DOI: 10.3745/KIPSTD.2009.16-D.2.185

시간 속성을 갖는 이벤트 집합에서 인터벌 연관 규칙 마이닝 기법

한 대 영[†] · 김 대 인[††] · 김 재 인[†] · 나 철 수[†] · 황 부 현[†††]

요 약

시간 속성을 갖는 이벤트 집합에서 동일한 이벤트 타입에 대한 이벤트 시퀀스는 하나의 이벤트로 요약될 수 있다. 그러나 정의된 시간 간격이 경과된 후 발생된 이벤트 타입은 하나 이상의 독립된 서브 이벤트 시퀀스로 요약하는 것이 바람직하다. 본 논문은 Allen의 시간 관계 대수에 기반하여 인터벌 이벤트를 요약하고, 요약된 인터벌 이벤트들로부터 인터벌 연관 규칙을 찾아내는 새로운 시간 데이터 마이닝 기법을 제안한다. 제안하는 기법은 독립적인 서브 시퀀스 개념을 도입하고 인터벌 이벤트 사이의 연관 규칙을 탐사함으로써 질적으로 우수한 정보를 제공한다.

키워드 : 시간 데이터 마이닝, 시간 속성, 연관 규칙, 인터벌 이벤트, 이벤트 타입

A Method for Mining Interval Event Association Rules from a Set of Events Having Time Property

DaeYoung Han[†] · Daeln Kim[††] · Jaeln Kim[†] · CholSu Na[†] · BuHyun Hwang[†††]

ABSTRACT

The event sequence of the same type from a set of events having time property can be summarized in one event. But if the event sequence having an interval, It is reasonable to be summarized more than one in independent sub event sequence of each other. In this paper, we suggest a method of temporal data mining that summarizes the interval events based on Allen's interval algebra and finds out interval event association rule from interval events. It provides better knowledge than others by using concept of an independent sub sequence and finding interval event association rules.

Keywords : Temporal Data Mining, Temporal Property, Association Rule, Interval Event, Event Type

1. 서 론

데이터 마이닝에 대한 최근의 많은 연구는 순차 패턴, 유사 시퀀스와 같은 시간 속성을 갖는 데이터로부터 시간 관계 규칙에 대한 지식을 발견하기 위한 시간 데이터 마이닝 문제에 초점이 맞추어져 있다. 시간 데이터 마이닝은 연관 관계, 분류, 특징 추출 등을 포함한 기존의 데이터 마이닝 기법을 확장하여 이벤트들 사이의 원인과 결과 관계에 대한 시간 연관 규칙을 찾아내는 새로운 기법이다. 이러한 시간 데이터 마이닝 기법에는 순환적으로 반복하는 연관 규칙을 발견하기 위하여 사용되는 순환 연관 규칙 탐사, 캘린더 형태로 표현된 시간 패턴에 대한 연관 규칙을 추출하기 위하여 사용하는 캘린더 연관 관계 탐사 기법 등이 있다. 그러나 이러한 연구들은 인터벌 데이터에 대한 시간 관계를 고려하지 않으며, 최근에 들어서 인터벌 데이터로부터 유용한 패턴을 마이닝하기 위한 기초적인 연구가 진행되고 있다. 실세계에는 환자 진료 차트, 상품 구매 이력, 웹 로그 등과 같은 인터벌 데이터에 대한 다양한 응용이 존재한다.

본 논문에서는 Allen[1]의 이론에 기반한 시간 데이터 마이닝 기법을 제안한다. 제안하는 시간 데이터 마이닝 기법의 기본 아이디어는 다음과 같이 요약된다.

- 이벤트들의 시퀀스를 독립 서브 시퀀스로 나누고, 각 서브 시퀀스를 인터벌을 갖는 이벤트로 요약한다.
- 인터벌을 갖는 이벤트들 사이의 시간 연관 규칙을 찾음으로서 이들 사이의 인과 관계를 파악하여 환자 치료에 사

※본 논문은 2007년도 전남대학교 연구년 교수 연수 연구비 지원에 의하여 연구되었음.
† 준 회 원 : 전남대학교 전자컴퓨터공학부 석사과정
†† 정 회 원 : 전남대학교 전자컴퓨터공학부 시간강사
††† 종신회원 : 전남대학교 전자컴퓨터공학부 교수
논문접수 : 2008년 11월 3일
수 정 일 : 1차 2009년 2월 4일
심사완료 : 2009년 2월 4일

용할 수 있는 유용한 지식을 추출한다.

본 논문의 구성은 다음과 같다. 2절에서는 시간 데이터 마이닝에 대한 관련 연구를 기술하고, 3절에서는 시간 데이터 마이닝에 대한 기본 개념 및 제안하는 알고리즘을 기술한다. 4절에서는 실험을 통하여 제안하는 알고리즘의 성능을 분석하고, 끝으로 5절에서는 결론 및 향후 연구에 대하여 기술한다.

2. 관련 연구

시간 속성을 갖는 데이터로부터 유용한 지식을 찾아내기 위한 시간 데이터 마이닝에 대한 많은 연구가 진행되고 있다[2, 3]. 그리고 이러한 연구들은 순차 패턴, 유사 시퀀스, 시간 규칙을 찾아내는 방법 등으로 크게 분류된다. 순차 패턴 마이닝은 기존의 연관 규칙 기법을 확장하여 트랜잭션에 포함된 특정한 아이템 집합들의 시퀀스를 찾아내는 기법이다[4, 5]. 즉, 순차 패턴인 시퀀스 (A, B, C)가 있는 경우 아이템 집합 A, B, C가 서로 다른 트랜잭션에 존재하더라도 이 트랜잭션들은 동일한 고객에 대한 트랜잭션 간주되어 처리되어 고객들의 행동 패턴을 찾아내는데 유용하게 사용된다. 순차 패턴 마이닝 문제는 사용자가 명시한 최소 지지도를 만족하는 모든 시퀀스 중에서 최대 길이를 갖는 시퀀스를 찾아내는 것으로 연관 규칙 탐사 알고리즘인 Apriori[6]에 기반한다.

그리고 기본적인 순차 패턴 마이닝 알고리즘을 개선하기 위하여 Apriori 알고리즘을 변형한 Episode 알고리즘이 제안되었다[7]. Episode는 사용자가 명시한 윈도우 크기를 갖는 시간 윈도우들 가운데서 최소 빈발 임계값을 만족하는 모든 에피소드를 찾아내는 것이다. 에피소드는 밀접히 관련된 이벤트들의 시퀀스이다. 예를 들어 이벤트 시퀀스 (A, B, D, A, C, B, D, A, D)에서 이벤트가 매 초마다 발생하며, 윈도우 사이즈가 3초이고, 최소 빈발 임계값이 50%라면 시퀀스 (A, B)는 3개의 윈도우 가운데 2개의 윈도우에서 발생하므로 빈발한 에피소드로 추출된다.

유사 시퀀스 탐색은 주식, 상품 가격, 판매량 등과 같은 시계열 데이터로부터 사용자 입력 질의 시퀀스를 만족하는 유사 시퀀스들의 모든 쌍을 찾아내는 마이닝 기법이다[8,9]. 예를 들어 두 주식의 관리 전략에서 유사성이 있다면 주식의 추세를 가리키는 시계열 데이터로부터 특정 주식에 대한 유사한 패턴을 발견할 수 있다.

시간 연관 규칙 탐사 기법은 연관 규칙 탐사, 분류, 특성화와 같은 기존의 데이터 마이닝 기법을 확장한 기법으로 시간 관계와 인과 관계에 대한 시간 연관 규칙을 탐사한다. 그리고 시간 연관 규칙 탐사 기법은 순환적으로 반복되는 연관 규칙을 발견하기 위한 순환 연관 관계 탐사[10]와 캘린더 형태로 표현된 시간 패턴에 대한 연관 규칙을 발견하는 캘린더 연관 관계 탐사[11, 12]를 포함한다.

그러나 이러한 대부분의 연구는 데이터의 특정 시점에만 한정되며 데이터의 인터벌을 고려하지 않으며 최근에 비로써 인터벌을 갖는 데이터로부터 유용한 패턴을 마이닝하기 위한 연구가 진행되고 있다[13]. 또한 인터벌 데이터에 대한 다양한 응용 분야가 존재하고 있지만 인터벌 연관 규칙 탐사 과정은 매우 복잡하므로 보다 효율적으로 인터벌 연관 규칙을 탐사할 수 있는 알고리즘 개발이 필요하다.

3. 인터벌 연관 규칙 탐사 알고리즘

본 절에서는 인터벌 연관 규칙 탐사에 대한 기본 개념과 제안하는 인터벌 연관 규칙 마이닝 기법을 기술한다.

3.1 기본 개념

환자 Cid가 주기적으로 진찰을 받는 경우, 환자(아이디 Cid)에 대한 한 번의 진찰은 하나의 트랜잭션 TCid로 정의되며, 환자의 증상은 하나의 이벤트 e로 표현할 수 있다. 또한 환자는 한 번의 진찰에 의하여 여러 증상을 나타낼 수 있으므로 하나의 트랜잭션은 다수의 이벤트를 포함한다. 본 논문에서는 환자와 고객은 서로 같은 의미로 기술한다. 또한 본 논문에서는 하나의 이벤트 시퀀스에 대한 인터벌 연관 규칙을 탐사하므로 고객 Cid의 트랜잭션에 포함된 이벤트 중 동일한 이벤트 타입을 갖는 이벤트 시퀀스만을 고려한다. 그리고 각 이벤트에 대한 시퀀스를 독립적인 서브 시퀀스들로 분할될 수 있는지 검사하여 서브 시퀀스를 생성한다. 그리고 각각의 시퀀스는 최종적으로 인터벌 이벤트로 요약한다. 그리고 인터벌 이벤트들에 대한 시간 연관성을 찾음으로써 이벤트에 대한 인터벌 연관 규칙을 찾아낸다. 이벤트 시퀀스와 인터벌 이벤트, 그리고 인터벌 관계는 다음과 같이 정의된다.

[정의 3.1.] (이벤트 시퀀스)

한 고객 Cid의 한 이벤트 타입 E에 대한 이벤트 시퀀스 ES(Cid,E)는 <e1, e2, ..., en>로 정의된다. 이 때 ei=(E,ti), ei ∈Ti, ti∈TS, ti ≤ti+1이다.

[정의 3.2.] (인터벌 이벤트)

인터벌 이벤트는 시간 간격을 갖는 이벤트를 의미한다. 인터벌 이벤트 e′는 (E, [vs,ve])로 표현하며, vs와 ve는 각각 인터벌의 시작 시점과 종료 시점을 의미한다. 그리고 인터벌 이벤트 e′의 시작 시점과 종료 시점은 e′.vs와 e′.ve로 기술된다. 그러므로 고객 Cid의 인터벌 이벤트 집합 IES(Cid)는 {e′1, e′2, ..., e′n}로 정의된다.

[정의 3.3.] (인터벌 관계)

IES(Cid)의 인터벌 이벤트들의 인터벌 관계 집합 IERS(Cid)는 다음과 같이 정의된다. IERS(Cid)={P(x,y)|x,y ∈ IES(Cid), P∈ IO}. IO={before, equals, meets, overlaps, during}, IO는 이진 인터벌 관계 연산자들의 집합을 의미한다. 그리고 P(x,y)는 이진 인터벌 관계를 표현하는 술어이며 <표 1>과 같이 정의된다.

〈표 1〉 P(x,y) 이진 인터벌관계 술어

Relation	Expression
before(x,y)	x.ve < y.vs
equals(x,y)	x.vs=y.vs ∧ x.ve=y.ve
meets(x,y)	x.ve=y.vs
overlaps(x,y)	x.vs<y.vs ∧ x.ve<y.ve
during(x,y)	x.vs<y.vs ∧ y.ve<x.ve

추출된 각각의 인터벌 이벤트는 이벤트에 대한 인터벌 연관 규칙을 찾아내기 위하여 IR-Graph로 표현된다. IR-Graph는 다음과 같이 정의된다.

[정의 3.4.] (IR-Graph)

인터벌 관계 집합에서 각각의 인터벌 이벤트는 IR-Graph의 노드가 되고, 두 인터벌 이벤트들의 관계는 하나의 에지로 표현된다. 그리고 각각의 에지는 카운트를 갖으며, 카운트는 동일한 인터벌 관계의 수를 표현한다. 그러므로 IR-Graph에서 각 이벤트들의 관계에 대한 지지도를 구하여 빈발한 인터벌 관계 집합을 추출할 수 있으므로 다양한 이벤트들 사이의 인과 관계를 파악하고 인터벌 관계 규칙을 탐사할 수 있다.

3.2 이벤트 시퀀스를 서브 시퀀스로 분할

하나의 이벤트 시퀀스를 단순하게 하나의 인터벌 이벤트로만 요약하는 것은 발마직하지 못하다. 예를 들어, 이벤트 시퀀스에 포함된 한 이벤트 ei의 발생시점 ti에서 다음 이벤트 ei+1의 발생시점 ti+1까지의 시간 간격이 응용에 따라 시스템에 정의된 임계값 ε보다 크다면 서브 시퀀스 <e1,e2, ..., ei>과 <ei+1, ei+2, ..., en>는 서로 독립적이라고 할 수 있다. 독립적이라는 것은 하나의 인터벌 이벤트가 다른 인터벌 이벤트에 영향을 끼치지 않는 것을 의미한다.

3.3 알고리즘

본 논문에서 제안하는 인터벌 연관 규칙 탐사 알고리즘은 다음과 같은 6개의 단계로 구성된다. 제안하는 알고리즘은 시점을 갖는 이벤트들로부터 인터벌 이벤트들과 인터벌 이벤트들 사이의 관계를 추출하고 이러한 관계들 사이에 존재하는 규칙을 찾아낸다. 제안하는 알고리즘의 각 단계를 설명을 위하여 예제와 함께 다음과 같이 기술할 수 있다.

단계 1) 빈발 이벤트 타입 계산

1. 데이터베이스에 있는 트랜잭션에 대하여 고객 아이디 Cid와 트랜잭션 발생 시점에 따라 정렬된 데이터베이스 DBsort를 구한다.
2. DBsort로부터 빈발 이벤트 타입을 계산하고, 빈발하지 않는 이벤트 타입은 DBsort에서 제거한다.

<표 2>의 (a)는 환자 진료 데이터에 대한 초기 DB이다. 초기 DB에서 T열은 달을 의미하며 14는 시작으로부터 14개월이 지난 시점을 의미한다. 지지도는 이벤트 타입 E에 대해 이벤트 타입을 포함하는 Cid를 계산한다. 예를 들어 이벤트 A는 Cid 101, 102, 103 104에 포함되므로 지지도가 4가 된다. 그리고 <표 2>의 (b)를 참조하여 지지도가 70% 이하인 빈발하지 않은 이벤트 타입들을 제거하므로 빈발한 이벤트 타입만을 포함한 데이터베이스는 <표 3>과 같이 추출된다.

단계 2) 이벤트 시퀀스 계산

1. DBsort에 있는 각각의 고객 Cid의 각 Ex에 대한 ES (Cid, Ex)를 구한다.
2. 전체 고객에 대하여 1의 과정을 반복한다.

〈표 2〉 초기 DB 및 이벤트 정보

Cid	T	E	Cid	T	E
101	2007/1	A	103	2007/1	B
	2007/3	A,B		2007/3	B,D
	2007/4	A,F		2007/4	B,D
	2007/5	B,F		2007/6	A
	2007/6	B		2007/7	A,B,C
	2007/9	D		2007/8	A
	2007/11	A,D		2007/11	C
	2007/12	A,D			
	2007/14	A			
102	2007/1	C	104	2007/1	B
	2007/2	C		2007/2	B
	2007/4	D		2007/3	E
	2007/6	A,D		2007/4	B
	2007/7	D		2007/5	D,E
	2007/8	A		2007/7	D
	2007/10	B,E		2007/8	A
	2007/11	B,E		2007/9	D,F
				2007/10	A
				2007/11	A,F

(a) 원본 DB

Event Type	Support
A	4
B	4
C	2
D	4
E	2
F	2

(b) 이벤트 타입 지지도

Event Type	Support
A	4
B	4
D	4

(b) 빈발 이벤트 타입 (지지도 70% 이상)

〈표 3〉 정렬된 빈발 항목 집합

C_{id}	Time-Point	Event Type	C_{id}	Time-Point	Event Type
101	2007/1	A	103	2007/1	B
	2007/3	A,B		2007/3	B,D
	2007/4	A		2007/4	B,D
	2007/5	B		2007/6	A
	2007/6	B		2007/7	A,B
	2007/9	D		2007/8	A
	2007/11	A,D			
	2007/12	A,D			
	2007/14	A			
102	2007/4	D	104	2007/1	B
	2007/6	A,D		2007/2	B
	2007/7	D		2007/3	E
	2007/8	A		2007/4	B
	2007/10	B		2007/5	D
	2007/11	B		2007/7	D
				2007/8	A
				2007/9	D
				2007/10	A
				2007/11	A

단계 3) 독립적인 서브 시퀀스 계산

1. 서브 시퀀스들의 집합 ESSsub=∅으로 초기화 한다.
2. DBsort 내에 있는 각 ES(Cid, Ex)에 대하여,
   a. ES(Cid, Ex) 내의 첫 이벤트 e1을 estart로 놓고
   b. estart부터 스캔하여 두 이벤트들 estart+i와 estart+i+1의 시점 tstart+i와 시점 tstart+i+1 사이가 임계값 ε보다 더 크면 서브 시퀀스 <estart, estart+1, ..., estart+i>에 시퀀스 식별자를 부여하고 서브 시퀀스 집합 ESSsub에 추가한다.
   c. start+i+1이 n보다 작으면 다시 estart+i+1을 estart로 놓고 스텝 b부터 시작한다.

단계 4) 이벤트 시퀀스 요약 및 인터벌 이벤트 계산

1. IES=∅으로 초기화 한다.
2. ESSsub에 있는 각 시퀀스에 대하여
   시퀀스의 첫 이벤트의 시작 시점을 인터벌 이벤트 e′의 e′.vs에, 마지막 이벤트의 종료 시점을 e′.ve에 넣는다. 그리고 인터벌 이벤트 (E′,[vs,ve])를 IES에 추가한다. 이 때 E′는 인터벌 이벤트 타입이다.

하나의 시퀀스를 독립적인 서브 시퀀스로 추출하기 위한 임계값 ε이 4라면, <표 4>와 같은 서브 시퀀스 집합과 인터벌 이벤트를 추출할 수 있다. 즉 시퀀스 <(A,1) (A,3) (A,4) (A,11) (A,12) (A,14)>만이 두 개의 독립적인 서브 시퀀스 <(A,1) (A,3) (A,4)>와 <(A,11) (A,12) (A,14)>로 나누어진다. 즉 이벤트 (A,4)와 (A,11) 사이의 시간 간격이 입계값인 4보다 크기 때문이다. <표 4>의 IES는 각 시퀀스를 인터벌 이벤트로 요약한 정보이다. 시퀀스를 요약하는 것은 각 시퀀스의 첫 이벤트의 발생 시점을 vs로 설정하고 마지막 이벤트의 발생 시점을 ve로 표현함으로써 이루어진다.

단계 5) 빈발 인터벌 관계 탐사

1. IES에 있는 각 고객 Cid의 인터벌 이벤트에 대한 인터벌 관계를 계산하여 인터벌 이벤트 관계 집합 IERS를 구한다.
2. IERS로부터 각 인터벌 관계의 지지도를 계산한다.
3. 각 인터벌 관계에 대하여 지지도가 주어진 임계값보다 작은 것들은 제거한 후 빈발한 인터벌 관계로 이루어진 IERSfreq를 구한다.

단계 6) 인터벌 연관 규칙 추출

1. IERSfreq에 있는 각 고객에 대하여 각 인터벌 관계를 IR-Graph에 추가한다.
   이 때 IR-Graph에서 에지의 카운트가 인터벌 관계의 주어진 지지도보다 작은 것은 제거한다.
2. IR-Graph의 각 컴포넌트 그래프로부터 인터벌 연관 규칙을 찾아낸다.

<표 6>을 적용하여 IR-Graph로 표현하여 40% 이상의 지지도를 만족하는 인터벌 관계만이 그래프에 포함되므로 (그림 1)과 같은 IR-Graph를 얻는다. (그림 1)에서 각 에지가 나타내는 괄호 숫자는 각 관계의 지지도를 의미한다. (그림 1)에서 이벤트 B는 이벤트 A와 D에 영향을 미치고 이벤트 D가 발생하는 중에 이벤트 A가 발생한다는 사실을 알 수 있다. 또한 이벤트를 나타내는 각 노드에서 에지들의 방향을 추적하면 하나의 이벤트가 다른 이벤트 발생에 어떻게 영향을 미치는지를 추론 할 수 있다. 예로 (그림 1)에서 이벤트 D가 발생하는 중에 이벤트 A가 발생했지만, 이벤트 B가 발생한 후에 이벤트 D가 발생하므로 이벤트 B는 이벤트 A를 발생시킬 수 있는 원인 요소가 될 수 있음을 알 수 있다.

〈표 4〉 서브 시퀀스 집합과 인터벌 이벤트
(임계값 ε=4인 경우)

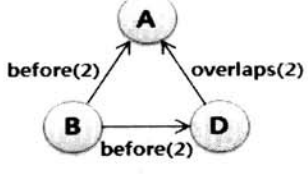
ES(Cid,E)	Sub Sequence Set	IES
ES(101,A)	{<(A,1)(A,3)(A,4)>, <(A,11)(A,12)(A,14)>}	{(A1,[1,4]), (A2,[11,14])}
ES(101,B)	{<(B,3)(B,5)(B,6>}	(B,[3,6])
ES(101,D)	{<(D,9)(D,11)(D,12)>}	(D,[9,12])
ES(102,A)	{<(A,6)(A,8)>}	(A,[6,8])
ES(102,B)	{<(B,10)(B,11)>}	(B,[10,11])
ES(102,D)	{<(D,4)(D,6)(D,7)>}	(D,[4,7])
ES(103,A)	{<(A,6)(A,7)(A,8)>}	(A,[6,8])
ES(103,B)	{<(B,1)(B,3)(B,4)(B,7)>}	(B,[1,7])
ES(103,D)	{<(D,3)(D,4)>}	(D,[3,4])
ES(104,A)	{<(A,8)(A,10)(A,11)>}	(A,[8,11])
ES(104,B)	{<(B,1)(B,2)(B,4)>}	(B,[1,4])
ES(104,D)	{<(D,5)(D,7)(D,9)>}	(D,[5,9])

〈표 5〉 인터벌 이벤트 관계 집합

Cid	IERS
101	{overlaps(A1,B), before(A1,D), before(A1,A2), before(B,D),before(B,A2), overlaps(D,A2)}
102	{overlaps(D,A), before(D,B), before(A,B)}
103	{during(B,D), overlaps(B,A), before(D,A)}
104	{before(B,D),before(B,A),overlaps(D,A)}

〈표 6〉 인터벌 관계 지지도

IER_{freq}	Support (40% 이상)
{during(A,D)}	2
{overlaps(D,A)}	2
{before(B,D)}	2



(그림 1) IR-Graph

4. 성능 평가

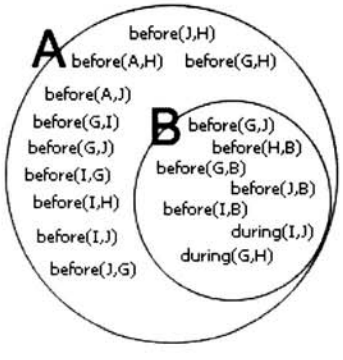
본 절에서는 하나의 이벤트 시퀀스 전체를 하나의 인터벌 이벤트로 요약하는 경우와 임계값을 고려하여 독립적인 서브 시퀀스로 분할하여 서브 인터벌 이벤트로 요약하는 경우 추출되는 정보의 정확성을 실험을 통하여 비교한다. 실험에

적용하는 인터벌 이벤트 관계에 대한 발생 빈도는 <표 7>과 같으며 환자 3,000명에 대하여 8,334건의 트랜잭션을 생성하여 적용하였다. <표 7>의 Occ. Ratio는 이벤트에 대한 관계가 발생할 빈도이며, Sub Sequence Occ. Ratio는 이벤트 시퀀스가 서브 시퀀스로 분할되는 빈도를 나타낸다.

<표 8>은 이벤트 타입에 대한 지지도를 기준으로 마이닝된 결과를 보여준다. 임계값을 이용하여 시퀀스를 서브 시퀀스로 분할하면 다른 시퀀스 사이에 영향을 주고 받는 관계 정보가 존재할 가능성이 커지게 된다. 따라서 추출되는 인터벌 이벤트 관계가 늘어난다.

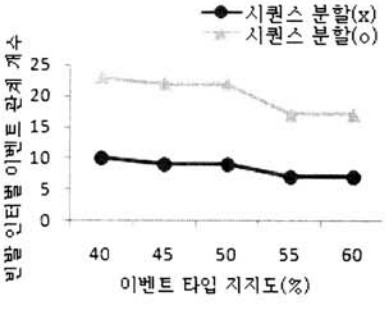
<표 8>은 빈발 인터벌 이벤트 지지도가 20%, 임계값 ε이 6인 경우 빈발 이벤트 타입을 결정하는 지지도를 기준으로 추출되는 빈발 인터벌 이벤트 관계 개수를 보여준다. 실험 결과 임계값을 두어 시퀀스를 서브 시퀀스로 분할하는 경우 분할하지 않는 경우보다 약 10%의 빈발 인터벌 이벤트 관계가 추출됨을 알 수 있다.

(그림 2)와 (그림 3)에서 이벤트 시퀀스를 서브 시퀀스로 분할하는 경우(A집합)에는 서브 시퀀스로 분할하지 않은 경우(B집합)보다 더 많은 빈발 인터벌 이벤트 연관 규칙을 추출 할 수 있음을 알 수 있다. 또한 서브 시퀀스로 분할되지 않아 빈발 인터벌 이벤트 관계에서 제외되었던 다수의 관계들이 서브 시퀀스로 분할되면서 빈발 인터벌 이벤트에 포함됨을 알 수 있었다. 즉 이벤트 시퀀스를 임계값을 고려하여 서브 시퀀스로 분할함으로써 보다 다양한 인터벌 연관 규칙 정보를 탐사할 수 있으며 분할되지 않는 경우 간과될 수 있



A집합: 서브 시퀀스로 분할한 경우
B집합: 서브 시퀀스로 분할하지 않은 경우

(그림 2) 빈발 이벤트 관계 다이어그램



(그림 3) 빈발 인터벌 이벤트 관계

〈표 7〉 데이터 생성 규칙

Relation	Event	Occ. Ratio (%)	Sub Sequence Occ. Ratio(%)
before	A,B	60	30
equals	C,D	40	30
meets	E,F	50	30
overlaps	G,H	70	30
during	I,J	80	30

〈표 8〉 이벤트 타입에 대한 지지도를 달리한 결과

Support (%)	ES	ES_{sub}	IERS	$IERS_{freq}$
40	18,268	-	375	10
40	18,268	21,056	388	23
45	15,830	-	228	9
45	15,830	18,258	237	22
50	15,830	-	228	9
50	15,830	18,258	237	22
55	12,830	-	117	7
55	12,830	14,804	124	17
60	12,830	-	117	7
60	12,830	14,804	124	17

는 인터벌 연관 규칙 정보를 추출할 수 있음을 알 수 있으므로 연속된 두 시점 사이의 인터벌이 긴 경우에는 서로 다른 독립적인 이벤트로 간주하는 것이 보다 합리적임을 알 수 있다.

5. 결론 및 향후 연구

본 논문에서는 이벤트 시퀀스를 시간 간격을 갖는 인터벌 이벤트로 요약하고 요약된 인터벌 이벤트들로부터 인터벌 연관 규칙을 찾아내는 새로운 시간 데이터 마이닝 기법을 제안하였다. 이벤트 시퀀스를 하나의 인터벌 이벤트로 요약하는 기본의 방법에 비하여 이벤트 시퀀스를 서브 시퀀스로 나누어 각 서브 시퀀스를 하나의 인터벌 이벤트로 요약하는 제안 방법은 보다 많은 인터벌 관계와 빈발 인터벌 관계를 추출할 수 있다. 그러므로 제안 기법은 인터벌 관계들 사이의 유효한 관계를 찾아줌으로서 이벤트 시퀀스를 분할하지 않는 다른 시간 데이터 마이닝 기법과 비교하여 보다 정확한 지식을 제공한다. 향후 연구 방향으로 인터벌 관계 규칙에서 인터벌 이벤트들 사이의 인과 관계를 보다 다양하게 추출하고 이를 추론 및 평가할 수 있는 방법에 대하여 연구하고자 한다.

참 고 문 헌

[1] A. Krokhin, P. Jeavons, and P. Jonsson, "Reasoning about Temporal Relations: The tractable Subalgebras of Allen's Interval Algebra," Journal of The ACM, Vol.50, Issue5, pp.591-640, 2003.

[2] C. Rainsford, J. F. Roddick, "Temporal Data Mining in Information Systems: A Model," Australasian Conference on Information Systems, 1996.

[3] S. Ye, J. A. Keane, "Mining Association Rules in Temporal

Databases," IEEE International Conference on Systems, Man, and Cybernetics, Vol.3, pp.2803-2808, Oct., 1998.
[4] R. Agrawal, R. Srikant, "Fast algorithms for mining association rules," The VLDB Conference, Santiago, Chile, Sep., 1994.
[5] H. Yun, D. Ha, B. Hwang, and K. Ryu, "Mining Association Rules on Significant Rare Data using Relative Support," Journal of Systems and Software, Vol.67, Issue3, pp.181-191, Sep., 2003.
[6] R. J. Swargam, M. J. Palakal, "The Role of Least Frequent Item Sets in Association Discovery," ICDIM '07. 2nd International Conference, Vol.1, pp.217-223, Oct., 2007.
[7] H. Mannila, H. Toivonen, and A. I. Verkamo, "Discovery of Frequent Episodes in Event Sequences," Data Mining and Knowledge Discovery, Vol.1, No.3, 1997.
[8] R. Agrawal, G. Psaila, E. Wimmers, and M. Zaot, "Querying Shapes of histories," The VLDB Conference, Zurich, Switzerland, 1995.
[9] R. Agrawal, K. Lin, Harpreet, S. Sawhney, and S. Kyuseok, "Fast Similarity Search in The Presence of Noise, Scaling, and Translation in Time Series Databases," The VLDB Conference, Zurich, Switzerland, 1995.
[10] B. Ozden, S. Ramaswamy, and A. Silberschatz, "Cyclic Association Rules," International Conference on Data Engineering, Orlando, USA, 1998.
[11] X. Chen, I. Petrounias, H. Heathfield, "Discovering Temporal Association Rules in Temporal Databases," International Workshop on Applications of Database Technology, 1998.
[12] S. Ramaswamy, S. Mahajan, and A. Silberschatz, "On The Discovery of Interesting Patterns in Association Rules," The VLDB Conference, New York, USA, 1998.
[13] Y. P. Huang, L. J. Kao, and F. E. Sandnes, "A Prefix Tree-Based Model for Mining Association Rules from Quantitative Temporal Data," IEEE International Conference on Systems, Man, and Cybernetics, Vol.1, pp.158-163, Oct., 2005.
[14] M. Ale, G. H. Rossi, 'An Approach to Discovering Temporal Association Rules', SAC'00, Italy, 2000.
[15] G. Berger, A. Tuzhilin, "Discovering Unexpected Patterns in Temporal Data Using Temporal Logic," Temporal Databases Research and Practice, Springer Verlag, 1998.
[16] S. Chakrabarti, S. Sarawagi, and B. Dom, "Mining Surprising Patterns Using Temporal Description Length," The VLDB Conference, New York, USA, 1998.
[17] J. Han, G. Dong, and Y. Yin, "Efficient Mining of Partial Periodic Patterns in Time Series Database," International Conference on Data Engineering, Sydney, Australia, 1999.
[18] J. Chung, O. Paek, J. Lee, and K. Ryu, "Temporal Pattern Mining of Moving Objects for Location-Based Service," International Conference on Database and Expert Systems Applications, 2002.
[19] C. Faloutsos, M. Ranganathan, and Y. Manolopoulos, "Fast Subsequence Matching in Time-Series Databases," The ACM SIGMOD Conference on Management of Data, Minneapolis, USA, 1994.

한 대 영
e-mail : nara9yo@gmail.com
2008년 전남대학교 전자컴퓨터공학부(학사)
2008년~현 재 전남대학교 전자컴퓨터공학부 석사과정
관심분야 : 데이터 마이닝, 멀티미디어, 임베디드

김 대 인
e-mail : dikim@chonnam.ac.kr
1998년 전남대학교 전산통계학과(이학석사)
2006년 전남대학교 전산통계학과(이학박사)
2004년~현 재 전남대학교 전자컴퓨터공학부 시간강사
관심분야 : 스트림 데이터, 데이터 마이닝, 디지털 콘텐츠

김 재 인
e-mail : sereno3@naver.com
2008년 전남대학교 전자컴퓨터공학부(학사)
2008년~현 재 전남대학교 전자컴퓨터공학부 석사과정
관심분야 : 스트림 데이터, 연속질의, USN 응용

나 철 수
e-mail : choulsu@daum.net
1998년 동신대학교(학사)
2007년~현 재 전남대학교 전자컴퓨터공학부 석사과정
관심분야 : 스트림 데이터 마이닝, 전자상거래

황 부 현
e-mail : bhhwang@chonnam.ac.kr
1978년 숭실대학교 전산학과(학사)
1980년 한국과학기술원 전산학과(공학석사)
1994년 한국과학기술원 전산학과(공학박사)
1980년~현 재 전남대학교 전자컴퓨터공학부 교수
관심분야 : 스트림 데이터 마이닝, 분산 시스템, 분산 데이터베이스