한국경제지리학회지 제10권 제2호 2006(44~63)

대용량 교통카드 트랜잭션 데이터베이스에서 통행 패턴 탐사와 통행 행태의 분석*

박종수**, 이금숙***

요약: 이 논문은 대용량의 교통카드 트랜잭션 데이터베이스에서 통행패턴을 찾아내는 데이터 마이닝 방법의 개발에 초점을 두었으며, 결과로 도출된 통행패턴의 공간적 특징과 시점 간 차이를 분석하였다. 특히 대용량 데이터베이스에서 요구하는 지식을 효과적으로 발굴해 내는 순회 패턴 탐사법을 원용하여 통행패턴분석에 적절한 데이터 마이닝 알고리즘을 개발하여 2004년 이후 2006년 까지 3개년의 하루 교통카드 자료에 적용하였다. 또한 통행 순차 데이터베이스에서 오전 출근 시간대, 낮 시간대, 저녁 퇴근 시간대의 출발 정류장과 도착 정류장에 대한 통행 수요를 산출하여 시간대별 통행패턴의 공간 특징을 분석하였다.

주요어: 데이터 마이닝, 순회패턴 탐사법, 통행패턴 분석, 알고리즘 개발, 시간대별 통행 수요, 공간구조.

1. 서론

정보 기술이 발달하면서 정보획득이 용이해져 다양한 부문에서 생성되는 정보자료의 양이 엄청나게 방대해 지게 되었다. 따라서 정보처리 부문에서는 이러한 방대한 자료를 효과적으로 처리하여 필요한 형태의 정보로 정리해 낼 수 있는 데이터 마이닝의 중요성을 인식하고 이에 대한 연구가 활발히 진행되고 있다. 특히 최근에는 Agrawal & Srikant (1995)에 의해 소개된 데이터베이스에서 요구하는 지식 발견(KDD; Knowledge Discovery in Databases)을 효과적으로 발굴해 내는 순회 패턴 탐사(Mining Traversal Patterns)에 대한 연구가 주목을 받고 있다(Han & Kamber 2006; Tan, et al., 2006).

서울을 중심으로 하는 수도권에서는 교통카드이용이 일반화되어 대중교통 이용자의 대부분이 교통카드를 이용하고 있어 서울 대중교통체계 상의 데이터 흐름(data stream)에 해당하는 교통카드의 거래 내역에 관한 대 용량의 데이터가 데이터베이스에 계속해서 저장되고 있는 상황이다. 현재 수도권 지역의 대중교통 이용자들이 움직이면서 생성하고 있는 교통카드데이터는 개개 통행자가 실제로 움직이는 궤적을 담은 통행 자료로서 하루에도 약 1000만 건에 달하는 교통 카드의 기록으로 구성된 트랜잭션 데이터베이스가 생성되고 있다. 이런 교통카드 데이터는 개개인의 통행에 대한 출발지점과 최종 목적지, 이용 교통수단, 환승에 대한 정확한 위치와 시간이 기록되어 있는 대용량 데이터베이스로서 도시 통행자의 통

* 이 논문은 2005년도 성신여자대학교 학술연구조성비 지원에 의하여 연구되었음.

** 성신여자대학교 컴퓨터정보학부 교수

*** 성신여자대학교 지리학과 교수

행행태를 고스란히 담은 귀중한 자료이다. 따라서 이 자료를 효과적으로 처리하면 수도권지역에서 실질적으로 움직이는 통행의 패턴에 대한 다양한 정보를 추출해 낼 수 있으며, 이는 다양한 교통정책이나 토지이용계획 및 시설계획에 귀중한 기초 자료를 제공할 수 있다.

이런 대용량 데이터베이스에서 숨겨져 있는 정보나 지식에 해당되는 여러 패턴이나 연관성을 찾아내는 것은 컴퓨터정보학에서 중요한 연구 토픽들 중의 한 분야이다. 특히 이를 지리정보체계의 데이터베이스와 결합하여 효과적으로 분석하면 수도권지역에서 지역 간의 실질적인 기능적 연계 및 도시의 공간구조를 분석 할 수 있고, 그 결과는 토지이용계획 및 시설계획 등 다양한 정책 수립에 귀중한 기초 자료를 제공할 수 있으므로 도시의 공간구조와 교통흐름을 연구하는 지리학과 다양한 교통관련 연구 분야에서 매우 관심을 가지고 있는 문제이다. 따라서 최근 교통카드 데이터를 기반으로 수도권 지역 대중교통 이용자의 통행패턴 분석은 교통 및 도시계획과 지리학 등 다양한 관련 학문 분야와 정책 기관에서 관심을 보이고 있으며 정보관련 분야에서도 많은 관심을 나타내고 있다. 이의 일환으로 저자들은 앞 서 컴퓨터정보학의 데이터 마이닝 기법을 적용하여 수도권 지역의 대중교통 이용자의 통행패턴 분석을 시도한 바 있다(이금숙 · 박종수, 2006). 그러나 선행연구는 이러한 분석의 가능성을 점검해보기 위한 초기 단계의 연구로서 많은 한계가 있었다.

본 연구의 목적은 선행 연구를 바탕으로 대용량 교통카드 데이터베이스에서 다양한 통행 행태를 효과적으로 탐사할 수 있도록 데이터 마이닝 알고리즘을 개발하고, 교통카드가 본격 도입된 이후 3년의 기간 동안 매년 하루치의 교통카드데이터베이스에 적용하여 얻어진 결과에 대해 통행행태와 통행패턴에 나타나는 다양한 특징을 비교분석하는 것이다. 특히 본 연구에서는 Agrawal & Srikant (1995)에 의해 소개된 후 후속 연구가 활발히 진행되고 있는 데이터베이스에서 요구하는 지식 발견을 효과적으로 발굴해 내는 순회 패턴 탐사법을 원용하여 (Park, et al. 1997; Chen et al. 1998; Han & Kamber 2006; Tan, et al., 2006) 수도권 교통카드 트랜잭션 데이터베이스에서 통행패턴을 탐사하는 저자들의 선행연구(이금숙 · 박종수, 2006)의 방법론을 좀 더 정교하게 체계화하여 통행패턴분석에 적절한 데이터 마이닝 방법을 개발하였다. 또한 하루 중 시간대에 따라 통행목적과 교통상황들이 달라질 수 있으므로 하루를 오전 출근 통행이 주를 이루는 출근 시간대, 그 밖의 업무들과 관련된 통행이 주로 나타나는 낮 시간대, 그리고 저녁에 퇴근 및 그 이후의 활동과 관련된 퇴근 시간대의 세 시점으로 구분하고, 각 시간대의 통행 수요의 공간적 구조를 비교 분석하였다.

본 논문은 제 2장에서 교통카드 트랜잭션 데이터베이스에서 각 승객이 통과한 승객 시퀀스를 찾아내는 방법에 대해 설명하고 이 과정에서 얻어지는 환승 횟수에 대한 결과도 설명한다. 3장에서 찾아내어진 승객들의 정류장이나 정거장들의 순차 데이터베이스에서 통행 패턴(trip pattern)을 찾는 방법을 설명한다. 4장에서 통행 순차 데이터베이스에서 출발 정류장과 도착 정류장에 대한 통계치를 찾아내어 대중교통 이용자들의 공간 특징을 시간대별로 설명하였다.

2. 통행 사슬 탐사과정 및 결과

이 장에서는 주어진 교통카드 트랜잭션 데이터베이스에서 승객들의 통행 사슬(trip chain)에 해당되는 승객 시퀀스(sequence) 데이터베이스를 만들어내는 과정을 설명하고 실험 결과를 분석하려 한다. 승객 시퀀스는 교통카드 트랜잭션에서 직접 찾아낼 수 있는 시퀀스가 아니고 서울 시내의 버스 노선들과 지하철 노선들의 정류장과 연결 정보가 있어야 추출해낼 수 있는 정보이다. 기본적으로 한 승객이 거쳐 지나가는 모든 정류장에 관한 정보는 승객들의 특성 분석과 노선 계획 등 교통 수요 예측에 필수적인 정보에 해당된다. 이전 연구(Lee & Park, 2005; 이금숙 · 박

표 1. 교통 카드 트랜잭션의 속성과 예제

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	366, 20041027103443, 009, 120, 0, 11110266, 6211번(신월동~상왕십리), 111007100, 중부운수주식회사, 111749763, 서울74사9763, 01, 일반, 20041027100520, 0009304, 연흥극장, 20041027104522, 0010010, 사육신묘, 1, 800, 0, 0, 0
트랜잭션 예제 2	1867, 20041027192720, 015, 200, 0, ,, 211100000, 한국철도공사, ,, 01, 일반, , 1006, 영등포, 20041027194230, 1803, 역곡, 1, 800, 0, 0, 0

종수, 2006)에 비해서 승객의 통행 사슬에 관한 개념적인 설명을 추가하였고, 환승역에 대하여 상세히 설명하고 있다.

서울시에서 주관하는 개선된 대중교통 시스템에서 대중교통 이용자들은 교통 카드를 사용하고 있다. 한 승객이 버스나 지하철을 이용하여 이동을 하면, 승차와 하차 시에 교통카드를 사용하여 요금을 지불하게 한다. 이와 같이 한 승객이 승차 시에 교통카드의 확인과 하차 시에 교통카드의 확인으로 한 트랜잭션을 구성한다. 그러면, 한국스마트카드(KSCC)에서 요금을 정산하고, 각 트랜잭션을 데이터베이스에서 관리하고 있다. 표 1은 트랜잭션의 속성의 이름들, 버스를 사용한 승객의 트랜잭션 내용, 그리고 지하철을 사용한 승객의 트랜잭션들의 일부분을 보여주고 있다. 승객에 대한 정보를 보호하기 위하여 카드번호로 일련번호를 사용하고 있다.

승객 시퀀스에 대한 설명을 위해 먼저 개념적인 대중 교통망을 설정하고, 승객들이 그 교통망을 사용하여 얻어지는 승객 시퀀스 트랜잭션의 구성에 대하여 설명한다(최계숙, 2006). 그렇게 되면, 승객 시퀀스의 의미를 이해하고, 그런 후에 다음 절에서 빈발 시퀀스를 구하는 기본적인 방법을 이 예제를 사용하여 설명한다. 〈그림 1〉은 서울 시내 대중 교통망의 특성을 설명하기 위하여 간단하게 버스 노선 2개와 지하철 노선 2개를 보여주고 있다. 버스 노선 2개는 각각 버스정류장ID의 집합으로 (101, 102, 103, 104, 105, 106)과 (201, 202, 103, 204, 205, 206, 207)을 가지고 있다. 지하철 노선 2개는 각각 지하철역ID의 집합으

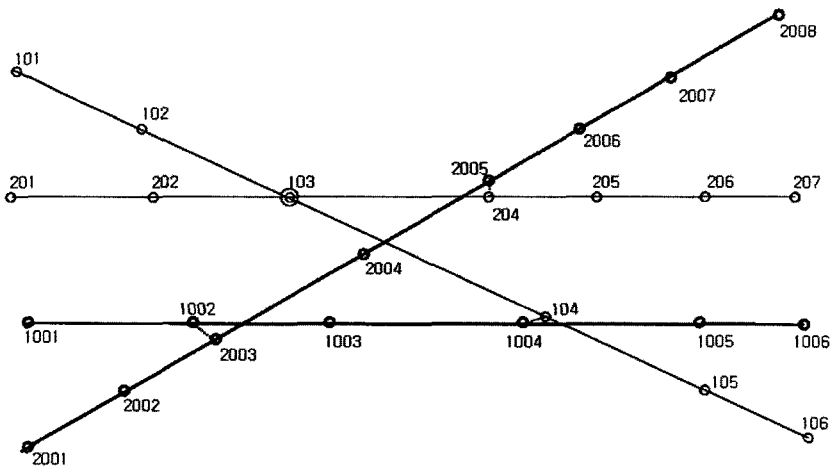

그림 1. 대중 교통망의 예제 노선(2개의 버스 노선과 2개의 지하철 노선)

표 2. 승객 시퀀스 트랜잭션 데이터베이스의 예제

CID	TID	환승회수	항목 개수	항목 리스트
1	1	0	4	102 103 104 105
1	2	1	6	101 102 103 204 205 206
2	1	0	4	202 103 204 205
2	2	1	6	201 202 103 204 2005 2004
3	1	1	7	10001 1002 1003 1004 104 105 106
3	2	1	7	2003 2004 2005 204 205 206 207
3	3	0	5	1002 1003 1004 1005 1006
4	1	2	7	2002 2003 1002 1003 1004 104 105
4	2	2	7	105 104 1004 1003 1002 2003 2002
5	1	2	7	1005 1004 104 103 204 205 206

로 (1001, 1002, 1003, 1004, 1005, 1006)과 (2001, 2002, 2003, 2004, 2005, 2006, 2007, 2008)을 가지고 있다. 서울 시내 대중 교통망에서와 같이 이 그림에서는 세 종류의 환승 정류장이 있다; 1) 버스 ↔ 버스 환승 정류장(정류장ID는 103번), 2) 지하철↔지하철 환승역(그림에서 지하철역ID 1002번과 2003번), 3) 마지막으로, 버스 ↔ 지하철 사이를 환승하는 버스 정류장과 지하철역(그림에서 두 곳: 204번과 2005번, 104번과 1004번). 두 번째 종류의 환승역은 지하철 안내도에 잘 정의되어 있지만, 첫 번째와 세 번째 종류의 환승 정류장은 명확히 정의되어 있지 않은 상태이다. 두 번째 환승역인 지하철역에서 지하철역을 갈아타는 시간은 8분의 이동시간으로 하여 최단 거리를 찾는 알고리즘에 적용하였다. 일일 교통카드 트랜잭션 데이터베이스를 이용하여 이런 종류들의 환승 데이터를 추출하여 분석하면 전체 교통망에서의 환승 정류장을 정의할 수 있고 이동 시간도 통계치로 얻을 수 있을 것이다.

〈그림 1〉의 노선을 사용한 결과로 얻어지는 〈표 2〉의 승객 시퀀스 트랜잭션 데이터베이스를 고려해 보자. 〈표 2〉는 10개의 승객 시퀀스 트랜잭션들을 포함하고 있고, 각 트랜잭션은 CID와 TID의 조합으로 식별할 수 있다. 전체 승객은 5명이고 각 승객은 트랜잭션ID(TID)에 따라 몇 개의 시퀀스들을 만들어내고 있다. 첫 번째 승객 시퀀스는 승객ID 1번이 트랜잭션 ID 1번으로 버스를 102번 정류장에서 승차하여 103번과 104번 버스 정류장을 거쳐서 105번 버스 정류장에서 하차하는 것을 나타내고 있다. 이 승객은 다른 버스나 지하철을 이용하지 않아서 환승회수가 0이 된다. 항목 개수는 이 승객이 출발지에서 목적지까지 도착하는 과정에 지나온 버스 정류장들의 개수인 4가 된다. 8번째 승객 시퀀스에 대해 설명하면 다음과 같다. 승객ID(CID) 4번은 지하철역 2002번에서 승차하여 2003번역에서 하차하여 환승역 1002번역으로 다른 지하철 열차를 승차하고 1004번역에서 하차한다. 그런 후에 버스 정류장 104번에서 버스를 승차하여 105번에서 하차한다. 표에서와 같이 2번 환승하며, 결과적으로 지하철 전철 2번과 버스 1번의 승/하차로 목적지에 도착하게 된다. 9번째 승객 시퀀스(CID=4, TID=2)는 같은 승객이 반대 방향으로 되돌아가는 과정의 정류장ID들을 보여주고 있다.

그러면,〈 표 1〉에서와 같은 트랜잭션이 10,000,000건 이상을 포함하는 대용량 트랜잭션 데이터베이스에서 승객 시퀀스를 구하는 것에 대해 고려해보자. 각 승객이 통과한 정류장 시퀀스에 대한 정보를 찾아내기 위해서는 주어진 트랜잭션의 데이터를 사용해야 하므로, 버스 승객은 버스 노선, 승차정류장, 그리고 하차정류장에 관한 정보를 추출한다. 지하철 승객

은 역시 승차역과 하차역의 정보를 추출하지만, 지하철 노선에 대한 정보는 이 트랜잭션 자료에서는 주어지지 않아서 이 논문에서 최단 거리 알고리즘에 의해 계산해낸다. 트랜잭션을 표현하는 형식은 필요에 따라 바뀔 수 있다. 〈표 1〉의 표현 형식은 2004년 10월 기준이고, 2006년도의 트랜잭션 형식에서는 승객의 이동 거리도 포함되어 있다. 승객 시퀀스를 구하는 과정에서 다음 세 가지 경우를 고려해야 한다.

1) 버스 승객

버스를 이용한 승객이 승차한 버스가 통과한 정류장ID들을 구하는 방법은 다음과 같다. 먼저 버스 노선에 대한 정보인 버스노선ID와 노선에 속한 정류장ID들을 배열과 해시 테이블을 사용하여 저장한다. 전체 버스 노선의 개수는 대략 800여개이고 전체 버스 정류장들의 개수는 15,000여개이다. 교통카드의 트랜잭션 중에서 버스를 사용한 승객의 트랜잭션에서 버스노선ID를 찾아내고, 그리고 그 노선 중에서 승차 정류장ID와 하차정류장ID가 둘 다 있으면 통과한 정류장들을 찾아낸다. 그러면, 이 승객이 버스를 승차하고 하차할 때 까지 통과한 정류장들로 하나의 통행 사슬(trip chain)이 만들어진다.

2) 지하철 승객

지하철을 사용하는 승객의 교통카드 트랜잭션에서 교통수단CD(code)는 200으로 한 숫자로 주어진다. 이런 승객의 통행 사슬을 만들기 위해서 승차역과 하차욕 사이에 통과하는 지하철역의 ID를 찾아내어야 한다. 그런데, 이 교통카드의 트랜잭션에서는 중간에서 다른 지하철 노선을 사용하는 환승 지하철역에 관한 정보는 주어지지 않는다. 이를 해결하기 위하여 9개의 지하철 노선과 400여개의 지하철역을 하나의 그래프로 간주하여 승차역과 하차역 사이의 최단 거리를 찾는 알고리즘을 적용하여 승객이 통과하는 지하철역을 찾아낸다. 승객이 지하철 내에서 열차를 환승하는 경우, 지하철역 사이를 이동하는 시간은 8분으로 두었고, 그이외의 지하철역 사이의 이동 거리는 2내지 3분으로 설정하였다.

3) 환승 승객

한 승객이 버스나 지하철을 환승하여 몇 번 정도는 갈아타서 이동한 거리에 따라 요금을 책정하도록 하였다. 그러므로 교통카드의 트랜잭션들 중에서 동일한 카드번호와 트랜잭션ID를 갖는 트랜잭션들을 모아서 하나의 통행 사슬을 만들어낼 수 있다. 그러면, 이 승객은 출발지와 도착지 사이를 통과한 버스 정류장들이나 지하철역들의 ID를 갖는 통행 사슬을 만들게 된다. 교통 카드에 의해 처리된 일일 통행 거래 데이터베이스에서 카드번호와 트랜잭션 ID가 같은 트랜잭션들을 찾아서 환승횟수 만큼의 승차 정차역과 하차 정차역을 앞의 두 방법에 의하여 각 트랜잭션에 따라 버스 통행과 지하철 통행에 따른 통행 사슬을 먼저 구한다. 그런 후에 이 통행 사슬들을 연결하여 하나의 시퀀스로 만들면, 이 승객이 지나가는 정류장이나 지하철역의 ID들을 순서대로 나열하여 하나의 통행 사슬을 완성하게 된다. 승객이 버스와 지하철을 사용하여 네 번까지 환승할 수 있으므로 다섯 번의 버스를 갈아타거나 그 중에서 한 번은 지하철을 사용할 수 있다.

앞에서 설명한 단계들을 거친 후의 결과인 통과 정류장들의 승객 시퀀스의 예제는 〈표 1〉의 교통카드 트랜잭션에서 얻어진 결과로 〈표 3〉에서 보여주고 있다. 〈표 3〉에서 지하철 정거장의 ID는 버스 정류장

표 3. 표1의 결과인 승객 시퀀스의 속성과 예제

속성(attribute)	카드번호, 트랜잭션ID, 환승횟수, 통과정류장개수, 승차정류장ID, 통과 정류장 ID들, 하차정류장ID
예제 1	366, 9, 0, 11, 9304, 9388, 9461, 9507, 9603, 9647, 9757, 9829, 9896, 9950, 10010
예제 2	1867, 15, 0, 8, 2001006, 2001999, 2001701, 2001813, 2001801, 200180, 2001821, 2001803

표 4. 교통카드 트랜잭션에서 구한 승객 시퀀스에 대한 출력 요약

	2004/10/27	2005/6/24	2006/5/17
버스노선개수	773	829	818
지하철역 개수	401	409	421
교통카드 트랜잭션개수	10,088,158	10,667,519	11,364,178
버스 승객	4,654,747	5,078,718	5,165,516
지하철 승객	4,819,015	4,909,316	5,297,030
처리된 트랜잭션 개수	9,473,762	9,988,034	10,462,546
미처리된 트랜잭션개수	614,396	679,485	901,632
출력된 승객 시퀀스 개수	7,463,638	7,651,578	7,832,979
승객 시퀀스에서 평균 정류장 개수	13.62	13.83	14.17
승객 시퀀스에서 최대 정류장 개수	245	220	210
승객 시퀀스에서 평균 환승횟수	0.2689	0.3054	0.3106

ID와 구별하기 위해서 원래 ID 값에다 2,000,000을 더해서 얻은 결과다.

〈표 4〉는 3년에 걸쳐 각 하루 동안 교통카드를 사용한 전체 승객들의 트랜잭션들을 입력으로 하여 얻어진 자료다. 2004년 10월 27일, 2005년 6월 24일, 2006년 5월 17일에 수집된 교통카드 트랜잭션 데이터베이스를 이용하여 통행 사슬에 해당하는 승객 시퀀스를 추출해내었다. 버스노선 개수는 교통카드 트랜잭션 데이터베이스를 얻는 시점에서의 전체 버스 노선들의 개수를 나타내고 있지만, 2006년에 대한 버스 노선 데이터는 온전히 자료를 수집하지 못하여 완전한 분석을 할 수는 없지만 그래도 일정한 경향을 분석하는데 도움을 줄 수 있기 때문에 자료로 포함하였다. 이에 따른 문제로 승객 시퀀스의 개수가 줄어들 가능성이 있다. 서울시는 버스 노선에 대한 개편 작업을 지속적으로 하고 있다. 지하철역의 개수도 그 시점에서 통행할 수 있는 역들의 개수이다. 하루에 처리되는 교통 카드 트랜잭션들의 개수는 약 일천만을 넘어서고 있고 약 1.7GB의 크기로 대용량 데이터베이스에 해당된다. 얻어진 승객들의 시퀀스 개수가 주어진 입력 교통카드 트랜잭션 카드보다 작은 것은 한 승객이 환승을 통해 몇 개의 교통카드 트랜잭션을 발생하기 때문이다. 미처리된 트랜잭션들의 개수도 전체의 약 6% 정도가 되는데, 그 원인은 주로 트랜잭션에 포함된 승하차 정류장ID가 그 버스노선ID에 포함되어 있지 않는 경우가 대부분을 차지한다. 이것은 승하차시에 확인하는 무선 시스템의 오류일 수도 있다.

서울 대중교통체계의 개편 후 3년간의 데이터 분석에서 어느 정도 대중교통 이용자들의 변화를 분석할 수 있다. 무엇보다 중요한 것은 사용하는 승객의 숫자가 늘어나고 있다는 것이다. 통행 사슬에서 평균 정류장들의 개수가 늘어나고, 그리고 환승하는 횟수도 조금씩 늘어나고 있음을 〈표 4〉에서 보여주고 있다. 〈그림 2〉는 승객들이 환승하는 횟수에 대한 분포를 보여주고 있다(Sen, et al., 2003). 대부분의 승객들은 환승을 하지 않고 있으나, 점차적으로 많은 승객들이 환승을 하여 원하는 도착지로 가고 있음을 보여주고 있다. 연도별로 환승하는 승객들의 퍼센티지는 20.9514%, 22.8673%, 23.3444%로 점차 증가하고 있다. 이것은 서울 대중교통체계를 더 많이 이해하여 효과적으로 그 시스템을 사용하는 것으로 해석할 수 있다. 여기서 구한 환승횟수에 대한 데이터에서 지하철을 사용하는 경우에 서로 다른 지하철 노선 사이를 환승하는 경우는 고려되지 않고 있다. 그 이유는 교통카드 트랜잭션 데이터에서 모든 지하철역들을 하

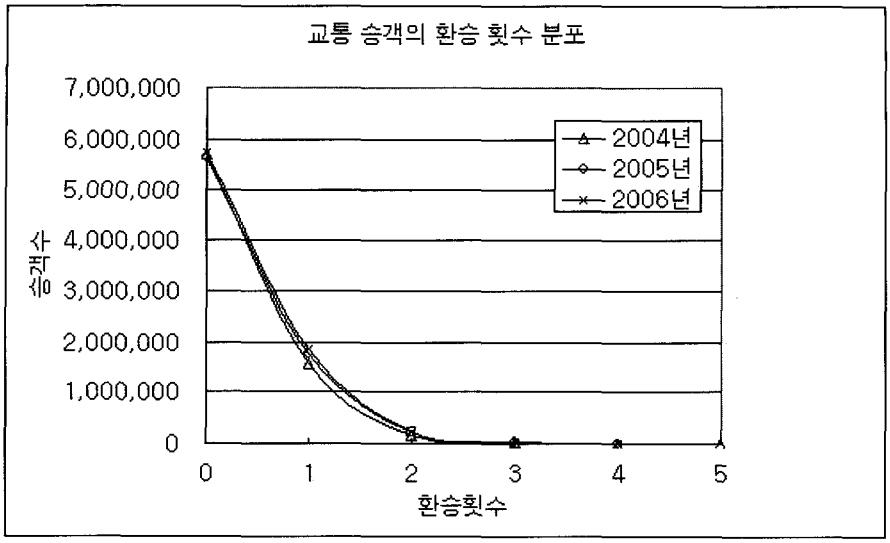

그림 2. 교통 승객의 환승 횟수 분포

나의 노선으로 간주하여 단지 승차역과 하차역에 관한 정보만을 제공하기 때문이다. 추후 연구과제로 승차역과 하차역 사이의 통행 시퀀스를 계산하면서 환승하는 지하철역을 찾아내어 통행 특성을 분석하는데 이용할 수 있을 것이다.

3. 통행 패턴 탐사 과정과 결과

트랜잭션 데이터베이스에서의 통행 패턴(trip pattern) 탐사 문제는 다음과 같이 정의할 수 있다(최계숙 2006; Han & Kamber, 2006; Tan, et al 2006). 버스 정류장이나 지하철역에 해당되는 객체(object)들은 서로 연결되어 있으며, 객체들 사이의 접근은 서로 연결된 경로를 따라 일정한 방향성을 가지며 접근한다. 이러한 환경 하에서 객체들을 순차적으로 접근하는 일정한 패턴을 발견하는 것을 통행 패턴 탐사라고 한다. 통행 패턴 탐사 문제는 순차 패턴(sequential pattern) 탐사(Han & Kamber, 2006)의 한 특별한 경우로써, 웹 환경 하에서의 순회 패턴(traversal pattern) 탐사 문제를 다룬 FS(Chen, et al, 1998)와 그 차이가 있다. FS의 경우 객체(웹 페이지)의 접근에 있어 역방향 접근의 문제를 다루는 반면, 본 논문에서 정의한 통행 패턴 탐사의 경우 객체인 정류장을 접근하는데 일정한 방향성을 가진 버스 노선이나 지하철 노선이 있다. 통행 패턴을 탐사하는 문제는 다음과 같이 표현될 수 있다. 본 논문의 통행 패턴 탐사 문제에서 각 항목은 한 개의 버스정류장 또는 지하철역을 나타낸다. 은 서로 다른 개의 항목들의 집합이다. 한 시퀀스(sequence)는 항목들의 순서 유지 리스트이다. 한 시퀀스 a는 $a=\{a_1 \rightarrow a_2 \rightarrow \cdots \rightarrow a_n\}$으로 나타낸다. 기존의 순차 패턴 탐사의 경우 a_i는 항목 집합이 될 수 있었지만, 통행 패턴 탐사의 경우 반드시 하나의 항목인 특정 정류장이 올 수 있다. k개의 항목을 가진 한 시퀀스를 $k-$시퀀스라고 부른다. 예를 들어, $(2 \rightarrow 1 \rightarrow 3)$는 $3-$시퀀스이다. i의 범위는 $1 \leq i \leq n$이고 $n \leq m$이고 범위가 $0 \leq K \leq m-n$을 만족하는 K가 상수이고 $1 \leq j \leq m$이고 $j=i+K$일 때 $\beta_j=\alpha_i$이면, 시퀀스 $\alpha=\{\alpha_1 \rightarrow \alpha_2 \rightarrow \cdots \rightarrow \alpha_i \rightarrow \cdots \rightarrow \alpha_n\}$는 다른 시퀀스 $\beta=\{\beta_1 \rightarrow \beta_2 \rightarrow \cdots \rightarrow \beta_j \rightarrow \cdots \rightarrow \beta_m\}$의 부분 시퀀

스이고 $\alpha \sqsubseteq \beta$로 표기한다. 또 한편으로, β는 α를 포함한다고 한다. 예를 들어, (4→ 5)는 (4→ 5→ 6)의 부분시퀀스이고, 즉, (4→ 5)⊑(4→ 5→ 6). 반면, (4→ 6)은 (4→ 5→ 6)의 부분시퀀스가 아니다. 즉, (4→ 6)⋢(4→ 5→ 6). 일반적인 순차 패턴의 정의에서는 (4→ 6)도 (4→ 5→ 6)의 부분시퀀스가 되지만, 통행 패턴 탐사의 경우는 항목의 순서 유지뿐만 아니라 연속적으로 발생해야 하기 때문에 부분시퀀스가 될 수 없다.

한 시퀀스의 지지도는 데이터베이스에서 그 시퀀스를 포함하는 트랜잭션들의 개수이다. 최소 지지도라고 불리는 사용자가 정의한 임계값이 주어졌을 때, 한 시퀀스가 최소 지지도보다 많이 발생한 경우를 빈발(frequent)하다고 한다. 빈발 k－시퀀스의 집합을 F_k로 표기하고, 후보 k－시퀀스의 집합을 C_k로 표기한다. 한 빈발 시퀀스가 다른 시퀀스의 부분시퀀스가 아닐 경우 최대 시퀀스(maximal sequence)라고 한다(Tan, et al, 2006). 통행 패턴 탐사는 이러한 빈발 시퀀스나 또는 최대 시퀀스를 찾는 것이다.

통행 패턴을 찾는 방법을 예제를 통하여 설명하고자 한다. 앞 절에서 〈표 2〉의 승객 시퀀스 트랜잭션 데이터베이스와 40%의 최소 지지도가 주어졌을 때, 통행 패턴에 해당되는 빈발 시퀀스를 찾아내는 과정을 〈표 5〉에서 〈표 8〉까지 설명하고 있다. 빈발 1－시퀀스인 F_1을 찾기 위하서는 각 시퀀스 트랜잭션을 읽어서 그것의 각 항목을 해시 테이블에 저장하면서 지지도를 증가시킨다. 후보 1－시퀀스인 C_1에는 〈표 2〉의 항목 리스트에 나오는 모든 항목들이 포함된다. C_1에서 지지도가 40이%인 4개 이상의 트랜잭션에 포함되는 항목이 F_1에 속하게 된다. 그 결과는 〈표 5〉에서 보여주고 있다. 빈발 2－시퀀스를 찾기 위해서는 먼저 후보 2－시퀀스인 C_2를 만드는데 통행 패턴에 해당하므로 $C_2 = F_1 \times F_1$에서 자기 자신과 같은 2－시퀀스를 제외한 42개의 후보 2－시퀀스가 만들어진다. 다시 〈표 2〉의 승객 시퀀스 트랜잭션들을 읽어서 각 2－시퀀스의 지지도를 계산하여, 지지도가 4이상인 2－시퀀스들을 찾아내면 〈표 6〉과 같이 된다. F_2에서 후보 3－시퀀스인 C_3를 결정하는 방법은 하나의 빈발 2－시퀀스의 두 번째 항목과 다른 하나의 2－시퀀스의 첫 번째 항목이 동일한 경우에만 두개의 2－시퀀스를 결합하여 한 개의 후보 3－시퀀스를 만든다. 〈표 6〉에서의 빈발 시퀀스에서 〈표 7〉의 후보 3－시퀀스를 만들어내고, 이를 다시 〈표 2〉의 트랜잭션들의 지지도를 계산한다. 그러면, 〈표 8〉에서와 같이 하나의 빈발 3-시퀀스를 얻게 되고, 후보 4－시퀀스를 만들 수가 없기 때문에 더 이상의 빈발 시퀀스를 찾을 수 없어서 중지한다. 결과적으로 〈표 5, 6, 8〉의 빈발 시퀀스인 통행 패턴을 찾아내게 된다.

표 5. 빈발 1－순차, F_1(최소지지도 : 40%)

빈발 1-시퀀스	지지도
103	5
104	5
204	5
205	4
1002	4
1003	4
1004	4

표 6. 빈발 2－순차, F_2

빈발 2-시퀀스	지지도
103 → 204	4
204 → 205	4

표 7. 후보 3－시퀀스, C_3

후보 3-시퀀스	지지도
103 → 204 → 205	4

표 8. 빈발 3－시퀀스, F_3

빈발 3-시퀀스	지지도
103 → 204 → 205	4

〈표 4〉에서 보여주는 교통카드 트랜잭션 데이터베이스에서 얻은 승객 시퀀스 트랜잭션 데이터베이스

를 입력으로 하여 통행 패턴을 찾아내는 것이 통행 패턴 탐사이다. 이를 위해 DHP(Park, et al, 1997)와 FS(Chen, et al, 1998) 알고리즘을 통행 패턴 탐사에 적용하여 새로운 알고리즘을 만들었다. 이 알고리즘의 이름은 DTP(Direct trimming for mining Trip Patterns)로 정하였다. DTP 알고리즘은 후보 시퀀스의 지지도 계산을 위해 트랜잭션 단위로 트랜잭션 데이터베이스를 읽어 들인다. 트랜잭션의 각 후보 $k-$시퀀스에 대해 지지도를 계산한 후, 다음 단계에 사용될 시퀀스만 데이터베이스에 기록한다. 다음 단계에 사용될 시퀀스의 조건은 후보 $k-$시퀀스를 포함해야 하고, 그 시퀀스의 길이가 $k+1$이어야 한다. 각 트랜잭션 중에서 다음 단계에서 빈발 항목집합의 구성원이 되지 못하는 항목들은 삭제되고, 구성원이 될 가능성이 있는 항목들만 데이터베이스에 기록하게 된다. 한 트랜잭션에서 남아있는 항목들의 개수가 보다 작으면, 다음 단계의 빈발 시퀀스를 구하는데 기여할 수 없으므로 그 트랜잭션은 삭제된다. 다음 단계에서 사용될 시퀀스만 데이터베이스에 기록할 경우, 그에 따른 데이터베이스의 사이즈는 점차 줄어들게 된다. 즉, DTP는 트랜잭션 데이터베이스의 크기를 줄이는데 있어서, 각 트랜잭션의 사이즈를 줄일 뿐만 아니라, 그 데이터베이스에서 트랜잭션의 개수도 줄인다. 각 단계별로 트랜잭션 데이터베이스의 사이즈가 점점 줄어들면, 후보 시퀀스의 지지도를 계산하기 위해 트랜잭션 데이터베이스를 스캔할 때 그만큼 디스크 입출력이 줄어들게 되므로 전체 트랜잭션 데이터베이스를 반복적으로 스캔하는 것보다 더 좋은 성능을 보이게 된다. 다음은 DTP 알고리즘의 각 단계를 상세히 설명하고 있다.

1) 단계 1 : 빈발 $1-$시퀀스(F_1)

다른 순차 패턴 탐사와 마찬가지로 트랜잭션 데이터베이스를 한번 스캔하는 것으로 F_1을 구한다. 즉 각 항목에 대해 해시 함수를 적용해서 대응되는 버킷에 해당 항목이 존재하면 같은 트랜잭션에서의 항목인지 중복을 확인 한 후, 중복 발생이 아닌 경우 그 지지도를 증가시킨다. 만약 대응되는 버킷에 해당 항목이 존재하지 않을 경우 해시 테이블에 항목을 추가한다. 모든 시퀀스들을 읽어서 각 항목의 지지도를 계산한다. 한 항목의 지지도가 최소 지지도보다 크거나 같으면, 그 항목은 빈발 $1-$시퀀스가 된다. 빈발 $1-$시퀀스들의 집합은 F_1이라 표기하고, 그에 속한 항목들의 개수는 $|F_1|$으로 표기한다.

2) 단계 2 : 빈발 $2-$시퀀스(F_2)

F_1의 각 항목을 인덱스로 사용하여 이차원 배열로 후보 $2-$시퀀스를 나열한 후 트랜잭션 데이터베이스를 스캔하여 지지도를 계산한다. 트랜잭션 데이터베이스를 스캔할 때 트랜잭션 단위로 스캔하며, 트랜잭션내의 각 후보 $2-$시퀀스에 대해 지지도 계산이 끝나면 다음 단계에 사용될 시퀀스만 데이터베이스에 기록한다. 트랜잭션 데이터베이스 스캔이 끝나면 각 후보 $2-$시퀀스에 대해 최소 지지도를 만족하는 후보 $2-$시퀀스만 F_2에 포함시킨다.

3) 단계 3 : 빈발 $k-$시퀀스(F_k)

F_2가 구해지면 그 다음부터는 더 이상 후보 시퀀스나 빈발 시퀀스가 발견되지 않을 때까지 반복 수행한다. 후보 $k-$시퀀스의 생성은 주어진 F_{k-1}에서 두 개의 시퀀스 중 한 시퀀스의 첫 번째 항목과 나머지 다른 시퀀스의 마지막 항목을 떨어뜨린 후, 남아있는 $(k-2)-$시퀀스가 동일한 경우 두 빈발 $(k-1)-$시퀀스를 조인시켜서 후보 $k-$시퀀스를 생성한다. 후보 $k-$시퀀스의 지지도 계산은 이전 단계에서 축소된 데이터베이스를 트랜잭션 단위로 읽어 들여, 각 k 길이의 시퀀스가 후보 $k-$시퀀스인지 확인해서 후보 시퀀스인 경우 중복 발생인지 확인한다. 중복 발생이 아닌 경우 그 지지도를 증가시키고, 다음 단계에 사용될 $(k+1)-$시퀀스인지를 계산하여 다음 단계에 사용될 $(k+1)-$시퀀스만 데이터베이스에 기록한다. 트랜잭션 데이터베이스 스캔이 끝나면 각 후보 $k-$시퀀스에 대해 최소 지지도를 만족하는 후보 $k-$시퀀스만 F_k에 포함시킨다.

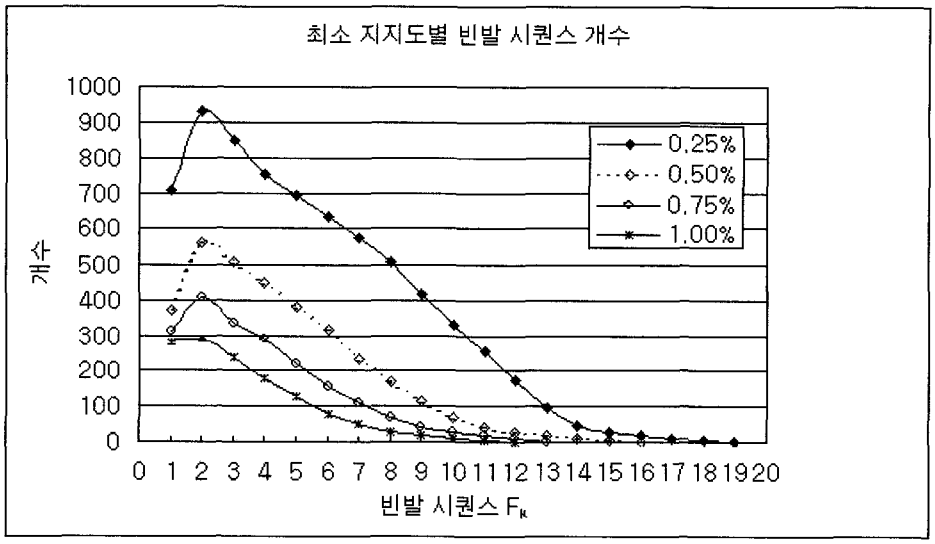

그림 3. 최소 지지도의 변하는 빈발 시퀀스 개수의 비교(2004년 데이터)

〈그림 3〉은 앞 절에서 얻은 출력 승객 시퀀스 트랜잭션들을 입력으로 하여 DTP 알고리즘을 적용하여 얻은 결과로 얻어진 통행 패턴에 해당하는 빈발 시퀀스들의 개수를 최소 지지도에 따라 보여주고 있다. 〈표 4〉에서 2004년도 승객 시퀀스 7,463,648개 중에서 최소 지지도가 주어진 퍼센티지 이상을 갖는 시퀀스들의 개수이다. $k_{\max}$는 빈발 시퀀스들 중에서 가장 길이가 긴 시퀀스의 항목들의 개수이다. 즉, 최대 빈발 시퀀스는 $F_{k_{\max}}$이고 빈발 $k_{\max}$－시퀀스의 집합이다. 〈그림 3〉에서 최소 지지도가 0.25%, 0.5%, 0.75%, 1.0%일 때, $k_{\max}$의 값은 각각 19, 16, 13, 12로 되어 최소 지지도가 높아질수록 최대 빈발 시퀀스의 길이는 짧아지고 있다. 최소 지지도별 빈발 시퀀스들의 전체 개수인 $\sum_{k=1}^{k_{\max}}|F_k|$는 각각 7032, 3264, 1978, 1298개로 나타내고 있다. 〈그림 3〉에서 F_2에 속한 빈발 2－시퀀스들의 개수가 가장 많고, 그 다음 단계의 빈발 시퀀스들의 개수는 점차 줄어들고 있다. 일반적으로 빈발 패턴 탐사 알고리즘에서 F_2를 효율적으로 얻을 수 있는 방법 연구가 주요 연구 초점이 되고 있다.

〈그림 4〉는 〈그림 3〉의 결과를 얻는데 실행된 시간을 초 단위로 보여주고 있다. 이 실험에 사용된 컴퓨터는 인텔 제온 3GHz CPU 두개, 8GB 메인 메모리, 맥스터 Atlas 15K 36GB SCSI 하드디스크, RHEL(Red Hat Enterprise Linux) WS4 OS로 구성되어 있고, 그것에서 DTP 알고리즘을 C++언어로 프로그래밍하여 교통카드 트랜잭션 데이터베이스를 입력으로 하여 실행하여 얻은 결과이다. 〈그림 3〉에서 얻어지는 빈발 시퀀스들의 개수가 작아지면 실행시간도 작아짐을 알 수 있다.

〈표 9〉는 연도별로 각 버스 정류장이나 지하철역의 지지도를 계산하여 상위 순서에 있는 정류장들을 보여주고 있다. 상위권은 대부분 지하철역들이 속하고, 300위 너머에서 버스 정류장들의 상위 6개를 표에서 나타내고 있다. 지하철역은 양 방향으로 오고 가는 승객들의 숫자를 나타내고, 버스 정류장은 한 방향을 나타내고 있다. 버스 정류장의 승객 숫자를 양 방향의 지지도를 합한 결과가 〈표 9〉의 마지막 행에 한 정류장에 대해서 보여주고 있다. 2004년 돈암동사거리 정류장의 양 방향의 승객수의 합은 97,661명으로 전체 순위는 243위에 해당된다.

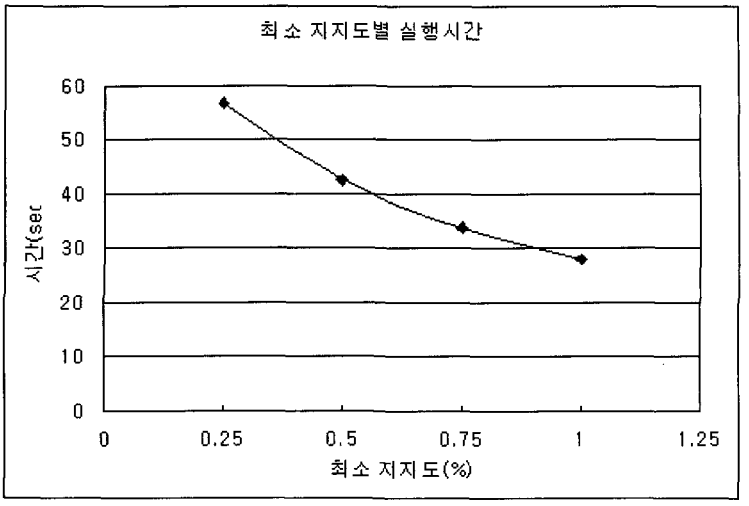

그림 4. 최소 지지도의 변화에 따른 실행시간의 비교

표 9. 연도별 상위 지지도를 갖는 지하철역명과 버스정류장명

2004년			2005년			2006년		
순위	정류장명	지지도	순위	정류장명	지지도	순위	정류장명	지지도
1	구로	556,187	1	구로	539,652	1	구로	594,691
2	신도림	550,076	2	신도림	535,734	2	신도림	561,946
3	교대	550,255	3	교대	526,911	3	교대	546,173
4	강남	473,652	4	강남	509,235	4	사당	533,133
5	동대문운동장	466,941	5	사당	467,852	5	강남	509,517
332	돈암동사거리 (성신여대입구, 미아리고개방향)	49,888	315	돈암동사거리 (성신여대입구, 미아리고개방향)	58,416	335	돈암동사거리 (성신여대입구, 미아리고개방향)	55,599
333	미아리고개 (삼선교 방향)	49,416	317	미아리고개 (삼선교 방향)	56,789	339	삼선교(한성대입구, 미아리고개 방향)	54,644
335	인공폭포, 수원지(당산역 방향)	48,338	318	미아리고개 (미아역 방향)	55,843	340	인공폭포, 수원지 (마포구청 방향)	54,557
337	돈암동사거리 (성신여대입구, 삼선교 방향)	47,773	319	삼선교 (한성대입구, 미아리고개 방향)	55,819	341	종로5가 (종로6가 방향)	54,479
339	미아리고개 (미아역 방향)	46,964	321	돈암동사거리 (성신여대입구, 삼선교방향)	55,622	345	종로2가 (광교 방향)	53,682
341	종로 2가 (동대문 방향)	46, 279	323	종로 2가 (종로 6가 방향)	54,919	347	미아리고개 (삼선교 방향)	53,090
243	*돈암동 사거리 (성신여대입구)*	*97,661*	*215*	*돈암동 사거리 (성신여대입구)*	*114,038*	*241*	*돈암동 사거리 (성신여대입구)*	*108,356*

표 10. 최대 지지도를 갖는 가장 긴 빈발 시퀀스의 예제(최소 지지도 = 1%)

년도	최대지지도를 갖는 가장 긴 빈발 시퀀스	지지도
2004	부평→부개→송내→중동→부천→소사→역곡→온수→오류동→개봉→구일→구로 (F_{12})	75,098
2005	송내→중동→부천→소사→역곡→온수→오류동→개봉→구일→구로→신도림(F_{11})	85,914
2006	부평→부개→송내→중동→부천→소사→역곡→온수→오류동→개봉→구일→구로(F_{12})	81,977

〈표 10〉은 최소 지지도를 1%로 주어졌을 때 연도별로 최대 지지도를 갖는 가장 긴 빈발 시퀀스를 표시하였다. 연도에 관계없이 거의 같은 구간에서 최대 빈발 시퀀스가 발생함을 알 수 있다.

4. 시점별 통행 수요와 통행 패턴에 나타나는 공간적 특징

통행(trip)은 사람들이 어떤 특별한 목적을 가지고 한 지점에서 목적지까지 이동하는 이동의 기본 단위를 일컫는 것으로 통근 · 통학, 업무, 구매 및 개인 용무, 사교 및 오락 등 통행 목적과 개개인의 사회 · 경제적 차이에 따라 통행행태에 차이를 보일 수 있으며 (허우긍, 1993), 그 지역의 토지이용패턴과 함께 주어진 교통망과 교통정책 등의 교통체계에 따라 영향을 받는다 (이금숙 · 박종수, 2006). 따라서 통행수요 및 통행패턴의 공간적 구조를 밝히는 작업은 한 도시의 기능적 공간구조를 분석하는데 필수적이며, 이러한 결과는 지역의 교통정책을 수립하거나 시설물의 입지계획에 의미있는 자료로 활용될 수 있기 때문에 교통관련 연구 분야에서는 일찍부터 이에 대한 관심이 높았으나 통행에 대한 자료의 한계로 표본 집단에 대한 설문조사를 실시하거나 지역의 속성을 나타내는 지리적 변수들을 이용한 모형을 개발하여 적용해 왔다 (Stern & Richardson 2005; Badoe & Chen 2004; Srinivasan & Ferreiva, 2003; Kitamura, et al., 1990; Pas & Koppleman 1987; Burnett & Thrift 1979; Marble, 1967).

그러나 서울에서는 2004년 대중교통체계 개편과 함께 정보 기술을 도입한 교통카드 사용이 활성화되어 대중교통 이용자들에 대한 통행 자료는 거의 전수에 가까운 자료를 확보하게 되었다. 따라서 이 자료를 효과적으로 처리하면 수도권지역에서 실질적으로 움직이는 통행의 패턴에 대한 다양한 정보를 추출해 낼 수 있다. 본 장에서는 대용량 데이터베이스에서

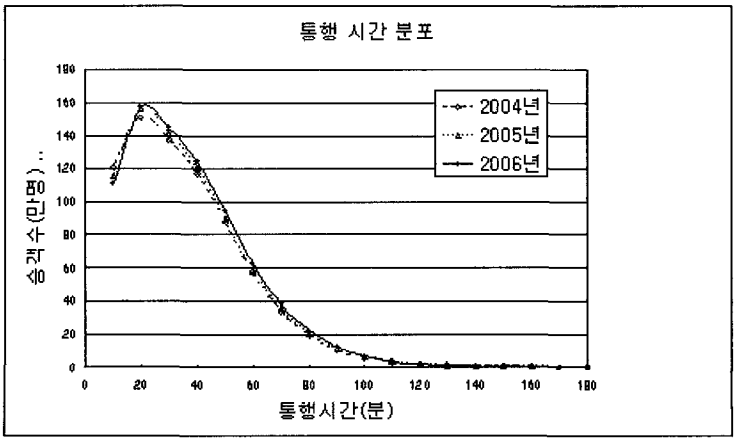

그림 5. 수도권 대중 교통 이용자의 통행 시간 분포

요구하는 지식을 효과적으로 발굴해 내는 순회 패턴 탐사법을 원용하여 개발된 통행패턴분석에 적절한 데이터 마이닝 알고리즘을 적용하여 얻어진 결과를 바탕으로 수도권 대중교통이용자의 통행행태와 통행수요 및 통행패턴의 공간적 특성을 분석하고자한다.

수도권 대중 교통 이용자들의 1회 통행에 소비하는 통행 시간을 분석해 보면 다음 〈그림 5〉와 같이 정리된다.

수도권 대중교통 이용자의 1회 통행에 소비하는 시간은 10분 이상 20분 이내에서 가장 높은 빈도를 보이며, 30분 이내가 전체 통행의 1/2 이상을 차지하고, 1시간 이내까지 확장하면 전체 통행의 거의 90% 정도를 차지하는 빈도분포를 보인다. 그리고 2004년 이후 수도권 대중교통 이용자가 1회 통행에 소비하는 시간으로 10분 이내는 그 절대량에서 감소하는 반면, 20분 이상 소요하는 통행자는 증가하고 있다. 특히 1시간 이상부터는 같은 통행시간대 2004년 통행수에 대한 2006년 통행수의 증가 비율이 10%를 넘게 되고, 이러한 현상은 통행시간이 길어질수록 그 증가율이 점차 더 크게 나타나고 있다. 결과적으로 총 통행에 대한 평균 통행 시간이 2004년 10월 31분 5초에서 2005년 6월에는 31분 54초, 그리고 2006년 5월에는 32분 38초로 점차 1회 통행에 걸리는 시간이 증가하고 있다.

통행 출발지와 도착지의 공간적 분포패턴은 대중교통에 대한 계획 수립에 기초가 되는 요소이다. 그러나 통행목적에 따라 통행 출발지와 도착지의 공간적 분포와 통행흐름의 방향은 각기 관련된 토지이용 패턴과 연관을 가지게 되므로(Sanchez, 2004) 시간대에 따라 통행패턴의 공간적 분포에 차이가 있을 것으로 예상된다. 본 연구에서는 하루를 출근과 관련된 통행이 주를 이루는 오전시간대와 업무나 기타 볼일

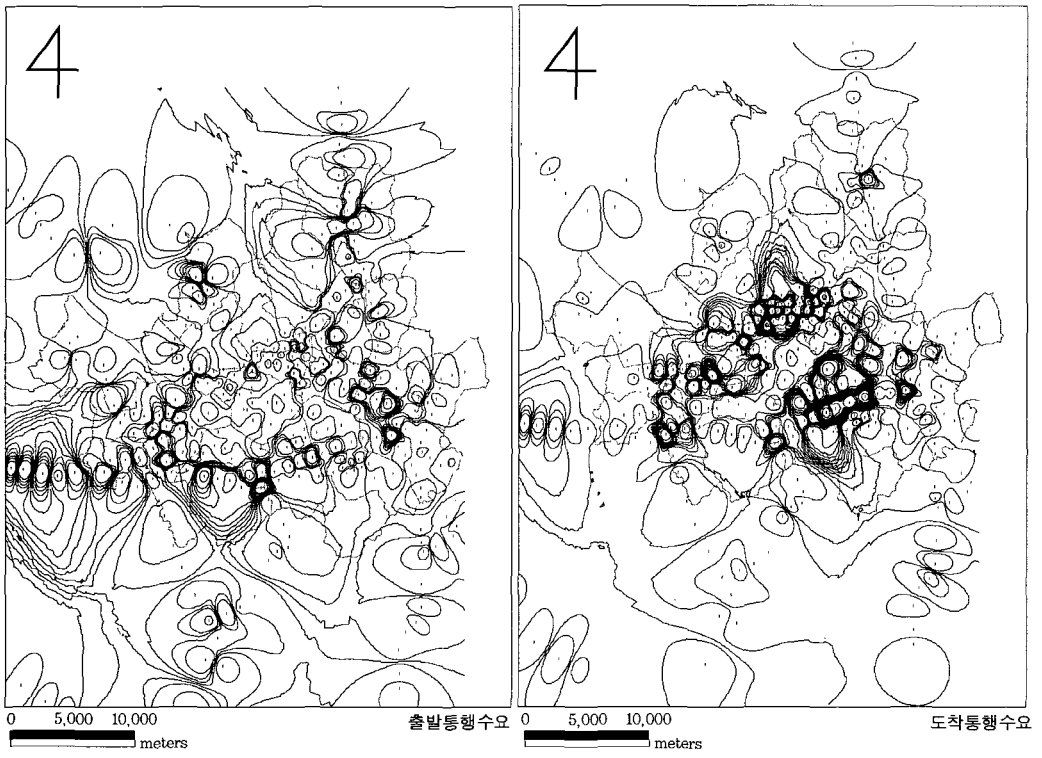

그림 6. 오전 시간대 출발-도착 통행 분포 패턴(2005년 6월 24일)

들과 관련된 통행이 주를 이루는 낮 시간대, 그리고 퇴근과 관련된 통행이 주를 이루는 저녁 시간대로 구분하여 각 시간대 별 통행 출발지와 도착지별 통행 수요를 산출하여 이들의 공간적 분포패턴을 분석하기 위하여 지리정보체계(Geographical Information System)를 이용하여 등치선도를 구축하였다. 다음 〈그림 6-8〉은 서울시의 대중교통체계 개편 이후 1년 가까이 시간이 경과하여 새로운 교통체계에 적응된 통행 행태가 어느 정도 정착되었을 것으로 여겨지는 2005년 6월 24일 통행 자료를 바탕으로 각 시간대 별 통행 수요의 공간적 분포를 나타낸 것이며, 〈표 11〉은 각각의 시간대별로 출발 통행 수요와 도착 통행 수요가 특히 높은 상위 20위까지의 지하철역이나 버스 정류장을 열거한 것이다.

〈그림 6〉에 나타나 있는 것처럼 오전 시간대에 출발 통행 수요는 대단위 거주 지역들이 분포하는 지역들에서 집중적으로 나타나고 있다. 특히 사당, 신림, 강변, 잠실, 서울대 입구 등 대단위 아파트 단지가 있는 지하철 2호선 역들과 부천, 송내, 부평, 역곡, 동암, 광명, 주안 등 지하철 1호선의 경인지역 역들, 그리고 도심 외곽의 연신내, 노원, 화곡 등 대단위 아파트 단지가 밀집되어 있는 지역의 지하철역에서 많은 통행 수요가 발생하고 있다. 이에 반해 오전 통행의 도착지는 강남의 업무지역과 강북 중심업무지역, 그리고 여의도를 중심으로 하는 영등포일대로 집중하는 패턴을 보이고 있다. 특히 강남, 선릉, 역삼, 삼성에 이르는 지하철 2호선 역과 시청, 을지로입구, 교대, 서울역, 여의도, 충무로, 종각, 종로3가, 을지로3가, 광화문 등 강북의 도심지 지하철역, 그리고 동대문, 동대문운동장, 회현 등 대형 도매시장에 접한 지하철 역들과 서울역, 고속터미널, 남부터미널 등 타 교통수단과 연결되는 터미널과 대학들이 밀집되어

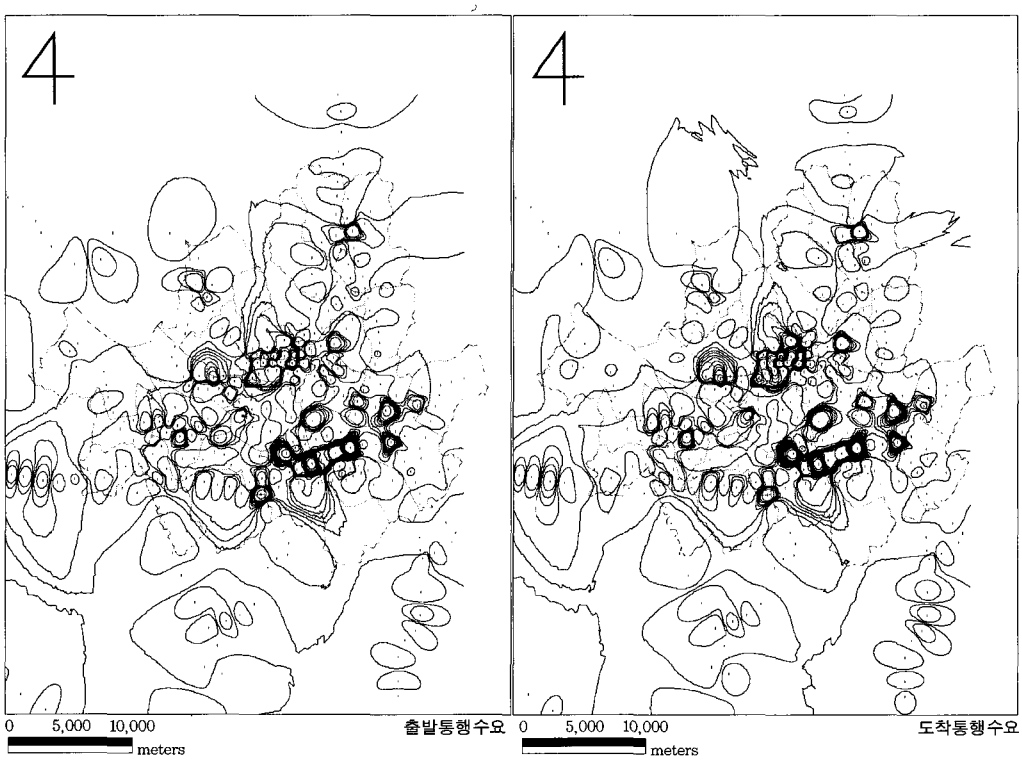

그림 7. 낮 시간대 출발-도착 통행 분포 패턴(2005년 6월 24일)

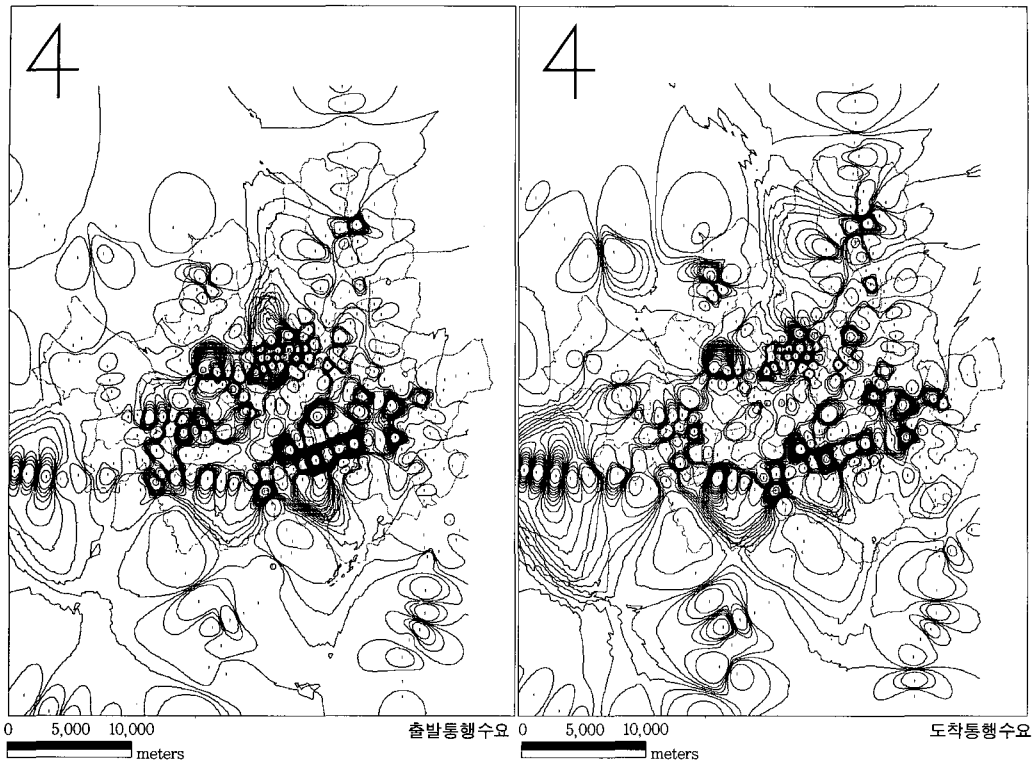

그림 8. 저녁 시간대 출발-도착 통행 분포패턴(2005년 6월 24일)

있는 신촌 등이 높은 수요를 보인다.

낮 시간대에 출발 통행량과 도착 통행량은 〈그림 7〉에 나타나는 것처럼 모두 강남과 강북의 중심업무지역들에서 집중적으로 발생한다. 특히 강남지역의 강남, 삼성, 선릉, 사당, 잠실, 역삼 등 지하철 2호선 역들과 강북의 을지로 입구, 종각, 종로3가, 시청 등과 같이 중심업무시설이 밀집되어 있는 지역들이 출발 통행량과 도착 통행량이 모두 높게 나타나며, 고속터미널, 강변, 서울역 등 고속버스나 기차의 터미널들, 그리고 동대문, 동대문운동장, 명동, 회현 등 대단위 도·소매시장이 위치하고 있는 지역이나 신촌과 같이 대학들이 몰려 있는 지역도 높은 출발 통행 수요를 나타낸다.

저녁 시간대의 출발 통행량은 오전 시간대의 도착 통행량의 분포와 유사한 패턴을 보이고 있으며, 낮 시간대의 도착 통행 분포가 결합된 분포를 보인다. 〈그림 8 참조〉. 따라서 출발통행의 경우 강남과 강북의 중심업무지구와 대단위 도매시장, 대학가들에서 출발통행량이 특히 높게 나타난다. 따라서 강남지역의 강남, 삼성, 선릉, 역삼 등의 지하철역들이 가장 높은 통행 수요를 보이며, 강북의 종각, 시청, 을지로 입구, 종로3가, 그리고 여의도 등의 중심업무지역과 동대문운동장, 명동, 충무로, 신촌 등의 대단위 도·소매 시장과 대학가 등이 또 다른 핵을 이룬다. 또한 고속터미널과 서울역 등도 많은 통행 수요가 나타난다. 그러나 저녁 시간대의 도착 통행량의 분포는 오전 시간대의 출발 통행 분포에 비해 매우 복잡한 분포를 보인다. 강남과 강북의 업무시설과 유흥시설이 밀집된 지역의 지하철역들로도 많은 도착 통행 수요가 있으며, 더불어 집으로 돌아가는 통행이 합해져 강남과 강북의 대단위 아파트 단지들을 있는 지하철역들도 비교적 많은 통행 수요를 보이고 있어 매우

표 11. 시간대별 상위 통행 수요 지점

순위	오전 시간대		낮 시간대		저녁 시간대	
	출발	도착	출발	도착	출발	도착
1위	사당	강남	강남	삼성	강남	강남
2위	부천	선릉	삼성	강남	삼성	사당
3위	송내	삼성	고속터미널	고속터미널	선릉	고속터미널
4위	신림	역삼	교대	종로3가	을지로입구	신촌
5위	부평	시청	선릉	잠실	역삼	신림
6위	강변	을지로입구	잠실	강변	고속터미널	건대입구
7위	잠실	서울역	강변	을지로입구	종각	부평
8위	역곡	종각	사당	동대문운동장	종로3가	홍대입구
9위	대림	여의도	서울역	명동	시청	부천
10위	동암	교대	종로3가	신촌	서울역	강변
11위	연신내	종로3가	동대문	교대	잠실	잠실
12위	서울대입구	충무로	을지로입구	선릉	교대	노원
13위	건대입구	광화문	동대문운동장	동대문	동대문운동장	송내
14위	까치산	압구정	종각	서울역	명동	동대문운동장
15위	총신대입구	을지로3가	역삼	종각	신촌	혜화
16위	노원	동대문	시청	혜화	사당	영등포
17위	광명	가리봉	신촌	역삼	혜화	삼성
18위	봉천	동대문운동장	회현	압구정	충무로	천호
19위	화곡	잠실	압구정	시청	압구정	종로3가
20위	주안	회현	명동	사당	여의도	총신대입구

복잡한 분포패턴을 보인다. 이는 저녁 시간대의 통행 패턴은 오전 출근시간대와는 달리 주거지로 돌아가기 전 여가나 쇼핑, 친구와의 만남, 자기개발 등 다양한 이유로 다른 지점들을 거쳐 가기 때문인 것으로 보인다. 따라서 음식점과 문화시설, 유흥시설들이 많이 사람들이 많이 모이는 신촌, 혜화, 홍대입구, 영등포 등이 또 다른 도착 통행량의 중심핵을 이룬다. 또한 이처럼 들러 가는 통행 수요 때문에 저녁 시간대의 통행량은 오전 출근시간대나 낮 시간대에 비해 출발 통행량과 도착 통행량 모두 크게 증가하고 있다.

위의 〈그림6-8〉에 나타난 수도권지역 대중교통 이용자의 시간대별 출발 통행량과 도착 통행량의 분포는 일차적으로는 수도권 통행흐름의 패턴을 나타내지만, 이에는 또한 수도권 지역의 주거지 분포, 일자리의 분포 등 도시 시설의 공간적 분포와 기능적 연계의 공간적 패턴을 보여 주고 있는 것이며, 그에 따른 지역 간 기능적 연계의 정도와 수도권 인구의 생활패턴도 반영된 것이다. 이는 〈그림 8〉의 도착 통행량의 분포에 특징적으로 반영되어 있다고 볼 수 있는데 수도권 사람들의 저녁 통행을 보면 일자리에서 곧바로 집으로 돌아가는 경우보다는 퇴근 이후 활동을 위해 일자리 주변이나 도심지역으로 다시 모여드는 야간 생활이 매우 활발한 양상을 반영하는 것이라고 볼 수 있다.

5. 결론

본 연구에서는 2004년 서울시에서 대중교통체계 개편의 일환으로 버스의 준공영제와 함께 교통카드를 이용한 환승제도를 도입한 이후 축적되고 있는 대용량의 교통트랜잭션 데이터베이스에서 수도권 지역의 대중교통이용자의 통행패턴을 분석하는 방법론을 개발하고 그의 결과를 분석하였다. 특히 본 연구에서는 데이터베이스에서 요구하는 지식 발견을 효과적으로 발굴해 내는 순회 패턴 탐사법을 원용하여 교통카드 통행 자료에서 통행 시퀀스(trip sequences)를 찾아내고, 이를 바탕으로 통행패턴을 찾아내는 데이터 마이닝 방법의 개발에 초점을 두었으며, 이를 2004년 이후 2006년 까지 3개년의 각기 다른 시점의 하루 교통카드 자료에 대해 데이터 마이닝 기법을 적용하여 결과로 도출된 통행패턴의 공간적 특징과 시점 간 차이를 분석하였다.

수도권 대중교통 이용자의 1회 통행에 소비하는 시간은 10분 이상 20분 이내에서 가장 높은 빈도를 보이며, 30분 이내가 전체 통행의 1/2 이상을 차지하고, 1시간 이내까지 확장하면 전체 통행의 거의 90% 정도를 차지하는 빈도분포를 보인다. 그리고 2004년 이후 수도권 대중교통 이용자가 1회 통행에 소비하는 시간으로 10분 이내는 그 절대량에서 감소하는 반면, 20분 이상 소요하는 통행자는 증가하고 있다.

하루 중 오전 시간대에 출발 통행 수요는 대단위 거주 지역들이 분포하는 지역들에서 집중적으로 나타나고 있다. 특히 부천, 송내, 부평, 동암, 역곡 등 지하철 1호선의 경인지역 역들과 사당, 신도림, 총신대입구 등 지하철 환승역, 그리고 강변, 잠실, 대림, 광명, 노원, 범계, 봉천, 구의, 창동 등 대단위 아파트 단지가 밀집되어 있는 지역의 지하철역에서 많은 통행 수요가 발생하고 있다. 이에 반해 오전 통행의 도착지는 강남의 업무지역과 강북 중심업무지역, 그리고 여의도를 중심으로 하는 영등포일대로 집중하는 패턴을 보이고 있다. 특히 강남, 선릉, 역삼, 삼성에 이르는 지하철 2호선 역과 시청, 을지로입구, 교대, 서울역, 여의도, 충무로, 종각, 종로3가, 을지로3가, 광화문 등 강북의 도심지 지하철역, 동대문운동장, 양재, 회현, 명동, 고속터미널, 남부터미널 등 대형 도매시장과 버스터미널, 대학들이 밀집되어 있는 신촌 등이 높은 수요를 보인다. 낮 시간대에 출발 통행량과 도착 통행량이 모두 강북의 중심업무지역과 강남을 중심축으로 하여 이들을 연결해 주는 지하철 2호선과 4호선, 그리고 3호선 상의 지하철역들에 집중되고 있다. 전반적으로 낮 시간대에는 출발 통행량과 도착 통행량이 모두 강북의 업무중심지나 대형 도매상가가 밀집되어 있는 지역에서 높게 나타나는 특징을 보이고 있으나, 도착 통행량에 있어서는 강남의 교대, 선릉, 사당, 삼성, 영등포 구청 등 업무시설이 밀집되어 있는 지역의 지하철 환승역들과, 잠실, 구의, 강변 등 대단위 상가와 아파트 단지의 지하철역들도 비교적 많은 통행 수요가 나타나고 있다. 저녁 시간대에 출발 통행량은 오전 시간대의 도착 통행량의 분포와 유사한 패턴을 보이고 있으나 도착 통행량은 집으로 돌아가는 통행과 함께 강남과 강북의 업무중심지와 그 주변 지역과 신촌일대와 여의도와 영등포 주변 등에 분포한 여러 지하철역들에 분산 집중하는 양상으로 매우 복잡한 분포패턴을 보인다. 이러한 시간대별 출발 통행량과 도착 통행량의 분포는 일차적으로는 수도권 통행흐름의 패턴을 나타내지만, 이에는 또한 수도권 지역의 주거지 분포, 일자리의 분포 등 도시 시설의 공간적 분포와 기능적 연계의 공간적 패턴을 보여 주고 있는 것이며, 그에 따른 지역 간 기능적 연계의 정도와 수도권 인구의 생활양식을 반영하여 보여 주고 있는 것이다.

도시 내의 교통흐름은 도시의 토지이용과 그에 따른 지역 간 상호작용과 밀접하게 관련되어 있으므로 한 도시의 교통흐름에 나타나는 공간적 특징을 파악하는 것은 도시공간구조 이해에 중요한 단서를 제공할 수 있다. 또한 이러한 통행패턴 분석 결과는 직접적으로 통행수요에 부합하는 버스노선 조정, 배차계획에 이용될 수 있으며, 서울시의 교통로 별 통행의 빈도 및 통행자의 이동거리의 빈도분포함수를 이용

하여 대중교통 요금체계 개선방향을 제시할 수 있음은 물론, 다양한 교통 관련 연구를 위한 매우 귀중한 자료로 활용될 수 있다. 그밖에도 주택정책이나 토지 이용 및 시설 입지 등, 도시계획과 공간 계획을 위한 중요한 기초 자료를 제공할 수 있을 것이다.

참고문헌

박준식 · 박창호 · 전경수, 2001, "오전 첨두시의 동적 교통관리를 위한 동적 통행배정모형에 관한 연구," 대한교통학회지 19(4), pp.97~108.

이금숙, 2005, "교통흐름과 미세먼지 분포의 공간적 특징," 2005년 대한지리학회 춘계학술대회 발표논문 초록집, p.7, (2005. 5.12-14, 여수).

이금숙 · 박종수, 2006, "서울시 대중교통 이용자의 통행패턴 분석", 한국경제지리학회지 9(3), pp.379~395.

최계숙, 2006, "교통 카드 트랜잭션 데이터베이스에서 순회패턴 탐사 알고리즘들의 비교 및 분석," 성신여자대학교 석사학위논문.

허우긍, 1993, "서울의 통근통행: 지리적 특성과 변화," 대한교통학회지 11(1), pp.5-21.

Agrawal, R. and Srikant, R., 1995, "Mining sequential patterns," Proc.11th Int'l Conf. Data Eng., pp.3-14, (Mar. 1995).

Badoe, D. A., and Chen, C., 2004, "Modeling trip generation with data from single and two independent cross-sectional travel surveys," *Journal of Urban Planning and Development* 130(4), pp.167-174.

Burnett, P. and Thrift, N., 1979, "New approaches to travel behavior," in D. Hensher & P. Stopher (Eds.) *Behavioral Travel Demand Modelling*, pp.116-136, London: Croom Helm.

Cervdesity, R. and Kockelman, K. M., 1997, "Travel demand and the three Ds: density, diversity and design," Transportation Research Part D: *Transport and Environment 2*, pp.199-219.

Chen, Ming-Syan, Park, Jong Soo, and Yu, Philip S., 1998, "Efficient data mining for path traversal patterns," *IEEE Transactions on Knowledge and Data Engineering* 10(2), pp.209-221.

Han, J., and Kamber, M., 2006, *Data Mining: Concepts and Techniques*, 2nd Ed., San Francisco: Morgan Kaufmann.

Lee, Keumsook and Park, J., 2005, "Traversal pattern analysis of transit users in the metropolitan seoul, Proceedings of International Forum on the Public Transportation Reform in Seoul, (July 7-8, 2005, Seoul).

Marble, D., 1967, "A theoretical exploration of individual travel behavior, in W. Garrison & D.Marble (Eds.), *Quantitative Geography, Part I (Economic and Cultural Topics)*, 57-93, New York: Plenum Press.

Park, J.S., Chen, M.-S., and Yu, P.S., 1997, "Using a hash-based method with transaction trimming for mining association rules," *IEEE Trans. on Knowledge and Data Eng.* 9(5), pp.813-825, Sept./Oct.

Pas, E. and Kopplelman, F., 1987, "An examination of the determinants day-to-day variability in individuals' urban travel behavior, *Transportation* 13, pp.183-200.

Sanchez, T. W., 2004, "Connecting mass transit and employment, in Hensher. D. A. et al.(eds.), *Handbook of Transport Geography and Spatial Systems*, Amsterdam: Elsevier, pp.111-124.

Srinivasan, S. and Ferreira, J., 2003, "Travel behavior at the household level: understanding linkages with residential choice," *Transportation Research Part D* 7, pp.225-242.

Srikant, R. and Agrawal, R., 1996, "Mining sequential patterns: Generalizations and performance improvements," In Proc. 5th Int. Conf. *Extending Database Technology (EDBT'96)*, pp.3-17, (Avignon, France, Mar. 1996).

Stern, E., and Richardson, H.W., 2005, "Behavioural modelling of road users: current research and future needs," *Transport Reviews* 25(2), pp.159~180.

Sen, Parongama and Dasgupta, Subinay and Chatterjee,

Arnab and Sreeram, P. A. and Mukherjee, G. and Manna, S. S., 2003, "Small-world properties of the Indian railway network", 67(3), pp.036106-036110, Mar, Phys. Rev. E, American Physical Society.

Tan, Pang-Ning, Steinbach, Michael, and Kumar, Vipin, 2006, *Introduction to Data Mining*, Boston: Addison Wesley.

교신: 이금숙, 서울특별시 성북구 동선동 3가 249-1, 성신여자대학교 지리학과, Tel: 02-920-7130, E-mail: kslee@sungshin.ac.kr

Correspondence: Keum-sook Lee, Departmanet of Geography, Sungshin Women's University, Seoul, 136-742, Korea, Tel: 02-920-7130, E-mail: kslee@sungshin.ac.kr

최초투고일 2007년 2월 13일
최종접수일 2007년 2월 23일

사사: 저자들은 본 논문에 삽입된 지도 작성에 애쓴 성신여자대학교 지리학과 대학원 석사과정 홍지연과 서위연에게 감사드립니다.

Journal of the Economic Geographical Society of Korea
Vol.9, No.1, 2007(44~63)

Mining Trip Patterns in the Large Trip-Transaction Database and Analysis of Travel Behavior

Jong Soo Park*, Keumsook Lee**

Abstract : The purpose of this study is to propose mining processes in the large trip-transaction database of the Metropolitan Seoul area and to analyze the spatial characteristics of travel behavior. For the purpose. this study introduces a mining algorithm developed for exploring trip patterns from the large trip-transaction database produced every day by transit users in the Metropolitan Seoul area. The algorithm computes trip chains of transit users by using the bus routes and a graph of the subway stops in the Seoul subway network. We explore the transfer frequency of the transit users in their trip chains in a day transaction database of three different years. We find the number of transit users who transfer to other bus or subway is increasing yearly. From the trip chains of the large trip-transaction database, trip patterns are mined to analyze how transit users travel in the public transportation system. The mining algorithm is a kind of level-wise approaches to find frequent trip patterns. The resulting frequent patterns are illustrated to show top-ranked subway stations and bus stops in their supports. From the outputs, we explore the travel patterns of three different time zones in a day. We obtain sufficient differences in the spatial structures in the travel patterns of origin and destination depending on time zones. In order to examine the changes in the travel patterns along time, we apply the algorithm to one day data per year since 2004. The results are visualized by utilizing GIS, and then the spatial characteristics of travel patterns are analyzed.The spatial distribution of trip origins and destinations shows the sharp distinction among time zones.

Keywords : data mining processes, trip chains, travel patterns, spatial structures of travel demand, Geographical Information System

* Professor, School of Computer Science & Engineering, Sungshin Women' s University
** Professor, Department of Geography, Sungshin Women' s University