www.calsec.or.kr
http://dx.doi.org/10.7838/jsebs.2013.18.3.211

스마트카드 빅데이터를 이용한 서울시 지하철 이동패턴 분석

Discovery of Travel Patterns in Seoul Metropolitan Subway Using Big Data of Smart Card Transaction Systems

김관호(Kwanho Kim)*, 오규협(Kyuhyup Oh)*,
이영규(Yeong Kyu Lee)**, 정재윤(Jae-Yoon Jung)***

초 록

지리적으로 인접되어 있으면서 이동관점에서 같은 역할을 수행하는 Zone의 파악은 사람들의 이동흐름을 이해하고 도시개발 및 이동편의성 개선 등을 위한 중요한 정보로 활용된다. 그러나 기존의 연구는 특정 지점간의 이동과 Zone 발견을 개별적으로 수행하여, 거시적 관점에서의 이동패턴을 이해하는 데에는 한계가 존재한다. 따라서 본 연구에서는 스마트카드 전자거래 빅데이터로부터 Zone들을 발견하고 동시에 Zone들 간의 관계를 설명하는 클러스터링 기반의 이동패턴 분석기법을 제안한다. 또한, 설명력과 종속성 관점에서 이동패턴을 정량적으로 평가하는 지표를 제안한다. 제안된 분석기법을 이용하여 서울시 지하철에서 수집된 실 데이터를 분석하여 서울시에서의 이동패턴을 밝혀내고 시각화하였다.

ABSTRACT

Discovering zones which a1re sets of geographically adjacent regions are essential in sophisticated urban developments and people′s movement improvements. While there are some studies that separately focus on movements between particular regions and zone discovery, they show limitations to understand people's movements from a wider viewpoint. Therefore, in this research, we propose a clustering based analysis method that aims at discovering movement patterns, which involves zones and their relations, based on a big data of smart card transaction systems. Moreover, the effectiveness of discovered movement patterns is quantitatively evaluated by using the proposed metrics. By using a real-world dataset obtained in Seoul metropolitan subway networks, we investigate and visualize hidden movement patterns in Seoul.

키워드 : 스마트카드 전자거래, 빅데이터 분석, 승하차 행위, 이동패턴 분석, Zone 발견
Smart Card Transaction, Big Data Analysis, Movement Behavior, Movement Pattern Analysis, Zone Discovery

---

본 논문은 2013년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (No. 2013R1A2A2A03014718).

* Department of Industrial and Management Systems Engineering, Kyung Hee University
** Seoul Metropolitan Rapid Transit Corp.
*** Corresponding Author, Department of Industrial and Management Systems Engineering, Kyung Hee University (E-mail : jyjung@khu.ac.kr)

2013년 07월 10일 접수, 2013년 08월 07일 심사완료 후 2013년 08월 20일 게재확정.

1. 서 론

오늘날 도시환경에서 지하철은 사람들의 출퇴근, 등하교, 여가생활 등의 활동에 매우 중요한 교통수단으로서의 역할을 수행하고 있다. 따라서 사람들의 지하철을 통한 이동은 도시환경에서 지역 간의 특징과 관계를 반영하고 있으며, 이를 통해서 지리적으로 인접되어 있으면서 이동관점에서 동일한 역할을 수행하는 Zone들을 발견하는 것은 이동 관점에서의 지역의 기능적 성격을 규명하고 연관된 지역과의 숨겨진 상호작용을 이해하는데 매우 중요하다[14]. 이와 더불어, Zone 분석은 도시개발 계획수립 및 이동편의성 개선을 위한 중요한 정보로 인식되고 있다. 여기서 Zone은 지리적으로 인접한 지역을 의미한다[4]. 본 논문에서도 유사한 의미로 인접한 지하철역들의 집합을 의미하기 위하여 Zone이라는 용어를 사용한다.

본 연구에서는 스마트카드 전자거래 빅데이터로부터 Zone들을 발견하고 두 Zone간의 연관성을 나타내는 이동패턴(MZP : Movement Pattern between Zones) 분석기법을 제안한다. 제안되는 분석기법은 상향식 접근법(Bottom-Up Approach)을 적용하여 데이터로부터 실생활이 반영되는 의미 있는 이동패턴을 찾는 데 연구의 주안점을 둔다. 첫째로, 수집된 이동 데이터로부터 이동패턴을 추출하기 위한 병합적 군집화 기법(Agglomerative Clustering Method)을 개발하여 어떤 인접지역들이 같은 기능을 수행하는지를 파악할 뿐만 아니라 이와 밀접한 관련을 갖는 인접지역을 동시에 밝히고자 한다. 둘째로, 밝혀진 이동패턴들을 설명력과 종속성 관점에서 밝혀진 이동패턴들을 정량적으로 평가하는 세 가지 지표들을 제안한다.

제시된 분석기법을 이용하여 서울시 지하철 5호선~8호선에서 수집된 승하차 이동 데이터를 분석하고 밝혀진 이동패턴을 제시하고 평가한다. 나아가, 도출된 이동패턴으로부터 서울시에서 나타나는 주요한 이동특성을 분석하고 시각화하였다.

2. 관련 연구

도시환경에서 지리적 이동패턴의 분석은 도시개발, 인구이동, 교통 등 다양한 관점에서 매우 중요한 주제로 인식되어왔다. 그럼에도 불구하고, 방대한 데이터의 수집과 분석기법의 한계로 인해서 그 동안 주로 설문조사를 통한 제한된 범위의 연구들이 수행되어왔다[1]. 하지만 근래에 급속히 보급된 스마트카드로 인해 대량의 이동정보가 실시간으로 축적될 수 있게 되었고, 이를 이용한 자동화된 분석기법이 제시되었다[2].

<Table 1>은 최근에 연구된 스마트카드 데이터를 이용한 이동패턴 분석연구를 나타내고 있다. 기존 연구에서는 지점간의 이동흐름과 Zone 발견을 개별적으로 접근하여, 거시적인 관점에서의 이동현상을 설명하는데 사용될 수 없는 한계가 존재한다. 또한, 제시된 이동패턴을 정량적으로 수치화 하지 못하여 실질적인 의사결정이나 정책수립에 활용되기에는 어려움이 존재한다. 이와 달리, 본 연구에서는 Zone 발견과 동시에 이들 간의 관계를 분석하는 데 목적을 두어 기존의 연구와는 차별화된다고 할 수 있다.

〈Table 1〉 Previous Studies on Movement Analysis Based on Smart Card Transactions

Research Topic	Analysis	Considered Feature	Reference(s)
Movement pattern	Movement pattern discovery between areas	Transfer points and individual behaviors	[6, 12]
Zone discovery	Identification of functional region	Economical and social factors	[4, 8, 10]
Movement prediction	Prediction of origin and destination	Amount of movements between areas	[11]
Transit performance	Estimation of transportation between regions	Usage statistics for routes and stations	[13]

3. 이동패턴 분석

3.1 승하차 데이터 속성

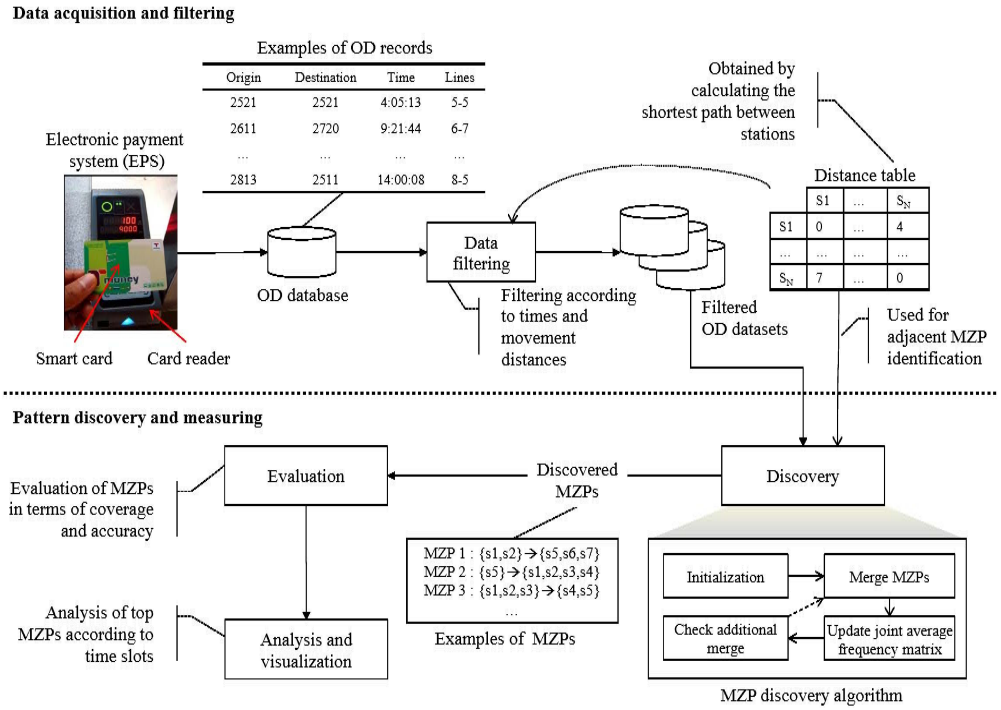
스마트카드 데이터는 수집방법과 활용목적 및 시스템의 성격에 따라 다양한 속성들을 갖는다. 하지만 이들 중 대부분의 속성들은 특정 목적이나 시스템에 종속적이어서 다른 분석환경에서 활용하는 데 제한적이다. 따라서 본 연구에서는 <Table 2>와 같이 스마트카드 데이터의 속성들 중에서 출발-도착(Origin-Destination)에 해당되는 네 가지의 기본적 승하차 속성들만을 활용하여 제시된 모델의 적용범위를 최대화하고자 하였다.

3.2 이동패턴 분석 프레임워크

본 연구에서의 이동패턴 분석은 <Figure 1>에서와 같이 데이터 수집 및 필터링(Data Acquisition and Filtering)과 패턴 발견 및 평가(Pattern Discovery and Measuring)의 두 부분으로 구분된다. 첫째, 데이터 수집 및 필터링 단계에서는 지하철의 스마트카드 시스템에서 기록된 승하차 이동 데이터를 수집하고, 이동거리 및 시간에 따라 분할하여 분석용 데이터베이스를 구축한다. 또한, 지하철 네트워크를 바탕으로 역간의 인접여부 및 역간 최단거리를 계산하여 향후 이동패턴 분석에 활용할 수 있도록 한다. 둘째, 이동패턴

〈Table 2〉 Considered Attributes of Origin-Destination Dataset

Attribute Name	Description
Origin station	Identifier of origin station (from which)
Destination station	Identifier of destination station (to which)
Time	Time to ride at origin station (when)
Lines	Subway lines on which origin and destination stations exist



〈Figure 1〉 Overall Research Framework for Data Acquisition and Filtering(upper side) and Pattern Discovery and Measuring(Lower Side)

분석 및 평가부분에서는 이동패턴 분석 알고리즘을 이용하여 데이터로부터 이동패턴을 분석한 후, 이들을 평가하여 이동패턴들을 정량적으로 비교 평가한다.

3.3 이동패턴 분석 알고리즘

본 연구에서는 개별 데이터로부터 전체적인 관점에서 의미 있는 이동패턴을 발견하기 위해서 병합적 군집화 기법[3]을 이용한다. 기존의 병합적 군집화 기법은 주어진 유사도 함수에 기반으로 주어진 데이터를 트리 형태의 군집화 수행에 목적이 있다[7]. 본 연구에서는 이동 데이터의 방향성과 병합 시 Zone에서의 평균 이동 횟수를 고려하여 이동 데이터들을 병합하도록 기존 기법을 수정하여 이동패턴 분석에 적용하였다. <Table 3>은 이동패턴 분석 알고리즘에 사용된 기호와 이의 설명을 나타내고 있다.

<Figure 2>는 제안된 이동패턴발견 알고리즘을 나타낸다. 알고리즘의 초기화 단계에서는 고유한 개별 이동 데이터를 각각 이동패턴으로 설정하여 총 T개의 이동패턴을 생성하게 된다. 생성된 이동패턴들은 다음 단계에서의 이동패턴 병합에 사용된다. 알고리즘의 k-번째 단계에서 알고리즘은 Zone간의 평균 이동수(Joint Average Movement Frequency) 행렬 $C^{(k)}$에서 가장 큰 값에 대응되는 두 이동 이동패턴을 찾아 이들을 하나의 이동패턴으로 병합한다.

〈Table 3〉 Summary of Notations

Notation	Description
O_i	The i-th origin zone.
D_j	The j-th destination zone.
$p_u = O_i \rightarrow D_j$	The u-th movement that explains movements from the i-th origin zone to the j-th destination zone.
$\rho_{u,v}$	The joint average movement frequency for the u-th and v-th movement patterns.
$f_{i,j}$	The number of movements from the i-th origin zone to the j-th destination zone.
M	The total number of observed movements in a given dataset.

1: **Initialize**
2: Set each station as a zone itself.
3: Build T MZPs for the distinct observed movements.
4: Calculate the joint average frequency matrix $C^{(1)}$ based on $\rho_{u,v}^{(1)}$ for u, v = 1, ···, T.
5: **Repeat**
6: Merge p_u and p_v, u, v = 1, ···, T-k, $u \neq v$, if $\rho_{u,v}^{(k)}$ is the largest value in $C^{(k)}$.
7: Update the joint average frequency matrix $C^{(k)}$.
8: **Until**
9: The highest value of $\rho_{u,v}^{(k)}$, u, v = 1, ···, T-k, is less than a threshold.
10: **Return** the remaining (T-k) MZPs.

〈Figure 2〉 Movement Pattern Discovery Algorithm

이동패턴 병합 후에는 다음 단계를 위해 행렬 $C^{(k)}$중의 일부를 재계산한다. 이때, 각 단계에서의 이동패턴 병합은 행렬의 병합된 이동패턴에 관련된 행과 열 쌍에만 영향을 미치게 되므로 전체 값들을 재계산 할 필요 없이 하나의 이동패턴에 대응되는 값들을 재계산하고, 나머지 하나의 이동패턴에 대응되는 값들을 행렬에서 삭제한다. 알고리즘에서의 이동패턴 병합은 더 이상 병합할 이동패턴이 없을 때까지 반복된다.

예를 들어, 역들의 순서 있는 집합$\{s_i, s_{i+1}, s_{i+2}, s_{i+3}, s_{i+4}, s_{i+5}\}$가 주워지고, 관측된 4개의 이동이 각각 $\{s_{i+1}\}\rightarrow\{s_{i+3}\}$, $\{s_{i+2}\}\rightarrow\{s_{i+4}\}$, $\{s_{i+3}\}\rightarrow\{s_{i+5}\}$, $\{s_{i+4}\}\rightarrow\{s_{i+5}\}$라고 가정하자. 알고리즘의 초기단계에서는 관측된 개별 이동들이 모두 이동패턴으로 설정되어, 총 4개의 이동패턴이 고려된다. 이 중, $\{s_{i+1}\}\rightarrow\{s_{i+3}\}$와 $\{s_{i+2}\}\rightarrow\{s_{i+4}\}$는 서로 인접한 출발역 또는 도착역을 가지고 있기 때문에 두 이동패턴은 병합될 수 있으며, $\{s_{i+3}\}\rightarrow\{s_{i+5}\}$와 $\{s_{i+4}\}\rightarrow\{s_{i+5}\}$도 마찬가지로 병합이 가능하다. 따라서 두 번의 이동패턴 병합 단계를 거친 후 $\{s_{i+1}, s_{i+2}\}\rightarrow\{s_{i+3}, s_{i+4}\}$와 $\{s_{i+3}, s_{i+4}\}\rightarrow\{s_{i+5}\}$가 이동패턴으로 생성되고, 더 이상의 이동패턴 병합이 이루어 질 수 없으므로 알고리즘의 수행은 종료된다.

따라서 이동패턴 병합이 반복될수록 점차 이동패턴의 수가 줄어들게 되며, 알고리즘이

종료된 후 남은 이동패턴을 최종 이동패턴으로 삼는다. 이동패턴이 병합되면서 Zone의 범위가 넓어지기 때문에 남은 이동패턴들은 기존의 것들과 비교해서 우수한 설명력을 보이게 된다. 또한, Zone간 평균 이동을 고려하기 때문에 Zone간의 의존도 측면에서도 양호한 이동패턴을 추출할 수 있게 된다.

제시된 알고리즘에서 $\rho_{u,v}$는 u-번째와 v-번째 이동패턴을 병합한 후의 Zone간 평균 이동 횟수를 나타낸다. 만일 p_u와 p_v가 각각 $O_i \rightarrow D_j$와 $O_{i'} \rightarrow D_{j'}$일 경우, 두 이동패턴들이 병합된 후의 Zone에 인접하지 않은 역이 존재할 경우는 $\rho_{u,v} = 0$으로 설정되고, 해당 Zone에 속하는 모든 역들이 서로 인접한 역을 가지는 경우에 $\rho_{u,v}$의 값은 다음과 같이 정의된다.

$$\rho_{u,v} = \frac{nr(O_i \cup O_{i'} \rightarrow D_j \cup D_{j'})}{|O_i \cup O_{i'}| \bullet |D_j \cup D_{j'}|} \qquad (1)$$

여기서 $nr(\bullet)$은 주어진 이동패턴이 설명할 수 있는 이동 횟수를 나타내고, $O_i \cup O_{i'} \rightarrow D_j \cup D_{j'}$는 Zone O_i 또는 Zone $O_{i'}$ 중 한 역에서 승차하고 Zone D_j 또는 Zone $D_{j'}$ 중 한 역에서 하차한 모든 이동에 대한 이동패턴을 의미한다.

4. 이동패턴 평가

본 연구에서는 이동패턴을 위해 다음과 같이 세 지표들을 제시한다. 제시되는 각각 지표들은 장바구니 분석에서 규칙의 평가를 위해 제안된 지표들인 지지도(Support), 향상도(Lift), 코사인(Cosine)을 기반으로 한다[5, 9]. 기존의 지표들은 동시에 발생하는 사건들에 대해 평가하는데 목적을 두고 있기 때문에, 이를 Zone 내의 임의의 역에서 발생하는 이동을 고려하여 평가할 수 있도록 수정하였다.

첫째, 이동패턴이 얼마나 많은 이동을 설명할 수 있는지를 나타내는 지표(v-value)를 제시한다. 구체적으로, Zone O_i 중 한 역에서 Zone D_j 중 한 역으로의 이동을 나타내는 이동패턴 $O_i \rightarrow D_j$의 v-value는 다음과 같이 정의된다.

$$v(O_i \rightarrow D_j) = \Pr(O_i \rightarrow D_j) = \frac{f_{i,j}}{M} \qquad (2)$$

여기서, $f_{i,j}$은 Zone O_i 중의 한 역에서 탑승하고 Zone D_j 중의 한 역에서 하차한 이동 횟수를 나타내며, M은 관측된 총 이동 횟수를 나타낸다.

둘째, 두 Zone들이 얼마나 큰 상호 종속성을 가지고 있는지를 고려하는 지표(a-value)를 제시하고, 이동패턴 $O_i \rightarrow D_j$의 a-value는 다음과 같이 계산된다.

$$a(O_i \rightarrow D_j) = \frac{\Pr(O_i|D_j)}{\Pr(O_i)} = \frac{\Pr(D_j|O_i)}{\Pr(D_j)} = M\frac{f_{i,j}}{f_{i,*}f_{*,j}} \qquad (3)$$

여기서, $f_{i,*}$과 $f_{*,j}$는 각각 Zone O_i 중의 한 역에서 탑승한 모든 이동 횟수와 Zone O_j 중의 한 역에서 하차한 모든 이동 횟수를 의미한다.

마지막으로, 앞서 제시된 두 지표를 같이 고려하는 복합지표(c-value)를 제시한다. 이동패턴 $O_i \rightarrow D_j$의 c-value는 다음 식을 통해 얻어진다.

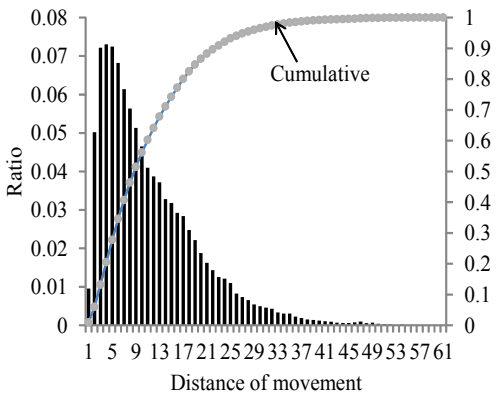
$$c(O_i \rightarrow D_j) = \sqrt{v(O_i \rightarrow D_j)a(O_i \rightarrow D_j)} \quad (4)$$
$$= \frac{f_{i,j}}{\sqrt{f_{i,*}f_{*,j}}}$$

5. 실험결과

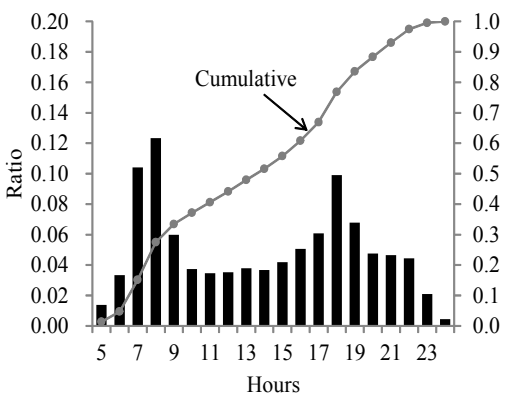
제시된 기법의 효과를 검증하기 위해서 서울도시철도공사에서 2012년 6월 18일부터 22일까지 5호선~8호선에서 수집된 총 5,405,736건의 승하차 이동 데이터 사용하였다. 승하차가 기록된 시간은 05시부터 23시까지이며, 총 역의 수는 148개였다. 수집된 데이터는 5, 6, 7, 8호선이 각각 32%, 25%, 27%, 14%의 분포를 보였다.

<Figure 3>의 (a)는 이동거리에 따른 데이터 분포를 나타내며, 이동 역 개수의 평균은 10.04, 중앙값은 8, 최빈값은 4로 나타났다. 가장 빈번하게 관측된 역간의 이동은 “광명사거리역”에서 “가산디지털단지역”이었으며, 전체 중 0.48%의 이동이 이에 해당되었다. 또한, 전체 21,904개의 출발역과 도착역의 쌍들 중 97개의 쌍이 전체 이동 데이터 중 10%가 넘는 이동에 해당되며, 2,265개의 쌍들은 이동이 관측되지 않았다. 이는 다수의 이동이 소수의 역에 집중되고, 다수의 역은 소수의 이동만이 발생함을 의미한다. 한편, <Figure 3>의 (b)는 시간에 따른 이동 데이터 분포를 나타내고, 출근 시간대인 8시와 퇴근 시간대인 18시에 가장 높은 이동이 발생함을 알 수 있다. 해당 출퇴근 시간에 전체 이동 중 22%가 발생하였으며, 다음 시간대에는 상대적으로 적은 이동이 이루어짐을 알 수 있다.

주요 이동패턴을 살펴보기 위해 수집된 데이터로부터 <Figure 2>에 제시된 알고리즘을 이용하여 이동패턴을 추출하였다. 밝혀진 이동패턴들은 제안된 세 지표로 평가되었다. <Table 4>는 복합지표(c-value)를 기준으로 상위 5개의 이동패턴을 나타내고 있다. 복합지표 관점에서 가장 뚜렷한 패턴을 보이는 Zone들은 7호선의 “철산역(Cheolsan)”-“온수역(Onsu)” 구간과 “상도역(Sangdo)”-“가산디지털단지역(Gasan Digital Complex)” 구간이



(a) Distribution of Movements According to Distances



(b) Distribution of Movements According to Hours

〈Figure 3〉 Distributions of the Collected Smart Cart Transaction Dataset

〈Table 4〉 Discovered Top 5 Movement Patterns for All Time Slots

Rank	Origin Zone	Destination Zone	Measures		
			υ-value (%)	a-value	c-value
1	Cheolsan - Onsu	Sangdo - Gasan Digital Complex	1.552	5.659	0.296
2	Jangseungbaegi - Gasan Digital Complex	Cheolsan - Onsu	1.280	6.122	0.280
3	Balsan - Kkachisan	Bangwha - Songjeong	0.794	7.193	0.239
4	Bangwah - Songjeong	Bansan - Kkachisan	0.782	7.261	0.238
5	Amsa - Mongchontoseong	Jamsil - Garak Market	0.580	7.934	0.215

었다. 해당 이동패턴은 전체의 1.55%에 해당되는 이동을 설명하며, 동시에 매우 높은 의존성을 보였다. 제시된 상위 5개의 이동패턴의 설명력의 합은 4.988%로서, 임의의 두 역간의 평균 설명력이 0.002%임을 감안할 때 분석된 이동패턴들은 Zone간의 종속성뿐만 아니라 이동량 측면으로도 매우 중요한 이동흐름을 표현하고 있다고 볼 수 있겠다.

또한, 3번째와 4번째 이동패턴들은 정확히 같은 Zone들을 가지고 서로 방향이 반대임을 알 수 있다. 이는 출근 및 퇴근과 같이 시간에 따라 이동방향이 바뀌는 현상 때문으로 판단되며, 해당 Zone들은 양방향 모두 높은 의존성을 보이고 있다. 따라서 두 Zone들은 각각 서로에 대한 유출지역과 유입지역의 역할을 수행하고 있는 것으로 볼 수 있다.

다음으로, 시간대별 이동패턴들의 특징을 살펴보기 위해 출근 시간대(08시~09시)와 퇴근 시간대(18시~19시)에 대해서 어떻게 이동패턴이 변화하는지를 살펴보았다. <Table 5>와 <Table 6>은 각각 출근과 퇴근시간대에 복합지표 기준으로 상위 5개의 이동패턴을 나타낸다. 또한, <Figure 4>는 추출된 시간대별 주요 이동패턴을 서울시 지도를 이용하여 시각화하고 있다. <Figure 4>의 (a)에서는 출근 시간대에는 외곽지역(거주지역)에서 중심지역(상업지역)으로의 이동패턴을 확인할 수 있다. 또한 <Figure 4>의 (b)는 18시에 보이는 상위 5개의 이동패턴을 나타낸다. 퇴근 시간대에는 출근 시간대와 유사한 Zone들에서 방향이 반대인 이동패턴이 관측됨을 알 수 있다. 특히 "모란(Moran)"과 "수진(Sujin)"은 퇴근 시간대에 매우 높은 a-value를 보여, 이 두 지역은 밀접하게 연관되어 있다고 할 수 있겠다.

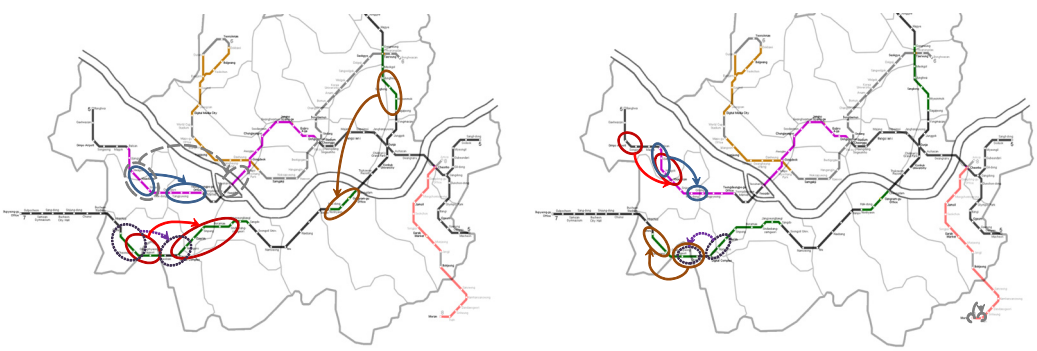
분석된 복합지표 기준의 상위 이동패턴을 살펴보면 모두가 같은 호선에서 비교적 가까운 Zone간의 강한 연관성을 보여주고 있음을 알 수 있다. 다른 호선의 Zone과의 강한 연관성이 나타나지 않은 이유는 다음과 같은 두 가지 측면으로 설명할 수 있겠다. <Figure 4>의 (a)에서 나타내는 바와 같이 전체 이동 중에서 탑승 후 3~5개의 역만을 통과한 후 하차하는 빈도가 가장 높기 때문에, 자연스럽게 두 Zone들 간의 거리 또한 비교적 가깝게 형성되고 있다고 볼 수 있다. 이와 더불어, 본 연구는 5~8호선에서 수집된 데이터로만

〈Table 5〉 Discovered Top 5 Movement Patterns at 08~09 Hours

Rank	Origin Zone	Destination Zone	Measures		
			v-value (%)	a-value	c-value
1	Cheolsan - Gwangmyeongsageori	Sindaebangsamgeori - Gasan Digital Complex	3.759	4.028	0.389
2	Cheonwang - Onsu	Namguro - Cheolsan	1.874	4.702	0.297
3	Hwagok - Sinjeong	Omokgyo - Yeongdeungpo-gu Office	1.755	4.141	0.270
4	Junghwa - Sagajeong	Cheongdam - Nonhyeon	3.757	1.546	0.241
5	Hwagok - Sinjeong	Yeouido - Mapo	1.365	2.404	0.181

〈Table 6〉 Discovered Top 5 Movement Patterns at 18~19 Hours

Rank	Origin Zone	Destination Zone	Measures		
			v-value (%)	a-value	c-value
1	Daerim - Gasan Digital Complex	Cheolsan - Gwangmyeongsageori	2.110	2.521	0.231
2	Gimpo Int'l Airport - Songjeong	Ujangsan - Kkachisan	1.746	2.581	0.212
3	Balsan - Hwagok	Mok-dong - Omokgyo	1.766	2.043	0.190
4	Moran	Sujin	0.194	16.223	0.177
5	Cheolsan - Gwanmyeongsageori	Cheonwang - Onsu	0.732	4.076	0.173



(a) Movement Patterns at 08~09 hours (b) Movement Patterns at 18~19 hours

〈Figure 4〉 Visualization of Top 5 Movement Patterns in Seoul According to Time

수행되어, 타 호선과 관련되고 멀리 떨어진 Zone간을 갖는 숨겨진 이동패턴이 분석되지 못한 한계가 존재할 것으로 판단된다.

6. 결론 및 추후 연구

본 논문에서는 스마트카드 빅데이터를 이용하여 이동패턴을 추출하여, 지리적으로 유사하면서도 동일한 기능을 수행하는 Zone을 발견하고 이들 간의 연관성을 파악하고자 하였다. 또한, 추출된 이동패턴을 정량적으로 평가하기 위한 지표를 제안하였다. 특히, 제안된 이동패턴 분석기법은 데이터 측면에서 Zone을 발견함과 동시에 Zone간의 관계분석을 수행하여, 오늘날 도시계획 및 수송계획에 필수적인 권역분석 및 권역 간의 이동분석에 직접적으로 활용될 수 있다는 측면에서 기존의 연구와는 차별화된다고 할 수 있다.

기존의 도시 내 출퇴근 패턴 분석이나 지하철 이동패턴에 대한 연구들이 주로 설문이나 센서스와 같은 데이터를 바탕으로 한 통계적 분석이 주를 이루고 있다. 그러나 최근에 방대한 양의 데이터를 대상으로 한 빠른 분석 도구들이 등장함에 따라 본 연구에서 대상으로 한 스마트카드 데이터와 같은 실제 이동데이터를 활용하는 연구들이 늘어나고 있다. 과거 비교적 적은 양의 데이터를 대상으로 한 연구에서 확인하기 힘든 다양한 주제들을 다룰 수 있다는 측면에서 빅데이터를 이용한 이동패턴 연구가 앞으로도 확산될 것으로 예상된다.

본 연구의 결과는 지하철뿐만 아니라 버스 및 택시 등의 다양한 대중교통에서 발생하는 대용량 이동 데이터를 바탕으로 보다 정확하고 의미 있는 이동분석을 가능하게 할 것이다. 또한, 설명력과 같은 지표들이 보다 의미 있는 지표로 활용되기 위해 데이터의 전처리 또는 승하차역간의 조합을 고려한 지표의 개선이 필요할 것으로 판단된다. 나아가서, 제시된 분석결과는 도시계획, 대중교통 서비스향상, 대체 이동수단 보완 등에 활용될 수 있을 것으로 기대된다.

References

[1] Bagchi, M. and White, P. R., “The Potential of Public Transport Smart Card Data,” Transport Policy, Vol. 12, No. 5, pp. 464-474, 2005.

[2] Blythe, P., “Improving Public Transport Ticketing Through Smart Cards,” Proceedings of the Institute of Civil Engineers, Municipal Engineer, Vol. 157, pp. 47-54, 2004.

[3] Day, W. and Edelsbrunner, H., “Efficient Algorithms for Agglomerative Hierarchical Clustering Methods,” Journal of Classification, Vol. 1, No. 1, pp. 7-24, 1984.

[4] Fusco, G. and Caglioni, M., “Hierarchical Clustering Through Spatial Interaction Data. The Case of Commuting Flows in South-Eastern France,” Lecture Notes in Computer Science, Vol. 6782, pp. 135-151, 2011.

[5] He, B., Ding, Y., and Yan, E., “Mining

Patterns of Author Orders in Scientific Publications," Journal of Informetrics, Vol. 6, No. 3, pp. 359-367, 2012.

[6] Jang, W., "Travel Time and Transfer Analysis Using Transit Smart Card Data," Journal of the Transportation Research Board, Vol. 2144, pp. 142-149, 2010.

[7] Jung, J.-Y., "PROCL : A Process Log Clustering System," Journal of Society for e-Business Studies, Vol. 13, No. 2, pp. 181-194, 2008.

[8] Karlsson, C., "Clusters, Functional Regions and Cluster Policies," JIBS and CESIS Electronic Working Paper Series, Vol. 84, 2007.

[9] Kim, J.-H. and Heo, H., "An Interpretation of Interoperability Definitions Using Association Rules Discovery," Journal of Society for e-Business Studies, Vol. 16, No. 2, pp. 39-91, 2011.

[10] Konjar, M., Lisec, A., and Drobne, S., "Method for Delineation of Functional Regions Using Data on Commuters," Proceedings of the 13-th AGILE International Conference on Geographic Information Science, Portugal, 2010.

[11] Park, J. Y. and Kim, D. J., "The Potential of Using the Smart Sard Data to Define the Use of Public Transit in Seoul," Journal of the Transportation Research Board, Vol. 2063, No. 1, pp. 3-9, 2008.

[12] Srinivasan, S. and Ferreira, J., "Travel Behavior at the Household Level : Understanding Linkages with Residential Choice," Transportation Research Part D, Vol. 7, No. 3, pp. 225-242, 2002.

[13] Trepanier, M., Morency, C., and Agard, B., "Calculation of Transit Performance Measures Using Smart Card Data," Journal of Public Transportation, Vol. 12, No. 1, pp. 79-96, 2009.

[14] Yuan, J., Zheng, Y. and Xie, X., "Discovering Regions of Different Functions in a City Using Human Mobility and POIs," Proceedings of the 18-th ACM SIGKDD International Conference on Discovery and Data Mining, Vol. 12, pp. 186-194, 2013.

저 자 소 개

김관호 (E-mail : kwanhokim@khu.ac.kr)
2006년 동국대학교 정보시스템전공 (학사)
2012년 서울대학교 산업공학과 (박사)
2012년~현재 경희대학교 산업경영공학과 Post-Doc
관심분야 통계적 기계학습, 빅데이터 분석

오규협 (E-mail : k8383@khu.ac.kr)
2010년 경희대학교 산업공학과/컴퓨터공학과 (학사)
2010년~현재 경희대학교 산업경영공학과 석박사통합과정
관심분야 유비쿼터스 프로세스 마이닝, 인터넷 비즈니스, 빅데이터

이영규 (E-mail : magpie@smrt.co.kr)
2002년 경희대학교 경영학부 (학사)
2002~현재 서울특별시 도시철도공사
관심분야 도시철도 중심의 교통량 분석 및 예측, 교통량 빅데이터

정재윤 (E-mail : jyjung@khu.ac.kr)
1999년 서울대학교 산업공학과 (학사)
2001년 서울대학교 산업공학과 (석사)
2005년 서울대학교 산업공학과 (박사)
2005년~2006년 네덜란드 아인트호벤공대 초빙연구원
2006년~2007년 유비쿼터스컴퓨팅 원천기술개발지원센터
2007년~현재 경희대학교 산업경영공학과 전임강사, 조교수
관심분야 비즈니스 프로세스 관리, 프로세스 마이닝, 빅데이터 분석