

시공간 이동 시퀀스 패턴 마이닝 기법*

한선영¹⁾ 용환승²⁾

요약

최근 모바일 컴퓨팅 시스템에서 위치 기반 서비스(Location Based System: LBS)에 대한 연구가 활발히 진행되고 있다. 시공간 이동 시퀀스 마이닝은 이동 경로 데이터로부터 사용자 이동 패턴을 추출하는 새로운 마이닝 기법이다. 시공간 이동 시퀀스 패턴 마이닝은 기존의 빈발 패턴 마이닝 기법과 유사하나 몇 가지 차이점이 있다. 빈발 패턴 마이닝은 장바구니 분석에서와 같이 고객이 구입한 아이템과 관련된 것이나 시공간 이동 시퀀스 패턴 마이닝은 사용자 이동 시퀀스 경로를 대상으로 한다. 또한 사용자의 관심도를 반영하기 위해 해당 위치에서의 소요시간을 고려한다. 본 연구는 대표적인 빈발 패턴 마이닝 기법의 하나인 Apriori 알고리즘에 이동 시퀀스 데이터를 적용하여 Apriori_msp 알고리즘을 제안하였으며 성능 평가를 수행한 결과를 제시하였다.

주요용어: 시공간 데이터 마이닝, 이동 시퀀스, 패턴 마이닝

1. 서론

최근 위치기반 서비스(Location Based Service: LBS)에 대한 연구가 활발히 진행되고 있다. 그러나 사용자에게 적합한 맞춤형 서비스를 위해서는 사용자의 정확한 위치 파악을 위한 연구에서 나아가 해당 위치에서 사용자가 원하는 서비스가 무엇인지를 파악하고 이를 바로 제공해 줄 수 있어야 한다. 시공간 이동 시퀀스 패턴 마이닝(Spatial-Temporal Moving Sequence Pattern Mining)이란 이러한 연구의 한 부분으로써 사용자의 이동 정보로부터 이동 시퀀스 패턴을 추출하는 것이다. 기존의 패턴 추출을 위한 대표적 마이닝 기법으로는 연관 규칙 마이닝의 선행 작업인 빈발 패턴 마이닝(Frequent Pattern Mining) 기법이 있다. 그러나 이러한 마이닝 기법은 빈발하게 발생하는 특정 항목집합에 대한 패턴 분석이므로 이동 순서가 중요한 이동 시퀀스 패턴에는 적합하지 않다. 또한 패턴 추출을 위한 값으로 항목집합의 발생 횟수만을 카운트하기 때문에 이러한 방법을 적용하면 건물 설계상의 문제로 반드시 지나쳐야만 하는 이동 경로 등은 항상 포함될 수밖에 없다.

본 연구에서는 이러한 문제점을 해결하고자 각 패턴에 대한 모든 시퀀스를 고려하고, 사용자가 해당 위치에 얼마만큼 관심 있게 머물렀는가 하는 사용자의 관심도를 반영하기

* 본 연구는 한국과학재단 목적기초연구(R01-2006-000-10609-0) 지원으로 수행되었음.

1) (463-844) 경기도 성남시 분당구 정자동 25-1 벤처타운, NHN(주), 엔터테인먼트 솔루션 개발팀 대리

E-mail: piousy@hanmail.net

2) (교신저자) (120-750) 서울시 서대문구 대현동 11-1, 이화여자대학교 컴퓨터학과, 교수

E-mail: hsyong@ewha.ac.kr

위해 유효소요시간을 도입하였다. 또한 아이템의 존재여부만을 중요시하는 빈발 패턴 마이닝에서는 고려하지 않았던 중복 값을 반영하기 위해 중복 방문하는 이동 경로에 대해 중복 방문 가중치를 부여하였다. 시공간 이동 시퀀스 패턴 마이닝 기법의 알고리즘은 대표적 빈발 패턴 마이닝 기법인 apriori로부터 파생되어 Apriori_msp라 지칭한다. Apriori_msp는 apriori와 같이 단계별로 후보 시퀀스 패턴을 생성하여 사용자로부터 주어진 최소 지지도 임계값과 최소 평균 유효소요시간 임계값을 기준으로 빈발 이동 시퀀스 패턴을 추출한다. 본 연구를 통해 추출된 사용자의 이동 시퀀스 패턴은 새로운 서비스를 위한 정보로 제공될 수 있다. 미술관이나 박물관 등에서 관람객의 이동 패턴을 분석하면 그 패턴에 따라 사용자의 주요 관심이 되는 작품들을 파악할 수 있고, 이를 이용해 전시품을 재배치하거나 또는 휴게실의 효율적인 배치 등에도 이용될 수 있다. 또한 쇼핑몰에서 고객의 이동 패턴에 따라 고객의 주요 이동 경로에 포함되는 제품을 함께 배치함으로써 동선을 줄일 수 있고, 고객들이 반드시 이동하는 장소에 새로운 상품을 소개하는 코너를 설치하는 등 신제품 판매에도 이용할 수 있다. 고객의 이동 패턴과 장바구니 분석을 통한 실제 구매 데이터를 비교한다면 고객이 관심 있게 살펴보는 코너임에도 불구하고 구입된 품목에 포함되지 않은 제품들을 발견할 수 있고 이에 대한 원인 분석이나 문제점 파악에도 이용될 수 있을 것이다.

논문의 구성은 다음과 같다. 2장에서는 시공간 이동 시퀀스 패턴 마이닝과 관련하여 기존의 패턴 마이닝 기법에 관해 살펴보고, 3장에서는 시공간 이동 시퀀스 패턴 마이닝을 위한 전처리 과정에 대해 기술한다. 4장에서는 시공간 이동 시퀀스 패턴 마이닝 기법을 정의하고 실험을 통해 알고리즘의 성능을 평가하며 5장에서는 결론 및 향후 연구에 대해서 기술한다.

2. 관련연구

본 연구와 관련된 기존의 패턴 마이닝에 관한 연구들로 Agrawal *et al.*(1993), Agrawal *et al.*(1994a), Han *et al.*(2000)의 연관 규칙(Association Rules) 마이닝과 Agrawal and Srikant(1994b)의 순차패턴(Sequential Patterns) 마이닝, Yavas *et al.*(2005)의 사용자 이동성 패턴(User Mobility Pattern:UMP) 마이닝 기법 등이 있다. 연관 규칙은 하나의 항목 집합과 다른 항목 집합 사이의 연관성을 나타내는 것으로 “빵을 구매하는 사람은 우유도 구매한다.”와 같은 연관성을 나타내는 규칙을 도출해내는 것이고, 순차패턴 마이닝은 이벤트나 행동의 순차적인 순서를 나타내는 규칙으로 “휴대폰을 구입한 30대 남성이 3개월 이내에 다시 핸드프리를 구입하고, 12개월 이내에 배터리를 구입하는 경향이 있다.”와 같이 정해진 시간 범위 내에서 발생하는 이벤트나 행동 사이의 연관성을 찾아내는 것이다. UMP mining은 PCS(Personal Communication Systems) 네트워크 안에서 사용자의 이동성 예측(mobility prediction)을 위한 이동성 패턴 마이닝 기법이다.

연관 규칙 마이닝에서 연관 규칙을 도출하기 위해서는 먼저 빈발 패턴 마이닝 작업을 수행한다. 이는 본 연구와 유사하나 다음과 같은 문제점들로 본 연구에 적용하기 어려운 문제점이 있다.

첫째, 패턴의 시퀀스 고려이다. 빈발 패턴 마이닝은 장바구니 분석과 같이 트랜잭션 내

에서 빈발하게 발생하는 동일한 항목집합의 패턴을 추출하기 위한 것으로 각 트랜잭션 내에서 발생하는 항목집합의 존재여부에만 관심을 갖는다. 따라서 고객의 이동 경로에 대한 모든 패턴을 고려할 수 없다. 예를 들어, $\langle 1, 2, 3 \rangle$ 과 $\langle 1, 3, 2 \rangle$ 라는 트랜잭션 데이터가 있을 때, 기존의 빈발 패턴 마이닝에서는 둘을 동일 패턴으로 취급한다. 두 개의 패턴 모두 $\{1, 2, 3\}$ 이라는 항목을 포함하기 때문이다. 그러나 본 연구에서는 다른 패턴으로 간주한다. $\langle 1, 2, 3 \rangle$ 은 $1 \rightarrow 2 \rightarrow 3$ 의 순서로 이동한 이동 시퀀스이고 $\langle 1, 3, 2 \rangle$ 는 $1 \rightarrow 3 \rightarrow 2$ 의 순서로 이동한 이동 시퀀스이기 때문이다. 순차패턴 마이닝에서도 시퀀스 패턴에 대한 마이닝을 수행하기는 하지만, 순차패턴의 시퀀스는 구매 제품에 대한 시간차에 따른 제품 아이템의 시퀀스이고 본 연구의 시퀀스는 이동 경로 패턴 분석을 위한 이동 경로 시퀀스이므로 그 의미가 다르다.

둘째, 중복 방문 시퀀스에 대한 반영 여부이다. 앞에서 언급했듯이 기존의 빈발 패턴 마이닝 기법에서는 각 트랜잭션 내에서 발생하는 항목집합의 존재여부에만 관심을 갖는다. 따라서 하나의 트랜잭션 내에 동일한 항목집합이 중복해서 발생하는 경우는 배제된다. 즉 $\langle 1, 2, 3, 2 \rangle$ 라는 트랜잭션 데이터는 발생할 수 없다. 장바구니 데이터 분석에서 만약 고객이 2라는 제품을 2개, 3개 또는 100개를 동시에 구입했다라도 2라는 제품을 구입했다는 사실만이 중요하다. 따라서 이러한 경우 $\{1, 2, 3\}$ 의 아이템셋 집합으로 취급한다. 그러나 본 연구에서는 $\langle 1, 2, 3, 2 \rangle$ 라는 패턴은 $1 \rightarrow 2 \rightarrow 3 \rightarrow 2$ 의 순서로 각 셀을 방문한 이동 시퀀스 경로이기 때문에 이동 경로의 순서가 중요하다. 따라서 방문 셀이 중복해서 나타나는 재방문의 경우도 반드시 고려되어야 한다.

셋째, 고객의 관심도 반영 문제이다. 쇼핑몰 내에서 고객의 이동 시퀀스 경로 패턴을 추출하기 위해 이동 경로 히스토리 데이터를 살펴보면, $\langle 1, 2, 3 \rangle$ 이라는 이동 경로가 발생했을 때 고객이 1, 2, 3이라는 매장 모두에 관심이 있는지 아니면 3이라는 매장을 방문하기 위해 1, 2라는 매장을 지나쳐 갔는지 알 수 없다. 따라서 고객의 관심 정도를 측정하기 위해 소요시간의 개념을 적용한다.

넷째, 각 단계의 후보집합 생성문제이다. 마이닝 수행 단계에서 apriori 속성을 적용하면 $k+1$ 번째 후보집합(C_{k+1}) 생성시 k 번째의 빈발 패턴 (L_k)으로부터 도출하게 된다. 예를 들어, 빈발 패턴 마이닝에서 L_2 가 $\{\{2,3\} \{2,5\} \{3,5\}\}$ 라는 집합이라고 하면, 이에 대한 C_3 는 $\{2, 3, 5\}$ 이다. 이는 이동 시퀀스 패턴 마이닝에서도 동일하다. 그러나 L_2 가 $\{\{2,3\} \{3,5\}\}$ 인 경우, apriori에서는 그 특성에 따라 항목집합-3인 후보집합은 존재하지 않는 반면, 이동 시퀀스 마이닝에서는 L_2 가 $\{\langle 2, 3 \rangle \langle 3, 5 \rangle\}$ 일 때, C_3 는 $\langle 2, 3, 5 \rangle$ 가 될 수 있다. L_2 에서의 $\langle 2, 3 \rangle$ 패턴이 $2 \rightarrow 3$ 으로 이동한 시퀀스이고, $\langle 3, 5 \rangle$ 가 $3 \rightarrow 5$ 로 이동한 시퀀스이므로, 이 두 개의 패턴으로부터 $2 \rightarrow 3 \rightarrow 5$ 로 이동한 이동 경로를 도출할 수 있는 것이다.

Yavas *et al.*(2005)의 UMP 마이닝은 이동 경로 패턴을 추출하기 위한 시퀀스 패턴 마이닝으로 이동성 예측을 위해 corruption 정도를 반영한 기존 연구에서와 다른 새로운 지지도(support) 계산 방법을 사용하였다. 직접(direct) 이동 경로와 간접(indirect) 이동 경로에 대한 지지도 값에 변화를 주어 직접 이동하는 경로에 대해 더 높은 지지도 값을 부여함으로써 다음 경로 예측에 중요도가 더 크게 작용함을 보여주었다. 그러나 본 논문에서는 사용자들이 보다 관심있게 머무르는 위치에 대한 이동 경로 패턴을 추출함을 목표로 하기 때

문에 직접 이동 경로에 대한 패턴만을 반영하며, 트랜잭션 당 중복 방문되는 이동 패턴에 대해 가중치를 부여하는 새로운 방식의 지지도 계산방법을 사용한다.

이 밖에도 관련 연구로 물리적, 시간적으로 유사한 특성을 가지는 그룹을 추출하기 위한 Wang *et al.*(2003), Hwang *et al.*(2005)의 그룹 패턴 마이닝이 있다. Wang(2003)의 AGP와 Hwang *et al.*(2005)의 ATGP는 Agrawal *et al.*(1993, 1994a)의 apriori로부터 Wang(2003)의 VG-growth, Hwang *et al.*(2005)의 TVG-growth는 Han *et al.*(2000)의 FP-growth로부터 파생된 알고리즘이다. 이들 알고리즘에서는 이동 디바이스로부터 얻어진 위치 데이터 로그에서 사용자 이동 데이터를 수집하여 이를 가지고 그룹 패턴을 추출한다. 그룹패턴을 추출하기 위해 최대 물리적 거리 임계값(max_dis)과 최소 지속시간 임계값(max_dur)을 파라미터로 가진다.

3. 시공간 이동 시퀀스 패턴 마이닝을 위한 전처리 과정

시공간 이동 시퀀스 패턴 마이닝을 위해 센서로부터 얻어진 이동 경로 히스토리 데이터를 이용한다. 센서가 구축된 환경에서 사용자가 태그를 가지고 이동함으로써 이동 경로 히스토리 데이터를 얻을 수 있다. 이동 경로 데이터는 센서로부터 수집된 수많은 점(사용자 위치)들의 집합이다. 사용자의 위치 정보는 x, y값으로 이루어진 좌표값과 각 위치에서의 시간정보가 함께 기록되므로 전처리 과정을 거쳐 소요시간을 동반한 형태의 마이닝 가능한 데이터로 변환할 수 있다. 실제 좌표값에 대한 마이닝을 수행한다는 것은 불가능하다. 따라서 본 논문에서는 좀더 효율적인 마이닝 수행을 위해 전체 공간을 셀(cell)이라는 작은 단위로 나누고자 한다. 원본 데이터로부터 전처리 과정을 거치면서 각각의 데이터의 좌표값은 해당 셀의 ID값으로 매핑된다.

표 3.1: 셀 정의 테이블

컬럼명	정의
Cell_ID	셀 아이디
Cell_Name	셀 이름
Min_X	셀 구간의 최소 X좌표값
Max_X	셀 구간의 최대 X좌표값
Min_Y	셀 구간의 최소 Y좌표값
Max_Y	셀 구간의 최대 Y좌표값
valid_time	유효시간 임계값

셀 정의를 위한 테이블은 표 3.1과 같다. 여기서 valid_time은 사용자가 조금이라도 관심을 보이는 데이터들만 마이닝에 반영하고자 하는 소요시간에 대한 임계값이다. 전처리 과정에서 좌표값을 셀 아이디로 매핑한 원본 데이터는 마이닝 과정에서 해당 셀의 유효시간 임계값에 따라 삭제처리 되거나 마이닝을 위한 데이터로 남는다.

그림 3.1과 같이 셀이 물리적으로 분할되어 있고 셀 아이디가 부여되어 있다고 가정하면, 다음과 같은 예제로 전처리 과정을 살펴볼 수 있다.

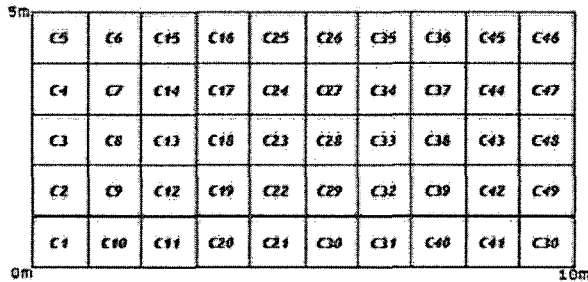


그림 3.1: 셀 분할

그림 3.2의 왼쪽 그림과 같이 사용자 이동 점들의 집합이 있을 때, 이 점들의 집합은 오른쪽 그림과 같이 사용자 아이디, 시간, 위치 정보로 나타낼 수 있다.

전처리 과정은 다음과 같이 두 단계를 거쳐 이루어진다.

[1단계] Cell ID매핑 → 소요시간 계산 → 소요시간 병합

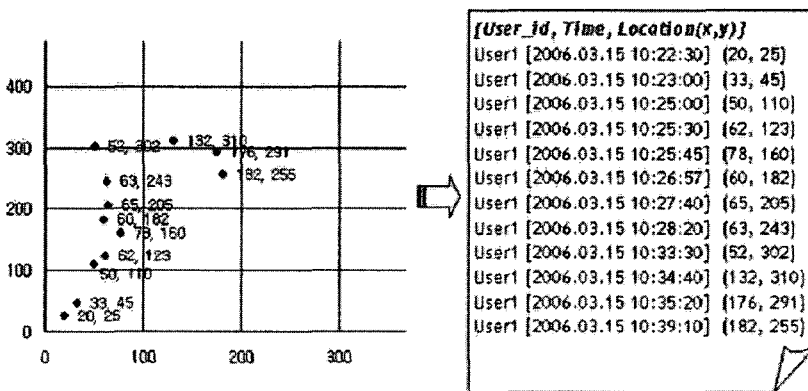


그림 3.2: 이동 경로 히스토리 데이터

첫 번째 단계는 이동 경로 히스토리 데이터로부터 미리 정의된 셀 테이블의 좌표값을 기준으로 해당 셀 아이디 값을 매핑하고, 각 위치의 시간 데이터로부터 소요시간을 계산한다.

$$\text{소요시간} = \text{해당 위치에서 사용자가 머무른 시간}$$

소요시간은 다음 위치의 시간으로부터 이전 위치의 시간을 빼준 값으로 정의한다. 이 때 만약 이전 위치의 셀 아이디와 다음 위치의 셀 아이디 값이 같으면 같은 셀 내에서 이

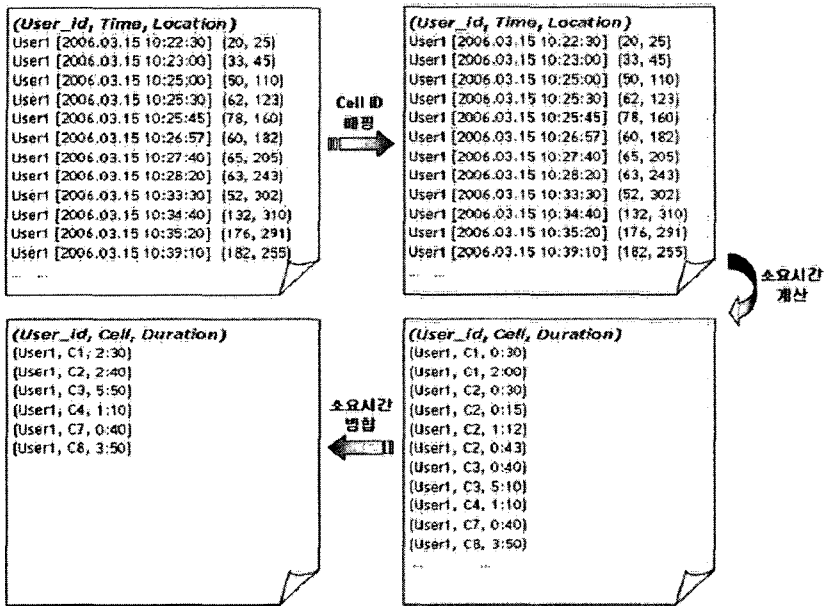


그림 3.3: 전처리 과정 1단계

동한 것으로 간주하고 소요시간을 합하여 준다. 그림 3.3와 같이 전처리 과정 1단계를 거친 User1의 트랜잭션 데이터는 다음과 같은 형태로 압축된다.

User1:<(C1, 2.5) (C2, 2.67) (C3, 5.83) (C4, 1.17) (C7, 0.67) (C8, 3.83)>

다른 사용자 ID의 데이터들도 같은 과정을 거치면 사용자 ID를 기준으로 하는 압축된 트랜잭션 데이터를 얻을 수 있다.

[2단계] 유효시간 filtering과 유효소요시간 산정단계

두 번째 단계는 1단계를 거친 트랜잭션 데이터들을 각 셀의 유효시간 임계값을 기준으로 임계값 이하인 이동 경로 데이터는 삭제 처리한다. 또한 임계값 이상인 데이터에 대해서는 임계값 이상으로 머물렀던 소요시간만을 고객이 관심있게 머물렀던 유효소요시간으로 간주하여 시간차이를 계산하여 트랜잭션 데이터를 재구성한다. 이는 각 셀에서의 소요시간을 평준화(equalization)하기 위한 것이다.

고객이 마트에서 제품을 구매할 때, 생선이나 고기를 구매할 때는 마트의 직원에게 생선의 종류나 고기의 부위 또는 신선도 등을 물어보고 구매하게 된다. 또한 과일이나 채소류의 경우도 크기, 신선도, 생산지 등을 모두 고려하여 구매여부를 결정한다. 그러나 가공식품의 경우는 대부분이 자신이 원하는 제품인지의 여부나 가격 등만을 보고 바로 구매여부를 결정한다. 또한 옷이나 신발 등의 경우는 입어보고 신어보고 여러 차례 다른 제품과 비교한 후 구매한다. 즉 어떤 제품을 판매하는 매장인가에 따라 고객이 구매하는데 걸리는 최소한의 시간, 또는 고객이 구매 의사를 표시하는데 걸리는 최소한의 시간이 다르다. 이처럼 각 매장별 특성을 고려하지 않는다면, 항상 소요시간이 오래 걸리는 특정 매장에 해당

하는 셀들이 포함된 이동 패턴만이 추출될 것이다.

따라서, 이를 반영하기 위한 요소로 각 셀에 대한 유효시간 임계값을 두고, 이를 이용한 유효소요시간이라는 개념을 다음과 같이 정의한다.

$$\begin{aligned} \text{유효시간 임계값} &= \text{고객의 구매활동을 위해 최소한으로 요구되는 요구시간} \\ \text{유효소요시간} &= \text{전체소요시간} - \text{유효시간 임계값} \end{aligned}$$

각 셀 별로 고객이 구매활동을 하는데 요구되는 최소한의 요구시간을 유효시간 임계값으로 정한다. 이는 고객들이 단순한 이동 경로로 지나치게 되는 이동 경로 패턴을 배제하고 고객들이 최소한의 관심도 보이지 않았을 것으로 추정되는 데이터를 배제함으로써 마이닝의 효율성을 높이고자 하는 것이다. 또한 유효시간 임계값 이상으로 머무른 시간 데이터를 실제 의미 있는, 사용자가 관심있게 살펴보는 시간으로 간주하여 고객의 관심도 적용에 유효한 유효소요시간으로 정의한다.

예제 3.1: 2단계 과정은 다음과 같은 예로 살펴볼 수 있다. 표 3.2의 셀 정의 테이블의 각 유효시간 임계값을 기준으로 전처리 과정 2단계를 수행하면 그림 3.4과 같은 결과를 얻을 수 있다.

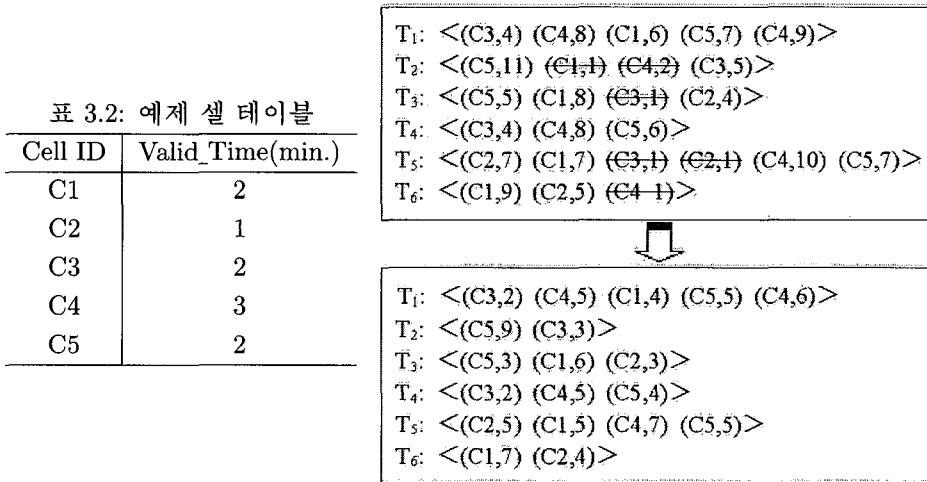


그림 3.4: 전처리 과정 2단계

위의 예제에서 T2를 살펴보면 T2:<(C5,11) (C1,1) (C4,2) (C3,5)>에서 시퀀스<C1>과 시퀀스<C4>에서의 소요시간은 유효시간 임계값 미만이므로 삭제되고 시퀀스<C5>와 시퀀스<C3>에 대해서는 임계값을 기준으로 트랜잭션 데이터를 수정한다.

$$T2:<(C5,11-2) (C1,1) (C4,2) (C3,5-2)> \Rightarrow T2:<(C5,9) (C3,3)>$$

따라서, T2는 두 개의 셀을 지나는 이동 경로 시퀀스인 <(C5,9) (C3,3)>로 수정된다. 이와 같이 전처리 과정을 통해 변환된 데이터를 가지고 시공간 이동 시퀀스 패턴 마이닝을 수행한다.

4. 시공간 이동 시퀀스 패턴 마이닝

이동 시퀀스 패턴을 추출하기 위해서 이동 경로 내의 모든 이동 시퀀스를 고려하고자 재방문 데이터를 반영한 새로운 지지도 계산 방법을 제안하고, 사용자의 관심도를 잘 반영하기 위한 요소로 유효시간과 유효소요시간의 개념에 대해서 소개한다. 또한 이를 적용한 시공간 이동 시퀀스 패턴 마이닝 알고리즘에 대해서도 기술한다.

4.1. 중복방문 시퀀스 패턴을 고려한 지지도 계산방법

기존의 패턴 마이닝 연구들은 공통적으로 패턴 추출을 위한 지지도 값으로 각 트랜잭션 내에 특정 항목이 존재하는지 존재하지 않는지의 여부만을 고려한다. 따라서 apriori나 fp-growth와 같은 기법에서는 하나의 트랜잭션 안에 중복 항목 발생을 아예 배제하거나 UMP 마이닝에서는 중복 발생되었다더라도 하나의 카운트 값으로 처리한다. 그러나 본 연구에서는 사용자 이동 경로의 모든 순서를 고려한 패턴 추출을 위해 중복되는 이동 경로들을 반드시 고려해야만 한다. 사용자가 1부터 5라는 곳으로 이동하는데 1, 2, 3, 4, 5의 경로를 따라 이동한 것과 1, 4, 2, 3, 5의 경로를 따라 이동한 것과는 의미가 다르다. 또한 1, 3, 4, 3, 5와 같이 방문했던 곳을 또 다시 방문할 수도 있기 때문에 중복 방문 시퀀스가 고려되어야만 정확한 이동 시퀀스 패턴을 추출할 수 있다.

[정의 1] 중복 방문 가중치(W)의 요건 : 한 트랜잭션 내에서 아무리 많이 중복 방문했다더라도 다른 트랜잭션 내에서 방문한 경우보다는 낮은 값으로 처리되어야 한다. 따라서 중복 방문에 대한 가중치와 이를 적용한 지지도 계산식은 다음과 같이 정의한다.

$$0 \leq \text{중복 방문 가중치}(W) \leq 1$$

$$\text{중복 방문 가중치}(W) = \frac{\text{트랜잭션 내 중복 방문 횟수}}{\text{트랜잭션 내 전체 후보 시퀀스 패턴 집합 수}}$$

$$\text{트랜잭션 당 시퀀스 패턴별 지지도} = 1 + W \text{ (방문횟수} > 0)$$

$$\text{전체 시퀀스 패턴별 지지도} = \sum(\text{트랜잭션 당 시퀀스 패턴별 지지도})$$

예제 4.1:

[트랜잭션ID: <(셀 ID로 이루어진 이동 시퀀스 경로)>]
T ₁ : <C1 C8 C5 C8 C5 C2 C8 C5 C6 >
T ₂ : <C7 C9 C2 C3 >
T ₃ : <C8 C2 C9 C7 C6 C9 C4 C1 >

예를 들어 위와 같은 트랜잭션들이 존재할 때 시퀀스 셀-1인 시퀀스 집합들의 지지도 값을 주어진 식에 따라 계산하면 표 4.1과 같다.

표 4.1: 시퀀스 패턴별 지지도(방문횟수)

시퀀스 패턴	시퀀스 패턴별 지지도(방문횟수)
C1	$1_{T1} + 1_{T3} = 2$
C2	$1_{T1} + 1_{T2} + 1_{T3} = 3$
C3	$1_{T2} = 1$
C4	$1_{T3} = 1$
C5	$(1 + 3/9)_{T1} = 1.33$
C6	$1_{T1} + 1_{T3} = 2$
C7	$1_{T2} + 1_{T3} = 2$
C8	$(1 + 3/9)_{T1} + 1_{T3} = 2.33$
C9	$1_{T2} + (1 + 2/8)_{T3} = 2.25$

그러면 중복 방문을 고려한 지지도 값을 어떻게 적용해야 할 것인가? 하나의 이동 경로 트랜잭션 내에서 중복 이동 시퀀스 패턴이 발생한 경우 이를 똑같이 하나의 방문 횟수로 처리하면 특정 트랜잭션에 치우친 결과를 초래할 수 있다.

<C5>, <C8>, <C9>는 중복 방문으로 중복 방문 가중치가 적용되었음을 볼 수 있다. <C8>을 예를 들어 살펴보면, T₁에서 세 번 중복해서 방문하였고, T₃에서 한 번 방문했으므로,

T₁의 <C8> 지지도: $1 + 3/9$

$[(T_1\text{에서 } \langle C8 \rangle \text{의 중복 방문 횟수}) / (T_1\text{내에 전체 시퀀스 패턴 후보집합 수})]$

T₃의 <C8> 지지도: 1(한 번만 방문)이다.

따라서 시퀀스 패턴 <C8>에 대한 지지도 값은 2.33이다.

4.2. 관심도 적용을 위한 유효소요시간

기존의 빈발 패턴 마이닝에 관한 연구들에서는 빈발 패턴을 추출하기 위한 기준으로 각 패턴의 발생 횟수만을 고려한다. 그러나 본 연구에서는 사용자들이 관심을 가지고 주로 이동하는 이동 시퀀스 패턴을 추출하고자 하므로 각 위치의 방문 횟수만으로는 사용자의 관심도를 정확히 반영할 수 없다.

쇼핑몰이나 박물관 같은 곳에서 고객들이 D라는 코너에 관심이 있다고 하자. 그런데 D를 가기 위해서는 반드시 A, B, C 코너를 지나쳐야만 한다. 이 때 각 코너의 방문 횟수만을 고려하여 패턴을 추출 한다면 D를 가기 위해 거쳤던 A, B, C 코너의 횟수가 높게 측정될 것이고, 상대적으로 실제 고객들이 관심을 가지는 D의 방문 횟수는 낮으므로 제거될 확률이 높다. 또한 건물 구조상의 설계 문제로 인해 E라는 코너를 가기 위해 고객들이 전혀 관심이 없는 F, G, H라는 코너를 거쳐 가야 한다면 이 코너들 역시 빈발 패턴으로 포함될 것이다.

고객들의 이동 경로 탐색 능력 또한 문제가 될 수 있다. 만약 동일한 쇼핑몰을 자주 방문하는 고객이 있다면 그 고객은 어떤 코너에 어떤 품목이 있는지 바로 찾아가서 구매할 수 있을 것이다. 그러나 그렇지 않은 경우 고객들은 원하는 코너로 이동하기 위해 잘못된 길

을 해마다 그 코너로 이동할 수도 있다. 이러한 경우 방문 횟수로만 관심도를 측정한다면 실제 관심도와 다른 왜곡된 패턴 결과를 추출하게 될 것이다.

따라서 본 연구에서는 이러한 문제점을 보완하기 위해 소요시간의 개념을 적용하고자 한다. 소요시간은 3장에서 언급한 바와 같이 고객이 해당 위치에서 머물렀던 시간으로 전처리 과정을 통해 고객이 관심있게 머물렀던 의미 있는 시간값인 유효소요시간 값을 다음과 같이 적용한다.

예제 4.2: 전처리 과정을 통해 계산된 유효소요시간을 적용한 트랜잭션 데이터베이스는 다음과 같은 형태이다.

[트랜잭션ID: <셀 ID, 유효소요시간>으로 이루어진 이동 시퀀스 경로>
T ₁ : <(C1,2) (C8,2) (C5,3) (C8,3) (C5,1) (C2,4) (C8,4) (C5,2) (C6,4)>
T ₂ : <(C7,1) (C9,1) (C2,3) (C3,2)>
T ₃ : <(C8,4) (C2,2) (C9,4) (C7,2) (C6,1) (C9,3) (C4,6) (C1,5)>

고객1이 이동한 이동 경로(T₁)는 C1→C8→C5→C8→C5→C2→C8→C5→C6이고 각 패턴에 대한 유효소요시간은 <C1>에서 2분, <C8>에서 2분, <C5>에서 3분, 다시 <C8>에서 3분 …… , <C6>에서 4분이다. 위의 예로 시퀀스 셀-1 집합들의 평균 유효소요시간을 구하면 표 4.2와 같다.

표 4.2: 시퀀스 패턴별 지지도(방문횟수)

시퀀스 패턴	시퀀스 패턴별 평균 유효 소요 시간
C1	$(2T_1 + 5T_3)/2 = 3.5$
C2	$(4T_1 + 3T_2 + 2T_3)/3 = 3$
C3	$2T_2/1 = 2$
C4	$6T_3/1 = 6$
C5	$(3 + 1 + 2)T_1/3 = 2$
C6	$(4T_1 + 1T_3)/2 = 2.5$
C7	$(1T_2 + 2T_3)/2 = 1.5$
C8	$((2 + 3 + 4)T_1/3 + 4T_3)/2 = 3.5$
C9	$(1T_2 + ((4 + 3)T_3)/2 = 2.25$

이 때 만약 마이닝을 위한 주요 패턴 추출의 기준으로 평균 유효소요시간만을 고려한다면, 최소 평균 유효소요시간을 3이라고 주었을 때, <C1>, <C2>, <C4>, <C8>이 주요 패턴에 포함될 것이다. 그러나 <C4>의 경우 하나의 이동 경로 트랜잭션(6T₃)에서만 한 번 방문한 데이터 임에도 불구하고 평균 유효소요시간이 가장 길기 때문에 패턴에 포함될 수밖에 없다. 이와 같은 경우 특정 트랜잭션 값에 치우치는(특정 고객의 성향에 치우치는) 왜곡된 결과를 초래할 수도 있다.

따라서 각 시퀀스 패턴의 방문 횟수나 평균 유효소요시간 중 어느 하나만을 기준으로 해서는 정확한 패턴을 추출할 수 없다.

4.3. Apriori_msp 알고리즘

본 절에서는 데이터 전처리 과정을 통해 얻은 트랜잭션 데이터를 가지고 시공간이동 시퀀스 패턴 마이닝(Spatial-Temporal Moving Sequence Pattern Mining)을 수행한다. 기본적으로 시공간 이동 시퀀스 패턴 마이닝은 apriori와 유사한 과정을 거치나 4.1절과 4.2절에서 언급한 중복방문 시퀀스 패턴을 고려한 지지도 계산 방법과 관심도 적용을 위한 유효소요시간을 반영하여 apriori의 변형된 형태인 apriori_msp로 정의한다.

apriori_msp가 apriori와 가장 크게 다른 점은 후보집합 생성문제에 있다. ‘모든 공집합이 아닌 빈발 항목집합의 부분집합은 반드시 빈번하다.’는 apriori 속성에 따르면 {1, 2}라는 집합이 빈번하면 이에 부분집합인 {1}, {2}도 또한 빈번하다고 할 수 있다. 반대로 {1}, {2}가 빈번하지 않으면 {1, 2} 집합은 빈번하지 않다. 따라서 length-3 후보집합을 생성시 {1, 2}, {2, 3}, {1, 3}이라는 세 개의 패턴이 빈발해야만 {1, 2, 3}이라는 length-3 후보집합을 생성할 수 있다.

```

apriori_msp_gen(Lk)
{
  for each l1 ∈ Lk
    for each l2 ∈ Lk
      if (l1[1] = l2[0]) ∧ (l1[2] = l2[1]) ∧ ⋯ ∧ (l1[k-1] = l2[k-2]) then {
        c = l1 ∪ l2
        add c to Ck+1
      }
  return Ck+1
}
    
```

그림 4.1: apriori_msp_gen 함수

그러나 apriori_msp에서는 주어진 패턴이 이동 시퀀스 패턴이기 때문에 <1 2>, <2 3> 두 개의 패턴만 존재하더라도 시퀀스 셀-3 후보집합에서 <1 2 3>이라는 후보 시퀀스 패턴을 생성할 수 있다. 각 단계에서 후보집합 생성을 위한 함수로 그림 4.1의 apriori_msp_gen을 호출한다.

그림 4.2의 apriori_msp는 이러한 속성을 기반으로 시퀀스 셀-1후보 시퀀스 집합으로부터 시작해서 각 단계별로 이동 시퀀스 빈발 패턴(large pattern) 집합을 구한다. 후보집합 C_{k+1}은 그 이전 단계의 빈발 패턴 집합인 L_k로부터 생성된다. 각 단계에서는 생성된 후보 항목집합들(C_k)에 대해, 4.1절과 4.2절에서 제안한 계산 방법에 따라 중복 방문 가중치가 고려된 지지도 계산 방법과 평균 유효소요시간을 계산한다. 계산된 지지도 값과 평균 유효소요시간이 주어진 최소 지지도 임계값(min_supp)과 최소 평균 유효소요시간(min_time)을 만족하면 그 패턴은 빈발한 이동 시퀀스 패턴으로 간주한다. 여기서 지지도 값은 얼마나 빈번하게 발생하였는지 빈번함의 정도를 나타내고, 평균 유효소요시간은 고객이 얼마만큼 관심있게 머물렀는가 하는 관심도의 정도를 의미한다.

```

apriori_msp()
Input: 1단계 마이닝 과정을 거친 트랜잭션 데이터베이스 D
        최소 지지도 임계값 (min_supp)
        최소 평균 유효소요시간 임계값 (min_time)
Output: 이동 시퀀스 빈발 패턴 집합  $\bar{L}$ 
 $C_1$  // 시퀀스 셀이 1인 후보 시퀀스 패턴 집합
 $k = 1$  // 각 단계에서의 시퀀스 셀의 개수
 $L = \emptyset$  // 이동 시퀀스 빈발 패턴 집합
while  $C_k \neq \emptyset$  {
  for each 이동 경로 시퀀스 트랜잭션  $t \in D$  {
     $C_k$ 의 각각의 후보 항목 집합( $s \in C_k$ )에 대해 D를 스캔
     $s.count$ (지지도)와  $s.avg\_time$ (평균 유효소요시간)을 계산
    if ( ( $s.count \geq min\_supp$ ) && ( $s.avg\_time \geq min\_time$ ) )
      add  $s$  to  $C_{k+1}$ 
  }
   $\bar{L} = \bar{L} \cup L_k$ 
  apriori_msp_gen( $L_k$ )
   $k = k+1$ ;
}
return  $\bar{L}$ ;

```

그림 4.2: apriori_msp 알고리즘

자세한 패턴 마이닝 수행 과정은 다음과 같은 간단한 예제로 살펴본다.

예제 4.3:

```

T1: <(C3,2) (C4,5) (C1,4) (C4,6) (C5,5) >
T2: <(C5,9) (C3,3) >
T3: <(C5,3) (C1,6) (C2,3) >
T4: <(C3,2) (C4,5) (C5,4) >
T5: <(C2,5) (C1,5) (C4,7) (C5,5) >
T6: <(C1,7) (C2,4) >

```

전처리 과정을 거친 위와 같은 트랜잭션 데이터베이스가 있고, 사용자로부터 주어진 임계값은 $min_supp = 2$, $min_time = 4$ (min.)이다.

주어진 데이터베이스를 트랜잭션 별로 간단히 도식화하면 그림 4.3와 같다. 각 노드는 셀을 나타내고 각 셀은 평균 유효소요시간을 값으로 갖는다. 먼저 시퀀스 셀-1 후보 시퀀스 패턴들을 기준으로 지지도와 평균 유효소요시간을 구하면 표 4.3과 같고, 이들 집합들 중에 주어진 최소 지지도 임계값과 최소 평균 유효소요시간에 만족되는 항목은 시퀀스 패턴 <C1>, <C2>, <C4>, <C5>이다.

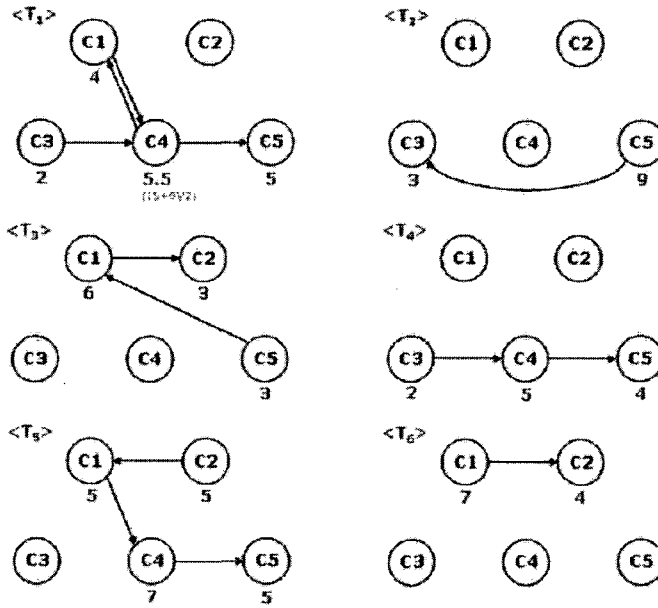


그림 4.3: 각 트랜잭션 별 이동 시퀀스 경로

표 4.3: C_1 : 시퀀스 셀-1 후보 시퀀스 집합

시퀀스 패턴	지지도(방문횟수)	평균 유효소요시간
<C1>	4	5.5
<C2>	3	4
<C3>	3	2.33
<C4>	3.4	5.83
<C5>	5	5.2

C_1 으로부터 추출된 $L_1 = \{<C1>, <C2>, <C4>, <C5>\}$ 으로부터 C_2 후보 시퀀스 패턴 집합을 구하면 표 4.4와 같다. 같은 방식으로 지지도 카운트와 평균 유효소요시간을 구하면 주어진 조건에 만족하는 이동 시퀀스 빈발 패턴 $L_2 = \{<C1 C2>, <C1 C4>, <C4 C5>\}$ 이다.

표 4.4: C_2 : 시퀀스 셀-2 후보 시퀀스 집합

시퀀스 패턴	지지도(방문횟수)	평균 유효소요시간
<C1 C2>	2	5
<C1 C4>	2	5.5
<C1 C5>	0	0
<C2 C1>	1	5
<C2 C4>	0	0
<C2 C5>	0	0
<C4 C1>	1	4.5
<C4 C2>	0	0
<C4 C5>	3	5.33
<C5 C1>	1	4.5
<C5 C2>	0	0
<C5 C4>	0	0

표 4.5: C_3 : 시퀀스 셀-3 후보 시퀀스 집합

시퀀스 패턴	지지도(방문횟수)	평균 유효소요시간
<C1 C4 C5>	2	5.33

L_2 로부터 생성된 후보 집합 C_3 는 표 4.5와 같고, 최소 지지도 임계값과 최소 평균 유효 소요시간 임계값을 만족하므로, 그림 4.4의 예제 트랜잭션 데이터베이스로부터 추출된 모든 이동 시퀀스 빈발 패턴 $L = \{<C1>, <C2>, <C4>, <C5>, <C1 C2>, <C1 C4>, <C4 C5>, <C1 C4 C5>\}$ 이다. L_3 가 하나의 집합만 존재하므로 더 이상 후보집합은 생성되지 않는다.

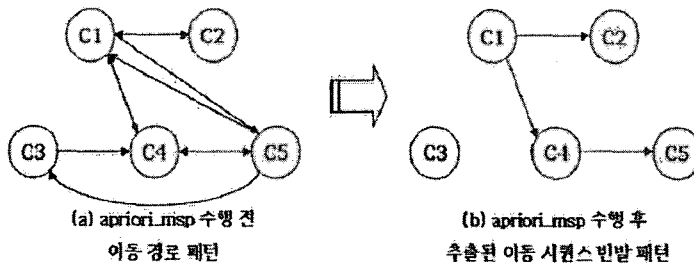


그림 4.4: apriori_msp 수행 전과 후

4.4. 성능평가

본 절에서는 시공간 이동 시퀀스 패턴 마이닝 기법의 구현과 성능평가에 대해서 살펴본다. 최소지지도 임계값(min_supp), 최소 유효소요시간 임계값(min_time)의 변화에 따른 실

행 결과를 비교하고, 데이터베이스의 크기, 트랜잭션 내 평균 방문 셀의 개수, 전체 셀의 개수 등에 따른 실행 결과를 알아본다.

모든 실험은 Windows XP 운영체제에 1GB 램(RAM)을 가진 2.40GHz 펜티엄 컴퓨터에서 수행되었으며, Java로 구현되었다. 실험을 위한 데이터는 다음과 같은 사실을 토대로 생성되었다. 데이터베이스 생성을 위한 파라미터 값은 표 4.6과 같고, 이를 기준으로 생성된 실험데이터는 표 4.7과 같다.

표 4.6: 데이터베이스 생성 파라미터

<i>D</i>	데이터베이스 내의 전체 트랜잭션수
<i>N</i>	한 트랜잭션 내 평균 방문 셀의 개수
<i>C</i>	전체 셀의 개수
<i>T</i>	최대 가능 유효 소요 시간

표 4.7: 실험을 위해 생성된 데이터

실행 데이터	<i>D</i>	<i>N</i>	<i>C</i>	<i>T</i>	Size
<i>D</i> 10_ <i>N</i> 10_ <i>C</i> 10_ <i>T</i> 10	10	10	10	10	1KB
<i>D</i> 100_ <i>N</i> 10_ <i>C</i> 10_ <i>T</i> 10	100	10	10	10	7KB
<i>D</i> 1000_ <i>N</i> 10_ <i>C</i> 10_ <i>T</i> 10	1000	10	10	10	63KB
<i>D</i> 10000_ <i>N</i> 5_ <i>C</i> 10_ <i>T</i> 10	10000	5	10	10	343KB
<i>D</i> 100000_ <i>N</i> 10_ <i>C</i> 5_ <i>T</i> 10	100000	10	5	10	634KB

그림 4.5에서는 기존의 빈발 패턴 마이닝에서 대표가 되는 두 알고리즘 apriori와 fp-growth를 apriori_msp와 비교하였다. 대상 데이터는 전체 트랜잭션이 10000개인 *D*10000_*N*10_*C*10_*T*10이다. apriori와 fp-growth에서는 유효소요시간의 개념이 존재하지 않으므로 이에 적합한 형태로 변형하여 수행하였다. 최소 지지도 임계값 변화에 따른 결과값으로 왼쪽 그림은 수행시간, 그림 4.6는 추출된 패턴 수에 대한 그래프이다.

그림 4.5와 4.6에서 볼 수 있듯이 이 apriori_msp는 모든 이동 순서 패턴을 고려하는 만큼 apriori나 fp-growth보다는 수행시간이 다소 오래 걸리는 경향이 있으나, 실행 결과 추출된 패턴은 현저히 줄어들 수 있다. 이는 앞에서도 언급한 것처럼 이동 순서를 고려하여 <1 2 3>과 <3 2 1>을 다른 패턴으로 간주하기 때문에 이를 같은 {1 2 3}의 아이템셋으로 취급하는 빈발 패턴 마이닝 달리 상대적으로 많은 후보 패턴이 고려되고, 또한 그 만큼 패턴별 지지도 값이 낮아지게 된다. 따라서 오른쪽 그림에서처럼 apriori_msp는 apriori와 fp-growth보다 낮은 지지도에서 많은 패턴이 추출됨을 알 수 있다.

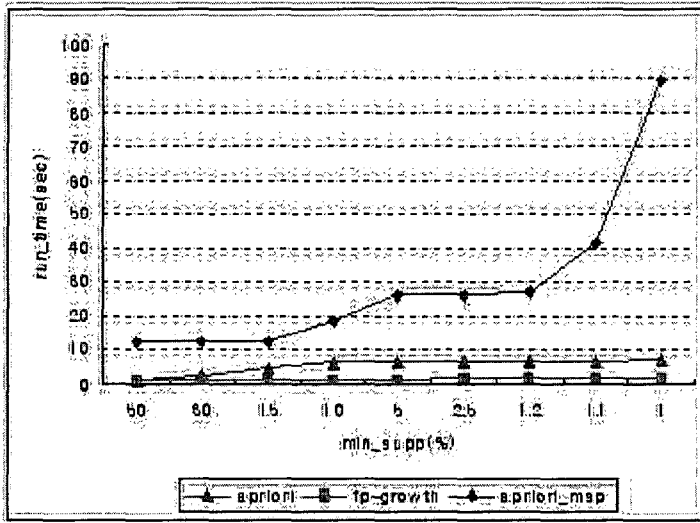


그림 4.5: apriori, fp-growth, apriori_msp 비교
최소 지지도 임계값 변화에 따른 수행시간 수행 전과 후

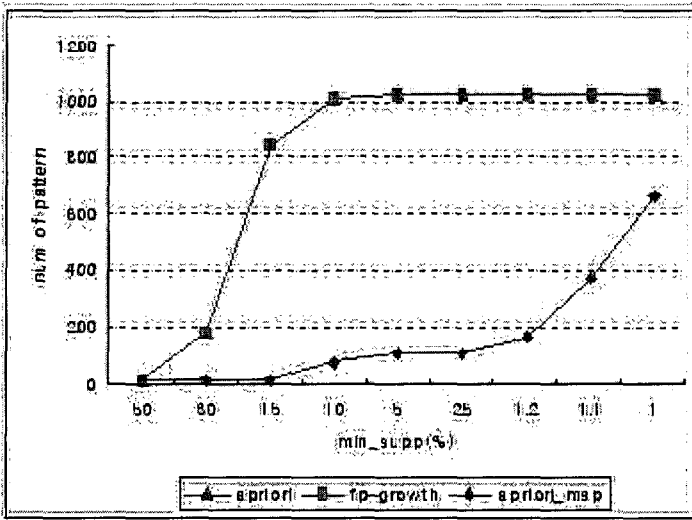


그림 4.6: apriori, fp-growth, apriori_msp 비교
최소 지지도 임계값 변화에 따른 추출된 패턴의 수

즉 apriori_msp 마이닝은 마이닝의 각 단계에서 이동 시퀀스 패턴의 가능한 모든 후보 집합을 고려하기 때문에 순서를 고려하지 않고 발생횟수만을 고려하는 기존의 빈발 패턴 마이닝의 수행보다는 실행 시간이 좀 더 걸리는 경향이 있으나, 실행 결과 방문횟수와 유

효소요시간의 적용에 따라 이동 순서를 고려한 정확한 패턴이 추출되었음을 살펴볼 수 있었다.

5. 결과 및 향후 연구

본 논문에서는 시공간 이동 시퀀스 패턴 마이닝 기법(apriori_msp mining)을 제안하였다. apriori_msp 마이닝은 일정한 공간 안에서 사용자가 관심있게 이동하는 이동 경로 패턴을 분석하는 마이닝 기법이다.

기존의 빈발 패턴 마이닝 기법에서는 빈발하게 발생하는 아이템 셋을 구하기 위해 트랜잭션 내에서 나타나는 아이템의 존재 여부만을 고려한 마이닝을 수행하였고, 순차패턴 마이닝 기법에서는 특정 시간 내에 발생하는 아이템의 시퀀스를 고려하기는 했지만, 이는 이동 시퀀스의 개념과 다르다. 또한 이동 시퀀스 패턴 마이닝을 통해 이동성을 예측하는 UMP마이닝에서는 사용자의 관심도와 중복방문에 대한 개념이 전혀 고려되지 않고 있다. 따라서 기존의 연구들과는 다른 사용자의 모든 방문 시퀀스를 고려한 이동 시퀀스 패턴 마이닝 알고리즘을 제안하였다. 마이닝 과정에서 중복 방문에 대한 패턴을 고려하기 위해 중복 방문 가중치를 적용한 새로운 지지도 계산방법을 도입하였고, 사용자들이 관심있게 이동하는 경로를 파악하기 위하여 유효시간 임계값과 유효소요시간의 개념을 사용하였다. 즉 apriori_msp 마이닝에서는 기존의 패턴 마이닝에서 고려하지 않았던 이동 시퀀스를 고려하였다는 점과 해당 위치에 대해 유효소요시간이라는 시간정보를 덧붙여 위치의 중요도를 나타내는 가중치를 부여하였다는 점에서 이동 시퀀스 패턴을 추출하는데 대한 기존 마이닝 기법의 문제점들을 보완하였다.

참고문헌

- Agrawal, R., Imieliński, T. and Swami, A. (1993). Mining Association Rules between Sets of Items in Large Databases, *Proc. Conf. on Management of Data*, ACM Press, New York, NY, USA.
- Agrawal, R. and Srikant, R. (1994). Fast Algorithms for Mining Association Rules, *Proc. of Int'l Conf. on Very Large Data Bases(VLDB)*.
- Agrawal, R., Srikant, R. (1994). Mining Sequential Patterns, *IBM Research Report RJ9910*, IBM Almaden Research Center. Oct.
- Han, J., Pei, H. and Yin, Y. (2000) Mining Frequent Patterns without Candidate Generation, *Proc. Int'l Conf. on Management of Data (ACM SIGMOD 2000, Dallas, TX)*. ACM Press, New York, NY, USA.
- Hwang, S., Liu, Y., Chiu, J. and Lim, E. (2005), Mining Mobile Group Patterns: A Trajectory-Based Approach, *Proc. of Int'l Conf. on Pacific Asia Knowledge Discovery in Databases*, 713-718.
- Wang, Y., Lim, E. and Hwang, S, (2003). On Mining Group Patterns of Mobile Users, *Proc. of Int'l Conf. on Database and Expert Systems Applications*, 287-296.

Yavas, G., Katsaros, D., Ulusoy, O. and Manolopoulos, Y. (2005). A data mining approach for location prediction in mobile environments, *Data & Knowledge Engineering*, **54**, 121-146.

[2006년 7월 접수, 2006년 8월 채택]

Spatial-Temporal Moving Sequence Pattern Mining*

Seon-Young Han¹⁾ Hwan-Seung Yong²⁾

ABSTRACT

Recently many LBS(Location Based Service) systems are issued in mobile computing systems. Spatial-Temporal Moving Sequence Pattern Mining is a new mining method that mines user moving patterns from user moving path histories in a sensor network environment. The frequent pattern mining is related to the items which customers buy. But on the other hand, our mining method concerns users' moving sequence paths. In this paper, we consider the sequence of moving paths so we handle the repetition of moving paths. Also, we consider the duration that user spends on the location. We proposed new Apriori_msp based on the Apriori algorithm and evaluated its performance results.

Keywords: Spatial-Temporal Data Mining, Moving Sequence, Pattern Mining

* This work was supported by grant No. (R01-2006-000-10609-0) from the Basic Research Program of the Korea Science and Engineering Foundation.

1) Developer, NHN Corp., 25-1 Jeongja-Dong, Bundang-Gu, Seongnam, Gyeonggi-Do 463-844, Korea
E-mail : syhan@ewhain.net

2) (Corresponding author) Professor, Department of Computer Science
Ewha Womans University, Seoul 120-750, Korea
E-mail : hsyong@ewha.ac.kr