한국경제지리학회지 제9권 제3호 2006(379~395)

서울시 대중교통 이용자의 통행패턴 분석*

이금숙** · 박종수***

요약: 본 연구의 목적은 도시성장과 함께 통행거리와 교통량이 크게 늘어나 심각한 교통체증현상을 보이고 있는 서울을 대상으로 교통흐름의 공간적 특징을 파악하고, 이를 수도권 지역의 토지이용 및 시설 분포와 연결시켜 수도권 지역의 기능적 연계의 공간구조를 밝히는 것이다. 이를 위해 본 연구에서는 교통카드를 이용하여 이동하는 대중교통 이용자들이 만들어 내는 통행 거래 자료를 바탕으로 통행행태와 통행흐름의 공간적 특징을 분석하였다. 교통카드 데이터는 하루 천만 건이 넘으며, 각 통행거래자료 마다 승·하차 및 환승의 위치와 시간, 이용 교통수단 등 방대한 정보를 담고 있어 이의 효율적 분석을 위하여 데이터마이닝 기법을 도입하였다. 특히 대중교통이용자의 통행행태를 파악하고 각 지점의 통행 수요를 파악하기 위한 새로운 알고리즘 개발하여 적용하였다. 분석결과의 공간적 특성을 파악하기위하여 지리정보체계를 이용하여 시각화시키고, 그에 입각하여 공간적 특성을 분석하였다. 또한 출발 교통량, 도착교통량, 총 지지도들 간의 관계성을 파악하기 위하여 상관관계분석을 시도하였다. 통행수요에 있어 강남의 2호선 지하철역을 따라 가장 많은 통행 수요가 나타나고 있으며, 그 다음으로 강북의 구도심지역이 또 하나의 중심축을 이룬다. 그밖에도 대단위 고층 아파트가 밀집되어 있는 주거지역들이 부수적인 중심축을 형성한다. 기·종점 수요와 함께 통과 수송량 까지 나타내는 총지지도는 강남의 구로-신도림 역이 가장 높게 나타나며 강남지역에 위치한 지하철 2호선 역들과 환승역들에서 높게 나타나고 있다. 이러한 통행패턴 분석은 일차적으로는 교통망 상의 교통흐름과 각 지점의 통행수요를 나타내며 지역 내에서 지점 간의 기능적 연계를 반영하고 있기 때문에 도시의 교통계획은 물론 지역의 토지이용 및 시설 입지 계획 수립에 필수적이다.

주요어: 통행거래 자료, 데이터마이닝, 통행패턴, 통행수요, 통행흐름, 공간구조, 지리정보체계

1. 서론

많은 인구와 다양한 경제·사회·문화 활동 및 시설이 응집되어 있는 도시공간에서는 많은 교통흐름이 발생하며, 하나의 도시가 제 기능은 하기 위해서는 이의 원활한 소통이 필연적이다. 그러나 교통의 발달로 개개인의 이동성이 확대되고 서비스산업의 발달 및 전문화로 기능들이 점점 세분되면서 사람과 물자의 공간 이동이 크게 늘어나게 되어 도시 교통흐름은 기하급수적으로 늘고 있다(Vasconcellos, 2001). 따라서 거의 세계 모든 도시들은 늘어나는 교통량으로 심각한 교통 혼잡 문제를 겪고 있으며, 교통 혼잡으로 야기되는 막대한 경제·사회·환경적 손실과(Evans, 1992; Vasconcellos, 2001), 시민 건강 문제를 야기하고 있다(Lee, 2004; Briggs & Gulliver, 2003; Sommer, *et al.*, 2000; Kuenzli, *et al.*, 2000;

* 이 논문은 2006년도 성신여자대학교 학술연구조성비 지원에 의하여 연구되었음.

** 성신여자대학교 지리학과 교수

*** 성신여자대학교 컴퓨터정보학부 교수

Koslowsky & Krausz, 1993). 특히 우리나라의 서울을 포함한 수도권지역은 우리나라 전체 인구의 46.6%가 거주하고 있으며, 경제 · 사회 · 문화 · 교육 등 중심기능들이 과도하게 집중되어 있어 세계 어느 도시 보다도 심각한 교통 혼잡 현상을 보이며, 이로 인해 유발되는 다양한 교통문제가 나타나고 있어 이의 개선을 위한 대책 마련이 시급한 상황이다(이금숙 2005; 박준식, 2001).

이러한 교통문제 해결을 위한 정책입안을 위해서는 도시의 교통흐름을 이해하고, 도시 내 시민들의 통행패턴과 통행행태에 대한 정확한 파악이 요구된다(허우긍 1993). 도시의 교통흐름은 인구 및 도시 시설의 공간적 분포와 그 도시의 교통망 구조와 직결되어 있다. 따라서 교통서비스의 변화나 도시의 성장으로 인구 및 도시기능이 확대되면 교통흐름의 양적 증가 뿐 만 아니라 그 공간적 패턴에도 변화가 나타난다. 또한 이러한 교통흐름의 변화는 도시 시설의 입지와 교통망에도 영향을 주게 되어 역으로 도시의 토지이용구조를 변화시킨다. 따라서 한 도시의 교통흐름은 교통계획, 도시계획, 입지계획, 사회학에서는 물론 지표의 공간구조를 연구하는 지리학에서도 많은 관심이 두고 있는 연구주제 중의 하나이다.

이처럼 도시의 교통흐름과 통행행태 자료의 중요성 때문에 교통관련 연구 분야에서는 일찍부터 이를 정확히 파악하려는 노력을 경주해 왔다. 그러나 이제까지 한 도시의 통행흐름을 정확히 나타내 주는 자료가 없었기 때문에 표본 집단에 대한 설문조사를 실시하거나 지역의 속성을 나타내는 지리적 변수들을 이용한 모형을 개발하여 적용해 왔다(박준식 외, 2001; 김명수 · 남궁문, 2000; 배영석 1996; 김익기, 1991; Stern & Richardson 2005;Badoe & Chen 204; Srinivasan & Ferreiva, 2003; Kitamura, *et al.*, 1990; Pas & Koppleman 1987; Adler & Ben-Akiva 1979; Burnett & Thrift 1979; Marble, 1967).

그러나 서울에서는 2004년 대중교통체계 개편과 함께 정보 기술을 도입한 교통카드 사용이 활성화되어 대중교통 이용자의 교통카드 이용율이 80%를 넘고 있어 적어도 대중교통 이용자들에 대한 통행 자료는 거의 전수에 가까운 자료를 확보하게 되었다. 이 자료는 실제로 수도권 지역의 대중교통 이용자들이 움직이면서 생성하고 있는 교통흐름에 대한 데이터로서 하루 10,000,000 건이 넘는 통행거래에 대한 실제 통행 자료가 생성되고 있는 것으로 전 세계 어느 도시에서도 이제까지 존재하지 않았던 도시통행패턴과 통행행태를 가늠해 볼 수 있는 귀중한 자료이다. 특히 그 안에는 개개 통행자의 통행에 대한 출발지점과 목적지, 이용 교통수단, 환승에 대한 위치와 시간 등에 대한 정확한 정보가 담겨있어 이를 적절히 분석하면 도시 내 통행흐름 및 통행행태에 대한 다양한 정보를 파악해 낼 수 있게 되었다. 또한 교통카드 자료는 과거 일회적으로 실시되던 표본조사의 자료와는 달리 매일 새로운 자료가 생성되어 저장되고 있어, 통행흐름과 통행행태, 통행수요에 대해 하루의 시간대별, 주별, 월별, 계절별, 연차별 통행의 변화를 시계열적 분석을 통해 파악할 수 있는 자료라는 점에서 관련 분야 연구자들에게 매우 매력적인 자료이다.

그러나 이러한 교통카드데이터는 정보량이 매우 많아 대용량의 데이터베이스의 형태로 존재한다. 따라서 이런 자료를 분석하는데 있어 관건은 엄청나게 방대한 자료에서 필요한 정보를 효과적으로 추출해 내는 문제이다. 최근 정보 기술의 발전으로 방대한 양의 데이터가 저장되고, 이를 효과적으로 활용하도록 관리하는 기술이 늘어감에 따라, 대용량의 데이터로부터 쉽게 드러나지 않는 유용한 정보들을 추출하는 데이터마이닝에 대한 관심이 증대되고 있다 (Chen, *et al.* 1998; Choi, *et al.* 1997).

본 연구의 목적은 도시성장과 함께 통행거리와 교통량이 크게 늘어나 심각한 교통체증 현상을 보이고 있는 서울을 대상으로 교통흐름의 공간적 특징을 파악하고, 이를 수도권 지역의 토지이용 및 시설 분포와 연결시켜 수도권 지역 기능적 연계의 공간구조를 밝히는 것이다. 이를 위해 본 연구에서는 수도권 교통카드 데이터베이스에서 데이터마이닝 기법을 적용하여 수도권 대중교통이용자의 통행패턴과 교통흐름

의 특징을 파악해 내고자 한다. 특히 본 연구에서는 다 시점의 교통카드데이터를 이용하여 수도권 지역 교통카드 이용자들의 통행 자료에서 통행 연속(trip sequences)을 찾아내어 통행의 특징을 분석하고, 수도권 교통망의 통행수요의 시·공간적 특징을 파악하고자 한다.

이러한 도시의 통행 특징과 통행흐름 및 통행수요의 공간적 특징에 대한 정확한 이해는 교통계획은 물론 지역의 토지이용계획 수립에 있어 가장 기본이며 필수적이다. 교통망 상의 교통흐름 즉, 통행수요는 지역 내에서 지점 간의 기능적 연계를 반영하고 있기 때문에 지역의 공간구조를 분석에 필수적인 내용이며, 또한 가로 상에 입지하고 있는 시설물이나 상점들에게는 그 지점에서 접할 수 있는 잠재적 이용자나 고객의 확보 가능성을 가늠할 수 있어 시설물의 입지계획에 중요한 자료로 활용될 수 있을 것이다. 한편 정보과학측면에서도 통행거래 자료와 지리정보 자료와 같이 대형 공간데이터베이스에서 필요한 정보를 효율적으로 찾아내기 위한 데이터베이스 처리를 위한 새로운 알고리즘 개발이라는 학문적 기여를 기대할 수 있다.

2. 수도권 지역의 교통관련 지리적 특성

도시인구 및 물자의 이동은 다양한 도시기능들의 공간적 분포와 이들을 잇는 교통망의 구조에 직접적인 영향을 받게 되므로 도시의 통행흐름과 토지이용의 공간적 특징은 상호 역동적 관계를 고려하여 종합적으로 분석되어야 한다(노정현 · 류재영, 1995; 김익기, 1994). 따라서 서울의 교통흐름을 정확히 이해하기 위해서는 서울과 일일 생활권을 이루고 있는 수도권지역의 교통망은 물론 인구 및 토지이용 상태 등의 지리적 속성에 대한 공간적 특징을 파악하여야 한다(허우긍, 1993).

수도권지역 여객통행에 대한 교통수단별 분담율은 시기에 따라 큰 변화를 보여 왔다. 1960년대까지는 도심 중심의 방사형 전차가 여객 통행의 대부분을 담당하다가 1960년대 이후 경제부흥과 함께 서울로 인구가 집중하면서 새로운 거주 중심지역들이 형성되면서 버스가 주요한 대중교통 수단 역할을 하게 되어 분담율이 85%까지 다다르기도 하였다. 그러나 1970년대 이후 지하철이 건설되고 승용차가 늘어나면서 지하철과 승용차의 분담율이 늘어나게 되어 1996년에 이르러서는 버스 30.1%, 지하철 29.4%, 그리고 승용차가 24.6%를 차지하는 상황이 되었다. 그러나 1990년대 말 IMF 사태를 거치면서 다소 상황에 바뀌어 2000년대 이후에는 도시철도, 버스, 자가용, 택시 순으로 분담율이 큰 것으로 나타나고 있다(음성직, 2004). 이중 특히 대중교통 수단인 지하철과 버스의 분담율이 높아 65% 가까이 차지하고 있다(표1 참조). 2004년 대중교통체계개편 이후 도시철도와 버스 이용자의 90% 가까이가 교통카드를 사용하고 있어 교통카드 자료를 바탕으로 분석한 결과가 수도권지역의 통행행태와 통행패턴과 함께 교통흐름의 공간적 특징을 비교적 잘 반영하고 있다고 간주할 수 있다.

표 1. 서울시 수송 분담율 변화

(%)

	도시철도	시내버스	소계	자가용	택시	기타
2001	36.5	27.6	64.1	18.7	8.4	8.8
2002	37.8	26.8	64.6	18.4	8.0	9.0
2003	35.0	27.6	62.6	25.0	7.3	5.1
2004	35.7	26.3	62.0	26.4	6.6	5.0
2005	35.9	26.8	62.7	26.3	6.2	4.8

자료출처 : 건설교통부 2005년도 자료

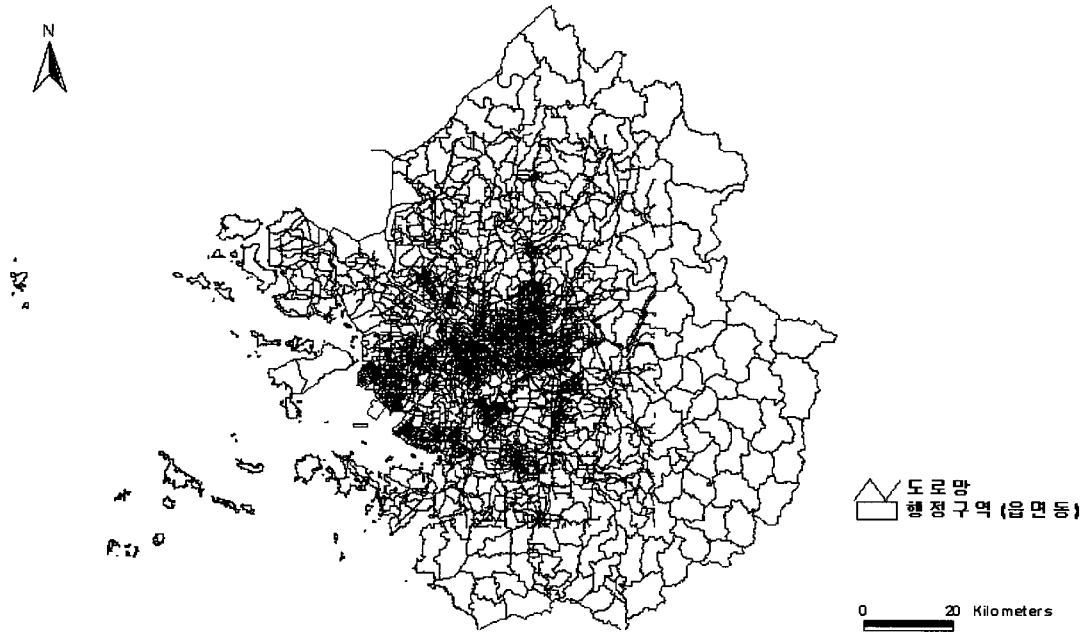


그림 2. 수도권 도로망

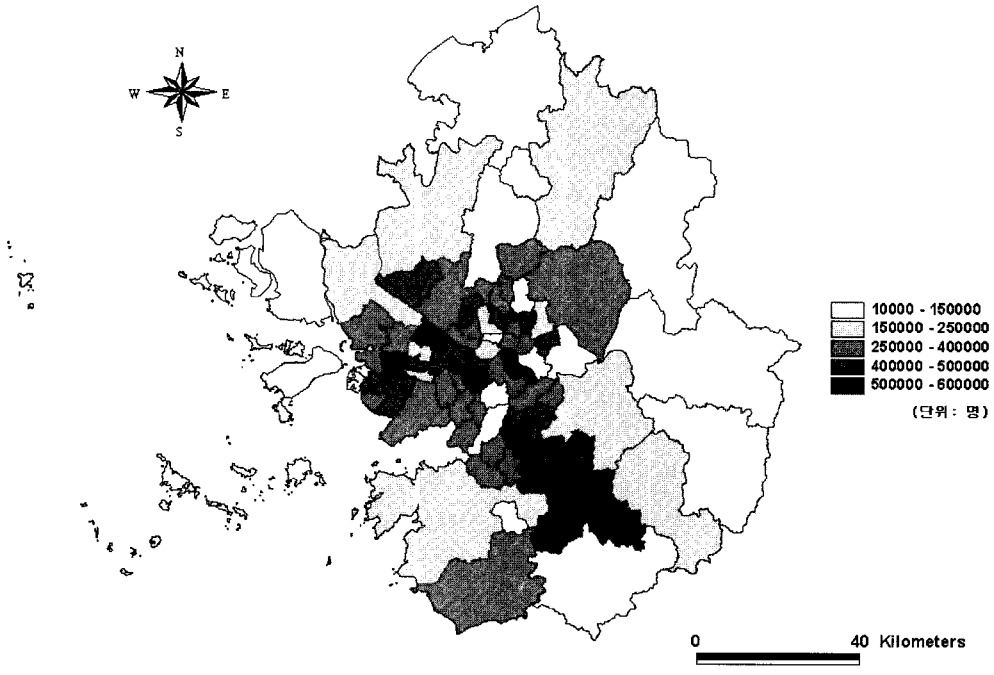


그림 3. 수도권 인구분포

도시의 교통흐름은 인구 및 도시 시설의 공간적 분포와 그 도시의 교통망 구조와 직결되어 있다. 따라서 한 도시의 교통흐름은 그 도시의 교통망 및 교통서비스, 그리고 인구분포 및 토지이용패턴 등의 지리적 특징들의 공간적 분포와 함께 다루어져야 한다. 다음 〈그림1-3〉은 각각 수도권지역의 도시철도망, 도로망, 인구분포를 보여 주고 있다.

통행(trip)은 사람들이 어떤 특별한 목적을 가지고 한 지점에서 목적지까지 이동하는 이동의 기본 단위로 통근 · 통학, 업무, 구매 및 개인 용무, 사교 및 오락 등 통행 목적과 개개인의 사회 · 경제적 차이에 따라 차이를 보일 수 있다. 또한 각 통행의 출발지와 목적지는 그 지점의 토지이용에 서 비롯되며, 이러한 토지이용은 주어진 교통수단 및 교통망으로 형성되는 접근성에 직접적으로 영향을 받게 되므로 통행패턴은 지역의 토지이용 구조 및 교통망과 교통정책과도 밀접히 연관되어 있다. 따라서 통행 연구에 대한 관심은 교통계획, 도시계획, 입지계획 등에서는 물론 지표의 공간구조를 연구하는 지리학에서도 많은 관심이 두고 있는 연구 주제이다.

3. 데이터마이닝 기법을 적용한 대용량 교통카드 데이터 분석

1) 데이터마이닝 적용을 위한 전처리 과정

서울 대중교통체계에서 관리되는 교통카드데이터베이스에는 교통카드를 이용하여 버스와 지하철 등의 대중교통을 이용하는 승객들의 실제 통행거래내역자료 (transaction data)이다. 이 원천 데이터베이스에는 각 승객이 버스나 지하철을 승차하고 하차한 다양한 정보(카드 ID, 카드 트랜잭션 번호, 환승횟수, 승차일시, 승차정류장, 하차정류장, 하차일시, 버스노선, 승차 금액 등)를 포함하고 있다. 수도권의 하루 통행거래 수는 천만을 넘고 있다. 데이터마이닝 기법을 적용하면 이러한 대용량의 데이터베이스에서 통행의 특징을 효과적으로 분석할 수 있다. 각 통행거래 자료에서 통행연쇄(Trip-chain)를 찾아내어 출발-도착(Origin-Destination)을 구축하고, 이를 바탕으로 통행 패턴을 찾아내어 대중교통 승객들의 행위특성인 이동 경로를 분석함으로 수도권지역 대중교통이용자의 통행패턴을 분석함으로서 통행의 공간적 특성은 물론 교통 정책분야에도 유용하게 활용할 수 있는 정보로 정리될 수 있다. 또한 승객 개개인에게 보다 효율적인 경로인 최단 경로를 제안할 수 있고, 승객들이 집중되는 노선의 재조정 및 배차 관리에도 이용될 수 있으며, 승객이나 교통 정책 당국자에게 통행 패턴에 관한 정보를 제시함으로써 대중교통의 활성화와 여러 연계 정책에 매우 유용한 정보를 제공할 수 있을 것이다.

본 연구에 이용된 원천 데이터는 한국스마트카드(KSCC)에서 집계한 2004년 10월 27일 데이터와 2005년 6월 24일 데이터로 사용자 개인의 실제 ID에 관한 사항 등은 제외시킨 자료이다. 2004년 10월 27일 총 카드 거래 기록 10,088,158 중 출발지-도착지의 정보가 완성되지 않은 정보 불분명 건 수 614,396

표 2. 교통카드 통행기록의 수송 수단별 증감 비교

(%)

구분	2004년 10월 27일	2005년 6월 24일	증감
총 카드 거래 기록 건수	10,088,158	10,667,519	579,361(5.7)*
-정보 불분명 건수	614,396	679,485	65,089(19.6)*
분석대상 통행 기록 건수	9,473,762(100)	9,988,034(100)	514,272(100)
-버스 이용 건수	4,654,747(49.1)	5,078,718(50.8)	423,971(82.4)
-지하철 이용 건수	4,819,015(50.9)	4,909,316(49.2)	90,301(17.6)

주 : * 2004년 6월 27일 통행거래 건수에 대한 증가율 임.

개를 제외한 9,473,762개 통행기록을 대상으로 분석하였다. 이중 버스 이용 거래 건수는 4,654, 747개이고 지하철 이용 거래 건수는 4,819, 015개 이다. 또한 2005년 6월 24일 자료의 경우 총 카드 거래기록 건수는 10,667,519개로 교통카드를 이용한 대중교통 이용 통행 건 수가 이 기간 동안 60만 건 정도 증가하였다. 그러나 출발-도착 기록이 불분명한 거래 건수 679,485개를 제외하면 실제 분석 대상 통행 기록 건수는 9,988,034개이며, 이중 버스 이용 거래 수는 5,078,718로 424,000 건 정도 증가하고, 지하철 기록 건수는 4,909,306개로 9만 건 정도 늘어 나 교통카드를 이용한 대중교통 통행 건 수 증가의 82%이상이 버스이고 지하철 이용의 경우는 17.6%로 상대적으로 버스 이용 통행이 크게 늘었음을 알 수 있다.(표2 참조)

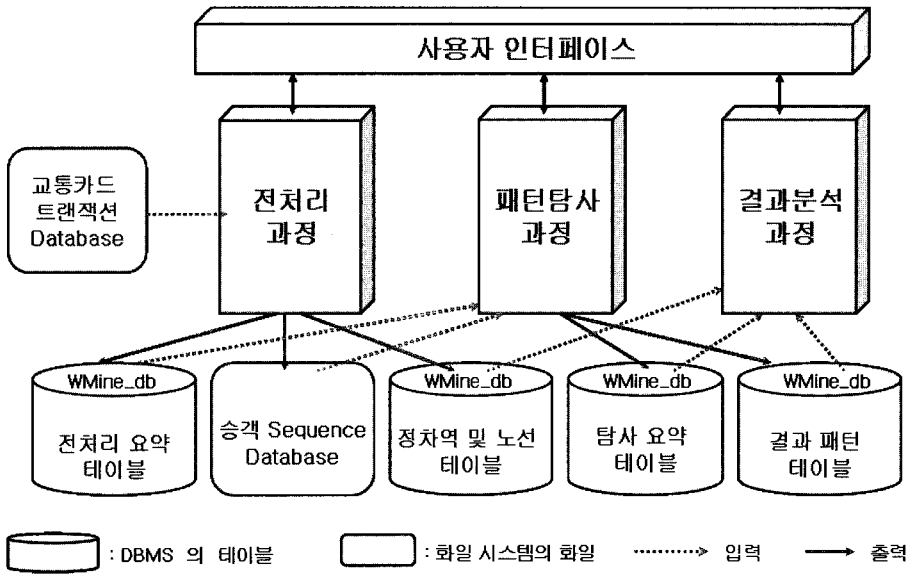
본 연구에서는 서울 대중교통체계에서 승객들의 교통 카드 거래로 저장되는 대 용량의 데이터베이스에서 승객들의 통행패턴(Trip Pattern)을 탐사하는 것이 목적이므로 모든 지하철 노선과 역에 관한 자료, 버스 노선과 정류장에 관한 자료 및 위치 정보 등이 필요하므로 원천자료에 이들을 추가하는 등 전처리 과정이 요구된다. 서울 대중교통체계에서 승객들의 교통 카드 거래로 저장되는 대 용량의 데이터베이스에서 승객들의 통행 패턴(Trip Pattern)을 탐사하기 위해서 데이터마이닝 기법을 적용하기 위해서는 다음 그림 4와 같이 전처리과정, 패턴탐사과정, 결과분석과정의 세 단계 과정을 거치게 된다.

전처리 과정에서는 일일 카드 거래 내역에서 카드번호에 따라 각 승객이 출발 정류장(Origin)에서 도착지(Destination)까지의 중간 버스 정류장 또는 지하철 정거장 ID를 순서대로 찾아내어 하나의 정차장 시퀀스(stop sequence)를 만든다. 통행 패턴 탐사과정에서는 전처리 과정에서 만들어진 승객들의 모든 정차장 시퀀스를 입력으로 하여 traversal pattern을 탐사하는 과정으로 통행 패턴(trip pattern)을 찾아낸다. 마지막으로 결과분석 과정에서는 지지도가 높은 패턴을 기준으로 정차장 ID에서 버스 정류장 이름이나 지하철 정거장 이름을 찾아내어 결과로 보여준다.

일일 카드 거래 내역에서 승객의 transaction 데이터는 승차 정류장 (Oi)와 하차 정류장(Dj)만을 보유하고 있으므로 각 승객의 transaction 데이터마다 출발지(Oi)와 최종 목적지(Dj) 사이에 통과하게 되는 중간 버스 정류장 또는 지하철 정거장 ID를 순서대로 찾아내어 하나의 정차장 시퀀스(stop sequence)를 만든다. 이를 위해서는 카드번호에 따라 출발 정류장(Origin)에서 도착지(Destination)까지의 승객이 통과하는 버스 정류장이나 지하철역에 관한 정보가 사전처리되어 배열과 해시 트리에 저장되어 있어야 한다. 전처리 단계에서 필요한 과정들은 다음과 같다.

1) 버스 노선 정보에서 버스 노선별로 각 정류장에 관한 정보인 정류장 ID와 정류장명을 배열과 해시 테이블을 사용하여 저장한다. 버스 노선은 800여개이고, 그 노선에서 통과하는 버스 정거장의 개수는 15,000여개이다[4].
2) 지하철을 사용하는 승객의 trip chain을 만들기 위해서 승차 정거장과 하차 정거장 사이에 통과하는 정거장의 ID를 찾아내어야 한다. 이것은 9개의 지하철 노선과 400여개의 지하철역을 하나의 그래프로 고려하여 최단 거리를 찾는 알고리즘을 적용하여 승객이 통과하는 지하철역을 찾아낸다.
3) 교통 카드에 의해 처리된 일일 통행 거래 데이터베이스에서 카드번호 SEQ와 TransID가 같은 트랜잭션을 찾아서 환승횟수 만큼의 승차 정차역과 하차 정차역을 찾아낸다. 버스로 승차와 하차가 이루어지면 해당 버스 노선에서 통과된 중간 정류장ID를 순서대로 저장한다. 지하철로 승차와 하차가 이루어지면 최단거리를 찾는 알고리즘을 적용하여 승객이 지나가는 지하철역ID를 또한 순서대로 저장한다. 한 승객이 같은 TransID로 환승한 모든 정차역(버스 정류장 또는 지하철역)을 순서대로 나열하여 하나의 시퀀스(sequence)로 한다.

이 과정을 통하여 얻어진 sequence database로부터 승객들이 대중교통을 이용하여 통행할 때 통과하는 정류장 수에 대해 다음 그림 5과 같은 특성을 읽어



교통 카드 DB에서 통행 패턴 탐사 시스템

그림 4. 교통 카드 DB에서 통행 패턴 탐사체계

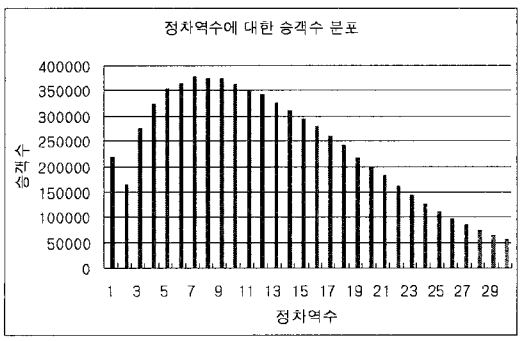
낼 수 있다.

수도권 대중교통 이용자들이 한 번 통행할 때 통과하는 정차역 수의 분포는 그림5와 같이 최빈값은 7을 중심으로 그 전 후에 대부분의 통행자가 분포하고 있어 전반적으로 왼쪽으로 치우친 분포 특성을 보인다. 하지만 오른쪽으로 꼬리가 길게 늘어지는 분포를 보여 승객들이 통행별 통과하는 평균 정차역 수는 2004년 10월 29일 자료와 2005년 6월 24일 데이터의 경우 각각 13.6개와 13.8로 나타나고 있다. 이러한 정차역 수 분포특성은 요금 책정 정책을 입안할 때 의미 있는 자료로 활용될 수 있을 것이다.

2) 통행패턴 탐사 데이터마이닝 알고리즘 개발

전처리 과정에서 각 승객의 sequence database가 생성되면 이러한 sequence database에서 traversal

2004년 10월 27일



2005년 6월 24일

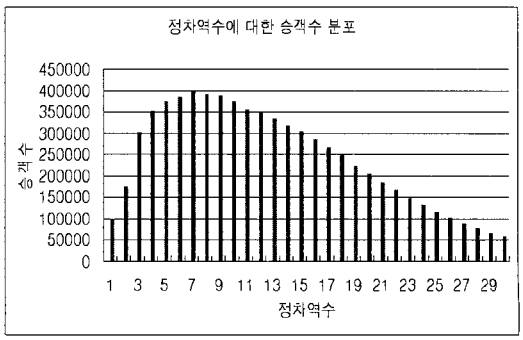


그림 5. 정차역 수에 대한 승객 분포도

pattern을 탐사하는 알고리즘 개발이 필요하다. 특히 본 연구에서는 Agrawal & Srikant(1995) 에 의해 소개된 후 후속 연구가 활발히 진행되고 있는 데이터베이스에서 요구하는 지식정보 (KDD; Knowledge Discovery in Databases)를 효과적으로 발굴해 내는 연속 패턴 탐사(Mining Traversal Patterns) 법(Park, *et al.* 1997; Chen, *et al.*, 1998; Zaki, 2001) 을 적용하여 통행분석에 맞게 새로운 알고리즘을 개발하고자 하였다.

본 연구에서는 교통 카드에 의한 카드 거래내역에 관한 데이터베이스가 전처리 단계를 거치게 되면 승객 단위의 O-D chain으로 나타내어진 승객 시퀀스 데이터베이스가 된다. 승객 시퀀스 데이터베이스에서 순회 패턴과 비슷한 방법으로 통행 패턴(trip pattern)을 탐사해낸다. 순회 패턴과 통행 패턴의 다른 점은 먼저 통행 패턴에서는 뒤로 향하는 패턴이 없다는 것이다. 승객이 반드시 원하는 목적지로 향한다고 가정한 것이다. 또한 어떤 버스 노선이나 지하철 노선에서도 정류장이나 지하철역을 건너뛸 수 없다는 것이다. 이는 기존의 웹 순회 패턴에 대한 알고리즘에서는 웹 순회의 경우 어떤 페이지를 순서에 관계없이 건너뛸 수 있음이 차이 나는 부분이다. 이런 측면에서 보면 통행 패턴을 효과적으로 찾아내는 알고리즘은 아직 개발된 것이 없는 실정이다. 순회패턴 탐사를 위해 개발된 알고리즘으로는 원 로그 데이터베이스에서 최대 순방향 참조들을 구한 후, 최대 순방향 참조 집합에서 빈발 참조 시퀀스를 결정하고, 빈발 참조 시퀀스에서 최대 참조 시퀀스를 찾는 FS 알고리즘 (Web Access Pattern 탐사): Compound Hash Tree를 이용하는 CHT_FS 알고리즘: 동질등급에서 탐사패턴을 탐색하는 TPADE 알고리즘 (Traversal PAttern Discovery using Equivalence classes) 등이 있다(Zaki, 2001).

다음은 교통카드데이터를 전처리 과정을 통해 sequence database로 변환시킨 후 통행 패턴을 찾아내는 과정을 정리한 것이다.

1) 승객 시퀀스 데이터베이스 DB를 읽으면서, 단일 빈발 항목을 찾아낸다. 여기서, 항목은 시퀀스 데이터베이스에 있는 정차역ID를 나타내고, 빈발 항목이란 사용자가 지정한 최소지지도 이상의 지지를 받는 항목을 나타낸다. 최소지지도는 보통 0.1%에서 1% 사이의 값이 지정된다. 예를 들면, 0.5%의 최소지지도가 주어지고, 시퀀스 데이터베이스는 5,000,000개의 승객 시퀀스로 구성되어 있다고 가정하자. 그러면, 최소지지도는 어떤 항목이 5,000,000 × 0.5 ÷ 100 = 25,000 승객 시퀀스들 이상에 포함되어 있을 경우 그 항목이 최소지지도를 만족하는 항목이 된다. 결론적으로, 먼저 시퀀스 데이터베이스에서 단일 빈발 항목 집합인 F1을 찾아낸다.
2) 승객 시퀀스 데이터베이스에서 F1에 속한 항목들로 이루어진 시퀀스인 데이터베이스 DB1을 만들어낸다. DB1에서 두 항목들로 이루어진 빈발 2-항목집합들인 F2를 구한다. F2를 구하는 방법은 F1에서 조합되는 모든 2-항목집합을 배열에 넣어서 지지도를 계산하여 최소지지도 이상인 2-항목집합들을 찾아내어 F2라 한다. 일반적으로, 연관규칙 탐사, 순차 패턴 탐사, 순회 패턴 탐사에서 F2를 구하는 데 가장 큰 실행 시간이 소비된다. 실행시간을 줄이는 방법의 연구가 일반적인 패턴 탐사에서 기본 과제이다. F1과 F2를 찾아내는 시간을 줄이기 위하여 자료구조를 개발하였다.
3) F2를 구한 후에는 빈발 3-항목집합들인 F3를 구한다. F3는 F2에 속한 항목들을 기반으로 하여 후보 3-항목집합들인 C3를 계산하여 시퀀스 데이터베이스 DB1을 사용하여 지지도를 계산하게 된다. C3에서 최소 지지도를 만족하는 빈발 3-항목집합들인 F3를 구한다.
4) 빈발 4-항목집합들인 F4와 이후의 Fn (n >= 5)을 구하는 방법은 F3를 구하는 방법에서 항목집합에 속하는 항목들이 늘어나는 점에서만 다르고, 후보 n-항목집합들을 찾아내고 시퀀스 데이터베이스를 사용하여 지지도를 계산하고, 그 다음으로 빈발 n-항목집합들을 찾아내는 방법은 동일한 과정이

적용될 수 있다.

본 연구에서는 기존의 순차 패턴 알고리즘을 수정하여 위의 과정을 구현하였다. 세 가지 종류의 알고리즘을 구현하였다. 첫 번째로 SPADE(Sequential PAttern Discovery using Equivalence classes(Zaki, 2001)와 TPADE를 결합하여 응용한 알고리즘(박종수 외, 2001), 두 번째로 GSP(Generalized Sequential Patterns(Srinkant & Agrawal, 1996)를 응용한 알고리즘, 그리고 마지막으로 FS(Full Scan)를 응용한 알고리즘(Chen, *et al.*, 1998)을 구현하여 빈발 항목집합들을 찾아내었다. 그리고 구현된 알고리즘들을 분석하여 순차 패턴보다는 통행 패턴에 더 적합한 자료구조와 알고리즘으로 정교화하고 있다.

앞 단계에서 탐사해낸 빈발 항목집합들을 지지도가 높은 패턴을 기준으로 정차장 ID에서 버스 정류장 이름이나 지하철 정거장 이름을 찾아내어 결과로 보여준다. 결과를 보여주는 과정에서 지리정보시스템과 접목하여 지도상에서 가장 빈발하게 사용하는 정거장들의 순서인 시퀀스를 디스플레이 할 수 있도록 한다.

4. 수도권 대중교통 이용의 공간적 특징

도시 내의 교통흐름은 도시의 토지이용과 그에 따른 지역간 상호작용과 밀접하게 관련되어 있으므로 한 도시의 교통흐름에 나타나는 공간적 특징을 파악하는 것은 도시공간구조 이해에 중요한 단서를 제공할 수 있다. 본 연구에서는 수도권 대중교통이용자들의 통행에 나타나는 공간적 특징을 파악하기 위하여 지리정보체계(Geographical Information System)를 이용하여 앞에서 개발된 통행패턴탐사 알고리즘을 적용하여 얻어진 각 지하철역과 버스정류장별 출발지지도, 도착지지도, 총통행지지도 값을 바탕으로 등치선도를 구축하였다(그림 6-8 참조).

앞의 데이터마이닝 알고리즘을 적용하여 통행패턴을 분석한 결과로 출발지지도와 도착지지도, 그리고 총지지도를 산출하였다. 출발지지도와 도착지지도는 각각 각 정류장이나 지하철역에서 통행이 시작되었거나 종착된 통행량을 의미한다. 따라서 이들은 대중교통수단을 이용하여 수도권 지역에서 발생하는 통행 수요의 정도를 알아볼 수 있는 척도로 간주할 수 있다.

출발지지도와 도착지지도의 경우 강남역에서 선릉-역삼-삼성-잠실까지 이어지는 지하철 2호선 노선을 따라 위치한 강남지역이 가장 높은 값을 보이는 중심핵을 이루며, 강북의 을지로입구-종각-동대문 지역으로 이어지는 구도심지역의 지하철 환승역들을 중심으로 또 하나의 중심핵을 이룬다. 그리고 강변과 고속터미널과 같이 고속버스와 연계되는 지하철역이나 신촌 일대처럼 대학교들이 밀집되어 있는 지역에 위치한 지하철역들도 작은 중심핵을 이루고 있다. 그 밖에도 외곽 지역 중 대단위 아파트들이 밀집되어 있는 거주지에 위치한 지하철역을 중심으로 국지적인 중심핵이 형성되어 있다.

물론 출발지지도의 경우 오전에는 주로 주거지가 밀집되어 있는 지역에서 높게 나타나지만 오후에는 일자리나 상가 등 중심업무 시설들이 밀집되어 있는 지점에서 높게 나타날 수 있다. 역으로 도착지지도의 경우는 오전과 오후에 이와 반대의 경향을 나타 날 것이다. 따라서 출발지지도와 도착지지도의 정도는 오전과 오후에 서로 역전되는 상황이 나타날 수 있다. 본 연구에서는 하루 전체의 통행패턴을 분석하였으므로 통행패턴에 나타나는 이러한 오전과 오후에 역전되는 특징이 뚜렷이 드러나지는 않고 있다.

그러나 본 연구에서 얻어진 하루 전체의 출발지지도와 도착지지도에도 공간적 특징이 어느 정도는 나타나고 있다. 일자리와 상업시설이 밀집해 있는 중심업무지구나 대형아파트단지가 밀집되어 있는 주거지구들이 높은 값을 보이며, 이들을 중심으로 크게는 두개의 중심핵과 2-3개의 작은 중심핵을 이루는 것을 알 수 있다. 특히 〈그림 9〉에 나타난 것과 같이 수도권지역 대중교통이용자들의 통행특징으로 강남지역의 신흥 업무지구가 강북의 구도심지역에 비해 월등히 높게 나타나고 있다.

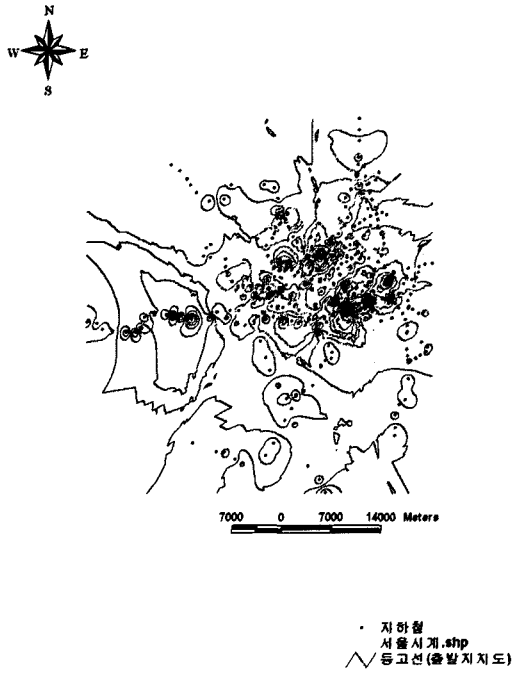


그림 6. 출발 지지도의 공간구조

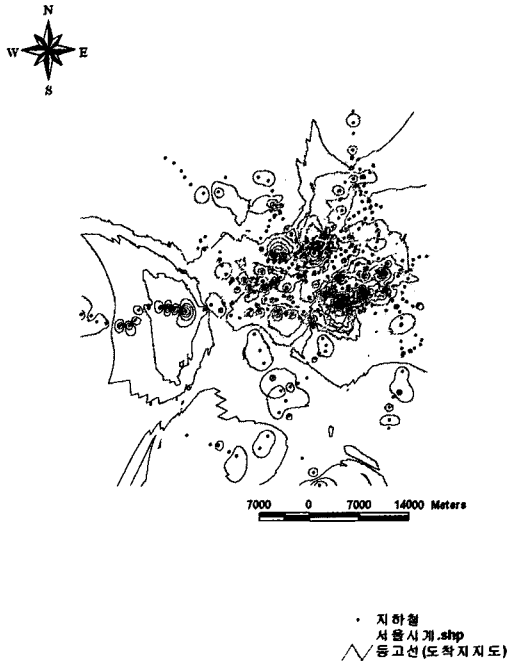


그림 7. 도착 지지도의 공간구조

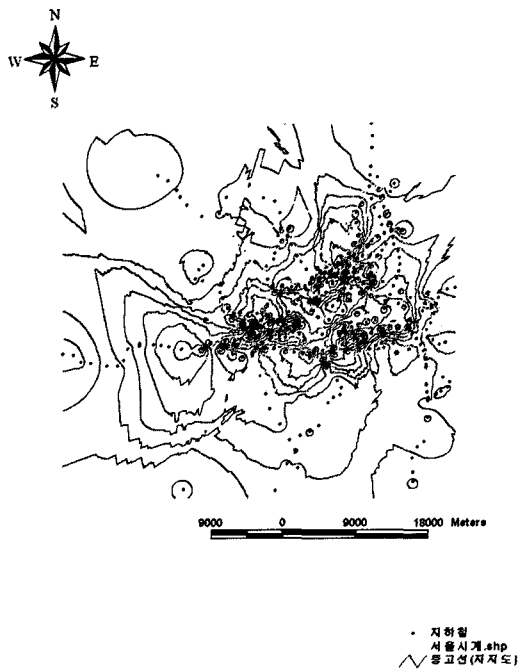


그림 8. 총 지지도의 공간구조

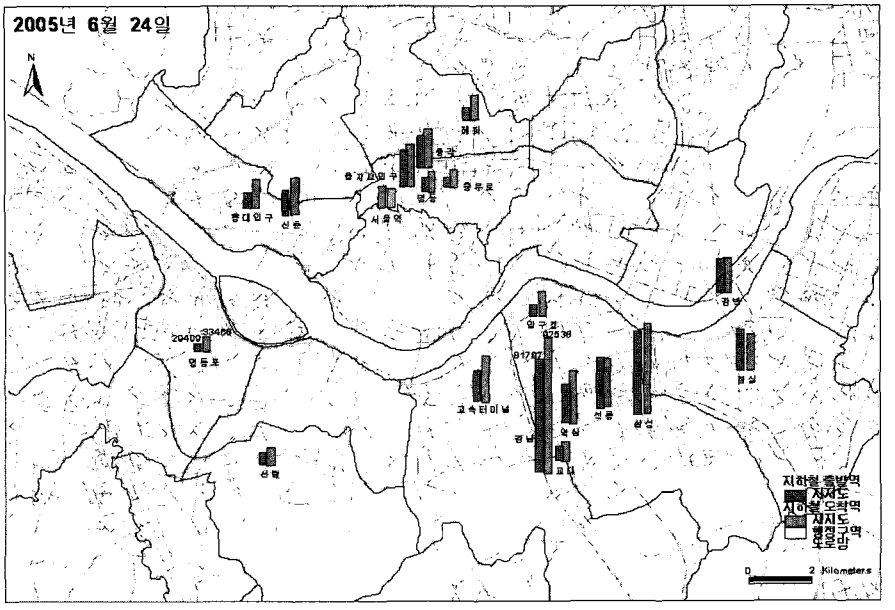


그림 9. 출발-도착 총지지도 상위 정류장의 분포

그러나 각 정류장지점에서 출발하거나 도착하는 통행은 물론 대중교통수단을 타고 있는 상태에서 그 지역을 경유하는 통행량까지 합한 총 통행 수를 의미하는 총지지도의 경우는 이들 출발지지도와 도착지지도와는 다소 다른 공간적 구조를 나타나고 있다. 수도권지역에서 대중교통을 이용하여 일어나는 통행으로 형성되는 총 지지도의 공간적 분포는 크게 구로-신도림-영등포-신길을 축으로 하는 영등포지역 일대, 사당-강남-교대-역삼-선릉-삼성 등 강남지역 일대, 혜화-동대문-동대문운동장을 축으로 하는 대학로-동대문지역 일대, 용산-서울역-충무로-종로로 이어지는 구도심지역의 네 개의 중심핵으로 구성되어 있다. 특히 지지도가 가장 높게 나타나는 지역은 지하철 2호선과 1호선이 교차하여 환승이 이루어지는 지하철역 들이며, 그 중에서도 지하철 2호선을 따라 강남 쪽에 위치한 지하철역들이 높은 값을 나타내고 있다.

출발 지지도와 도착 지지도사이의 상호 상관계수는 0.976로 매우 높은 상관관계를 보이고 있으며, 총지지도와는 출발지지도와 도착지지도사이의 상관계수는 각각 0.609와 0.598로 어느 정도 상관은 있는 것으로 나타나고 있다.[1)]

모든 분석 결과 지하철역들의 지지도가 압도적으로 높아 상위 330위 정도까지는 모두 지하철역이 차지하고 있다. 또한 지하철의 상위 순위의 역들이 분포하는 지역과 버스의 상위 정류장의 공간적으로 차별화 되어 있어 지하철 이용자의 총지지도가 높은 상위 지점들은 주로 강남지역에 집중되고 있음에 반해 버스 이용자의 총지지도가 높게 나타나는 지점들은 강북지역에 집중 분포함을 알 수 있다. 〈그림10〉은 지하철의 총 지지도가 높은 상위 50개 지점과 버스의 총지지도가 높게 나타나는 상위 50개 버스 정유장의 분포를 보인 것이다. 수도권 대중교통이용의 공간적 특징의 하나는 강남의 경우 지하철 이용이 두드러짐에 비해 강북지역의 경우 버스이용이 상대적으로 높게 나타나고 있다는 것이다. 특히 버스이용에 의한 총지지도에서 성신여대-미아삼거리를 중심으로 미아로를 따라 높은 지지도를 보이고 있다. 특히 출발지지도와 도착지지도에 있어 강남지역의 지하철역을 따라 높게 나타나고 있어 수도권 지역 생활의 중심지

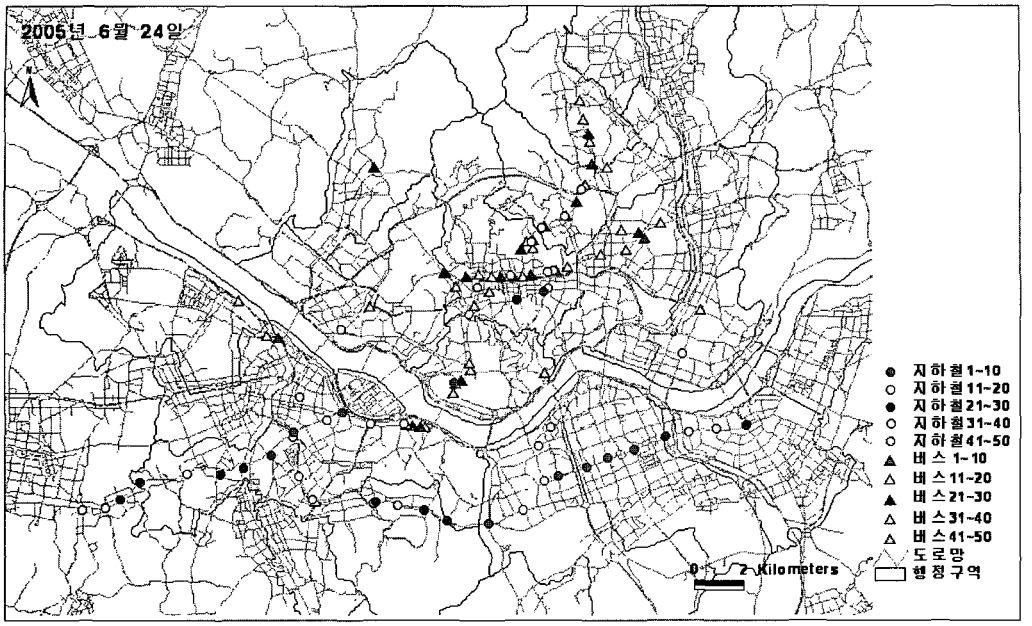


그림 10. 지하철과 버스 총지지도 상위 정류장의 분포

가 구도심에서 강남 지역으로 이전되었음을 확인할 수 있다.

5. 결론

현재 수도권 지역의 대중교통 이용자들이 움직이면서 생성하고 있는 교통카드데이터는 개개 통행자가 실제로 움직이는 궤적을 담은 통행자료로서 하루에도 약 1000만 건에 달하는 교통 카드의 기록으로 구성된 트랜잭션 데이터베이스가 생성되고 있다. 그 안에 있는 막대한 통행에 대한 정보 자료를 적절히 분석하면 서울시를 포함한 서울 광역지역에서 대중교통 이용자의 정확한 통행행태와 통행패턴을 파악할 수 있다.

본 연구에서는 컴퓨터 시스템의 데이터마이닝 기법을 도입하여 교통 카드에 의한 대 용량 데이터베이스에서 사용자들의 통행 특성을 파악할 수 있는 이동경로를 탐색하는 방법론과 그들의 공간적 특징을 파악해 보았다. 이를 위하여 각 승객의 승하차 트랜잭션에서 그 승객이 통과한 모든 정류장이나 정거장을 찾아내어 한 시퀀스를 만들고, 이런 승객들의 시퀀스들로 이루어진 데이터베이스에서 일정한 지지도 이상의 순회 패턴을 탐사하였다. 순회패턴을 탐사하는 알고리즘들을 이 데이터베이스에 적용하여 통행패턴을 찾아내고, 각 알고리즘의 성능을 측정하였으며, 분석에서 얻어진 결과의 공간적 특성을 분석하였다. 특히 본 연구에서는 Agrawal & Srikant에 의해 소개된 후 후속 연구가 활발히 진행되고 있는 데이터베이스에서 요구하는 지식정보 (KDD; Knowledge Discovery in Databases)를 효과적으로 발굴해 내는 연속 패턴 탐사(Mining Traversal Patterns) 법을 적용하여 통행분석에 맞게 새로운 알고리즘을 개발하였다.

컴퓨터 시스템의 데이터마이닝 분야에서 또한 컴퓨터 과학의 관점에서 볼 때 서울 대중교통체계 상의 교통카드의 거래 내역에 관한 데이터베이스는 데이터 흐름(data stream)에 해당하는 대 용량의 데이터가 계속해서 저장되고 있는 상황이다. 이런 대용량 데이터베이스에서 숨겨져 있는 정보나 지식에 해당되는 여러 패턴이나 연관성을 찾아내는 것은 중요한 연구

토픽들 중의 한 분야이다. 특히 이를 지리정보체계의 데이터베이스와 결합하여 효과적으로 분석하면 수도권지역에서 지역간의 실질적인 기능적 연계 및 도시의 공간구조를 분석 할 수 있고, 그 결과는 토지이용계획 및 시설계획 등 다양한 교통정책 수립에 귀중한 기초 자료를 제공할 수 있으므로 도시의 공간구조와 교통흐름을 연구하는 지리학과 다양한 교통관련 연구 분야에 매우 관심을 가지고 있는 문제이다.

또한 본 논문에서는 승객 트랜잭션 데이터베이스에서 또한 수도권 대중교통이용자들의 통행에 나타나는 공간적 특징을 파악하기 위하여 지리정보체계를 이용하여 앞에서 개발된 통행패턴 탐사 알고리즘을 적용하여 얻어진 각 지하철역과 버스정류장별 출발 지지도, 도착지지도, 총통행지지도 값을 바탕으로 등치선도를 구축하였다. 출발지지도, 도착지지도는 각각 수도권 지역에서 발생한 교통카드데이터에서 각 정류장이나 지하철역에서 통행이 시작되었거나, 종착지로 유입하는 통행량을 나타내며, 총 통행지지도는 이들과 함께 대중교통수단을 타고 있는 상태에서 그 지역을 경유하는 통행량까지 합한 총 통행 수를 나타내는 것이다. 따라서 이들은 대중교통수단을 이용하여 수도권 지역에서 발생하는 통행 수요의 정도를 알아볼 수 있는 척도로 간주할 수 있으며, 수도권 지역의 토지이용과의 연관성을 담고 있다.

출발지지도와 도착지지도의 경우 강남역에서 선릉-역삼-삼성-잠실까지 이어지는 지하철 2호선 노선을 따라 위치한 강남지역이 가장 높은 값을 보이는 중심핵을 이루며, 강북의 을지로입구-종각-동대문지역으로 이어지는 구도심지역의 지하철 환승역들을 중심으로 또 하나의 중심핵을 이룬다. 그리고 강변과 고속터미널과 같이 고속버스와 연계되는 지하철역이나 신촌 일대처럼 대학교들이 밀집되어 있는 지역에 위치한 지하철역들도 작은 중심핵을 이루고 있다. 그 밖에도 외곽 지역 중 대단위 아파트들이 밀집되어 있는 거주지에 위치한 지하철역을 중심으로 국지적인 중심핵이 형성되어 있다. 위의 분포도는 하루 전체의 통행에 대한 것으로 출발지지도의 경우 오전에는 주로 주거지가 밀집되어 있는 지역에서 높게 나타나지만 오후에는 일자리나 상가 등 중심 시설들이 밀집되어 있는 지점에서 높게 나타날 수 있다. 역으로 도착지지도의 경우는 오전과 오후에 이와 반대의 경향을 나타 낼 것이다. 따라서 출발지지도, 도착지지도의 정도는 오전과 오후에 역전되는 상황이 나타날 수 있다. 물론 본 연구에서는 하루 전체의 출발, 도착, 총통행의 지지도를 산출하였으므로 시간대별에 이러한 특징이 뚜렷이 드러나지는 않고 있지는 않지만 주택지구와 일자리가 밀집해 있는 장소 또는 상업시설 등 중심시설이 응집되어 있는 지역들이 핵을 이루는 것을 알 수 있다.

이러한 통행패턴 분석 결과는 직접적으로는 서울광역시 각 지역별 교통수요를 파악할 수 있으며, 구체적으로 각 교통로 및 교통수단별 통행행태와 통행수요를 산출할 수 있다. 이러한 결과는 직접적으로 통행수요에 부합하는 버스노선 조정, 배차계획에 이용될 수 있으며, 서울시의 교통로 별 통행의 빈도 및 통행자의 이동거리의 빈도분포함수를 이용하여 대중교통 요금체계 개선방향을 제시할 수 있음은 물론, 다양한 교통 관련 연구를 위한 매우 귀중한 자료로 활용될 수 있다. 그밖에도 주택정책이나 토지이용 및 시설 입지 등, 도시계획과 공간 계획을 위한 중요한 기초 자료를 제공할 수 있을 것이다. 또한 도시 내의 교통흐름은 도시의 토지이용과 그에 따른 지역간 상호작용과 밀접하게 관련되어 있으므로 한 도시의 교통흐름에 나타나는 공간적 특징을 파악하는 것은 도시공간구조 이해에 중요한 단서를 제공할 수 있다.

그러나 본 연구에서는 아직 GIS작업을 위해 필요한 정류장의 위치에 대한 좌표정보를 모두 확보하지 못하였으며, 지역의 사회 · 경제지표에 대한 자료가 정류장 주변지역과 같이 미세 대한 토지이용과 교통수요의 통합적 분석까지 수행하지는 못하였다. 그러나 교통흐름과 토지이용은 상호 역동적으로 영향을 주고받게 되므로 이들을 결합시킨 분석이 요구된다. 추후 모든 정류장에 대한 위치 정보와 정류장 주변지역과 같은 미세한 지역 단위로 도시공간에 대한 지

리정보 데이터베이스가 좀 더 다양하고 상세한 구축되면 교통카드자료의 데이터베이스와 적절히 결합시켜 분석하여 지역간의 실질적인 기능적 연계 및 도시의 공간구조 분석이 뒤따라야 할 것이다.

감사의 글

저자들은 본 논문에 삽입된 지도 작성에 애쓴 성신여자대학교 지리학과 대학원 석사과정 서위연, 오가영과 학부생 노영희에게 감사드립니다.

주

1) 본 연구에서는 SPSS의 pearson 분석방법을 사용하였으며, pearson 분석방법은 산점도에서 대략적으로 파악할 수 있는 두 변수의 관계를 하나의 수로 나타내는 방법으로서 공식은 다음과 같다.

$$r=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

$$=\frac{n(\sum_{i=1}^{n}x_i y_i)-(\sum_{i=1}^{n}x_i)(\sum_{i=1}^{n}y_i)}{\sqrt{n(\sum_{i=1}^{n}x^2{}_i)-(\sum_{i=1}^{n}x_i)^2}\sqrt{n(\sum_{i=1}^{n}y^2{}_i)-(\sum_{i=1}^{n}y_i)^2}}$$

참고문헌

김대웅 · 유영근 · 김문주, 1994, "도시버스교통 이용자 개인특성에 관한 연구," 국토계획 29(4), pp.197~209.

김명수 · 남궁문, 2000, "통행행동을 바탕한 동적 교통행태 모형에 관한 연구," 대한토목학회논문집 20(5), pp.495~502.

김익기, 1991, "행태의 다양성을 고려한 토지이용-교통모형의 개발," 대한교통학회지 9(2), pp.87~100.

___, 1994, "토지이용-교통모형의 이론적 비교분석," 국토계획 29(4), pp.135~155.

노정현 · 류재영, 1995, "토지이용-교통 통합모형을 이용한 공간계획 평가기법의 개발," 국토계획 31(2), pp.205~222.

박종수, 하미라, 김연호, 2001, "웹 로그 파일에서 순회 패턴 탐사 알고리즘", 2001년 데이터마이닝 WORKSHOP 발표자료.

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

배영석, 1996, "개별 행태 모형을 이용한 통근인구의 교통행동분석에 관한 연구," 대한교통학회지 14(4), pp.31-48.

음성직, 2004, "서울시 도시교통정책," 2004년도 한국지리연구소 추계 초청 강연회 발표자료.

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

Adler, T. and Ben-Akiva, M., 1979, "A theoretical and empirical model of trip chaining behavior," *Transportation Research B* 13, pp.243-257.

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

Axhausen, K. and Garling, T., 1992, "Activity-based approaches to travel analysis: Conceptual frameworks, models, and research problems," *Transport Reviews* 12(4), pp.323-341.

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.

Briggs, D. and Gulliver, J., 2003, *Modelling exposure to air pollution using GIS*, WHO-HEARTS Working Paper.

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

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, M.-S., Park, J. S., and Yu, P. S., 1998, "Efficient data mining for path traversal patterns," *IEEE Transactions on Knowledge and Data Engineering* 10(2), pp.209-221.

Evans, A.W., 1992, "Road congestion pricing: When is it a good policy · " *Journal of Transport Economics and Policy 26*, pp.213-244.

Kitamura, R., Kazuo, N., and Goulias, K., 1990, "Trip chain behavior by central city commuters: A causal analysis of time-space constraints," in P. Jones (Ed.), *Development in Dynamic and Activity-Based Approaches to Travel Analysis*, pp.145-170, Avebury: Aldershot.

Koslowsky, M. and Krausz, M., 1993, "On the relationship between commuting, stress, and attitudinal measures: A LISREL application," *Journal of Applied Behavioral Science* 29, pp.485-492.

Kuenzli, N., Kaiser, R, and Medina, S., 2000, "Public-health impact of outdoor and traffic-related air pollution: a european assessment," *Lancet 2000* 356, pp.795-801.

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).

Lee, Keumsook, 2004, "Spatial relationships between respiratory disease and the local environment in Korea," Proceeding of IGC-UK (2004. 8. 15-20, Glasgow).

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

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.

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.

Pei, J., Han, J., Mortazavi-Asl, B. and Zhu, H., 2000, " Mining access patterns efficiently from web logs (PDF)," Proc. 2000 Pacific-Asia Conf. on Knowledge Discovery and Data Mining (PAKDD' 00), Kyoto, Japan, April 2000.

Sommer, H., Kunzli, N., Seethaler, R., et al., 2000, *Economic Evaluation of Health Impacts Due to Road Traffic-related Air Pollution*, Expert Workshop on Assessing the Ancillary Benefits and Costs of Greenhouse Gas Mitigation Strategies, March 27-29, 2000, Washington, D.C.

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), 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.

Vasconcellos, E., 2001, *Urban Transport, Environment and Equity*, London and Sterling: Earthscan Publication Lid.

Wu, K.L., Yu, P.S.and Ballman, A., 1998, "Speed tracer: A web usage mining and analysis tool," IBM Systems Journal 37(1), pp.89-105.

Zaki, Mohammed J., 2001, "SPADE: an efficient algorithm for mining frequent sequences," in *Machine Learning Journal*, special issue on Unsupervised Learning (Doug Fisher, ed.), pp.31-60.

교신 : 이금숙, 서울특별시 성북구 동선동 3가 249-1, 성신

여자대학교 사회과학대학 지리학과, Tel: 02)920-7138, E-mail: kslee@sungshin.ac.kr
Correspondence : Keumsook Lee 249-1 Dongseon-dong 3-ga, Seongbuk-gu, Seoul 136-742, Korea, Tel: 02-920-7138, Fax : 02-920-2041, E-mail: kslee@sungshin.ac.kr

최초투고일 2006년 11월 15일
최종접수일 2006년 12월 4일

Journal of the Economic Geographical Society of Korea
Vol.9, No.3, 2006(379~395)

Travel Patterns of Transit Users in the Metropolitan Seoul

Keumsook Lee* · Jong Soo Park **

Abstract : The purpose of this study is to analyze the spatial characteristics of travel patterns and travel behaviors of transit users in the Metropolitan Seoul area. We apply the data mining techniques to explore the travel patterns of transit users from the T-money card database which has been produced over 10,000,000 transaction records per day. The database contains the information of locations and times of origin, transfer, and destination points for each transaction as well as the informations of transit modes taken via the transaction. We develop an data mining algorithm to explore traversal patterns from the enormous information. The algorithm determines the travel sequences of each passenger, and produce the volumes of support on each points (stops) of transportation networks in the Metropolitan Seoul area. In order to visualize the spatial patterns of travel demands for transit systems we apply GIS techniques, and attempt to investigate the spatial characteristics of travel patterns and travel demand. Subway stops located in the Gangnam area appear the highest peak for the travel origin and destination, while the CBD in the Gangbuk stands at the second position. Two or three sub-peaks appear at the densely populated residential areas developed as the high-rise apartment complex. Subway stations located along the Subway Line 2, especially from Guro to Samsung receive heavy travel demand (total support), while bus stops located at the CBD in the Gangbuk stands the highest travel demand by bus.

Keywords : transaction records, data mining, travel patterns, travel demand, traffic flows, spatial characteristics, geographical information system

* Professor, Department of Geography, Sungshin Women's University

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