, this is not possible from the information encoded within the cliques. For example, in C1, the actual origin is estimated to be residential owing to the presence of a residential area around "Ota." By contrast, there are cultural facilities and beaches around "Shioya." However, it is not clear whether these places are related to trips of the actual users. We intend to clarify these issues in future research.

Discussion

Based on the obtained results, several of the 127 extracted cliques were observed to be composed of identical passenger types. However, it was suggested that the similarity between different passenger types could be quantitatively clarified based on cliques containing different combinations of day, time, and passenger type.

The prediction of such usage vertex combinations in advance may be difficult. For example, in this study, a clique consisting of the "Saturday × 23:00–23:59 hrs × Child" usage vertex and the "Public holiday × 24:00–24:59 hrs × Adult'" usage vertex was extracted. This combination cannot be identified via basic analysis, such as aggregate analysis. Even though one existing additional method can be used to identify similar combinations of day and time of travel corresponding to each passenger type, it is not practicable because the number of combinations become significantly high for efficient processing with the increase in the number of considered categories. Therefore, we conclude that a combination comprising "Saturday × 23:00–23:59 hrs × Child" and "Public holiday × 24:00–24:59 hrs × Adult" cannot be extracted by any method other than the proposed one, which is capable of considering multiple attributes simultaneously.

This study also provided an understanding of the travel patterns of passengers of different types at different times of the day and on different days of the week. Characteristic travel behaviors of smart card users and origin station and destination station combinations could be successfully extracted for each clique.

The results demonstrate that the proposed method is capable of effectively extracting travel patterns of IruCa users based on graphs comprising day × time × passenger

type × origin station × destination station tensors. The travel patterns based on the extracted cliques provide conclusions about the behaviors of adult commuters and students traveling in the morning and returning home in the afternoon. These patterns are intuitive, which corroborate the success of pattern extraction performed. Moreover, the performance of the proposed method is not dependent on the number of samples, as approximately half of all cliques were observed to be related to small groups of smart card users, such as children, the elderly, and people with disabilities.

With respect to the promotion of public transportation in society, it is important to enhance the utilization frequency of not only users who avail it regularly but also those who avail it sporadically. However, it is difficult to identify their travel patterns via data analysis because of the low number of associated samples. This study demonstrates that data polishing is effective even in case of users who avail public transport sporadically, and the distribution of the number of samples is biased.

The operational plans of public transportation are usually formulated to effectively serve the needs of commuters and students who constitute the majority of the users in most cases. In such a scenario, elderly commuters or children with small sample sizes may be inadvertently inconvenienced. Thus, ensuring the availability of proper public transportation for all people without automobiles remains a critical problem in regional traffic. The conclusions of this study are expected to be beneficial in the formulation of public transport policies that can effectively serve the needs of low-frequency users such as elderly people and children, besides the commuters and students who constitute the majority of commuters.

Conclusions

This study proposed a method for extracting travel patterns from smart card data using data polishing. In particular, we presented a method comprising five steps—(1) construction of the co-occurrence graph, (2) construction of the similarity graph, (3) application of data polishing to the similarity graph, (4) enumeration of cliques, and (5) extraction of combinations of origin and destination stations associated to each clique. We used this method to estimate the applicability of data polishing to pattern recognition based on smart card data.

Data collected from the IruCa smart card, used on the Kotoden rail system in the Kagawa Prefecture in Japan, were used in this study. To analyze the data, we constructed a graph representing the relationships between various categories of the five attributes—day (8 categories), time (20 categories), passenger type (11 categories), origin station (52 categories), and destination station (52 categories)—and applied the proposed method to this graph.

Usage vertices with highly similar interrelationships were grouped together by applying data polishing to the similarity graph. The groups were extracted as cliques, revealing the similarities between behaviors of card users in the extracted cliques and clarifying user groups with similar behaviors. Then, the origin and destination station combination corresponding to the largest number of users in each usage vertex was identified. Via these processes, this study provided a comprehensive account of the travel patterns of passengers of different types at different times of the day and on different days of the week.

In future research, we intend to develop an efficient algorithm that eliminates complex calculation and requires fewer processing steps. Further study is needed to clarify the occurrence factors and similarity factors in the extracted travel patterns.

Acknowledgements
The authors especially thank Takamastu-Kotohira Electric Railroad Co. Ltd. in Takamatsu City, Kagawa Prefecture, Japan for providing the smart card data used in this paper.

Authors' contributions
MK: Conceptualization, Methodology, Writing-Review & Editing and Supervision. TM: Validation, Methodology and Supervision. MH: Methodology, Software, Formal analysis, Data curation, Writing-Original Draft, and Project Administration. All authors read and approved the final manuscript.

Funding
This study was supported by JSPS Grants-in-Aid for Scientific(KAKENHI) Grant Numbers 20H02277 and Grant-in-Aid for JSPS Research Fellow Grant Numbers 20J15417.

Availability of data and materials
The data that support the conclusions of this study are available from the Takamastu-Kotohira Electric Railroad Co. Ltd., but restrictions apply to the availability of these data, which were used under license for the current study. Thus, they are not publicly available. They can still be accessed from the authors upon reasonable request and with the permission of the Takamastu-Kotohira Electric Railroad Co. Ltd.

Competing interests
The authors declare that they have no competing interests.

Received: 22 September 2020 Accepted: 14 December 2020
Published online: 07 January 2021

References

1. Hofleitner A, Herring R, Bayen A. Arterial travel time forecast with streaming data: a hybrid approach of flow modeling and machine learning. Transp Res Part B Methodol. 2012;46(9):1097–122.
2. Krause C, Zhang L. Short-term travel behavior prediction with GPS, land use, and point of interest data. Transp Res Part B Methodol. 2019;123:349–61.
3. Jenelius E, Koutsopoulos HN. Travel time estimation for urban road networks using low frequency probe vehicle data. Transp Res Part B Methodol. 2013;53:64–81.
4. Espinoza C, Munizaga M, Bustos B, Trépanier M. Assessing the public transport travel behavior consistency from smart card data. Transp Res Procedia. 2018;32:44–53.
5. Morency C, Trépanier M, Agard B. Measuring transit use variability with smart-card data. Transp Policy. 2007;14(3):193–203.
6. Ma X, Liu C, Wen H, Wang Y, Wu Y. Understanding commuting patterns using transit smart card data. J Transp Geogr. 2017;58:135–45.
7. Agard B, Morency C, Trepanier M. Mining public transport user behavior from smart card data. IFAC Proc. 2006;39:399–404.
8. Ordóñez Medina SAO. Inferring weekly primary activity patterns using public transport smart card data and a household travel survey. Travel Behav Soc. 2018;12:93–101.
9. Nazem M, Lemone A, Chu A, Spurr T. Analysis of travel pattern changes due to a medium-term disruption on public transit networks using smart card data. Transp Res Procedia. 2018;32:585–96.
10. Li YT, Iwamoto T, Schmocker J, Nakamura D, Uno T. N. Analyzing long-term travel behavior: a comparison of smart card data and graphical usage patterns. Transp Res Procedia. 2018;32:34–43.
11. Zhang Y, Martens K, Long Y. Revealing group travel behavior patterns with public transit smart card data. Travel Behav Soc. 2018;10:42–52.
12. Sun L, Axhausen KW. Understanding urban mobility patterns with a probabilistic tensor factorization framework. Transport Res Part B. 2016;91:511–24.
13. Han Y, Moutarde F. Analysis of large-scale traffic dynamics in an urban transportation network using non-negative tensor factorization. Int J ITS Res. 2016;14(1):36–49.
14. Vazifehdan M, Moattar MH, Jalali M. A hybrid bayesian network and tensor factorization approach for missing value imputation to improve breast cancer recurrence prediction. J King Saud Univ Comput Inf Sci. 2019;31(2):175–84.
15. Yao D, Yu C, Jin H, Ding Q. Human mobility synthesis using matrix and tensor factorizations. Inf Fusion. 2015;23:25–32.
16. Şimşekli U, Virtanen T, Cemgil AT. Non-negative tensor factorization models for Bayesian audio processing. Digit Signal Process. 2015;47:178–91.
17. Taneja A, Arora A. Cross domain recommendation using multidimensional tensor factorization. Expert Syst Appl. 2018;92:304–16.
18. Wang L, Bai J, Wu J, Jeon G. Hyperspectral image compression based on lapped transform and Tucker decomposition. Signal Process Image Commun. 2015;36:63–9.
19. Correa FE, Oliveira MDB, Gama J, Corrêa PLP. LP, Rady J. Analyzing the behavior dynamics of grain price indexes using Tucker tensor decomposition and spatio-temporal trajectories. Comput Electron Agric. 2016;120:72–8.

20. Favier G, Fernandes CAR, de Almeida ALF. Nested Tucker tensor decomposition with application to MIMO relay systems using tensor space–time coding (TSTC). Signal Process. 2016;128:318–31.
21. Briand AS, Come E, Trepanier M, Oukhellou L. Analyzing year-to-year changes in public transport passenger behaviour using smart card data. Transport Res Part C. 2017;79:274–89.
22. Ma X, Wu YJ, Wang Y, Chen F, Liu J. Mining smart card data for transit riders' travel patterns. Transport Res Part C. 2013;36:1–12.
23. Faroqi H, Mesbah M, Kim J, Tavassoli A. A model for measuring activity similarity between public transit passengers using smart card data. Travel Behav Soc. 2018;13:11–25.
24. Uno T, Maegawa H, Nakahara T, Hamuro Y, Yoshinaka R, Tatsuta M. Micro-clustering: finding small clusters in large diversity. 2016. arXiv preprint arXiv:1507.03067v2.

Publisher's Note