LETTER *Special Section on Advanced Log Processing and Office Information Systems*

Urban Zone Discovery from Smart Card-Based Transit Logs*

Jae-Yoon JUNG[†], ***Member***, **Gyunyoung HEO**[†], ***and*** **Kyuhyup OH**[†a)], ***Nonmembers***

SUMMARY Smart card payment systems provide a convenient billing mechanism for public transportation providers and passengers. In this paper, a smart card-based transit log is used to reveal functionally related regions in a city, which are called zones. To discover significant zones based on the transit log data, two algorithms, minimum spanning trees and agglomerative hierarchical clustering, are extended by considering the additional factors of geographical distance and adjacency. The hierarchical spatial geocoding system, called Geohash, is adopted to merge nearby bus stops to a region before zone discovery. We identify different urban zones that contain functionally interrelated regions based on passenger trip data stored in the smart card-based transit log by manipulating the level of abstraction and the adjustment parameters.

key words: *smart card-based transit log, zone discovery, public transportation, Geohash*

1. Introduction

Public transportation systems in modern cities often use smart card payment systems because they are convenient and provide reasonable billing mechanisms to transit providers based on a passenger's travel distance [3]. Many smart card-based transit logs are accumulated during this process, and these data can be used to analyze the movement patterns of citizens and identify functional regions in a city [4], [5], [10].

In this paper, smart card-based transit logs were used to identify functionally interrelated regions in a city, called zones. Zone discovery is important to urban system designers, public service providers, and commercial marketers, who need to understand people's behaviors and the functional regions in a city [10]. The discovered zones can be interpreted by characterizing facilities such as commercial streets, schools, stadiums, and parks. The urban zones can also be combined with civil planning such as transportation network redesigns and new urban development [1], [9].

To discover the interrelated regions in a city from the smart card-based transit log, the maximal spanning tree (MST) algorithm and agglomerative hierarchical clustering (AHC) algorithm were modified by adding geographical characteristics such as distance and adjacency between regions. In addition, a grid-based hierarchical geocoding system, called Geohash, was adopted to preprocess the geographical locations of bus stops [6]. The geocoding system helped to group close bus stops in the same region based on the required scale.

In this research, urban zones were discovered from the smart card-based transit logs of bus networks in Seoul based on different levels of abstraction. In particular, Geohash codes with length of 5 or 6 were used to find the zones in the city. Such zones can be used to understand citizen behavior and to make city plans based on real bus trips in the city.

2. Smart Card-Based Transit Log

Smart card-based transit log data contains useful information such as ride and alight time, transportation type, transfer, bus and station ids, and fare amount, as shown in Table 1. The sample log data include four passengers' travels, among whom the first and the forth, '1464' and '6474,' transferred twice and once, respectively, which is apparent in the FREQ field. The transferred records are connected before analysis to investigate the relationship between the origin and final destination in this research. In the table, the bus route id, the vehicle id, and the station id are included, and the passenger's class and the fare are also recorded; for example, adults of class '01' paid 1050 KRW, and the student of class '02' paid 720 KRW.

The purpose of this research was to extract the origins and destinations of passengers from the smart card-based transit log (considering transfers) and then identify the interrelated zones in a city. Bus networks are more complicated than train or subway networks because there are many kinds of buses and a variety of bus routes that are managed

Table 1 A sample of smart card-based transit data from bus services.

CARD_ID	RIDE_DTIME	TRANS_ID	FREQ	BUS_ROUTE_ID	VEHC_ID	RIDE_STA_ID	PASGR_CLS	RIDE_AMT
1464	20120314191514	005	0	11110241	111749741	0060028	01	1050
1464	20120314192851	005	1	11110251	111745505	0071985	01	1050
1464	20120314194912	005	2	11110803	111708544	9008817	01	0
1585	20120314134405	009	0	11110241	111743704	0060028	02	720
4440	20120314081537	011	0	41110071	111742454	0074426	01	1050
6474	20120314161353	011	0	11110017	111752103	0011174	01	1050
6474	20120314165349	011	1	11110342	111744273	0071456	01	0

Manuscript received November 11, 2016.
Manuscript revised March 31, 2017.
Manuscript publicized July 21, 2017.

[†] The authors are with Kyung Hee University, Yongin, Gyeonggi-do, 446–701 Republic of Korea.

[*] This research was financially supported by a grant from Kyung Hee University in 2015 (KHU-20150653).

a) E-mail: k8383@khu.ac.kr (Corresponding author)

DOI: 10.1587/transinf.2016OFL0004

Table 2 OD data summarized from a smart card transition log.

Day	Hour	Origin	Destination	Num. of trips
12	7	wydqh	wydmg	48
12	7	wydqh	wydmu	2
12	7	wydqh	wydq5	55
12	8	wydhz	wydjn	6
12	8	wydhz	wydjp	364

by different companies. Moreover, several bus stops are often located near one another to support different groups of bus routes depending on the bus type, direction, destination group, and other factors. To reduce this complexity of the bus network, we used the Geohash geocoding system, which uniquely provides a hierarchical encoding scheme to a specific location in the world based on the degree of abstraction. A Geohash code can be represented in one to 12 lengths of 32 bases, which include 10 digits 0–9 and 22 alphabetic characters [6]. A Geohash region with length k is decomposed into 32 Geohash regions with length $k-1$. A 5-length Geohash grid ("wydm6") is divided into 32 6-length Geohash grids. Such a hierarchical geocoding system can be useful for solving variable scales of geographical problems in the same manner. For instance, the zone discovery algorithms proposed in this paper can be utilized to discover zones in a city through 5-length Geohash encoding, and they can also be performed in country-wide areas though 4-length encoding.

In this paper, 5-length and 6-length Geohash encoding schema were adopted to assign all bus stops in Seoul to the grids on the map. For instance, a location with a longitude of 37.504593 and a latitude of 126.987901 is assigned a 12-length Geohash code 'wydm3etb0dcf.' Therefore, the location will be mapped to 'wydm3e' and 'wydm3' as 6-length and 5-length Geohash codes, respectively.

Now, the smart card-based transit log data can be summarized as origin-destination (OD) data between Geohash regions, as shown in Table 2. The number of trips from an origin to a destination Geohash region was aggregated from the transit log based on the date and hour. In this paper, the OD data are mainly used to discover zones based on passenger trips in a public transportation network.

3. OD Data and Geographical Data

In previous studies, zones were identified to meet different goals such as travel pattern recognition [4] and tariff system planning [8]. In our research, a zone is considered a group of adjacent regions that are functionally related based on passenger trips. The functional relationship between two regions can be interpreted as movements among residence, workplace, school, shopping district, and so on [10].

We now clearly define some terms for zone discovery. A *zone* can be defined as a set of regions that are directly adjacent or connected through other adjacent in-between regions in the same zone. Two zones U and V can be said to be *adjacent* if there exist grids $u \in U$ and $v \in V$ such that u and v are adjacent. The adjacency relationship is used to merge two zones in the course of zone discovery. The adjacency distinguishes the zone discovery problem from typical clustering problems because a zone must not contain disconnected regions.

Three matrices were prepared for the zone discovery algorithms: an OD matrix, an adjacency matrix, and a distance matrix. For N given regions and OD data, an *OD matrix* T is an $N \times N$ matrix in which an element t_{uv} is the number of travels from origin u to destination v during a specific time period. The diagonals of the OD matrix are assumed to be zero because internal travels within zones are not considered in this research. An *adjacency matrix* J is an $N \times N$ matrix in which an element j_{uv} is 1 if two regions u and v are adjacent and zero otherwise. A *distance matrix* D is an $N \times N$ matrix in which an element d_{uv} is the distance between the centers of two regions u and v.

The definitions of the OD matrix and adjacency matrix for regions can be easily extended for zones in the same way. However, the distance matrix for zones needs to be created in a different manner based on the zone discovery algorithms, which will be explained in the next section.

4. Zone Discovery Algorithms

In this research, two zone discovery algorithms were proposed for the purpose of zone discovery from smart card-based transit log data. The algorithms are extensions of the maximal spanning tree (MST) and the agglomerative hierarchical clustering (AHC) algorithms. However, zone discovery is different from typical clustering problems because it must consider geographical information like adjacency and distance to maintain geographically connected regions in a zone [2]. The adjacency matrix and distance matrix defined in Sect. 3 were used to accomplish this. Moreover, OD-based zone discovery algorithms in this paper also reflect the number of travels (which are aggregated in the OD matrix) for the purpose of discovering functionally related zones based on a person's number of trips.

There exist three types of hierarchical clustering algorithms depending on the distance between two clusters: single-linkage, complete-linkage, and average-linkage clustering. A single-linkage clustering algorithm involves the application of Kruskal's MST algorithm. Therefore, the first algorithm presented in this paper is an extension of the single-linkage clustering algorithm, while the second algorithm is an extension of the average-linkage clustering algorithm. Those algorithms are dependent on proximity measurements between two observations (or weights between two nodes in a graph). In this research, connectivity measurements between two zones will be defined and used for the two proposed zone discovery algorithms instead of the proximity measurements.

The structures of the MST-based and AHC-based zone discovery algorithms are exactly the same, the only difference lies in how connectivity is measured. The inputs of the algorithms are the OD matrix, adjacency matrix, and distance matrix, which were defined in Sect. 3. Initially, every region is considered an individual zone. The distance between two adjacent zones is calculated based on their connectivity, and the two nearest zones are iteratively merged into a new zone. This process will terminate when the highest connectivity between two zones is lower than a threshold θ.

```
Algorithm: Zone Discovery
Input: OD matrix $T$, Adjacency matrix $J$, Distance matrix $D$
Output: Zone clusters
Set every region as a single zone.
repeat
    Find two zones with the highest connectivity.
        $(U^*, V^*) = \underset{(U,V)}{argmax}\, connectivity(U, V)$
    Merge $U^*$ and $V^*$ into a new zone $K$.
        $K = U^* \cup V^*$
    Update three matrices, $T$, $J$, and $D$.
until $max\ conn(U, V) < \theta$
```

We first describe the connectivity measurement for the MST-based algorithms, named ZD-MST. Kruskal's algorithm was adopted to find a specific number of connected trees that will become the discovered zones. The MST connectivity between two zones U and V is calculated:

$$connectivity^{MST}(U, V) = \begin{cases} \max\limits_{u \in U, v \in V} \left(\dfrac{t_{uv} + t_{vu}}{d_{uv}{}^{q}} \right), & \textit{if U and V are adjacent} \\ 0 & , \textit{otherwise} \end{cases}$$

To discover zones with unseparated regions, the measurement between zones U and V is zero if they is not adjacent. If two zones were adjacent, the measurement is a positive value that reflects the maximum number of trips between the two regions in the zones, $u \in U$ and $v \in V$, which is proportional to the distance between the two regions. The proportional relationship helps two zones that are near each other merge earlier than others with longer distances, although they have the same numbers of trips. Parameter q can be used to adjust the effect of the distance.

Next, the connectivity measure for the AHC-based algorithm, named ZD-AHC, considers the average number of trips between regions in two zones and the average distance between two groups of regions. The two averages were calculated by dividing only by the numbers of grids in the two zones, because every grid with the same length has the same area in the Geohash coding system. The effect of the average distance between two zones can also be adjusted using parameter q.

$$connectivity^{AHC}(U, V) = \begin{cases} \dfrac{avg\text{-}traffic(U, V)}{avg\text{-}dist(U, V)^{q}}, & \textit{if U and V are adjacent} \\ 0 & , \textit{otherwise} \end{cases}$$

$$\text{where} \quad avg\text{-}traffic(U, V) = \frac{\sum_{u \in U, v \in V} (t_{uv} + t_{vu})}{|U||V|}$$

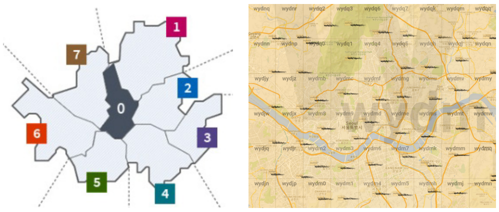
$$\text{and} \quad avg\text{-}dist(U, V) = \frac{\sum_{u \in U, v \in V} d_{uv}}{|U||V|}$$

The main difference between the two algorithms is that the MST-based algorithm focuses on the traffic on a single arc, i.e., u and v, while the AHC-based algorithm investigates overall arcs between two zones, i.e., U and V.

5. Experimental Results

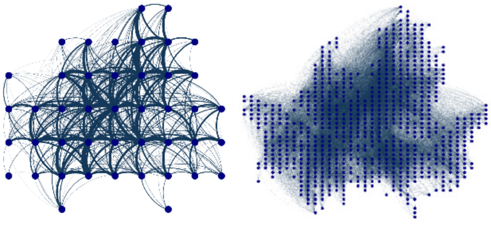
Smart card-based transit data were collected from the bus network in Seoul from Monday, March 12, 2012 to Friday, March 16. The transit data include 5,269,112 trip records which cover 6,078 bus stops in Seoul. The data was converted to OD data as shown in Table 2 to aggregate the number of movements between two regions. Seoul city can be divided into 8 groups of administrative districts [7]. The regions in Seoul can be mapped to 44 Geohash codes with length 5 and 732 Geohash codes with length 6, as shown in Fig. 1. The sizes of 5-length and 6-length Geohash regions in Seoul are in average 4,886 m by 3,876 m (longitude by latitude) and 611 m by 969 m, respectively. All bus stops in Seoul were mapped to Geohash codes based on their GPS location information using the two Geohash encoding methods.

Figures 2 (a) and 2 (b) depict the traffic between every two regions based on 5-length and 6-length Geohash regions, respectively. In the original traffic data, it is difficult to find meaningful zones. OD data and geographical information of regions were used to create OD matrices, adjacent matrices, and distance matrices for both 5-length and 6-length Geohash regions, and then the three matrices were utilized as inputs to the two proposed zone discovery algo-



(a) Seoul districts [7] (b) Geohash

Fig. 1 Seoul city districts and mapping to Geohash codes.



(a) 5-length Geohash (b) 6-length Geohash

Fig. 2 Original trips in public transportation networks for a week.

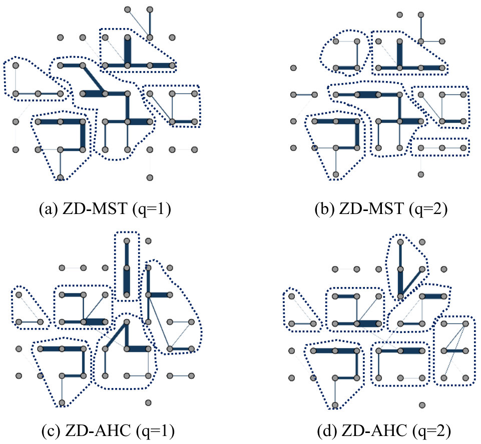

(a) ZD-MST (q=1)
(b) ZD-MST (q=2)
(c) ZD-AHC (q=1)
(d) ZD-AHC (q=2)

Fig. 3 Zone discovery results from 5-length Geohash regions.

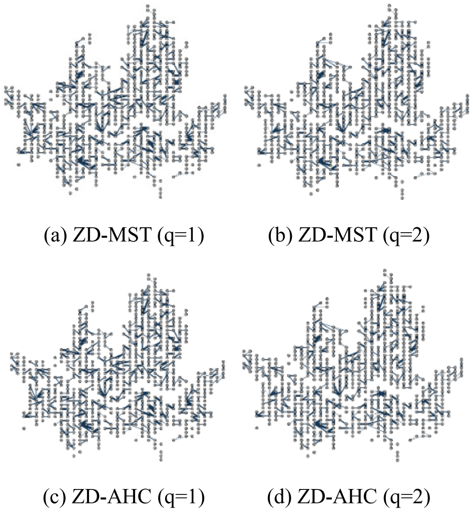

(a) ZD-MST (q=1)
(b) ZD-MST (q=2)
(c) ZD-AHC (q=1)
(d) ZD-AHC (q=2)

Fig. 4 Zone discovery results from 6-length Geohash regions.

rithms.

The results of zone discovery based on 5-length and 6-length Geohash codes are illustrated in Figs. 3 and 4, respectively. The figures show the results of both MST- and AHC-based zone discovery algorithms with various q values. Here, q is the parameter used to adjust the effect of the inverse distance in the connectivity measurement. As illustrated in Fig. 3, as q value increases, the diameter of zones decreases in the results of both algorithms. It can be said that the q value helps clusters to retain their circular shapes in iterative steps.

In the experiments for the 6-length Geohash, it is difficult to investigate the effectiveness of zone discovery in Fig. 4 at a glance, so the results are summarized in Table 3. The results of the 6-length Geohash regions also showed that the mean diameters of the clusters decreased as q value increased.

Table 3 Summary of zone discovery results from 6-length Geohash regions.

	MST-based algorithm				AHC-based algorithm			
	q=0.5	q=1	q=2	q=3	q=0.5	q=1	q=2	q=3
Num. of clusters	114	114	113	115	131	130	131	132
Mean num. of regions in a cluster	6.421	6.421	6.478	6.365	5.588	5.630	5.588	5.545
Mean diameter of clusters (km)	0.919	0.918	0.904	0.901	0.914	0.896	0.873	0.869

The results of zone discovery in Seoul can be compared to Seoul districts in Fig. 1 (a). The result zones of ZD-AHC (q = 2) in Fig. 3 (d) was nearly same as the 8 districts in Seoul, except that District 0, the old downtown, was merged to District 7. The characteristic of the ZD-MST results is that the zones in the center crossed the Han River in Seoul, differently from the result of ZD-AHC. It is because ZD-MST considers only the highest connectivity between two adjacent regions, while ZD-AHC considers all the pairs from two clusters and reflects their average traffic. The result of ZD-MST (q = 1) says that District 0 in the downtown was more related to Districts 4 (around Gangnam) and 7 (around Mapo). In addition, the north area of District 6 is clustered with District 5, not the south area of District 6. It implies that the Guro Digital Complex in the north of District 6 has many trips from District 5, in which the large residence towns are located.

6. Conclusions

In this paper, zone discovery algorithms from smart card-based transit logs were presented to identify zones based on passenger trips. A grid-based geocoding system, Geohash, was adopted to map regions near bus stops to the proper level of geographical regions. MST- and AHC-based zone discovery algorithms were evaluated through smart card-based transit data in Seoul bus networks. It is expected that the results of zone discovery algorithms can provide insights based on reality data to understand and improve civil planning and transportation network for different purposes.

References

[1] J. Antikainen, The concept of functional urban area, Findings of the Espon Projects, vol.1, no.1, 2005.

[2] G. Fusco and M. Caglioni, “Hierarchical clustering through spatial interaction data. The case of commuting flows in south-eastern France,” Lect. Notes Comput. Sci., vol.6782, pp.135–151, 2011.

[3] W. Jang, “Travel time and transfer analysis using transit smart card data,” J. Transp. Res. Board, vol.2144, pp.142–149, 2010.

[4] K. Kim, K. Oh, Y.K. Lee, S.H. Kim, and J.-Y. Jung, “An analysis on movement patterns between zones using smart card data in subway networks,” Int. J. Geogr. Inf. Sci., vol.28, no.9, pp.1781–1801, 2014.

[5] M. Konjar, A. Lisec, and S. Drobne, “Method for delineation of functional regions using data on commuters,” Proc. AGILE Int. Conf. Geogr. Inf. Sci., 2010.

[6] R. Moussalli, M. Srivatsa, and S. Asaad, “Fast and Flexible Conversion of Geohash Codes to and from Latitude/Longitude Coordinates,” Proc. Int. Symp. Field-Prog. Cust. Comp. Mach., pp.179–186, 2015.

[7] Public transportation of Seoul Metropolitan Government, http://english.seoul.go.kr/policy-information/traffic/public-transport/

1-bus-operation-system/

[8] A. Schöbel, Optimization in Public Transportation, Springer, 2006.

[9] N.J. Yuan, Y. Zheng, and X. Xie, "Discovering regions of different functions in a city using human mobility and POIs," Proc. ACM SIGKDD Int. Conf. Disc. Data Mining, 2012.

[10] N.J. Yuan, Y. Zheng, X. Xie, Y. Wang, K. Zheng, and H. Xiong, "Discovering Urban Functional Zones Using Latent Activity Trajectories," IEEE T. Knowl. Data En., vol.27, no.3, pp.712–725, 2015.