DOI : 10.5392/JKCA.2010.10.12.091

교통카드 트랜잭션 데이터베이스에서 지하철 탑승 패턴 분류

Classification of Subway Trip Patterns from Smart Card Transaction Databases

박종수*, 김호성**, 이금숙***
성신여자대학교 IT학부*, 성신여자대학교 미디어커뮤니케이션학과**, 성신여자대학교 지리학과***

Jong Soo Park(jpark@sungshin.ac.kr)*, Ho Sung Kim(hkim@sungshin.ac.kr)**,
Keumsook Lee(kslee@sungshin.ac.kr)***

요약

서울 수도권 지하철 승객들의 탑승 패턴의 특성을 이해하는 것은 효율적인 수도권 지하철 시스템을 입안하는 데 중요하기 때문에 대용량 교통카드 트랜잭션 데이터베이스에서 유용한 패턴을 탐사하거나 귀중한 패턴의 분류에 대한 연구가 진행되어오고 있다. 본 논문에서 새로운 지하철 탑승 분류를 정의하고 하루 약 천만 건 트랜잭션들로 구성된 교통카드 트랜잭션 데이터베이스로부터 지하철 승객들의 11 가지 탑승 패턴을 분류하는 알고리즘을 제안하였다. 제시된 알고리즘을 구현하여 탑승 패턴들을 분류하기 위하여 하루 동안의 교통카드 트랜잭션 데이터베이스에 적용하였다. 실험 결과에서 왕복-탑승 패턴, 통근 패턴, 예상치 못한 흥미로운 패턴들에 초점을 맞추어 분석하였다. 각 분류된 패턴에 대해서 시간대별로 승객수를 지하철 트랜잭션의 승차시간과 하차시간 기준으로 그래프로 설명하여 유용한 패턴의 특성을 이해하도록 하였다.

■ 중심어 : | 분류 | 통행 분석 | 교통카드 트랜잭션 데이터베이스 | 패턴 탐사 |

Abstract

To understand the trip patterns of subway passengers is very important to making plans for an efficient subway system. Accordingly, there have been studies on mining and classifying useful patterns from large smart card transaction databases of the Metropolitan Seoul subway system. In this paper, we define a new classification of subway trip patterns and devise a classification algorithm for eleven trip patterns of the subway users from smart card transaction databases which have been produced about ten million transactions daily. We have implemented the algorithm and then applied it to one-day transaction database to classify the trip patterns of subway passengers. We have focused on the analysis of significant patterns such as round-trip patterns, commuter patterns, and unexpected interesting patterns. The distribution of the number of passengers in each trip pattern is plotted by the get-on time and get-off time of subway transactions, which illustrates the characteristics of the significant patterns.

■ keyword : | Classification | Trip Analysis | Smart Card Transaction Database | Pattern Mining |

I. 서 론

서울시를 중심으로 한 수도권에서 인구 밀집 현상이 일어나서 효율적인 일반 대중교통 서비스가 절실하다. 2004년에 서울시가 일반 대중교통 서비스의 질을 향상시키기 위하여 교통카드의 도입으로 버스와 지하철을

* 본 연구는 2009학년도 성신여자대학교 학술연구조성비 지원에 의하여 연구되었음.
접수번호 : #100824-003
접수일자 : 2010년 08월 24일
심사완료일 : 2010년 11월 17일
교신저자 : 박종수, e-mail : jpark@sungshin.ac.kr

연계하고 환승하게 하여 대중교통요금을 통합하였다. 대중교통을 사용하는 비율이 60% 이상이 되고 지하철 승객들은 이중에서 40% 정도에 이른다. 대중교통을 이용하는 한 승객이 지하철을 사용했을 경우, 승차시간, 승차지하철역 이름, 하차시간, 하차지하철역 이름 등의 정보가 실시간으로 한 트랜잭션에 저장되고 있다. 수도권에서 하루 천만 건이 넘는 트랜잭션들이 처리되어 대용량의 트랜잭션 데이터베이스로 관리되고 있다. 2010년 1월 기준으로 수도권 지하철은 1호선에서 9호선까지, 분당선, 인천1호선, 중앙선으로 12개 노선이 연계 운영되고 있으며 교통카드 이용률은 2006년도부터 2008년도까지 해마다 79.5%, 80.3%, 81.6%로 지속적으로 증가하고 있다[1].

본 논문에서는 지하철 승객들의 통행 특성을 연구하고 기존의 방법론들을 분석한 후에 수도권 지하철을 이용하는 승객들의 탑승 패턴들을 논리적으로 분류하여 실험적으로 이 패턴들에 속하는 승객들의 숫자와 탑승 시간 등을 여러 관점에서 분석하고자 한다. 본 논문에서의 패턴 분류 방법은 다음과 같은 세 가지 관점에서 연구하였다:

1) 지하철 승객들 중에서 얼마나 많은 사람들이 처음 출발지로 되돌아오는가?
2) 지하철 승객들 중에서 얼마나 많은 사람들이 통근 수단으로 지하철을 이용하는가?
3) 예상하지 못한 흥미로운 패턴을 발견할 수 있는가?

앞의 세 가지 관점에서 지하철을 이용하는 승객들의 탑승 패턴을 각 승객의 탑승 트랜잭션들의 개수에 따라 경우의 수를 분석하여 11개 패턴들로 분류하도록 하였다. 교통카드 트랜잭션 데이터베이스로부터 각 승객의 탑승 패턴을 11개 패턴들 중의 하나로 분류하는 알고리즘을 구현하여 각 패턴의 지지도와 승객들의 지하철 이용 시간 등을 계산해내었다.

대용량 데이터베이스에서 패턴을 탐사하는 연구는 연관규칙 탐사와 순차패턴 탐사가 이루어지고 있고[2] 웹에서 이동 상황을 파악하는 순회패턴 탐사 등의 연구가 진행되어 오고 있다[3]. 교통 카드의 도입으로 하루 천만건 이상의 트랜잭션 데이터베이스에서 가치 있고 흥미로운 패턴이나 지식을 탐사하는 연구도 많이 이루어지고 있다. 하루 동안의 교통카드 트랜잭션 데이터베이스에서 순차 패턴의 알고리즘을 적용하여 승객들이 이동한 지하철역들의 시퀀스를 탐사하는 알고리즘이 연구되었고[4], 승객들의 통행 패턴에 대한 일반적인 분석을 수행하기 위하여 1년에 하루치씩 3년간의 교통카드 트랜잭션 데이터베이스로부터 승객 시퀀스의 평균 정류장 개수와 환승 횟수 등을 연도별로 비교하였다[5]. 수도권 지하철 시스템의 네트워크의 구조적 특성을 분석하는 연구도 시도 되었다[6]. 지리학의 지리 교통 분야에서의 통행에 관한 연구는 주로 승객의 기종점 통행 행렬(Origin-Destination trip matrix) 분석을 수행하였다[7]. 이러한 기종점 분석은 출발역의 승차 승객의 수와 도착역의 하차 승객의 수에 대한 분석이 주를 이루었다. 서울 수도권 지하철에서 기종점 통행에 의한 승객 흐름의 분포가 멱함수 법칙(power law)임을 보여주는 연구가 있다[8]. 서울 수도권의 지하철 교통망에서 승객들의 흐름을 찾아내어 지하철의 모든 링크상의 승객 흐름을 분석하였다[9]. 승객 흐름을 시각화하여 시간대별로 승객들의 이동 흐름을 보여주는 연구[10]를 하였고, 교통카드 트랜잭션 데이터베이스에서 통근 패턴을 탐사하는 것에 대한 기초적인 연구[11]가 있다.

본 논문의 구성에 대해 설명하면, Ⅱ절에서는 교통카드 트랜잭션 데이터베이스에 대해 속성의 종류와 특성에 대해 설명하고 데이터베이스의 크기에 대해서도 설명한다. Ⅲ절에서는 지하철 탑승 패턴을 분류하기 위하여 승객의 트랜잭션을 기준으로 경우의 수를 고려한 분류 방법을 서술하고 분류 알고리즘을 설명한다. Ⅳ절에서는 실험결과를 기술하고 탑승 패턴들의 특성을 상술하고, Ⅴ절에서는 결론을 서술하고 추후 연구 내용을 설명한다.

Ⅱ. 교통카드 트랜잭션 데이터베이스

수도권 지역의 승객이 지하철을 이용하면서 교통카드를 사용하는 경우에 (주)한국스마트카드(KSCC)에서 요금을 정산한다. KSCC에서 관리하는 트랜잭션은 승

객의 승차와 하차에 관련된 여러 가지 속성들을 가지고 있다. 승객들에 대한 개인 정보를 제외하고 정제된 승차와 하차에 관한 데이터만 [표 1]의 속성 부분에서 24가지 속성들을 나타내었다. 속성 중에서 카드번호는 승객의 카드 번호가 아니고 임의로 부여된 일련번호에 해당된다. 트랜잭션ID는 한 승객이 하루 동안에 발생한 트랜잭션들의 순서를 나타낸다. [표 1]의 트랜잭션 예제 1-4는 각 승객이 지하철을 사용했을 때 승차와 하차에 관한 데이터를 각각의 트랜잭션으로 보여주고 있다. 트랜잭션 예제 5와 6은 버스 승객의 승차와 하차에 관한 데이터이다.

트랜잭션 예제 1과 2는 한 승객이 두 번 지하철을 사용한 것을 나타내는 것으로 카드번호 10329는 동일하고 트랜잭션ID는 015와 016으로 서로 다르다. 두 트랜잭션은 아침에 독산역에서 지하철을 승차하여 신길역에 하차한 후 약 3시간의 업무를 보거나 또는 머문 후에 다시 신길역에서 승차하고 독산역에서 하차한 트랜잭션을 보여주고 있다. 이 승객의 탑승 유형은 출발지에서 도착지로 가서 일정 시간 후에 다시 출발지로 되돌아오는 패턴으로 왕복-탑승 패턴에 해당된다. 트랜잭션 3과 4는 어떤 승객이 출발지에서 도착지에 도착한 후에 일정 시간이 지나고 다른 곳에서 승차하여 처음 출발지로 되돌아가는 것이 아니라 다른 도착지로 가는 유형에 속한다. 이 승객은 아침 9시 6분 55초에 연신내역에서 승차하여 9시 44분 19초에 압구정역에 하차하여 약 11시간 후인 오후 8시 55분 49초에 압구정역에서 승차하여 처음 출발역이 아닌 녹번역에 하차한 것을 보여주고 있다.

트랜잭션 예제 5와 6은 카드번호가 같은 한 승객이 오전 7시 4분에 쑥고개시장 버스정류장에서 승차하고 7시 26분에 동작구청앞 버스정류장에서 하차한 것을 보여주고, 다시 오후 2시 51분에 동작구청앞에서 승차하고 3시 7분에 봉원중학교 버스정류장에서 하차한 것을 나타낸다.

교통카드 트랜잭션 예제는 승차와 하차에 대한 정보를 내포하고 있기 때문에 승객이 대중교통에 승차하는 시점에 교통카드를 사용하고 하차하는 시점에 교통카드를 사용한 결과가 하나의 승하차 트랜잭션이 완성된다. 그러므로 하루에 천백만 트랜잭션들은 2200만 건의 승차와 하차 교통카드를 처리하고 정제한 결과이다. 대략 하루에 처리되는 교통카드 데이터는 10GB에 해당되고 1년에 3TB에 이르는 거대 데이터베이스를 만들어가고 있다. 각 개인의 정보를 제외하고 교통카드 트랜잭션에 일련번호로 코딩된 천백만 트랜잭션 데이터베이스의 크기는 약 2GB에 이른다. 대중 교통정책의 입안이나 승객들의 통행 특성을 파악하려면 대용량 교통카드 트랜잭션 데이터베이스에서 유용한 패턴이나 통찰을 추출해내는 연구가 많이 이루어져야 한다.

표 1. 교통 카드 트랜잭션의 속성과 예제(2005년 기준)

속성 (attribute)	카드번호,승차일시,트랜잭션ID,교통수단CD,환승횟수,버스노선ID,버스노선명,교통사업자ID,교통사업자명,차량ID,차량등록번호,사용자구분코드,사용자구분명,운행출발일시,승차정류장ID,승차정류장명,하차일시,하차정류장ID,하차정류장명,이용객수_다인승,승차금액,승차위반금액,하차금액,하차위반금액
트랜잭션 예제 1	10329,20050624093758,015,200,0,,,2111 00000,한국철도공사,,,01,일반,,1714,독산,20050624095917,1032,신길,1,800,0,0,0
트랜잭션 예제 2	10329,20050624132650,016,200,0,,,2111 00000,한국철도공사,,,01,일반,,1032,신길,20050624134640,1714,독산,1,800,0,0,0
트랜잭션 예제 3	10418,20050624090655,005,200,1,,,2110 00000,서울지하철공사,,,01,일반,,0311,연신내,20050624094419,0326,압구정,1,300,0,200,0
트랜잭션 예제 4	10418,20050624205549,006,200,0,,,2110 00000,서울지하철공사,,,01,일반,,0326,압구정,20050624212748,0313,녹번,1,800,0,100,0
트랜잭션 예제 5	1020176,20050624070423,005,120,0,111 10242,5517번(서울대~중앙대),111007800,한남여객운수(주),111748516,서울74사8516,01,일반,20050624065552,0009970,쑥고개시장,20050624072620,0009929,동작구청앞,1,800,0,0,0
트랜잭션 예제 6	1020176,20050624145136,006,115,0,111 10063,641번(문래동~양재동),111007100,중부운수주식회사,111707974,서울70사7974,01,일반,20050624142035,0009926,동작구청앞,20050624150744,0008450,봉원중학교,1,800,0,0,0

본 논문에서는 교통카드 트랜잭션 데이터베이스로부터 지하철을 사용하는 승객들만의 트랜잭션을 찾아내고, 그 트랜잭션에서 탑승 패턴 탐사에 사용되는 속성들인 카드번호, 승차일시, 트랜잭션ID, 승차정류장ID,

교통수단CD(코드), 하차일시, 하차정류장ID를 추출해 내어 탑승 패턴의 분류에 이용한다.

Ⅲ. 지하철 탑승 패턴 분류 및 알고리즘

1. 승객의 탑승 유형 분석 및 탑승 패턴 분류

앞서 연구된 통행패턴 분석[4][5]을 살펴보면, 데이터 마이닝의 한 기법인 순회 패턴과 순차 패턴을 탐사해내는 알고리즘을 승객들의 승차역과 하차역 사이의 통과 역들의 시퀀스들에 적용하여 빈발 시퀀스를 찾아내는 것이었다. 승객들이 이동하는 통과역들의 시퀀스를 탐사해내는 방법 이외에 또 다른 통행 특성을 분석해내는 방법을 고찰해보자. 예를 들어 승객들이 지하철을 이용하여 통근을 한다고 가정하면, 통근 패턴으로 나타날 수 있는 여러 유형들을 살펴보고 이중에서 승객들이 이용할 가능성이 있는 패턴들의 유형을 분류 집합에 포함시킨 후에 실제로 주어진 각 패턴의 지지도를 살펴보면 원하는 탑승 패턴을 찾아내고 해석할 수 있을 것이다.

본 논문에서는 대중교통 이용자 중에서 지하철을 이용하는 승객들의 탑승 패턴을 논리적으로 추정하여 발생할 수 있는 모든 조합의 경우의 수에 해당되는 패턴들을 고려하고 그 중에서 많은 승객들이 지지할 것으로 예측되는 패턴들을 지정하여 하나의 분류 집합으로 한다. 승객이 출발역에서 승차하여 도착역에 하차하고 난 후에 일정 시간이 지나고 다시 출발역으로 되돌아오는 경우와 같이 여러 종류의 왕복 패턴이나 되돌아오지 않는 패턴 등을 고려해볼 수 있다. 본 논문의 연구에 의하면, 승객들이 지하철을 한 번 승차하는 비율은 38% 정도이고 두 번 승차하는 비율은 50%이다. 두 비율을 합하면 88%가 되어 대부분의 지하철 승객들이 두 번 이내에 지하철을 승차하고 하차하는 것임을 알 수 있다. 한 번에서 세 번 까지 탑승하는 승객들을 합한 비율은 전체의 약 97%가 된다. 이런 분석에 근거하여 승객들이 승차하고 하차하는 모든 조합의 경우의 수를 따져보고 탑승 패턴에 적절한 패턴의 유형을 유도해낸다. 다음 설명에서 알파벳 기호는 지하철역 이름을 나타낸다. 화살표를 포함한 표시 a ⇒ b는 한 승객이 지하철의 출발역 a에서 탑승하여 도착역 b까지 가는 한 번 탑승을 나타내고, [표 1]에서 지하철 승객의 하나의 교통카드 트랜잭션으로 표현할 수 있다.

한 번 탑승하는 경우에는 두 가지 유형이 있다:

case 11: a ⇒ b

case 12: a ⇒ a

첫 번째 경우는 한 역에서 출발하여 다른 역에 도착하는 것이다. 두 번째 경우는 출발역으로 되돌아오는 경우의 패턴이다.

두 번 탑승하는 조합을 열거하기 위해서 일반적인 승차와 하차를 표시하면 다음과 같다:

A ⇒ B, C ⇒ D

처음 출발역에 올 수 있는 지하철역을 A로 하여 경우의 수로 1가지, 처음 도착역 B에 올 수 있는 경우의 수는 출발역 A와 다른 지하철역으로 2가지, 두 번째 출발역 C에 올 수 있는 경우의 수는 A, B, 그리고 또 다른 지하철역으로 3가지이다. 마지막으로 두 번째 도착역 D에 올 수 있는 경우의 수는 A, B, C와 또 다른 지하철역으로 4가지가 된다. 두 번 탑승하는 경우의 모든 조합가능한 수는 $1 \times 2 \times 3 \times 4 = 24$ 가지가 된다. 두 번 탑승 트랜잭션들로 나올 수 있는 24가지의 경우의 수에서 일반적으로 많은 지하철 승객들이 이용할 것으로 예상되는 4가지 탑승 유형과 그 이외의 유형을 고려해 볼 수 있다:

case 21: a ⇒ b, b ⇒ a

case 22: a ⇒ b, b ⇒ c

case 23: a ⇒ b, c ⇒ a

case 24: a ⇒ b, c ⇒ d

case 25: 위의 네 경우가 아닌 두 번 탑승 경우

첫 번째 case 21의 경우는 출발지에서 승차하고 도착지에서 하차한 후에 일정 시간이 지난 후에 도착지에서 다시 승차하여 처음 출발지로 되돌아와서 하차하는 두 번 지하철을 탑승하는 것으로 전형적인 출퇴근 사례이다. case 24는 첫 트랜잭션과 두 번째 트랜잭션의 승하차역이 모두 다른 경우를 나타낸다. case 25는 앞의 네 가지 경우를 제외한 무의미할 수 있는 20가지 경우를 대표하는 것이다. 그것들 중의 하나의 예로 a ⇒ b, a ⇒ b를 살펴보면, 두 번째 출발역은 처음 출발역으로 가

서 탑승하고 다시 처음 도착역에서 하차하는 경우로 일반적인 패턴으로 볼 수 없다.

한 승객이 지하철을 세 번 이상 탑승하는 경우에 발생할 수 있는 조합의 수는 너무 많기 때문에, 일반적으로 지하철을 탑승하는 방식과 본 논문에서 분석하고자 하는 유형을 결합하여 다음과 같이 세 가지 유의미한 탑승 유형과 그 이외의 나머지 조합의 수를 대표하는 경우를 고려해보기로 한다:

case 31: a ⇒ b ... b ⇒ a

case 32: a ⇒ b ... y ⇒ a

case 33: a ⇒ b ... y ⇒ z

case 34: 위의 세 경우가 아닌 세 번 이상 탑승 경우

먼저 세 가지 탑승 유형을 설명하면, case 31은 한 승객이 세 번 이상 지하철을 탑승하는데 첫 번째 탑승과 마지막 탑승한 지하철의 승차역과 하차역에 관심을 두어 분석하고 있다. 마지막 탑승하는 승차역은 첫 번째 승차의 도착역에서 다시 출발하여 처음 출발역으로 되돌아가는 경우를 나타낸다. case 32는 마지막 출발역은 첫 번째 탑승의 하차역과 다르지만 마지막 도착역은 처음 출발역으로 되돌아가는 경우를 설명하고, case 33은 마지막 탑승한 지하철의 출발역과 도착역이 처음 출발역과 도착역과 전혀 다른 경우를 나타내고 있다. case 34는 세 번 탑승 트랜잭션들로 조합할 수 있는 수많은 패턴들 중에서 앞의 세 경우를 제외한 경우들의 패턴을 나타낸다. 앞에서 설명한 11가지 탑승 패턴을 표 2에 요약하였다.

표 2. 지하철 탑승 패턴의 분류

패턴유형	승하차 지하철역	탑승회수	본문설명
pattern 1	a ⇒ b	한 번	case 11
pattern 2	a ⇒ b, b ⇒ a	두 번	case 21
pattern 3	a ⇒ b, b ⇒ c	두 번	case 22
pattern 4	a ⇒ b, c ⇒ a	두 번	case 23
pattern 5	a ⇒ b, c ⇒ d	두 번	case 24
pattern 6	a ⇒ b ... b ⇒ a	세 번 이상	case 31
pattern 7	a ⇒ b ... y ⇒ a	세 번 이상	case 32
pattern 8	a ⇒ b ... y ⇒ z	세 번 이상	case 33
pattern 9	*pattern 1이 아닌 경우*	*한 번*	*case 12*
pattern 10	*pattern 2,3,4,5가 아닌 경우*	*두 번*	*case 25*
pattern 11	*pattern 6,7,8이 아닌 경우*	*세 번 이상*	*case 34*

2. 왕복-탑승 패턴

[표 2]에서 11가지 탑승 패턴 중에서 처음 출발역으로 되돌아오는 왕복-탑승 패턴들(round-trip patterns)은 패턴 2, 4, 6, 7이다. [표 2]에서 패턴 9(case 12: a ⇒ a)는 출발역에서 탑승하여 다시 출발역에서 하차하는 것으로 일반적인 지하철 사용 방법에서 예외적인 상황이라 왕복-탑승 패턴 부류에서 제외한다. 전체 승객들 중에서 얼마나 많은 사람들이 다시 처음 출발역으로 되돌아오는 이런 패턴에 속하는 지를 실험결과에서 분석한다.

3. 집-직장 통근 패턴

어떤 승객이 회사에 출근하기위하여 지하철을 이용한다고 가정하면, 집에서 가장 가까운 역에서 승차하여 시간거리가 가장 짧은 최단 거리 알고리즘으로 회사 근처의 지하철역에서 하차할 것이다. 승차역에서 하차역까지 가는 노선이 한 노선이 아니고 환승을 하더라도 서울 수도권 지하철 시스템에서는 한 번의 탑승 트랜잭션으로 교통카드에서 처리된다. 왕복-탑승 패턴들 중에서 본 논문에서 관심을 가지고 있는 패턴은 집과 직장 간의 통근 수단으로 지하철을 사용하는 패턴 2번과 4번이다. 패턴 6과 7은 어떤 승객이 지하철을 이용하여 집과 직장 사이를 이동하더라도 출근과 퇴근을 전후하여 다른 지하철역에 한 번 이상 이동하는 경우를 나타내는 것으로 생각할 수 있다. 그래서 패턴 2번과 4번을 집-직장 통근 패턴들(home-office commuter patterns)로 정의한다. 실험 결과에서 이 패턴으로 얼마나 많은 비율의 승객들이 이용하는 지를 분석해본다. 패턴 2와 4번을 이용하는 승객들이 직장에서 근무 시간은 첫 번째 탑승의 하차 시간과 두 번째 탑승의 승차 시간 사이의 잔류시간으로 볼 수 있다. 잔류시간의 히스토그램을 분석하여 상근 직장인에 추정되는 통근 패턴의 비율을 알아본다.

4. 예기치 못한 흥미로운 패턴

[표 2]에서 패턴 1에서 8번까지는 어느 정도 쉽게 예측할 수 있는 일반적인 패턴들이다. 패턴 9는 한 승객이 출발역에서 개찰하였다가 바로 나오는 경우나 다른 역

들에 들렀다가 다시 출발역으로 되돌아오는 경우를 추측해볼 수 있다. 이런 패턴은 분석적인 관점에서 아주 흥미로운 패턴이라 여겨진다. 패턴 10은 두 번 탑승에서 일반적인 네 가지 경우를 제외한 20가지 경우의 패턴을 대표하는 것으로 승객들의 통행 행위로는 일반적이지 아닌 경우로 생각할 수 있다. 패턴 11은 세 개 이상의 트랜잭션들에서 일어날 수 있는 확률이 높은 패턴들 중의 하나다. 그래서 패턴 9와 10은 본 논문에서 예기치 못한 흥미로운 패턴들(unexpected interesting patterns)로 보고 실험결과에서 어느 정도의 승객들이 이용하는 지를 분석한다.

5. 탑승 패턴 분류 알고리즘

[그림 1]은 지하철 탑승 패턴을 분류하고 찾아내는 알고리즘을 간략히 설명한 것이다. 먼저 지하철 네트워크를 빠른 데이터 접근이 용이하도록 해시 테이블로 구성하였다. 지하철 네트워크는 지하철역들로 이루어진 정점집합(vertex set) V와 지하철역들 사이의 연결선으로 구성된 간선집합(edge set) E로 정의되는 그래프 G = (V, E)에 기반을 둔다. 그리고 교통카드 트랜잭션 데이터베이스에서 트랜잭션(smart_card_trnx)을 하나씩 읽어서 버스 승객 트랜잭션을 제외하고 지하철 승객 트랜잭션(subway_trnx)의 승객이 하루 동안 이용한 모든 지하철 트랜잭션들을 묶어서 앞에서 설명한 탑승 패턴들 중의 하나로 분류한다. 분류가 끝나면 그 패턴에 관한 시간 정보를 추출하여 시간-히스토그램에 저장한다.

```
ClassifyTripPatterns()
{
  BuildSubwayNetwork();
  while (smart_card_trnx in TransactionDB) {
      if (smart_card_trnx == subway_trnx) {
        1. 카드번호가 동일하고 트랜잭션ID가 다른 트랜
           잭션들을 하나로 묶어서 한 승객의 하루 탑승
           패턴의 후보로 한다.
        2. 승객의 탑승 패턴을 표 2에 따라 11개의 탑승
           패턴들 중 하나로 분류한다.
        3. 첫 탑승과 마지막 탑승 트랜잭션의 출발시간과
           도착시간을 시간-히스토그램에 저장한다.
      }
  }
  WritePatternHistogram();
}
```

그림 1. 탑승 패턴 분류 알고리즘

IV. 실험 결과 및 분석

1. 입력 데이터 및 실험 환경

먼저 실제 교통카드 트랜잭션 데이터베이스와 그 시점의 서울 수도권 지하철에 대해 설명한다. 수도권 지하철 시스템의 기준 시점은 2005년 6월이고, 입력 데이터는 2005년 6월 24일 금요일 하루 동안에 지하철과 버스를 이용한 승객들의 교통카드 트랜잭션 데이터베이스이다. 기준 시점의 수도권 지하철 시스템은 지하철 1-8호선, 분당선, 인천1호선으로 구성된다. 기준 시점에서 지하철역들의 개수는 357개이고, 이 역들 사이의 간선들의 개수는 400개이다. 지하철역들 중에는 서로 다른 노선으로 환승할 수 있는 환승역들의 개수는 54개이다. 하루 동안 기록된 전체 트랜잭션의 개수는 10,667,518개이고, 이 중에서 5,737,479개는 버스 승객 트랜잭션들이고 4,930,039개는 지하철을 사용한 승객들의 트랜잭션들이다. 지하철 트랜잭션들 중에서 탑승 패턴 분류에 기여한 트랜잭션은 4,909,316개이고, 나머지 20,723개는 트랜잭션 내용 중에서 정확한 속성을 가지고 있지 않은 경우 등으로 탑승 패턴 분류에 사용할 수 없는 것으로 간주되었다. 지하철 트랜잭션들 4,909,316개에서 탑승 패턴으로 분류된 승객들의 숫자는 2,746,517명이다.

실험에 사용된 컴퓨터는 CPU(Intel i7 920 2.67GHz), 메인메모리(DDR3 1333MHz 12GB), 2개의 하드디스크(SAS 15Krpm 147GB)로 구성되었다. 운영체제는 마이크로소프트 회사의 Windows 7 Enterprise K 64 bit를 사용하였고, 프로그램 개발 환경은 MS Visual Studio 2008 C++ 언어를 사용하였다. 10,667,518개의 교통 카드 트랜잭션들을 처리하여 주어진 탑승 패턴으로 분류하는데 수행된 시간은 113.41초이며, 실행 시간은 입력 교통카드 트랜잭션 개수에 근사적으로 선형적인 특성을 보여준다.

2. 지하철 탑승 회수 비율과 패턴 분류 결과

[표 3]은 지하철 이용자 2,746,517명이 지하철을 탑승하는 회수에 따라 승객수와 비율을 나타낸 것이다. 11회 이상의 탑승회수의 승객수는 합해서 48명으로 전체

승객수에 비해 아주 미미하여 제외하였다. 가장 큰 비율이 하루 두 번 탑승하는 승객들로 약 50%임을 보여준다. 대부분의 지하철 승객들이 하루에 지하철을 한 번 또는 두 번 탑승하고 있고, 그런 승객들의 비율은 88%가 된다. 본 논문의 III절에서 지하철 승객들의 탑승 유형을 추정할 때 주로 한 번 탑승하는 경우와 두 번 탑승하는 경우를 분석하여 탑승 패턴들을 결정하였는데 그 실증적인 근거가 [표 3]에 의해서 나타나고 있다.

표 3. 지하철 탑승 회수에 따른 승객 비율

탑승회수	승객수	비율(%)
1	1,032,187	37.5817
2	1,386,643	50.4873
3	236,197	8.5998
4	70,292	2.5593
5	15,623	0.5688
6	3,986	0.1451
7	1,053	0.0383
8	337	0.0122
9	109	0.0039
10	42	0.0015

[표 4]는 [표 2]의 지하철 탑승 패턴을 주어진 입력 데이터로 분류된 결과를 보여준다. 결과의 정확성을 검증하면, [표 3]의 탑승회수 1회 승객수 1,032,187명은 [표 4]의 패턴 1과 9의 합과 같고, 표 3의 탑승회수 2회 승객수는 [표 4]의 패턴 2, 3, 4, 5, 10의 승객수의 합과 같음을 보여주어 패턴 분류가 정확하게 이루어졌음을 보여준다. 이 표에 의하면 약 37%의 승객들이 패턴 1에 속하는 한 번만 지하철을 타고 내리는 방법으로 지하철을 이용하고 있음을 보여준다. 그 다음으로 많은 승객들이 패턴 2에 해당하는 것으로 지하철을 왕복으로 어떤 지점에 갔다 오는 것으로 볼 수 있다. [표 4]에서 패턴 1에서 8까지 이용한 승객들의 비율은 97.4%로 대부분의 지하철 승객들이 분석하고자 하였던 주요 패턴들에 속함을 알 수 있다.

표 4. 지하철 탑승 패턴별 승객수

탑승패턴	승객수	비율(%)	본문설명
pattern 1	1,025,057	37.3221	case 11
pattern 2	748,702	27.2601	case 21
pattern 3	349,144	12.7122	case 22
pattern 4	151,010	5.4982	case 23
pattern 5	123,708	4.5042	case 24
pattern 6	23,030	0.8385	case 31
pattern 7	135,752	4.9427	case 32
pattern 8	119,370	4.3462	case 33
pattern 9	7,130	0.2596	case 12
pattern 10	14,079	0.5126	case 25
pattern 11	49,535	1.8036	case 34

[그림 2]는 [표 4]에서 가장 많은 이용 비율을 보이는 패턴 1(case 11: a ⇒ b)에 해당되는 승객들의 승차시간과 하차시간을 보여주고 있다. 출근시간대에 승차와 곧 뒤따르는 하차 승객들이 각각 약 9만 명 내외임을 보여주고, 오후 7시경의 퇴근 시간대에 승•하차 승객이 집중됨을 보여주고 있다.

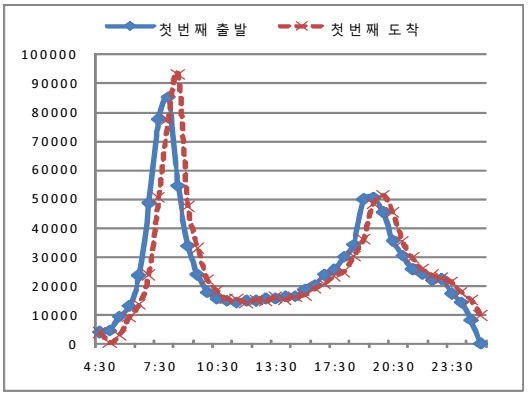

그림 2. pattern 1의 시간에 따른 승객수

3. 왕복-탑승 패턴 분석

[표 5]는 처음 출발했던 지하철역으로 되돌아오는 여부에 따른 이용자 수와 비율을 보여주고 있다. Round-trip 유형은 [표 4]의 탑승 패턴 2, 4, 6, 7에 해당되는 승객들로 그 비율이 약 39%임을 보여주고 있으며, 나머지 승객들은 버스나 택시 등의 다른 교통수단을 이용하여 되돌아가거나 단일 트랜잭션으로 지하철을 탑승하고 있음을 알 수 있다.

표 5. Return type에 따른 승객 비율

RETURN-TYPE	승객수	비율(%)	패턴 타입
One-way	1,617,279	58.8847	pattern 1,3,5,8
Round-trip	1,058,494	38.5395	pattern 2,4,6,7

4. 집-직장 통근 패턴 분석

[표 4]에서 지하철을 이용하여 집과 직장 사이를 통근하는 패턴으로 패턴 2(case 21; a ⇒ b, b ⇒ a)와 4(case 23: a ⇒ b, c ⇒ a)로 정의하였는데, 두 패턴들의 비율의 합은 32.76%로 지하철 승객들의 약 삼분의 일이 이 패턴들에 속한다. [그림 3]은 패턴 2의 시간대별 승차와 하차 승객수를 보여준다. 주로 출•퇴근을 하는 승객들의 대표적인 통근 패턴으로 첫 번째 지하철 승·하차시간은 출근 시간대인 8시경이 최고치이고, 두 번째 지하철 승차는 퇴근 시간대인 오후 7시경에 최고치를 보여준다. 첫 번째 탑승 지하철의 하차시간의 최대 승객수를 갖는 8시30분경과 두 번째 탑승 지하철의 승차시간의 최대 승객수는 오후 7시경이 되어 그 시간 차이는 10시간 30분 정도가 된다. [표 4]에서 세 번째로 많은 승객수를 갖는 패턴 3(case 22: a ⇒ b, b ⇒ c)의 승객들의 시간대별 승객수 분포도 [그림 3]과 비슷함을 실험결과에서 알 수 있었다.

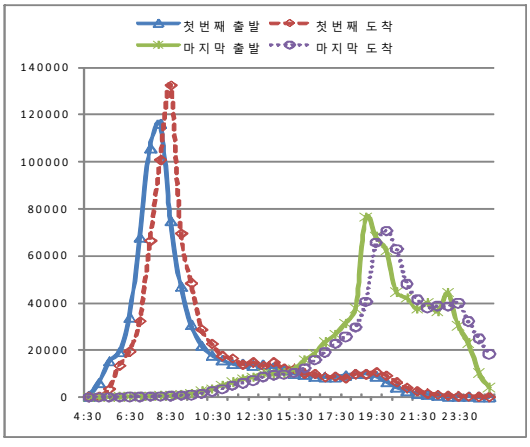

그림 3. pattern 2의 시간에 따른 승객수

[그림 2]와 [그림 3]에서 볼 수 있듯이 출근시간대는 집중되어 있는 반면에 퇴근 시간대는 넓게 퍼져 있어 퇴근시간이 일정치 않음을 알 수 있다.

[그림 4]는 패턴 2와 4에 속하는 승객들의 첫 번째 도착역의 하차시간과 두 번째 출발역의 승차시간 사이의 잔류시간의 분포를 보여주고 있다. 집-직장 통근 패턴을 지지하는 승객들 중에서 잔류시간 7시간 이상을 잔류하는 상근 직장인으로 추정되는 승객수는 626,675명이고 패턴 2와 4의 승객들의 70%에 해당된다. 하루 전체 승객 275만여명의 23%에 이르는 수치이다. 이런 패턴을 집-직장-상근 출퇴근 패턴으로 부를 수 있다. 잔류시간 분포에서 최대 승객수 65,397명은 10시간 30분간을 잔류하는 것으로 나타났다.

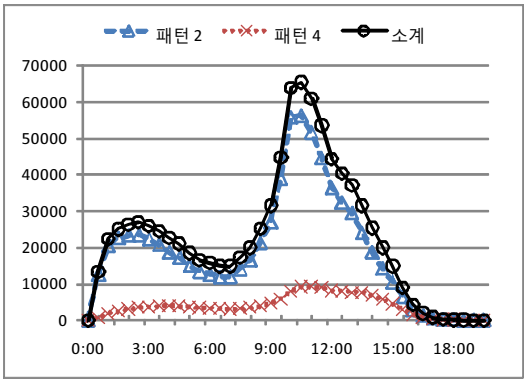

그림 4. pattern 2와 4의 잔류 시간에 따른 승객수

5. 흥미로운 패턴 분석

[표 2]에서 패턴 9와 10을 예기치 못한 흥미로운 패턴으로 고려하였는데 실험 결과 [표 4]에서는 패턴 10보다 패턴 9가 상대적으로 예상치 못한 큰 비율을 보여주어서 이에 대한 분석을 한다. 패턴 9(case 12: a ⇒ a)는 처음 출발역에서 다시 출발역으로 되돌아오는 단일 트랜잭션을 갖는 것으로, 이에 속하는 승객들이 7,130명이고 전체승객에 대한 비율이 0.26%나 된다. [그림 5]는 패턴 9에 속하는 승객들의 승차시간과 하차시간에 따른 승객수를 보여준다. 패턴 9가 잘못된 탑승을 했다고 가정하면, 오전 5시경에 최소값을 보였다가 점차적으로 오후 7시까지 많아지고 있음을 보여준다. 이는 오후에

주의력이 떨어져 실수를 많이 한다고 해석할 수 있다.

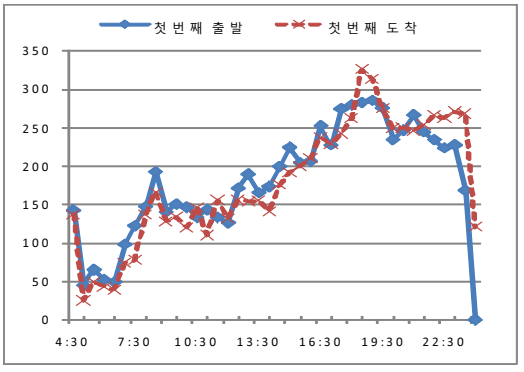

그림 5. pattern 9의 시간에 따른 승객수

[그림 6]은 패턴 9에 속하는 승객들의 승차시간과 하차시간 사이의 탑승시간을 보여준다. 이 그림에서 이 패턴의 7130명의 44%인 3,102명이 20분 내에 하차한 것으로 보아 실수가 개입된 것으로 추정할 수 있다.

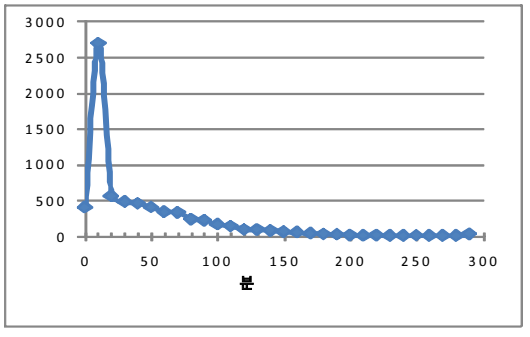

그림 6. pattern 9의 탑승 시간에 따른 승객수

V. 결 론

본 논문에서는 수도권 지하철 사용자들의 지하철 탑승 패턴을 연구하였다. 지하철 승객들의 탑승 회수를 바탕으로 하여 11가지의 탑승 패턴들로 분류하였다. 본 논문에서는 탑승 패턴을 분류하는 알고리즘을 제안하고 이를 구현하여 2005년도의 하루치의 교통카드 트랜잭션 데이터베이스에 적용하여 각 패턴에 속하는 지하철 승객들을 분류하고 해당 승객들의 승차시간과 하차시간에 대한 정보를 구하였다. 관심을 가질 수 있는 주요 분석 결과는 다음과 같다. 첫째, 처음 출발지로 되돌아오는 승객들의 왕복-탑승 패턴에 속하는 비율은 전체 승객의 39%에 이른다. 둘째, 집-직장 통근 패턴으로 되돌아오는 승객들 중에서 두 번째 탑승하기 전까지 7시간 이상 잔류하는 승객의 비율은 전체 승객의 23%에 해당되고, 이런 패턴은 집-직장-상근 출퇴근 패턴으로 볼 수 있다. 셋째, 중요하다고 예상하지 못한 흥미로운 패턴들 중에서 패턴 a ⇒ a는 처음 출발역으로 되돌아오는 단일 트랜잭션을 갖는 승객들의 숫자가 7,130명으로 전체 승객의 0.26%에 이른다는 것이다.

본 연구 결과로 지하철 대중교통 이용자들의 탑승 패턴을 이해할 수 있고 이를 기반으로 수도권 교통 정책을 입안할 때 중요한 참고 자료로 사용할 수 있다. 분류 패턴인 11가지 탑승 패턴은 지하철 승객들에 대한 분석이지만, 앞으로 연구과제는 수도권 대중교통 이용자인 버스 승객, 지하철 승객, 환승 승객들에 대한 전체 승객들의 탑승 패턴을 분류하고 흥미로운 패턴을 탐사하는 것이다.

참 고 문 헌

[1] 서울특별시>시정소식>보도자료(담당부서: 서울메트로), “서울메트로 2007년 상반기 수송실적(2007/08/07)”, “지난해, 지하철 얼마나 타고 얼마나 내렸나(2009/01/21)”, http://spp.seoul.go.kr/main/news/news_report.jsp

[2] P.-N. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining*, Pearson Addison Wesley, Boston, 2006.

[3] M.-S. Chen, J. S. Park and P. S. Yu, "Efficient data mining for path traversal patterns", IEEE Transactions on Knowledge and Data Engineering, Vol.10, No.2, pp.213-221, 1998.

[4] 이금숙, 박종수, “서울시 대중교통 이용자의 통행 패턴 분석”, 한국경제지리학회지 제9권, 제3호,

pp.379-395, 2006.

[5] 박종수, 이금숙, "대용량 교통카드 트랜잭션 데이터베이스에서 통행패턴 탐사와 통행행태 분석," 한국경제지리학회지 제10권, 제1호, pp.44-63, 2007.

[6] 박종수, 이금숙, "서울대도시권 지하철망의 구조적 특성분석," 한국경제지리학회지, 제11권, 제3호, pp.459-475, 2008.

[7] 임강원, 임용택, *교통망 분석론*, 서울대학교출판부, 2003.

[8] K. Lee, W.-S. Jung, J. S. Park, and M. Y. Choi, "Statistical analysis of the Metropolitan Seoul Subway System: Network structure and passenger flows," Physica A: Statistical Mechanics and its Applications, Vol.387, Iss.24, pp.6231-6234, 2008.

[9] 박종수, 이금숙, "서울 수도권 지하철 교통망에서 승객 흐름의 분석", 한국정보과학회 논문지: 컴퓨팅의 실제 및 레터, 제16권, 제3호, pp.316-323, 2010.

[10] 김호성, 박종수, 이금숙, "서울 수도권 지하철 교통망 승객 흐름의 시각화", 한국콘텐츠학회논문지, 제10권, 제4호, pp.397-405, 2010.

[11] 박종수, "대용량 교통카드 트랜잭션 데이터베이스에서 통근 패턴 탐사", 한국정보과학회 2010 한국컴퓨터종합학술대회 논문집, Vol.37, No.1(A), pp.38-39, 2010.

저 자 소 개

박 종 수(Jong Soo Park) 정회원

- 1981년 2월 : 부산대학교 전기기계공학과(학사)
- 1983년 2월 : 한국과학기술원 전기및전자공학과(석사)
- 1990년 2월 : 한국과학기술원 전기및전자공학과(박사)
- 1983년 ~ 1986년 : 국방부 군무설계기좌
- 1994년 ~ 1995년 : IBM Watson 연구소 객원 연구원
- 1990년 ~ 현재 : 성신여자대학교 IT학부

<관심분야> : 데이터베이스, 데이터마이닝, 교통지리

김 호 성(Ho Sung Kim) 종신회원

- 1982년 2월 : 한양대학교 전자공학과(공학사)
- 1984년 2월 : KAIST 전기및전자공학과(공학석사)
- 1988년 8월 : KAIST 전기및전자공학과(공학박사)
- 1987년 ~ 현재 : 성신여대 미디어커뮤니케이션학과
- 1993년 ~ 1994년 : 워싱턴대학교(시애틀)방문연구원
- 2000년 ~ 2004년 : 열린사이버대학교 학술정보처장
- 2004년 ~ 2005년 : 성신여대 정보통신처장

<관심분야> : 시각화, 영상처리, e-Learning

이 금 숙(Keumsook Lee) 정회원

- 1979년 2월 : 성신여자대학교 지리학과(학사)
- 1981년 2월 : 성신여자대학교 지리학과(석사)
- 1987년 5월 : Boston University 지리학과(박사)
- 1992년 ~ 현재 : 성신여자대학교 지리학과
- 1994년 ~ 1998년 : 성신여대 한국지리연구소 소장
- 2000년 ~ 2001년 : 워싱턴대학교(시애틀)방문연구원
- 2008년 ~ 2009년 : Boston University Center for Transportation Studies 방문연구원

<관심분야> : 교통지리, 경제지리, 입지분석, GIS