빈발 인터벌 이벤트 관계에 기반한 이벤트 발생 예측 방법

송명진, 김대인, 황부현
전남대학교 전자컴퓨터공학부
e-mail:audwls0324@nate.com

A Method for Predicting Event Occurrence based on the Relations of Frequent Interval Events

Myung-Jin Song, Dae-In Kim, Bu-Hyun Hwang
Dept of Electronics and Computer Engineering,
Chonnam National University

요 약

시간 속성을 갖는 이벤트들의 집합에서 이벤트들 사이의 인과관계를 보다 정확히 파악할 수 있는 방법의 개발은 의료 분야 등의 응용에서 미리 발생할 이벤트에 발생 시점 예측을 위하여 필요하다. 본 논문은 이벤트들의 시퀀스를 독립적인 서브 시퀀스로 나누고 각 서브 시퀀스를 인터벌을 갖는 이벤트로 요약하여 인터벌 이벤트들 사이의 관계를 표현한다. 그리고 인터벌 이벤트 관계에서 원인 인터벌 이벤트가 결과 이벤트에 미친 영향 정도의 측정 방법을 개발하고 실험을 통하여 사용한 척도의 의미와 정확성을 파악한다. 실험 결과는 제안 방법이 지지도 기반의 평가보다 보다 우수함을 입증한다.

1. 서론

데이터 마이닝은 축적된 데이터를 분석하여 의사결정을 위한 가치 있는 지식을 추출하는 기법이다. 일반적으로 이벤트들은 발생 시점이라는 속성을 갖는다. 그리고 이벤트들을 고객 단위 또는 환자 단위로 축적하여 놓은 데이터베이스가 있다면 이 데이터베이스에서 유용한 정보를 탐사할 수 있다. 특히 이벤트들의 원인과 결과에 대한 인과 관계 규칙을 찾아낸다면 과거의 정보를 바탕으로 미래를 예측할 수 있는 중요한 정보로 이 규칙을 사용할 수 있다. 시간 관계 규칙 마이닝은 연관 관계, 분류, 특징 추출 등을 포함하는 기존의 데이터 마이닝 기법을 확장하여 이벤트들 사이의 시간적 관계 즉, 원인과 결과 관계를 표현하는 시간 연관 규칙을 찾아내는 새로운 기법이다. 기존의 시간 데이터 마이닝 기법에는 순환 연관 규칙 탐사, 캘린더 연관 관계 탐사 등이 있다. 그러나 이러한 연구들은 인터벌 데이터에 대한 시간 관계를 고려하지 않고 단지 이벤트 발생 시점만을 고려하는 한계가 있다.

어떤 이벤트의 원인이 다수라면 각각의 원인이 결과에 어느 정도 영향을 미쳤는가를 파악하는 것은 발생 가능한 이벤트에 대한 가장 중요한 원인 요소를 분석함으로써 미래에 발생 가능한 비상 상황에 대처할 수 있는 중요한 정보로 활용될 수 있다. 본 논문에서는 이벤트들 사이의 발생에 대한 영향력을 미친 정도를 표현할 수 있는 척도를 제안하고 타당성을 검증한다.

본 논문의 구성은 다음과 같다. 2장에서는 연관 규칙에 대한 관련 연구를 논의하고 3장에서는 인터벌 관계와 인터벌 관계 그래프에 대하여 정의한다. 4장에서는 이벤트들 사이의 인터벌 관계를 통한 인과 관계 정도를 측정 방법을 기술하고, 5장에서는 시뮬레이션을 통하여 제안하는 방법의 우수함을 보인다. 마지막으로 6장에서 결론 및 향후 연구에 대하여 기술한다.

2. 관련연구

시간 속성을 갖는 데이터로부터 유용한 지식을 찾아내기 위한 시간 데이터 마이닝에 대한 많은 연구가 이루어지고 있다[1][2]. 그리고 이러한 연구들은 순차 패턴, 유사 시퀀스, 시간 규칙을 탐사하는 것으로 분류된다.

순차 패턴 마이닝은 트랜잭션 집합에서 트랜잭션에 포함된 특정한 아이템 집합들의 시퀀스를 찾는 기법이다[3][4]. 즉, 순차 패턴 시퀀스 $\langle A,B,C\rangle$가 있는 경우 아이템 A, B, C가 서로 다른 트랜잭션에 존재하더라도 동일한 고객에 대한 트랜잭션으로 간주되어 고객들의 행동 패턴을 찾는다. 그리고 순차 패턴 마이닝 문제는 사용자가 명시한 최소 지지도를 만족하는 모든 시퀀스 중에서 최대 길이를 갖는 시퀀스를 탐사하는 것으로 연관 규칙 탐사 알고리즘인 Apriori[5] 방법에 기반한다.

유사 시퀀스 탐색은 시계열 데이터로부터 유사한 데이터 패턴을 발견하기 위한 마이닝 기법으로 전체 시퀀스 매칭 기법과 서브 시퀀스 매칭 기법으로 분류된다[6][7].

시간 연관 규칙 탐사 기법은 시간 관계와 인과 관계를 갖는 시간 연관 관계 규칙을 탐사할 수 있다. 이 기법은 순환 연관 관계 탐사[8]와 캘린더 형태로 표현된 시간 패

턴에 대한 연관 규칙을 발견하는 캘린더 연관 관계 탐사[9]를 포함한다.

기존의 연구들은 이벤트 발생 시점만 고려하며 이벤트 발생 시점을 확장한 인터벌에 대한 고려가 부족하다.

Wu와 Zhang은 긍정적 연관 규칙(Positive Association Rules)과 부정적 연관 규칙(Negative association Rules)을 마이닝하기 위한 방법을 제안하였다[10]. 긍정적 연관 규칙은 연관 규칙에 긍정적인 영향을 미치는 것을 의미하며, 부정적 연관 규칙은 주어진 연관 규칙의 영향력을 부정적으로 하는 연관 규칙을 의미한다. 그러나 지지도 기반의 연관 규칙만 사용하여 이벤트들 사이에 존재하는 인과 관계를 분석하고 영향력을 판단하는 데에는 한계가 있으며 이벤트 발생에 대한 인과 관계 정보를 측정할 수 있는 새로운 알고리즘 개발이 필요하다.

3. 인터벌 관계, 인터벌 관계 그래프

환자가 주기적으로 진찰을 받는다고 할 때, 한 번의 진찰은 하나의 트랜잭션으로, 그리고 특정 시간에 일어난 증상은 이벤트로 정의 할 수 있다.

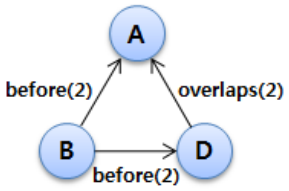
동일한 타입의 이벤트가 진행되어 가는 과정을 관찰하는 것은 증상 치료에 중요한 정보를 제공한다. 그리고 동일한 타입의 이벤트들을 시간적인 순서로 나열한 것은 이벤트 시퀀스로 정의 할 수 있다. 이벤트 시퀀스는 시작 시점과 종료 시점을 갖는 인터벌을 갖는 이벤트로 요약할 수 있으며, 이벤트 발생 시간이 시스템에 정의된 시간 간격보다 크다면 두 인터벌 이벤트는 지속된(continuous)이라고 볼 수 없으며 독립적(independent)인 관계로 보는 것이 합리적이다. 그러므로 이벤트 시퀀스를 독립적인 서브 시퀀스로 나누는 것이 타당하다.

서로 다른 인터벌 이벤트에 대하여 인터벌 관계를 계산함으로써 이벤트들 사이의 원인과 결과를 찾아 낼 수 있다. 표 1은 두 개의 인터벌 이벤트 x, y에 대한 관계를 표현하는 이진 연산자들이다.

<표 1> 이진 인터벌 관계 연산자

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

빈발 인터벌 관계는 주어진 지지도를 만족하는 인터벌 관계로 인터벌 관계를 갖는 고객의 수를 계산하여 빈발하게 발생하는 이벤트들 사이의 인과 관계를 파악할 수 있다. 또한 빈발 인터벌 관계를 그래프로 표현함으로서 인터벌 이벤트들 사이의 인과 관계를 한눈에 파악할 수 있다. 그림 1과 같이 원인과 결과를 표현하는 인터벌 관계 그래프의 각 노드는 인터벌 이벤트이고 하나의 에지는 두 인터벌 이벤트 사이의 인터벌 관계를 표현한다.



(그림 1) 인터벌 관계 그래프

그림1에서 각 에지가 나타내는 괄호 숫자는 각 관계의 지지도를 의미한다. 그림 1에서 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 추론할 수 있다. 그림 1에서 이벤트 D가 발생하는 중에 이벤트 A가 발생하지만, 이벤트 B가 발생한 후에 이벤트 D가 발생하므로 이벤트 B는 이벤트 A를 발생시킬 수 있는 원인중의 하나임을 알 수 있다. 이러한 정보는 이벤트 발생에 대한 원인 요소를 발견함으로써 특정 증상을 보이는 환자 치료를 위한 증상 원인을 발견할 수 있는 정보가 될 수 있다.

4. 인터벌 관계에서 이벤트 영향력 계산 절차

Input: 환자 진찰 기록 데이터베이스
Output: 환자별 인터벌 이벤트 관계 규칙에서의 영향력

Step 1. 빈발 이벤트 타입 계산
Step 2. 빈발 이벤트로 구성된 시퀀스 계산
Step 3. 이벤트 시퀀스로부터 독립적인 서브 시퀀스 계산
Step 4. 이벤트 시퀀스 요약
Step 5. 인터벌 이벤트들 사이의 인터벌 관계 계산
Step 6. 각 환자에 대하여 빈발 인터벌 관계 계산
Step 6. 각 환자별 인터벌 관계 그래프 생성
Step 7. 결과 이벤트에 영향을 끼친 이벤트들의 영향력 계산

데이터베이스의 트랜잭션들을 환자 식별자와 트랜잭션 발생 시점으로 정렬되어있다고 가정한다. 그리고 각 이벤트에 대한 지지도를 계산하고 정의된 지지도 이하로 발생된 이벤트 타입은 제거한다. 다음 표 2는 설명을 위하여 환자 200명에 대한 진찰 기록 데이터베이스에서 이벤트들의 지지도를 구한 결과이다.

<표 2> 빈발 이벤트 타입 지지도(80명 이상)

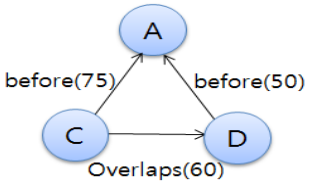
이벤트타입	지지도
A	150
~~*B*~~	~~40~~
C	80
D	90
~~*E*~~	~~50~~

빈발 이벤트 타입만으로 구성된 데이터베이스에서 각 환자별로 이벤트 타입에 대한 이벤트 시퀀스를 구한다. 이벤트 시퀀스들은 시퀀스의 시작 시점과 종료 시점을 사용하여 인터벌 이벤트로 요약한다. 예를 들어 <(*B*,3)(*B*,5)(*B*,6)> 시퀀스는 인터벌 이벤트 (*B*,[3,6])으로 요약된다. 그리고 인터벌 이벤트에서 인터벌 관계를 계산하여 인터벌 이벤트 관계들의 집합을 구하고, 각 인터벌 관계의 지지도를 계산한다. 인터벌 이벤트 관계 집합이 추출되면 각 지지도 이하의 인터벌 관계들을 제거함으로써 표 3과 같은 빈발 인터벌 관계 집합을 구한다.

<표 3> 빈발 이벤트 관계

빈발 인터벌 관계 규칙	지지도
{before(*C*,*A*)}	75
{before(*D*,*A*)}	50
{overlaps(*C*,*D*)}	60

표 3의 빈발 인터벌 관계 규칙은 그림 2와 같은 인터벌 관계 그래프로 표현된다. 그리고 인터벌 관계 그래프를 사용하여 인터벌 이벤트들 사이에 존재하는 관계 정보를 알 수 있다.



(그림 2) 인터벌 관계 그래프 예

본 논문에서는 다음과 같은 척도를 정의하여 이벤트들 사이의 원인 및 발생 정도를 측정한다.

$$\frac{sup\ (A \rightarrow X)}{sup\ (A)} \quad (1), \qquad \frac{sup\ (A \rightarrow X)}{sup\ (X)} \quad (2)$$

위 식에서 $sup(X)$는 이벤트 X에 대한 지지도를 의미하며 기호 →는 한 이벤트로부터 다른 어떤 이벤트로의 관계를 의미한다. 예들 들어 $X \rightarrow_b Y$는 X before Y 관계를 의미한다. 식 1은 이벤트 A가 발생할 때 이벤트 X가 발생할 가능성을 표현하며, 식 2는 이벤트 X가 이벤트 A의 영향력을 얼마나 받는가를 표현한다. 그리고 식 1과 2를 사용하여 최종적으로 이벤트 A가 이벤트 X에 대한 영향력을 식 3과 같이 측정할 수 있다.

$$EFF(A \rightarrow X) = \alpha\left(\frac{sup\ (A \rightarrow X)}{sup\ (A)}\right) + \beta\left(\frac{sup\ (A \rightarrow X)}{sup\ (X)}\right) \quad (3)$$
$$, \text{ where } \alpha+\beta=1$$

식 3에서 α와 β는 가중치이며, $0 \leqq EFF(A \rightarrow X) \leqq 1$의 값을 갖는다. $EFF(A \rightarrow X)=0$인 경우는 A는 X에 전혀 영향을 끼치지 않았다는 것을 의미한다. $EFF(A \rightarrow X)=1$인 경우는 100% A가 X에 영향을 미쳤다는 것을 의미한다. 즉 이벤트 A가 발생하면 이벤트 X가 반드시 발생하며 또한 이벤트 X는 이벤트 A가 발생하는 경우에만 발생한다느 것을 의미한다.

그리고 α와 β의 값을 조정하여 최적의 영향력을 끼치는 정도를 정할 수 있다. 일반적인 경우 α와 β를 0.5로 설정하며 응용에 따라 α와 β를 조절하여 사용함으로써 다양한 이벤트들 사이에 존재하는 발생 영향력의 정도를 측정할 수 있다.

예를 들어 $sup(A)=150$, $sup(C)=80$, $sup(C \rightarrow_b A)=75$, $\alpha=0.5$, $\beta=0.5$일 때, 그림 2의 인터벌 관계 그래프에서의 영향력은 다음과 같이 해석된다.

$EFF(C \rightarrow_b A)=0.5(75/80)+0.5(75/150)=0.47+0.25=0.72$

이는 이벤트 C가 종합적으로 72%정도의 영향력을 이벤트 A에 끼쳤다고 볼 수 있다.

5. 시뮬레이션

본 절에서는 식 3에 기반하여 인터벌 이벤트 관계 사이에 존재하는 이벤트들 사이의 발생 영향력을 측정한다. 실험에 적용하는 인터벌 이벤트 관계에 대한 발생 빈도는 표 4와 같으며 환자 5,000명에 대하여 38,579건의 트랜잭션을 생성하여 적용하였다. 표 4의 Occ. Ratio는 이벤트에 대한 관계 발생 빈도이며, Sub Sequence Occ. Ratio는 이벤트 시퀀스가 서브 시퀀스로 분할되는 빈도를 나타낸다.

<표 4> 데이터 생성 규칙

Relation	Event	Occ. Ratio (%)	Sub Sequence Occ. Ratio(%)
before	B,A	80	10
equals	C,F	30	30
meets	E,G	30	30
overlaps	D,A	60	10
during	B,D	70	10

<표 6> 빈발 이벤트 타입(2,000명 이상)

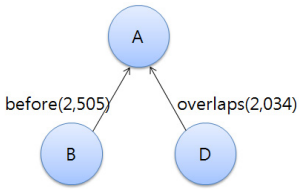
이벤트타입	지지도
A	4,631
B	4,729
D	4,455

지지도 40% 이상을 만족하는 빈발 이벤트 타입으로 이벤트 타입 A, B, D가 발견되었다.

<표 5> 발견된 빈발 인터벌 이벤트 관계

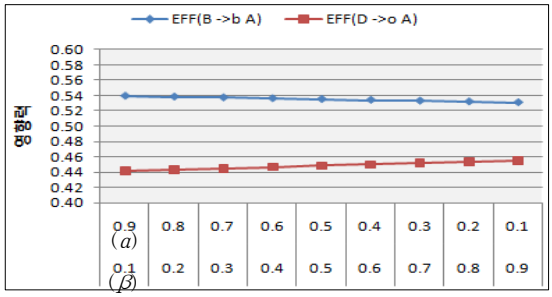
빈발 인터벌 관계 규칙	지지도
{before(B,A)}	2,505
{overlaps(D,A)}	2,034

그리고 빈발 이벤트 타입들 사이에 존재하는 빈발 인터벌 관계 규칙으로 before(B,A)와 overlaps(D,A)가 발견되었다. 표 5의 빈발 인터벌 관계 규칙으로 구한 인터벌 관계 그래프는 그림 3과 같다.



(그림 3) 인터벌 관계 그래프

그림 3의 인터벌 관계 그래프의 에지 방향으로 이벤트 A가 이벤트 B, D로부터 영향을 받음을 알 수 있다.



(그림 4) 가중치 변화에 따른 영향력 변화

식 3에서의 α, β는 가중치로서 관심 있는 척도에 대한 정도를 달리함으로서 영향력 지수를 측정 할 수 있다. $EFF(B \rightarrow_b A)$에서 β값의 감소에 따라 영향력이 감소하는 것으로 보아 이벤트 A가 이벤트 B에 상대적으로 크게 영향력을 받지 않음을 의미한다. 즉 이벤트 B는 이벤트 A로 인하여도 발생하지만 상대적으로 이벤트 A가 아닌 다른 이벤트에 의하여서도 발생함을 의미한다. 그러나 $EFF(D \rightarrow_o A)$에서는 α값의 증가에 따라 영향력이 커지는 것으로 보아 이벤트 D가 이벤트 A의 발생에 대한 많은 영향력을 줌을 의미한다.

그림 4에서와 같이 본 연구에서 제안한 방법은 인터벌 이벤트 관계 사이에서 원인 이벤트가 결과 이벤트에 끼친 영향 정도를 파악할 수 있다.

6. 결론 및 향후 연구

본 논문에서는 시간 속성을 갖는 이벤트에 대한 데이터베이스에서 인터벌 이벤트들 사이의 관계 규칙을 찾아내고, 원인 인터벌 이벤트가 결과 이벤트에 미친 영향 정도의 측정 방법을 제안하였다. 제안한 방법은 이벤트간의 영향력의 척도를 제공함으로서 의학 분야 등의 응용 분야에서 특정 증상에 대한 발생 요인의 척도를 측정할 수 있다. 향후 연구로 긍정적인 영향력의 관계 뿐 만아니라 부정적인 영향력의 관계와 같은 다양한 이벤트 간 영향력에 대한 연구를 진행하고자 한다.

참고문헌

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

[2] 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.

[3] R. Agrawal, R. Srikant, "Fast algorithms for mining association rules," The VLDB Conference, Santiago, Chile, Sep. 1994.

[4] 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, Issue 3, pp.181-191, Sep. 2003.

[5] 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.

[6] R. Agrawal, G. Psaila, E. Wimmers, and M. Zaot, "Querying Shapes of histories," The VLDB Conference, Zurich, Switzerland, 1995.

[7] 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.

[8] B. Ozden, S. Ramaswamy, and A. Silberschatz, "Cyclic Association Rules," International Conference on Data Engineering, Orlando, USA, 1998.

[9] X. Chen, I. Petrounias, H. Heathfield, "Discovering Temporal Association Rules in Temporal Databases," International Workshop on Applications of Database Technology, 1998.

[10] X. Wu, C. Zhang, and S. Zhang, "Efficient mining of both positive and negative association rules,", ACM Transactions on Information Systems, Vol. 22(3), pp.381-405, July, 2004.