사용자 행동 데이터의 시퀀스 패턴 마이닝 기술 동향
Technology trend on sequential pattern mining of user behavior data

임지연(한국전자통신연구원 휴먼증강연구실)

차 례

■ keyword : | 데이터 마이닝, 라이프로그 데이터, 시퀀스 데이터 |

1. 시퀀스 데이터란?

1.1 시퀀스 데이터의 의미와 종류

시퀀스 데이터는 특정기간 동안 대상의 행동을 나타내는 선후 관계가 있는 항목들의 집합이다. 이러한 시퀀스 데이터는 시간 정보를 포함하는 데이터를 수집하고 저장하여 활용하거나 분석하는 다양한 분야에서 활용된다. 마케팅 분야에서는 고객이 쇼핑을 하는 순서를 분석하여 물건을 배치하고, 의료 분야에서는 환자가 받는 치료 및 처방 순서를 이용해 특정 질병의 증상을 알아내기도 한다. 다량의 데이터가 발생하여 데이터베이스 구축하여수집하는 경우를 살펴보면, 지진 기록, 기상 등 시공간(spatio-temporal) 데이터, 로봇의 동작과 같은 엔지니어링 프로세스 데이터, 주식 시장의 거래 정보 데이터, 전화를 걸고 받는 패턴, 사용자들의 클릭 스트림과 같은 웹로그, 프로그램 실행 순서 등을 들 수 있다. 또, 생명 과학 분야에서는 DNA 시퀀스들과 유전자 표현 및 구조 데이터를 분석하기 위한 노력을 오랫동안 기울여 왔다.[1]

웹에서 수집된 사용자의 서비스 이용 트랜잭션 데이터들과 같은 시퀀스 데이터는 통상적으로 그 크기가 매우 큰 경우가 많다.

1.2 시퀀스 데이터의 정의

시퀀스 데이터는 아래와 같이 정의할 수 있다. $I=\{i_1,i_2,i_3,...,i_n\}$를 항목 집합이라고 하자. 이때, 항목집합 X는 I의 부분 집합이고 $X \subseteq I$과 같이 나타낼 수 있다. 시퀀스는 항목(또는 원소나 이벤트)의 순서가 있는 리스트로 이루어진 집합이다. 항목들은 알파벳 순으로 정렬된 순서가 없는 원소로 정의할 수 있으며, 항목을 구성하는 원소들은 중복을 허용하지 않는다. 그러나 시퀀스의 하나의 항목을 구성하는 원소가 다른 항목에 중복되어 나타는 것은 가능하다. 이 때, 시퀀스를 구성하는 항목의 개수는 시퀀스의 길이가 된다. 시퀀스 길이가 1인 경우 $1-sequence$ 라고 명명한다. 예를 들어 $s=\langle a(ce)(bd)(bcde)f(dg)\rangle$의 시퀀스는 7개의 서로 다른 항목으로 이루어져 있으며 6개의 원소로로 이루어져 있다. 이때, 시퀀스의 길이는 12이다.

시퀀스 데이터베이스에는 시퀀스들이 그룹으로 묶여 저장되어 있다. 만약 시퀀스 s가 t의 부분시퀀스(subsequence)이고, s가 t에서 and나 or로 중복되는 항목들을 제거한 전사 (projection)라고 가정하자. 그러면 $\langle a(c)(bd)f\rangle$는 s의 부분 시퀀스가 될 수 있다. 또, $k=2,3...n$이고 $j_k-j_{k-1} \le \delta$ 와 $s_1 \subseteq t_{j1}, s_2 \subseteq t_{j2}, s_3 \subseteq t_{j3}, ...s_n \subseteq t_{jn}$를 만족하는 정수 $j_1 < j_2 < ... < j_n$가 존재한다고 할 때, 시퀀스 s가 t의 길이 δ 인 부분 시퀀스라고 하자. 그렇다면 t 내부의 인접한 항목(adjacent element)들이 가질 수 있는 최대 길이는 δ 가 된다.

2. 시퀀스 패턴 마이닝 기법 소개 및 적용 분야

2.1 시퀀스 패턴 마이닝 개요

시퀀스 패턴 p가 주어졌을 때 패턴 p의 지지도값(support)은 데이터베이스에 존재하는 패턴 p를 포함한 시퀀스의 개수이다. 지지도 임계치 min_sup 이상의 값을 지지도로 가지는 패턴은 빈발 패턴 또는 빈발 시퀀스 패턴이라고 부른다. 또, 길이가 l인 패턴은 $l-pattern$으로 표시한다. 이처럼 시퀀스 패턴 마이닝은 주어진 시퀀스들의 집합에서 완전한 빈발 부분 시퀀스 집합(complete set of frequent subsequences)을 탐색하는 방

법이다.

시퀀스 패턴 마이닝 알고리즘이 제대로 작동하기 위해 갖추어야 하는 조건은 다음과 같이 3가지로 요약할 수 있다. (1) 최소 지지도 이상의 완전한 패턴 집합을 탐색하고 (2) DB를 탐색하는 횟수를 최소화하는 효율적이고 확장가능한 방식이어야 하며 (3) 연구자가 설정한 제한 조건을 반영할 수 있어야 한다.

2.2 Apriori 기반 시퀀스 패턴 마이닝

시퀀스 패턴 마이닝 중 가장 먼저 제안된 방식은 빈도 기반 apriori-based method이다. 그 중에서도 Srikant와 Agrawal(1996)이 제안한 generalize sequential Pattern(GSP)이 가장 대표적인 방법이다.[2]

2.3 깊이 탐색 기반 시퀀스 패턴 마이닝

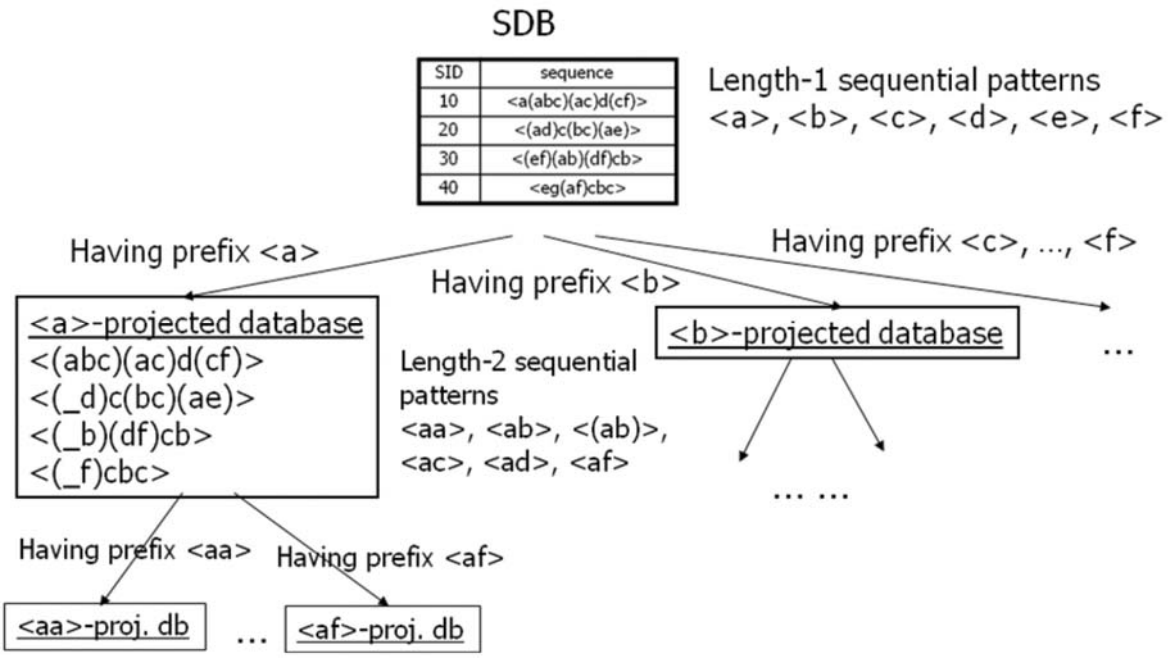
깊이 탐색을 통해 빈발 패턴을 찾는 Vertical format-based method들 중에서 Zaki(2001)가 제안한 Sequential PAttern Discovery using Equivalent class(SPADE)가 대표적이다.[3] 이 알고리즘은 통계 패키지 R의 arulesSequences 라이브러리로도 개발되어 배포되어있다. 그러나 apriori 기법은 모든 빈발 시퀀스 패턴을 탐색하려고 하기 때문에 시간이 많이 소요되는 단점이 있다. 이러한 apriori의 단점을 개선하기 위해서 pattern growth based method이 제안되었다. 대표적인 방법으로는 FreeSpan (Frequent pattern projected Sequential pattern mining) [4]과 PrefixSpan (Pei, et al., 2001) [5]이 있다. 이들 방법에서는 규모가 큰 데이터베이스의 시퀀스 데이터의 부분 집합을 이용하여 효율적인 빈발 패턴 탐색을 꾀한다는 공통점이 있다. PrefixSpan의 경우에는 python 라이브러리가 공개되어 있다.[6]

2.4 제약식 기반 시퀀스 패턴 마이닝

지금까지 소개된 시퀀스 패턴 마이닝 방법은 조정 가능한 파라메터가 min_sup값 하나 뿐이었다.그런 이유로 min_sup 값 이상인 시퀀스를 모두 탐색하여 비슷해 보이는 수많은 시퀀스가 결과로 도출되는 경우가 많다. 이런 경우, 식별하기 어려운 시퀀스들이 너무 많아 해석이 어려운 효과성 문제가 발생한다. 또, 완전한 시퀀스 패턴을 탐색할 때 데이터베이스의 크기가 큰 경우 컴퓨팅 자원을 필요로 하여 효율적이지 못하다는 점도 문제가 되었다. 이러한 문제점들을 해결하기 위해서 제안된 방식 중 하나가 constraint based method이다. 시퀀스 패턴 마이닝에서 사용 가능한 제약식(constraint)은 anti-monotonicity, monotonicity, succinctness, convertiblilty 및 inconvertibility를 들 수 있다. 다음은 이들 제약식을 활용한 시퀀스 패턴 마아닝 방법들에 대해서 살펴 보겠다.

먼저, 시간 제약식, sliding time windows 및 사용자 정의 분류법 등을 적용한 apriori 기반의 시퀀스 패턴 마이닝 알고리즘이 Srikant & Agrawal(1996)에 의해 제안되었다.[7] Mannila 등 (1997)은 시퀀스를 구성하는 항목들이 acyclic graph 형태로 표현될 수 있다고 가정하여 탐색 방법을 제한하는 알고리즘을 제안하기도 했다.[8]

길이가 긴 시퀀스들로 이루어져 밀도가 높은(dense) 데이터베



▶▶그림 1. PreFixSpan[1]

이스에서 시퀀스 패턴 마이닝을 수행하면 성능이 낮아지는 경우가 많은데 이런 문제를 해결하기 위해서 제약식을 가중치로 사용하는 Wspan이 제안되었다[Yun, 2008]. 또, 크기가 매우 큰 데이터베이스에서 효율적으로 빈발 시퀀스를 도출하기 위해 시간 제약 조건을 적용한 그래프 구조에서 시퀀스 패턴을 탐색하는 Graph for Time Constraints (GTC)가 제안되기도 했다[9]. 그 밖에도 사용자의 목적과 필요에 따라 직접 제약 조건을 생성할 수 있도록 하는 user-defined tough aggregate constraint 방식은 시퀀스의 모든 항목을 점검하는 시간을 줄이고 필요없는 데이터베이스를 전사하지 않도록 했다.[10] 이와 같은 다양한 제약식 유형은 표1. 과 같이 7가지로 정리할 수 있다. [11]

표 1.

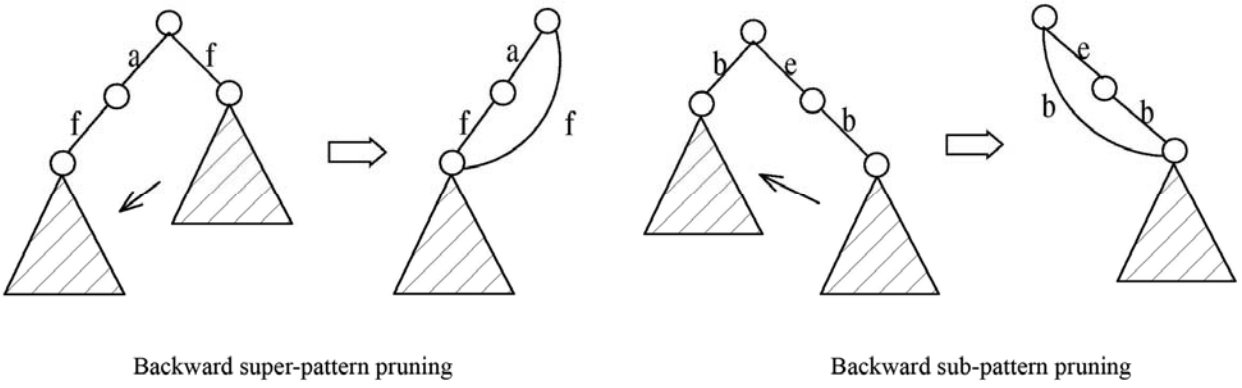
제약식 종류	의미
항목 (Item)	항목의 부분집합이 포함되어야 하거나 포함되어서는 안되는 조건
시퀀스 길이 (Length)	패턴의 길이에 대한 조건으로 특정 항목이 등장하는 횟수 또는 트랜잭션의 개수
슈퍼 패턴 (Supter-pattern)	특정 한 패턴을 적어도 하나 이상을 부분집합으로 가지는 (슈퍼 패턴) 조건
합계 (Aggregate)	아이템의 합계에 대한 제약조건으로 총합, 평균, 최대값, 최소값, 표준편차 등을 예로 들 수 있음
정규식 (Regular expression)	문자열이나 문자 집합에 쓰이는 연산들(disjunction 또는 Kleene closure) 등을 이용한 제약 조건
지속 기간 (Duration)	time-stamp가 있는 시퀀스 데이터베이스에만 적용될 수 있으며, 시퀀스의 지속시간 또는 항목 지속 시간에 대한 제약조건
간격 (Gap)	time-stamp가 있는 시퀀스 데이터베이스에만 적용될 수 있으며, 트랜잭션 사이의 시간 간격에 대한 제약조건

2.5 Closed 시퀀스 패턴 마이닝

CloSpan[12]으로 시간 간격이 있는 반복되는 시퀀스 패턴을 탐색할 때 사용된다. closed sequence pattern s는 수퍼 패턴 s'가 없는 시퀀스를 의미한다. 이 때 $s' \supset s$를 만족하며 s와 s'는 같은 지지도 값을 가진다. 이렇게 유사한 시퀀스 패턴을 부분 집합으로 쪼개서 탐색하는 방식을 활용하면 무수히 많은 유사한 시퀀스들로 구성된 시퀀스 패턴 마이닝 결과를 해석하느라 애쓰는 사태를 방지할 수 있다. CloSpan은 이와 같은 결과로 시퀀스 패턴을 요약하는 효과를 가지기 위해 후방 부분 패턴과 후방 수퍼 패턴을 가지치기(prune)를 수행하는 방식을 채택하였다. PrifixSpan과 유사하지만 데이터베이스의 크기에 따라 부분 패턴과 수퍼 패턴을 고려하여 더 좋은 결과를 보여주는 알고리즘으로 평가받고 있다.

2.6 실세계 데이터에 적용된 방법론

온라인 서비스 이용 패턴과 같은 실세계 데이터는 고객 등급 등 패턴이 속한 분류가 여러개로 나눠지는 경우가 많다. 예를 들어 인터넷 서비스 이용 패턴 P_1={100시간 무료 이용권 사용 패키지 → 15시간/월 정액 패키지 → 30시간/월 정액 패키지 → 무제한 패키지}이 있다고 하자. 또 이 패턴은 35세 이하 고객군에서 70% 발견되는 빈발 패턴이라고 하자. 하지만 다른 고객군에서는 P_2={100시간 무료 이용권 사용 패키지 → 30시간/월 정액 패키지} 패턴이 빈발 패턴일 수 있다. 이러한 경우, 연령 및 고객 분류별로 빈발 패턴이 달라지는 것이 명백하여 다른 고객 분류에서도 이처럼 결과가 달라지는지 확인 할 필요도 있음을 시사하고 있다. 이와 같이 시퀀스 패턴을 구성하는 항목을 분류하는 차원을 고려하여 제안된 알고리즘이 다차원 시퀀스 패턴 마아닝(multi-dimensional sequential pattern mining)이다. 빈발 패턴을 찾는 분석을 시행하기 전에 각 시퀀스 별로 차원을 표현할 수 있는 코드를 삽입하여 효율적으로 빈발 패턴을



▶▶그림 2. CloSpan[1]

찾을 수 있도록 설계되어 있다.

실제 데이터베이스를 구성하고 있는 시퀀스 패턴은 여러번 반복되는 경우가 많다. 이에 대한 예로 프로그램 실행 코드, 텍스트 데이터의 단어 시퀀스, 그리고 신용카드 사용 이력 등을 들 수 있다. 만약 $S_1 = AABCDABB$, $S_2 = ABCD$ 와 같은 두 개의 시퀀스가 있다고 하면 패턴 AB는 패턴 CB 보다 더 자주 발생하는 시퀀스로 볼 수 있을까? 이런 질문에 답을 하기 위해서는 반복적으로 발생하는 패턴 개념에 대한 정의가 필요하다. 이때, sup(P)는 max{|INS|: INS는 P를 구성하는 중복되지 않는 항목 집단}으로 정의할 수 있다. 또 P'은 P의 수퍼 패턴이며 sup(P')는 sup(P)보다 큰 값을 가질 수 없다. 이렇게 정의된 문제 해결을 위해서 제안된 방법이 탐욕적 항목성장 알고리즘(greedy instance-growth algorithm)이다. 각 항목이 확장해 나갈 때 가장 가까운 가능한 이벤트부터 확장해나간다는 것이 기본 아이디어이다. 이 알고리즘은 Ding 등(2009)에 의해 Closed Repetitive Gapped subsequence라는 이름으로 제안되었다.[13]

기존의 시퀀스 패턴 마이닝은 시간이 흘러도 시퀀스의 행동 방식이 변하지 않는다는 가정을 하고 있어 실제 상황을 제대로 반영하지 못하는 한계를 가진다. 실제 상황에서 데이터가 생성될 때의 환경은 대부분 동적으로 변화한다. 이러한 변화를 반영하기 위해 최신성(recency)과 간결성(compactness) 개념을 정의하고 이를 이용하여 빈발 시퀀스 패턴을 탐색하는 방법도 등장했다.[14]

2.7 시퀀스 패턴 마이닝 활용 사례

시퀀스 패턴 마이닝은 넓은 분야에서 활용되고 있다. 시퀀스 형태의 데이터가 존재한다면 어떤 분야에서든지 패턴을 분석하여 활용하여 시스템의 활용성을 높이고, 향후 일어날 일들을 예측하고, 상태나 이벤트를 탐지하는 일 등에 활용할 수 있다. 이렇게 시퀀스 패턴 마이닝이 활용되는 분야는 헬스케어, 교육, 웹 사용 패턴, 텍스트 마이닝, 바이오정보 및 보안 분야까지 다양하다.

헬스케어 분야에서 시퀀스 패턴 마이닝은 매우 많이 활용되는 기법이다. 그 중에서도 환자의 의료 기록을 분석하는데 많이 활용된다. 환자의 진단과 처방 그리고 치료 내역에 대한 시퀀스 패턴을 분석하면, 특정 질환에 대한 증상 패턴을 도출 할 수 있다. 실제로 감염병 학자들이 소화불량과 관련된 증상이 시간이 지남에 따라 어떤 패턴을 갖는지 연구하는데 시퀀스 패턴 마이닝을 적용하였다. 노약자들이 혼자 생활하는 경우, 스마트 홈 시스템을 이용하여 데이터를 수집하고 이 데이터를 분석할 때 시퀀스 패턴 마이닝을 이용하여 위급한 상황을 미리 예측하는 연구도 있다.

프로그램의 코딩패턴을 분석하는데에도 시퀀스 패턴 마이닝이 활용될 수 있다. Ishio 등(2008)은 자바 프로그램의 코딩 패턴을 분석하는데 시퀀스 패턴 마이닝을 활용하였다.[15] 코딩 패턴은 복사하고 붙이는 활동, 크로스 컷팅 고려 그리고 자주 사용하는 코딩 형태나 패턴을 의미하는 관용구(idioms) 등으로 구성되어 있다. 코딩 시퀀스 패턴 마이닝을 적용하면 소스 코드를 구성하는 코딩 패턴을 파악할 수 있다. 또, TRAC 이라는 오픈소스 프로젝트 도구를 이용해 팀워크를 하는 시퀀스에 대한 데이터를 습득하여 과제를 성공적으로 수행해내는 그룹과 그렇지 않은 그룹 사이의 활동 패턴 차이를 보여주기도 했다.[16]

웹 사용 데이터에 시퀀스 패턴 마이닝을 적용하면 폭넓은 분야에서 활용이 가능하다. 예를 들면, 웹 사용 데이터의 트랜젝션을 분석하여 사용자를 식별해 내기도 하고 [17] 고객의 웹 사용 시퀀스 패턴을 분석하여 개인 맞춤 서비스를 제공하는 마케팅 계획을 세울 수도 있다.[18,19] 웹 사용 시퀀스 패턴 분석 결과는 웹 사이트 디자인에서도 활용되기도 한다. Berendt(2000)는 독일어 교육 잡지 데이터베이스의 사용자 검색 행위 패턴을 도출하여 웹 사이트 디자인에 적용하였다. 예를 들면 특정 항목 에 대한 검색어가 반복적으로 수정되어 발생한다면 웹 사이트에 나타난 해당 항목은 직관적이지 못함을 의미하여 개선 할 필요가 있다.[20]

시퀀스 패턴 마이닝을 이용하여 테스트 데이터를 분석하면 텍스트가 어떤 분류에 속하는지[21], 어떤 트렌드를 보이는지[22] 판단할 수 있다. 예를 들어 어떤 기업이 주력 분야를 바꾸고 싶을 때 해당 분야에 대한 시장 조사의 일환으로 소셜 미디어에서의 관련 단어나 구절들이 언급되는 빈도 등에 대한 트렌드 분석을 할 수 있다.

바이오인포매틱스 분야에서는 특정 유전자가 구성된 규칙을 찾거나 단백질 구조를 예측하거나 유전자 표현식을 분석하거나, 단백질 접힘 인식 또는 DNA 시퀀스에서의 motif discovery에 시퀀스 패턴 마이닝이 활용된다.[1]

3. 사용자 행동 데이터의 시퀀스 패턴 마이닝

시퀀스 패턴 마이닝 기법이 적용될 수 있는 다양한 분야가 있어 왔지만 실제 사용자의 행동 데이터를 분석하는 영역은 데이터 수집 환경이 뒷받침 되어야 했기 때문에 웹 서비스 영역으로 한정되어 있었다. 웹서비스 및 마케팅 분야에서의 사용자 행동 방법론에 대한 연구는 매우 활발하게 연구되어 있으며, 이와 관련한 자바 스크립트 기반의 오픈소스 분석 도구가 공개되어 있

기도 하다.[23]

최근 통신 및 센서 기술 등의 발전으로 스마트 워치, 스마트 패브릭 등 신체 부착형 디바이스의 사용화가 가속화 되면서 스스로의 생활을 자동으로 기록하고 분석하고자 하는 수요가 늘어나고 있다. 상용화된 서비스로는 헬스케어 분야가 가장 활발한데 사용자의 운동 패턴을 트래킹하여 운동과 움직임을 장려하는 서비스가 가장 일반적이고, 모바일 센서를 이용한 심전도 측정을 통해 건강 상태를 모니터링하는 것 과 같은[24] 전문 의료 서비스와 관련된 서비스도 출시되었다.

데이터베이스에서 추출한 데이터를 분석하여 결론을 도출하는 시퀀스 패턴 마이닝과는 별개로 심리학 분야에서도 사람의 행동 시퀀스에 대한 데이터를 모아 분석하여 범죄 예방 등의 분야에 활용하고 있다.[25] 범죄자의 행동 예측과 같이 그 대상이 범죄자인 경우에는 온라인 데이터를 수집하기 어렵기 때문에 전문 프로파일러 등이 탐문 수사 등 당사자에게 인터뷰를 수행하여 관찰한 결과를 바탕으로 데이터를 생성하고 분석을 수행한다. 하지만 선후 관계를 고려한 빈도 기반의 분석 방법 그리고 Markov Chain 등의 확률 기반 예측 방법 등이 적용되어 지금까지 소개한 시퀀스 패턴 마이닝과 유사한 분석이 이루어지고 있다.[26]

따라서 품질과 용량을 확보할 수 있는 데이터 획득이 이루어진다면 사용자의 실제 생활 및 행동 데이터를 습득하여 사용자의 시간의 흐름에 따른 패턴을 인식하고 예측하는 연구를 수행할 수 있다. 하지만 실제 상황에서 시간의 흐름에 따른 사용자의 행동 패턴을 분석하는 것은 도메인이 특정되어 있다고 해도 데이터 수집부터가 쉽지 않다. 또, 복잡한 인간의 행동 양식을 모델링하여 시간의 흐름에 따라 변화하는 패턴을 인식할 때, 내생적 요인과 외생적 요인 모두를 고려하는 것이 쉽지 않다. 하지만 이러한 행동 패턴 예측은 기업활동에 도움을 주는 마케팅 분야 뿐 아니라 보안, 헬스케어 등 삶의 질을 개선해 줄 수 있는 다양한 분야에서 활용될 수 있는 도구다. 따라서 관련 연구 동향을 파악하고 필요한 기술을 습득하여 적절한 분야에 적용하는 연구가 필요하다.

본 연구는 한국전자통신연구원 연구운영지원사업의 일환으로 수행되었음. [20ZS1100, 자율성장형 복합인공지능 원천기술 연구]

참고문헌

[1] Gupta, Manish, and Jiawei Han. "Approaches for Pattern Discovery Using Sequential Data Mining." Data Mining: Concepts, Methodologies, Tools, and Applications. IGI Global, 2013. 1835-1851.

[2] Srikant, R., & Agrawal, R. (1996, March). Mining sequential patterns: Generalizations and performance improvements. In International Conference on Extending Database Technology (pp. 1-17). Springer, Berlin, Heidelberg.

[3] Zaki, M. J. (2001). SPADE: An efficient algorithm for mining frequent sequences. Machine learning, 42(1-2), 31-60.

[4] Han, J., Pei, J., Mortazavi-Asl, B., Chen, Q., Dayal, U., & Hsu, M. C. (2000, August). FreeSpan: frequent pattern-projected sequential pattern mining. In Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 355-359).

[5] Han, J., Pei, J., Mortazavi-Asl, B., Pinto, H., Chen, Q., Dayal, U., & Hsu, M. (2001, April). Prefixspan: Mining sequential patterns efficiently by prefix-projected pattern growth. In proceedings of the 17th international conference on data engineering (pp. 215-224). IEEE Washington, DC, USA.

[6] https://pypi.org/project/prefixspan/

[7] Srikant, R., & Agrawal, R. (1996, March). Mining sequential patterns: Generalizations and performance improvements. In International Conference on Extending Database Technology (pp. 1-17). Springer, Berlin, Heidelberg.

[8] Mannila, H., Toivonen, H., & Verkamo, A. I. (1997). Discovery of frequent episodes in event sequences. Data mining and knowledge discovery, 1(3), 259-289.

[9] Masseglia, F., Poncelet, P., & Teisseire, M. (2003). Incremental mining of sequential patterns in large databases. Data & Knowledge Engineering, 46(1), 97-121.

[10] Chen, E., Cao, H., Li, Q., & Qian, T. (2008). Efficient strategies for tough aggregate constraint-based sequential pattern mining. Information Sciences, 178(6), 1498-1518.

[11] Kum, H. C., Chang, J. H., & Wang, W. (2007). Bench

marking the effectiveness of sequential pattern mining methods. Data & Knowledge Engineering, 60 (1), 30-50.

[12] Yan, X., Han, J., & Afshar, R. (2003, May). CloSpan: Mining: Closed sequential patterns in large datasets. In Proceedings of the 2003 SIAM international conference on data mining (pp. 166-177). Society for Industrial and Applied Mathematics.

[13] Ding, B., Lo, D., Han, J., & Khoo, S. C. (2009, March). Efficient mining of closed repetitive gapped subsequences from a sequence database. In 2009 IEEE 25th International Conference on Data Engineering (pp. 1024-1035). IEEE.

[14] Chen, Y. L., & Hu, Y. H. (2006). Constraint-based sequential pattern mining: The consideration of recency and compactness. Decision Support Systems, 42(2), 1203-1215.

[15] Ishio, T., Date, H., Miyake, T., & Inoue, K. (2008, October). Mining coding patterns to detect crosscutting concerns in Java programs. In 2008 15th Working Conference on Reverse Engineering (pp. 123-132). IEEE.

[16] Perera, D., Kay, J., Yacef, K., & Koprinska, I. (2007, July). Mining learners' traces from an online collaboration tool. In AIED07, 13th International Conference on Artificial Intelligence in Education, Educational Data Mining Workshop (pp. 60-69).

[17] Cooley, R., Mobasher, B., & Srivastava, J. (1999). Data preparation for mining world wide web browsing patterns. Knowledge and information systems, 1 (1), 5-32.

[18] Büchner, A. G., & Mulvenna, M. D. (1998). Discovering internet marketing intelligence through online analytical web usage mining. ACM Sigmod Record, 27(4), 54-61.

[19] Mobasher, B., Dai, H., Luo, T., & Nakagawa, M. (2002, December). Using sequential and non-sequential patterns in predictive web usage mining tasks. In 2002 IEEE International Conference on Data Mining, 2002. Proceedings. (pp. 669-672). IEEE.

[20] Berendt, B., & Spiliopoulou, M. (2000). Analysis of navigation behaviour in web sites integrating multiple information systems. The VLDB journal, 9(1), 56-75.

[21] Jaillet, S., Laurent, A., & Teisseire, M. (2006). Sequential patterns for text categorization. Intelligent Data Analysis, 10(3), 199-214.

[22] Garboni, C., Masseglia, F., & Trousse, B. (2005, November). Sequential pattern mining for structure-based XML document classification. In International Workshop of the Initiative for the Evaluation of XML Retrieval (pp. 458-468). Springer, Berlin, Heidelberg.

[23] http://www.philippe-fournier-viger.com/spmf/

[24] Apple, https://support.apple.com/en-us/HT208955

[25] Marono, A., Clarke, D. D., Navarro, J., & Keatley, D. A. (2017). A behaviour sequence analysis of nonverbal communication and deceit in different personality clusters. Psychiatry, Psychology and Law, 24(5), 730-744.

[26] Keatley, D. A., Golightly, H., Shephard, R., Yaksic, E., & Reid, S. (2018). Using behavior sequence analysis to map serial killers' life histories. Journal of interpersonal violence, 0886260518759655.

저자소개

● 임 지 연(Jiyoun Lim)

■ 2005년 2월 : KAIST 산업및시스템즈공학과 (공학사)

■ 2007년 2월 : KAIST 산업및시스템즈공학과 (공학 석사)

■ 2013년 8월 : KAIST 산업및시스템즈공학과 (공학 박사)

■ 2011년 3월 ~ 2013년 7월 : 한국기술교육대학교 산업경영학과 대우교수

■ 2013월 3월 ~ 현재 : 한국전자통신연구원 선임연구원

〈관심분야〉 지식서비스, 경영정보시스템, 데이터마이닝, IoT, 센서데이터, 인간행동분석