인터벌 이벤트를 고려한 시간 데이터 마이닝 기법

한대영*, 김재인*, 나철수*, 김대인*, 황부현*
*전남대학교 전자컴퓨터공학과
e-mail:abyo0111@naver.com

Temporal Data Mining for considering Interval Event

Dae-Young Han*, Jae-In Kim*, Chul-Su Na*, Dae-In Kim*, Bu-Hyun Hwang*
*Department of Computer Science, Chonnam National University

요 약

환자 이력, 구매자 이력, 웹 로그 이력 데이터에 대한 시간 데이터 마이닝에 대한 연구에서 시간 간격 관계 규칙을 찾아내는 것은 가변적인 시간 간격의 데이터를 하나의 이벤트로 요약하는 것은 합리적이지 못하다. 이는 그 이벤트가 가변적인 시간 간격 내에서 서로 독립적인 이벤트일 수 있기 때문이다. 그러므로 이벤트들의 시퀀스를 독립적인 서브 시퀀스로 나누어 각 서브 시퀀스별로 시간 간격을 갖는 인터벌 이벤트로 요약하는 것이 합리적이다. 본 논문은 이벤트 시퀀스를 시간 간격을 갖는 인터벌 이벤트로 요약하고 요약된 인터벌 이벤트들로부터 인터벌 관계 규칙을 찾아내는 새로운 시간 데이터 마이닝 기법을 제안하고 있다. 이 기법은 인터벌 관계들 사이의 규칙을 찾아줌으로서 기존의 데이터 마이닝 기법과 비교하여 질적으로 우수한 지식을 제공한다.

1. 서론

시간 데이터 마이닝은 연관 관계, 분류, 특징 추출을 포함하는 기존의 데이터 마이닝 기법을 확장하여 이벤트들 사이의 시간적 원인과 결과 관계를 표현하는 시간 연관 규칙을 찾아내는 새로운 기법이다. 이러한 시간 데이터 마이닝 기법은 유용한 연관 규칙을 발견하기 위하여 사용되는 순환 연관 규칙 탐사, 캘린더 형태로 표현된 시간 패턴을 만족하는 연관 규칙을 추출하기 위하여 사용하는 캘린더 연관 관계를 포함한다.

본 논문에서는 Allen의 시간 관계 대수 이론에 토대를 둔 시간 데이터 마이닝 기법을 제안한다. 시간 데이터 마이닝 기법의 기본 아이디어는 다음과 같이 요약할 수 있다.

- 이벤트들의 시퀀스를 인터벌을 갖는 이벤트로 요약하는 것이다. 특히 본 논문에서는 하나의 이벤트 시퀀스를 독립적인 서브 시퀀스들로 나누어 마이닝의 질을 향상시키고자 한다.
- 인터벌을 갖는 이벤트들 사이의 인터벌 관계 규칙을 찾아내는 것이다.

본 논문의 구성은 다음과 같다. 2절에서는 시간 데이터 마이닝에 대한 관련 연구를 논의하고, 3절에서는 시간 관계와 시간 간격(인터벌)을 정의한다. 그리고 시간 속성을 갖는 데이터 시퀀스를 인터벌을 갖는 데이터로 요약하는 알고리즘과 요약된 데이터로부터 인터벌 사이의 관계 규칙을 찾아내는 알고리즘을 기술한다. 4절에서는 3절에서 기술한 알고리즘을 예를 통하여 설명하고, 5절에서는 시뮬레이션을 통하여 제안하는 방법의 우수함을 보인다. 마지막으로 6절은 결론 및 향후 연구에 대하여 기술한다.

2. 관련 연구

시간 속성을 갖는 데이터로부터 유용한 지식을 찾아내기 위한 시간 데이터 마이닝에 대한 많은 연구가 이루어지고 있다[2,3]. 이러한 연구들은 순차 패턴, 유사 시퀀스, 시간 규칙을 찾아내는 것으로 분류된다.

순차 패턴 마이닝은 트랜잭션에 포함된 특정한 아이템 집합들의 시퀀스를 찾아내는 기법이다[4]. AprioriAll과 AprioriSome 같은 전형적인 순차 패턴 마이닝 알고리즘들은 연관 규칙 탐사 알고리즘인 Apriori에 토대를 두고 있다.

유사 시퀀스[5,6]는 주식, 상품 가격, 판매량 등과 같은 시계열 데이터로부터 유사한 데이터 패턴을 발견하기 위한 마이닝 기법이다.

시간 연관 규칙 탐사 기법은 시간 관계와 인과 관계를 갖는 시간 연관 관계 규칙을 탐사할 수 있다. 이 기법은 순환적으로 반복하는 연관 규칙을 발견하기 위한 순환 연관 관계 탐사[24], 캘린더의 형태로 표현된 시간 패턴을

만족하는 연관 규칙을 발견하는 캘린더 연관관계 탐사[7,8]를 포함하고 있다.

그러나 이러한 연구들은 시점을 갖는 데이터로 한정이 되어 있고 데이터들 간의 인터벌은 고려하지 않고 있다. 데이터들 사이에는 다양한 인터벌이 존재하며 이러한 인터벌을 고려한 시간 관계 규칙 탐사는 매우 복잡하며 보다 효율적으로 시간 관계 규칙을 탐사할 수 있는 알고리즘 개발에 대한 연구가 필요하다.

3. 인터벌 이벤트 연관 규칙

환자가 주기적으로 진찰을 받는다고 할 때, 한 번의 진찰은 하나의 트랜잭션으로, 그리고 특정 시간에 일어난 하나의 증상은 이벤트로 정의할 수 있다. 그러므로 트랜잭션은 한명의 환자에 대하여 한 시간에 일어난 이벤트들의 집합이다. 그리고 트랜잭션은 한 환자가 어느 한 시점에서 진찰받은 결과 및 증상들의 집합이며 트랜잭션에 포함된 모든 이벤트들의 시점은 동일하다. 그리고 이벤트 시퀀스는 한 환자의 한 이벤트 타입에 대한 이벤트들의 시퀀스로 정의하며, 시퀀스는 발생 시점을 기준으로 시간대별로 정리된다.

한 환자에 대하여 동일한 타입을 갖는 이벤트들의 시퀀스는 인터벌 이벤트로 요약되며, 인터벌 이벤트는 시간 간격에 대한 이벤트를 의미한다. 그리고 시간 데이터 마이닝에서 의미 있는 인터벌 이벤트로 간주되기 위하여 시퀀스 내의 이벤트들은 주어진 인터벌 내에서 균등한 간격으로 발생한다는 것이 전제가 되어야 한다. 그렇지 못하다면 서브 시퀀스로 나누는 임계값을 정의하고 정의된 임계값보다 긴 인터벌을 갖는 이벤트는 서로 다른 서브 시퀀스로 정의된다.

이러한 인터벌 이벤트들 사이에서 다음과 같은 연산자를 이용하여 인터벌들 사이의 관계를 표현할 수 있다. 다음은 인터벌들 사이의 관계를 표현하는 이진 연산자들이다.

- 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

before(x,y)는 이벤트 x의 종료가 이벤트 y의 시작 전에 이루어졌다는 것을 의미한다. equals(x,y)는 두 이벤트 x와 y의 시작시점과 종료시점이 동일하다는 것을 의미한다. meets(x,y)는 이벤트 x가 끝나는 동시에 이벤트 y가 시작된다는 것을 의미한다. overlaps(x,y)는 이벤트 x가 시작한 후 그리고 x가 종료하기 전에 이벤트 y가 시작되어 y가 종료하기 전에 x가 종료한다는 것을 의미한다.

4. 빈발 인터벌 이벤트 관계를 찾아내는 알고리즘

데이터베이스 내에 있는 트랜잭션들을 환자 식별자와 트랜잭션 발생 시점에 따라 정렬된 데이터베이스 DB_{sort}를 추출한다. 이 때 정의된 지지도 이하로 출현하는 이벤트 타입은 DB_{sort}에서 제거한다.

그림 1은 식별자 101을 갖는 환자의 트랜잭션을 발생 시점을 기준으로 정렬한 예이다.

그림 1. DB_{sort}

환자 식별자	트랜잭션 발생 시점	이벤트 타입
101	2007/1	A
101	2007/3	A,B
101	2007/4	A,F
101	2007/5	B,F
101	2007/6	B
101	2007/9	A
101	2007/10	A
101	2007/11	D
101	2007/12	A,D

그림 2. 후보 이벤트 타입 지지도

이벤트타입	지지도
A	4
B	4
C	2
D	3
E	2
F	2

그림 3. 빈발 이벤트 타입 지지도(70%이상)

이벤트타입	지지도
A	4
B	4
D	3

그리고 각 이벤트 타입의 지지도가 그림 2와 같은 경우 그림 1의 DB_{sort}는 빈발한 이벤트만을 포함한 DB_{sort}로 그림 4와 같이 정의된 지지도 이상의 빈발한 이벤트 타입만을 포함하는 이벤트들로 정의된다.

그림 4. 빈발 이벤트 타입만을 포함한 DB_{sort}

환자 식별자	트랜잭션발생 시점	이벤트타입
101	2007/1	A
101	2007/3	A,B
101	2007/4	A,F
101	2007/5	B,F
101	2007/6	B
101	2007/9	A
101	2007/10	A
101	2007/11	D
101	2007/12	A,D

DB_{sort} 내의 이벤트 타입은 빈발한 이벤트 타입만이 포함되어있으며, 각 환자의 각 이벤트 타입에 대한 이벤트 시퀀스를 구한다.

그림 5. 환자 101에 대한 이벤트 타입별 시퀀스

환자, 이벤트	시퀀스
101,A	<(A,1)(A,3)(A,4),(A,9)(A,10)(A,12)>
101,B	<(B,3)(B,5)(B,6>
101,D	<(D,11)(D,12)>

한 환자에 대하여 동일한 타입을 갖는 이벤트들의 시퀀스를 인터벌 이벤트로 요약하고, 이벤트 시퀀스를 스캔하여 연속된 두 시점의 사이가 정의된 임계값 ε보다 더 크면 서로 다른 서브 시퀀스로 나눈다. 그리고 나누어진 이벤트 시퀀스들은 시퀀스의 시작 시점과 종료 시점을 사용하여 인터벌 이벤트로 요약하여 인터벌 이벤트 집합 IES에 추가한다.

그림 6은 임계값 ε이 4인 경우 그림 5의 시퀀스에 대한 서브 시퀀스와 요약된 인터벌 이벤트를 보여준다.

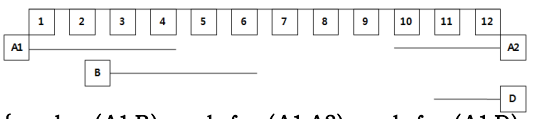
그림 6. 서브 시퀀스에서 산출된 인터벌 이벤트 (IES)

환자, 이벤트	서브 시퀀스 집합	인터벌 이벤트
101,A	{<(A,1)(A,3)(A,4)>,<(A,9)(A,10)(A,12)>}	{(A,[1,4]), (A,[9,12])}
101,B	{<(B,3)(B,5)(B,6>}	(B,[3,6])
101,D	{<(D,11)(D,12)>}	(D,[11,12])

IES에 있는 각 이벤트들의 인터벌 관계를 계산하여 인터벌 이벤트 관계들의 집합 IERS를 구하고, IERS로부터 각 인터벌 관계의 지지도를 계산한다. 인터벌 이벤트들의 관계를 구할 때 두 인터벌 이벤트 시간 간격이 임계값 ε보다 더 크다면 두 인터벌 이벤트 사이에 관계가 형성되지 못하므로 제외된다.

그림 7은 그림 6의 인터벌 이벤트들 사이에서 인터벌 이벤트 관계를 표현하는 연산자를 사용하여 인터벌 이벤트 관계를 구한 것을 도식화하고 있다.

그림 7. 인터벌 이벤트 관계 집합



{overlaps(A1,B), before(A1,A2), before(A1,D), before(B,A20, before(B,D), during(A2,D)}

인터벌 이벤트 관계 집합이 추출되었으면, 각 인터벌 관계에 대하여 지지도가 주어진 임계값보다 작은 것들은 제거한 후 빈발한 인터벌 관계들로 이루어진 빈발 인터벌 관계 집합 $IERS_{freq}$를 구한다.

$IERS_{freq}$로부터 산출된 인터벌 이벤트 관계들에 대한 지지도는 인터벌 이벤트 관계를 포함하고 있는 환자의 수를 의미한다. 각 인터벌 이벤트 관계들에 대한 지지도를 구한 후 주어진 지지도 보다 작은 것을 $IERS_{freq}$로부터 제거한다. $IERS_{freq}$의 인터벌 이벤트 관계들을 그룹화하고, 각 그룹의 지지도를 구함으로서 빈발 이벤트 관계 규칙을 구할 수 있다.

그림 8. 빈발 이벤트 관계 규칙

빈발 인터벌 관계 규칙	지지도
{during(A,D)}	2
{before(D,B)}	2
{before(A,B)}	3

마이닝의 과정을 통하여 추출된 최종 결과는 그림 8과 같으며 추출된 빈발 이벤트 관계 규칙을 통하여 다음과 같은 추론을 할 수 있다. “증상 A는 증상 D와 함께 발생하며 증상 A와 증상 D는 증상 B의 발병의 원인이 될 수 있다”라고 판단할 수 있다.

5. 시뮬레이션

본 절에서는 하나의 이벤트 시퀀스 전체를 하나의 인터벌 이벤트로 정의할 때와 임계값을 고려하여 독립적인 서브 시퀀스로 분할하여 서브 인터벌 이벤트로 정의하는 경우 추출되는 정보의 정확성을 모의실험을 통하여 비교한다. 모의실험에 적용하는 인터벌 이벤트 관계에 대한 발생 빈도는 그림 9와 같다.

그림 9. 데이터 생성 규칙

관계	이벤트	발생 빈도 (%)	서브 시퀀스 발생 빈도 (%)
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

그리고 모의실험을 위하여 환자 3,000명에 대하여 8,334건의 트랜잭션을 생성하여 적용하였다.

그림 10은 이벤트 타입에 대한 지지도를 달리 했을 때의 마이닝 결과를 보여준다. 임계값을 이용해 시퀀스를 서브 시퀀스로 분할하면 다른 시퀀스들과 서로 영향을 주고받는 가능성이 커지게 된다. 따라서 추출되는 인터벌이벤트 관계가 늘어남을 알 수 있다.

그림 10: 이벤트 타입에 대한 지지도를 달리한 마이닝 결과
(인터벌 이벤트 관계 지지도: 20%, 임계값 ε: 6)

지지도 (%)	시퀀스	서브 시퀀스	인터벌 이벤트 관계	빈발인터벌 이벤트 관계
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

그림 11은 빈발 인터벌 이벤트의 지지도가 20%, 임계값 ε이 6인 경우 빈발 이벤트 타입을 결정하는 지지도에 대한 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 11. 빈발 인터벌 이벤트 관계

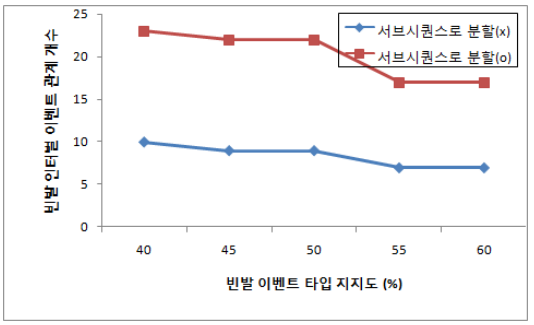
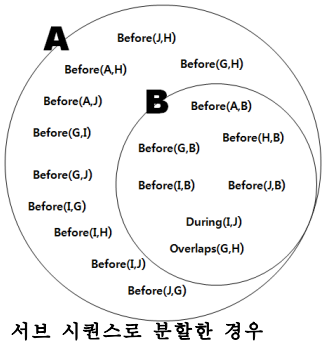


그림 12. 빈발 인터벌 이벤트 관계 다이어그램



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

그림 12에서 이벤트 시퀀스를 서브 시퀀스로 분할하는 경우(A집합)에는 서브 시퀀스로 분할하지 않은 경우보다 더 많은 빈발 인터벌 이벤트 관계를 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. 따라서 연속된 두 시점 사이의 인터벌이 긴 경우 서로 독립적인 이벤트로 간주하는 것이 보다 합리적이다.

6. 결론 및 향후 연구

본 논문에서는 이벤트 시퀀스를 시간 간격을 갖는 인터벌 이벤트로 요약하고 요약된 인터벌 이벤트들로부터 인터벌 관계 규칙을 찾아내는 새로운 시간 데이터 마이닝 기법을 제안하고 있다. 이 기법은 인터벌 관계들 사이의 관계를 찾아줌으로서 다른 데이터 마이닝 기법과 비교하여 질적으로 우수한 지식을 제공한다. 향후에는 시간 데이터 마이닝의 시간 복잡도 및 공간 복잡도를 고려하여 성능을 더욱 개선하는 연구를 진행하고자 한다.

참고문헌

[1] D. H. Kim, K. H. Ryu, H. S. Kim: A Spatiotemporal database model and query language, The journal of systems and software, Vol. 5(2000)
[2] X. Chen, I. Petrounias: A framework for temporal data mining, Int'l Conf. on Database and Expert Systems Applications(1998)
[3] S. Ye, J.A Keane: Mining association rules in temporal database, International Conference on Systems, Man and Cybernetics(1998)
[4] R. Agrawal, R. Stikand: Mining sequencial patterns, Int'l Conf. on Data Engineering, Taipei, Taiwan(1995)
[5] R. Agrawal, G. Psaila, E. Wimmers, M. Zaot: Quering shapes of histories, the VLDB Conference, Zurich, Switzerland(1995)
[6] R. Agrawal, King-lp Lin, Harpreet S. Sawhney, Kyuso다 Shim: Fast similarity search in the presence of noise, scaling, and translation in time series database, the VLDB Conf., Zurich, Switzerland(1995)
[7] B. Dzden, S. Ramaswamy, and A. Silberschatz: Cyclic association rules, Int'l Conf. on Data Engineering, Orlando, USA(1998)
[8] X. Chen I. Petrounias, H. Heathfield: Discovering temporal association rules in temporal database, Int'l. Workshop on Issues and Applications of Database Technology(1998)