다차원 스트림 데이터 요약 및 인과 관계 탐사를 위한 실시간 데이터 마이닝 기법

송명진, 김대인, 황부현
전남대학교 전자컴퓨터공학부
audwls0324@nate.com, dikim007@naver.com, bhhwang@chonnam.ac.kr

A Method of Realtime Mining for Summarization and Discovery of a Casual Relationship based on Multidimensional Stream Data

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

요 약

실시간 데이터 마이닝 기법은 다양한 종류의 센서에서 수집된 다차원 스트림 데이터들 사이에 존재하는 의미있는 정보를 탐사할 수 있다. 전통적인 데이터베이스 시스템에서의 마이닝 기법은 정적인 데이터베이스에 기초하므로 실시간으로 수집되는 스트림 데이터는 시간 속성을 갖는 인터벌 이벤트로 요약되어야 한다. 이 논문은 다차원 스트림 데이터 환경에서 스트림 데이터를 요약하고 이들 사이에 존재하는 인과 관계를 탐사하는 실시간 데이터 마이닝 기법을 제안한다. 제안 기법은 센서에서 수집되는 데이터의 대부분이 객체의 정상적인 상태 데이터임을 고려하여 의미있는 이상 이벤트를 선별하여 전송한다. 그리고 스트림 데이터의 연속성을 고려하며 스트림 데이터를 세 가지 상태의 이벤트로 요약하고 인과 관계 규칙을 탐사한다. 인과 관계 규칙은 시간에 따라 이벤트 발생에 영향력을 미치는 원인 이벤트를 발견함으로써 이벤트의 발생을 미리 예측할 수 있다.

1. 서 론

USN(Ubiquitous Sensor Network) 환경은 다양한 종류의 센서를 사용하여 다차원 스트림 데이터를 수집한다. 그리고 수집된 다차원 스트림 데이터는 센서 감지 시점과 같은 시간 속성을 가지므로 시간 데이터마이닝 기법을 적용하여 유용한 예측 정보를 탐사할 수 있다 데이터 마이닝은 데이터베이스에 잠재된 정보를 추출하는 기술로써 최근에는 데이터의 시간 속성을 고려하여 인과 관계 규칙을 탐사하는 시간 데이터 마이닝 기법들이 연구되고 있다[1]. 그러나 기존의 연구들은 정적인 데이터베이스에 기초하여 시간 규칙을 탐사하므로 연속적으로 발생하는 스트림 데이터에는 적용할 수 없다 그리고 센서에서 연속적으로 수집되는 스트림 데이터의 크기는 무한하므로 모든 스트림 데이터를 저장하고 처리하는 것은 불가능하다. 또한 센서에서 수집된 대부분의 스트림 데이터는 객체의 정상적인 상태를 의미한다 그러므로 수집된 모든 스트림 데이터를 저장하여 인과 관계를 탐사하는 것은 비효율적이며 이 논문에서는 객체에 영향을 줄 수 있는 의미있는 이상 이벤트만을 선별하여 데이터 마이닝을 실시한다 그리고 수집 데이터의 시간 속성을 이용하여 세 가지 상태의 이벤트로 요약하고 이벤트들 사이에 존재하는 인과 관계 규칙을 탐사한다시간 속성을 갖는 스트림 데이터에 기초하여 인과 관계 규칙을 탐사 하는 것은 미래 발생 가능한 위험 이벤트 발생을 예측할 수 있으므로 중요하다

이 논문의 구성은 다음과 같다 2절에서는 기존의 스트림 데이터 처리와 데이터 마이닝에 대한 관련 연구를 기술하고 3절에서는 다차원 스트림 데이터의 요약 기법을 기술한다. 4절에서는 다차원 스트림 데이터의 연관 관계 탐사 알고리즘을 기술한다 끝으로 5절에서는 결론 및 향후 연구에 대하여 기술한다

2. 관련연구

USN 환경이 발전함에 따라 다양한 종류의 센서에서 감지하는 스트림 데이터는 연속적이며 그 크기는 무한하므로 모든 데이터를 손실 없이 저장하고 처리하는 것은 불가능하다. 따라서 스트림 데이터 저장 및 요약 기법에 대한 연구가 진행되고 있다[2,3]. 그러나 [2,3]에서 제안한 기법들은 한 종류의 센서에서 수집된 단일 스트림 데이터 요약만을 고려하므로 다차원 스트림 데이터에는 적용할 수 없다.

연속적으로 발생하는 스트림 데이터는 무한한 크기를 가지므로 모든 이벤트를 서버로 전송하는 것은 저장 비

용과 처리 비용을 증가시킨다[4]. 또한 센서에서 수집되는 스트림 데이터는 객체의 정상적인 상태를 나타내는 이벤트가 대부분이며 데이터의 발생 빈도가 데이터의 중요도를 의미하지 않는다[5]. 그러므로 수집된 스트림 데이터 중 상대적으로 의미있는(significant) 이상 이벤트를 선별하여 전송하는 연구가 필요하다.

인터벌 데이터 마이닝은 데이터의 연속성과 시간 속성을 바탕으로 인터벌 이벤트를 요약하고 이벤트들 사이에 존재하는 연관 규칙을 탐사한다[1]. 그리고 이벤트의 연속성을 고려하여 보다 합리적인(reasonable) 인터벌 이벤트를 구성하고 인터벌 이벤트의 발생빈도를 기준으로 빈발 이벤트들 사이에 존재하는 연관 규칙을 탐사한다. 그러나 [1]의 기법은 정적인 데이터베이스를 바탕으로 연관 규칙을 탐사하므로 연속적으로 무한하게 수집되는 스트림 데이터에 적용하기에는 한계가 존재한다.

이 논문에서는 다양한 종류의 센서에서 수집되는 다차원 스트림 데이터의 시간 속성을 이용하여 세 가지 상태의 이벤트로 요약하고 이벤트들 사이의 인과 관계를 탐사하는 실시간 데이터 마이닝 기법을 제안한다. 제안하는 방법은 다차원 스트림 데이터 중에서 객체의 비정상적인 상태를 의미하는 이상 데이터를 선별하여 이벤트화하고 이벤트의 발생 시점에 따라 세 가지 상태의 이벤트로 요약한다. 그리고 제안 방법을 통하여 이벤트 사이의 인과 관계 정보를 탐사함으로써 시간에 따라 이벤트 발생에 영향을 미치는 원인 이벤트를 사전에 예측할 수 있다.

3. 다차원 스트림 데이터의 요약 기법

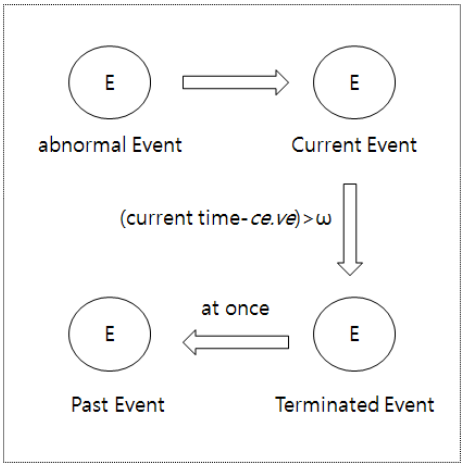
다차원 스트림 데이터는 둘 이상의 센서에서 실시간으로 수집되는 스트림 데이터를 의미한다[6]. 그러나 수집된 다차원 스트림 데이터는 무한한 크기를 가지므로 모든 이벤트를 전송 및 처리하는 것은 한계가 존재한다. 또한 센서에서 수집되는 스트림 데이터는 객체의 정상적인 상태를 나타내는 이벤트가 대부분이므로 모든 이벤트를 처리하는 것은 비효율적이다. 따라서 다차원 스트림 데이터 중 이상 이벤트를 선별하여 이벤트 사이의 인과 관계를 탐사하는 것이 필요하다. 이상 이벤트란 객체의 상태 변화에 영향을 미치는 의미있는 데이터를 의미 한다[5]. 제안 방법은 센서에서 이상 이벤트 감지 시점의 다차원 스트림 데이터를 서보로 전송함으로써 데이터 전송 비용과 처리 비용을 줄일 수 있다. 그리고 객체에 영향을 주는 의미있는 이상 이벤트에 기초하여 인과 관계를 탐사하고 객체에 영향을 미칠 수 있는 원인 이벤트 발생을 예측한다.

센서에서 수집된 다차원 스트림 데이터의 인과 관계를 탐사하기 위하여 다차원 스트림 데이터를 이벤트E로 요약한다. 그리고 센서 감지 시점과 같은 시간 속성t를 정용하여 이벤트 E는 (E,t)로 기술한다. 또한 이벤트 E의 연속 여부를 판별하기 위하여여 이벤트의 발생 시작 시점 vs와 종료 시점 ve를 갖는 인터벌 이벤트 $(E, [vs,ve])$로 요약한다. 이때 이벤트 E의 발생 시점 및 종료 시점은 각 $E.vs$와 $E.ve$로 기술한다[1].

다차원 스트림 데이터는 센서를 통하여 실시간으로 수집되므로 기존의 정적인 데이터베이스에서 사용한 마이닝 기법의 이벤트 정의를 사용할 수 없다. 따라서 이 논문에서는 이벤트 발생 시점 및 이벤트 연속성을 고려하여 세 종류의 이벤트를 정의한다. 그리고 이벤트를 정의하기 위하여 사용하는 이벤트 임계값과 이벤트들 사이의 인과 관계 탐사를 위한 임계값에 대한 정의는 정의3.1과 같다. 그리고 다차원 스트림 데이터에서 인과 관계를 추출하기 위하여 요약하는 세 종류의 이벤트 정의는 정의 3.2와 같다.

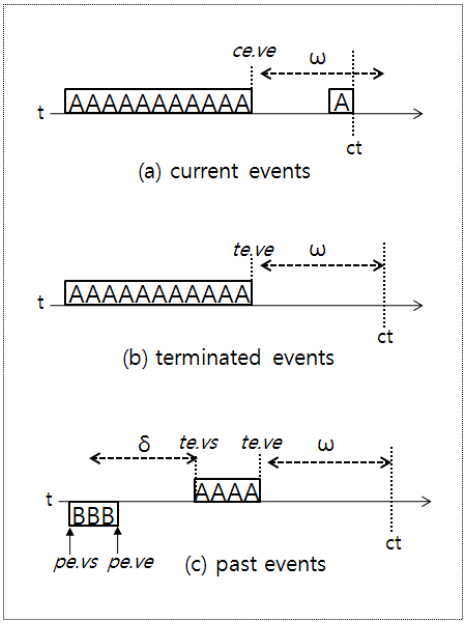
[정의 3.1] (임계값)
- 이벤트 임계값(ω) : 동일한 이벤트 타입의 이벤트 지속성을 판단하기 위한 임계값이다.
- 관계 임계값(δ) : 서로 다른 이벤트들 사이의 연관성을 판단하기 위한 임계값이다.

[정의 3.2] (이벤트)
- 현재 이벤트 (Current Events : *ce*) : 동일한 이벤트 E의 마지막 발생 시점과 현재 시간과의 시간 간격이 이벤트 임계값 ω안에서 연속적으로 발생하는 이벤트이다.
- 종료 이벤트 (Terminated Events : *te*) : 동일한 이벤트 E의 마지막 발생 시점과 현재 시간과의 시간 간격이 이벤트 임계값 ω동안에 발생하지 않는 이벤트이다.
- 과거 이벤트 (Past Events : *pe*) : 가장 빠른 시작 시점을 갖는 현재 이벤트의 시작 시점과 기 완료된 종료 이벤트의 종료 시점의 시간 간격이 관계 임계값δ보다 작은 이벤트이다.



(그림 1) 동일한 타입의 이벤트 상태 변화도

이 논문에서 스트림 데이터는 세 가지 상태의 이벤트로 정의된다. 이벤트 타입 Ei의 이벤트 E는 발생 시점에 따라 그림 1과 같이 상태 변화한다. 그리고 이상 이벤트의 발생 시점과 현재 시간(current time)과의 시간 간격이 이벤트 임계값 ω안에서 연속적으로 발생하면 현재 이벤트로 요약되고 현재 시간을 기준으로 이벤트가 이벤트 임계값 ω동안 발생하지 않는다면 현재 이벤트는 종료 이벤트로 상태 전이된다. 연속적으로 종료 이벤트는 시작 시점과 종료 시점을 갖는 과거 이벤트로 즉시 상태 전이된다.



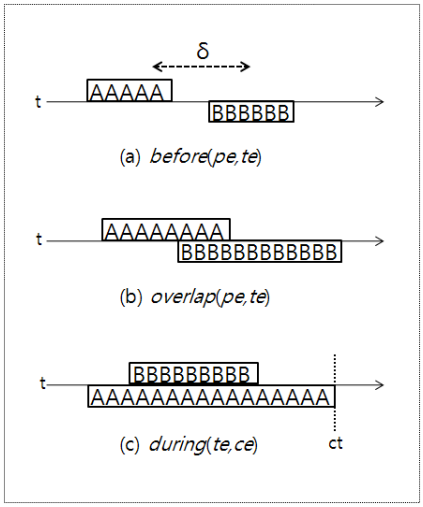
(그림 2) 정의된 이벤트의 표현

그림 2에서 t는 시간(time)의 흐름이고, ct는 현재 시간(current time)을 의미한다. 그림 2의 (a)에서 현재 이벤트는 이벤트 *E*의 마지막 발생 시점 *ce.ve*와 현재 시간에 발생한 동일한 이벤트 *E*의 시간 간격이 정의된 이벤트 임계값 ω보다 크지 않고 연속적으로 이벤트 *E*가 발생하는 이벤트이다. *E*의 발생이 현재 시간을 기준으로 *ce.ve*와의 시간 간격이 ω보다 크다면 이벤트는 그림 2의 (b)와 같은 종료 이벤트가 된다. 연속적으로 종료 이벤트는 과거 이벤트가 되고 과거 이벤트는 가장 빠른 시작 시점을 갖는 현재 이벤트의 시작 시점 *ce.vs*와 기 완료된 종료 이벤트의 종료 시점 *te.ve*과의 시간 간격이 관계 임계값 δ보다 작은 이벤트 이다. 과거 이벤트 집합의 과거 이벤트들은 연속적으로 다른 타입의 종료 이벤트와 시간 간격을 검사한다. 그리고 과거 이벤트의 종료 시점 *pe.ve*와 종료 이벤트의 시작 시점 *te.vs*와의 시간 간격이 관계 임계값 δ보다 크다면 과거 이벤트는 과거 이벤트 집합에서 삭제된다. 이러한 이유는 두 이벤트 사이의 시간 간격이 관계 임계값 δ보다 크다면 두 이벤트의 발생은 서로 연관성이 없는 독립적인 관계이기 때문이다.

이 논문에서는 세 종류의 이벤트에 근거하여 이벤트들 사이에 존재하는 인과 관계를 탐사한다. 이벤트들 사이에는 서로 다른 이벤트 타입의 이벤트 발생 상태에 따라 표 1과 같은 인과 관계가 존재한다. 제안 방법에 적용하는 인과 관계는 Allen의 연산자에 근거하며 표 1과 같다.

<표 1> 인터벌 관계

관계	조건
before(*pe*,*te*)	$pe.ve \le te.vs$
overlap(*pe*,*te*)	$pe.vs < te.vs < pe.ve$
during(*te*,*ce*)	$ce.vs \le te.vs$



(그림 3) 인터벌 관계의 표현

표 1과 같은 인터벌 관계는 서로 다른 이벤트 타입의 이벤트들 사이에 그림 3과 같이 존재한다. 인터벌 관계 *before*(*pe*,*te*)는 과거 이벤트 *pe*의 발생이 종료 되고 시간 간격이 관계 임계값 δ보다 작은 종료 이벤트 *te*가 발생했을 때의 두 이벤트 관계를 의미한다. 인터벌 관계 *overlap*(*pe*,*te*)는 과거 이벤트 *pe*가 발생하고 과거 이벤트 *pe*가 완료하기 전 종료 이벤트 *te*가 발생하고 과거 이벤트 *te*가 완료됨을 의미한다. 인터벌 관계 *during*(*te*,*ce*)는 현재 이벤트 *ce*가 연속적으로 발생하고 있을 때 과거 이벤트 *te*의 발생시점을 기준으로 탐사한다.

4. 다차원 스트림 데이터의 연관 관계 탐사 절차

이 절에서는 다차원 스트림 데이터를 이벤트로 요약하고 이벤트 사이에 존재 하는 인과 관계 규칙을 탐사하는 알고리즘을 기술한다. 첫 번째 단계는 다양한 센서에서 발생한 다차원 스트림 데이터 중에서 객체의 정상 범위를 벗어나는 의미있는 이상 이벤트만을 선별하여 서버로 전송한다. 두 번째 단계는 이벤트 임계값과 관계 임계값을 적용하여 이상 이벤트를 현재 이벤트, 종료 이벤트, 과거 이벤트로 요약한다. 그리고 표 1에 기초하여 이벤트들 사이에 존재하는 인터벌 관계를 탐사한다.이상 이벤트 발생 시점의 이벤트 전송 과정은 알고리즘1과 같다.

Input : 센서에서 발생한 다차원 스트림 데이터
Output : 이상 이벤트가 발생한 시점의 스트림 데이터

For each 센서 *s*에 대하여
  if (시점 *t*에 발생한 스트림 데이터 *Strs* ∉ 객체의 정상 조건)
  then 서버로 이벤트 타입 *Ei*와 감지 시점 *t* 전송

알고리즘 1. 이상 이벤트 선별 알고리즘

다차원 스트림 데이터에서 선별된 의미있는 이상 이벤트는 세 종류의 이벤트로 요약된다. 이상 이벤트의 종료 시점과 현재 시간의 시간 간격이 정의된 이벤트 임계값 보다 작다면 현재 이벤트, 크다면 종료 이벤트로 요약된다. 그리고 과거 이벤트는 연속적으로 다른 이벤트 사이의 인과 관계를 탐사하기 위하여 과거 이벤트의 유효성이 검사된다. 과거 이벤트의 종료 시점과 종료 이벤트의 시작 시점 사이의 시간 간격이 관계 임계값 보다 크다면 다른 이벤트 발생에 영향을 주지 않는 이벤트로 간주하여 과거 이벤트 집합에서 삭제한다. 이벤트 분류 과정과 과거 이벤트의 유효성 판단하는 과정은 다음 알고리즘2와 같다.

Input : 이상 이벤트,
이벤트 임계값 ω, 관계 임계값 δ
Output : 현재 이벤트 집합 *CES*,
종료 이벤트 집합 *TES*, 과거 이벤트 집합 *PES*

For each 센서 s에서 발생한 이벤트 *Ei*의 종료 시점 *Ei.ve*에 대하여
If ($Ei.ve - ct \leq \omega$)
then 동일한 이벤트 *Ei*의 새롭게 발생한 종료 시점 *ve*와 함께 (*Ei*, [*vs*, *ve*])로 요약하여 현재 이벤트 집합 *CES*추가
else 이벤트 *Ei*의 마지막 종료 시점 *ve*와 함께 (*Ei*, [*vs*, *ve*])로 요약하여 종료 이벤트 집합 *TES*에 추가
For each 과거 이벤트 집합 *PES*의 모든 과거 이벤트 *PEi*에 대하여
if ($PEi.ve - CEj.vs > \delta$)
then 이벤트 *Ei*는 과거 이벤트 집합 *PES* 삭제

알고리즘 2 이벤트 요약

이상 이벤트는 이벤트 발생 시점과 현재 시간의 시간 간격에 따라 현재, 종료, 과거 이벤트로 요약되고 이를 바탕으로 이벤트 사이의 인과 관계 규칙을 탐사한다. 이벤트의 시작 시점과 종료 시점을 기초로 이벤트 사이의 연관 관계 규칙을 탐사 하는 방법은 다음 알고리즘3과 같다.

Input : 현재 이벤트 집합 *CES*, 종료 이벤트 집합 *TES*,
과거 이벤트 집합 *PES*, 관계 임계값 δ
Output : 인터벌 관계 집합 *IRS*

For each 현재 이벤트 집합 *CES*, 종료 이벤트 집합 *TES*,
과거 이벤트 집합 *PES*의 이벤트에 대하여
For each 서로 다른 타입의 이벤트들 사이의 시점 관계를 관계 임계값 δ과 비교 하여 인터벌 이벤트 관계 집합 *IERS* 구성

알고리즘 3. 인터벌 연관 규칙 탐사

다차원 스트림 데이터의 이상 이벤트를 선별하여 발생 시점에 따라 세 가지 상태의 이벤트로 요약하여 인과 관계를 탐사하였다. 인과 관계 규칙을 탐사하는 것은 시간에 따라 이벤트 발생에 영향력을 미치는 원인 이벤트를 미리 예측할 수 있으므로 중요하다.

5. 결론 및 향후 연구

이 논문에서는 다차원 스트림 데이터에서 의미있는 이상 이벤트를 선별하고 세 종류의 이벤트로 요약하여 이벤트 사이의 인과 관계 탐사하는 데이터 마이닝 기법을 제안하였다. 제안한 방법은 다차원 스트림 데이터에서 의미 있는 이상 이벤트를 선별하여 전송함으로써 센서에서 서버로의 데이터 전송을 최소화하고 서버는 수신한 이상 이벤트를 세 종류의 이벤트로 정의하여 인과 관계 규칙을 탐사하였다. 추출된 인과 관계 규칙은 이벤트 발생에 영향을 미치는 원인 이벤트를 발견함으로서 이벤트의 발생을 사전에 예측 가능하게 한다. 향후 연구로 다차원 스트림 데이터의 요약기법을 사용한 인과 관계 탐사 알고리즘의 성능을 평가하고 이벤트 사이에 영향을 받은 정도를 파악할 수 있는 연구를 진행하고자 한다.

참고문헌

[1] 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.

[2] G. S. Manku, and R. Motwani, "Approximate Frequency Counts over Data Streams," In Proc. of Very Large Data Bases, pp. 346-357, 2002

[3] D. Kim, P. park, and B. Hwang, "Mining Association Rule for the Abnormal Event in Data Stream Systems," Journal of Korea Information Processing Society, Vol.14-D, No.5, pp.483-490, 2007

[4] H. Li, S. Lee, and M. Shan, "Online Mining (Recently) Maximal Frequent Itemsets over Data Streams," In Proc. of Research Issues in Data Engineering: Stream Data Mining and Applications 2005, pp.11-18, 2005.

[5] D. Han, D. Kim, J. Kim, C. Na, and B. Hwang, "A Method for Mining Interval Event Association Rules from a Set of Events Having Time Property," Journal of Korea Information Processing Society, Vol.16-D, No. 2, pp.185-190, 2009

[6] D. Kim, P. Park, H. Kim, and B. Hwang, "Mining Association Rules in Multidimensional Stream Data," Journal of Korea Information Processing Society, Vol.13-D, No.6, pp.765-774, 2006