韓國컴퓨터情報學會 論文誌
第16卷 第5號, 2011. 05.

2011-16-5-4-5

USN 기반의 화재감시 응용을 위한 센서 데이터 처리 시스템

박 원 익*, 김 영 국**

A Sensor Data Management System for USN based Fire Detection Application

Won-Ik Park*, Young-Kuk Kim**

요 약

오늘날 센서 기술의 발전 및 보급으로 인해 USN 기반의 실시간 모니터링 응용에서의 센서 데이터 처리 시스템에 대한 연구가 활발히 진행 되고 있다. 센서 데이터는 시간에 따라 빠르게 변화하고 연속적인 저수준 상태의 방대한 양의 데이터를 생성하는 특성을 갖는다. 하지만 엔드유저는 상대적으로 고수준 상태의 데이터에 관심이 있기 때문에 빠르게 변화하고 연속적인 대량의 저수준 센서 데이터를 효과적으로 처리 하는 시스템이 필수적이다. 본 논문에서는 USN 기반의 화재감시 응용에서 OLAP(On-Line Analytical Processing) 기술을 이용한 다차원 분석 질의 처리 기능과 학습기반 분류기를 통한 이상치 탐지 기능을 제공하는 센서 데이터 처리 시스템을 제안한다. 실험 시나리오를 통해 우리의 센서 데이터 처리 시스템에 대한 타당성을 검증하며 실험에 필요한 다양한 센서 데이터는 자체 개발한 센서 데이터 생성기를 이용한다.

▸ Keyword : 스트림 데이터, 화재감시, 다차원 분석, 이상치 탐지

Abstract

These days, the research of a sensor data management system for USN based real-time monitoring application is active thanks to the development and diffusion of sensor technology. The sensor data is rapidly changeable, continuous and massive row level data. However, end user is only interested in high level data. So, it is essential to effectively process the row level data which is changeable, continuous and massive. In this paper, we propose a sensor data management system with multi-analytical query function using OLAP and anomaly detection function using learning based classifier. In the experimental section, we show that our system is valid through

• 제1저자 : 박원익 • 교신저자 : 김영국
• 투고일 : 2010. 12. 28, 심사일 : 2011. 01. 21, 게재확정일 : 2011. 02. 07,
* 충남대학교 컴퓨터공학과 박사과정 (Dept. of Computer Engineering, Chungnam National University)
** 충남대학교 컴퓨터공학과 교수 (Dept. of Computer Engineering, Chungnam National University)

the some experimental scenarios. For the this, we use a sensor data generator implemented by ourselves.

▸ Keyword : USN, Stream data, Fire Detection, OLAP, Multi-dimensional analysis, Anomaly detection

I. 서 론

USN(Ubiquitous Sensor Network)은 모든 사물에 센서를 부착하여 사물에 대한 인식 및 주변 환경정보를 탐지하여 네트워크 통해 실시간으로 제공 및 관리가 가능하도록 센서 네트워크를 확장한 것이다. 센서 기술의 발전 및 보급으로 인해 화재 감시 시스템, 수질 관리 시스템, 농작물 재배 시스템등과 같은 다양한 실시간 모니터링 응용분야에 널리 활용되고 있다. 이로 인해 데이터의 흐름이 시간에 따라 빠르게 변화하고 연속적인 특성을 갖는 센서 데이터 즉, 스트림 데이터가 방대한 양으로 생성된다. 하지만 대부분의 센서 데이터는 저수준 상태로 생성이 되는 반면 엔드유저는 상대적으로 고수준 상태의 데이터에 관심이 있다. 따라서 센서 데이터 처리 시스템은 연속적이면서 대량의 저수준 센서 데이터로부터 고수준 상태의 지식을 실시간으로 발견하기 위한 기능이 필요하다. 본 논문에서는 센서 데이터를 위한 다차원 분석 질의 처리기능과 학습 기반의 분류기를 통한 이상치 탐색 기능을 갖는 센서 데이터 처리 시스템의 구현 방법을 제안한다.

USN 기반의 실시간 모니터링 응용을 위한 다차원 분석을 위한 질의는 저수준의 센서 데이터로부터 분석되어지는 것이 아니라 고수준의 센서 데이터로부터 분석되어져야 한다. 이로 인해 다차원 분석 질의의 응답시간은 높게 된다. 따라서 모든 차원에 따른 결과값을 미리 저장하여 제공하는 것이 일반적이며 데이터 웨어하우스 시스템에서의 OLAP이 이와 같은 방법을 통해 다차원 분석 질의를 효과적으로 제공한다. 하지만 기존의 OLAP은 저장된 고수준의 데이터의 크기가 저수준 데이터의 크기보다 크고 다중 스캔이 필요하기 때문에 대량의 저수준 데이터를 처리해야 하는 스트림 데이터에 적용하기는 어렵다. 또한 끊임없이 발생되는 스트림 데이터의 특징으로 인해 기존 데이터 마이닝 기법을 이용한 이상치 탐지 기법은 적용이 어렵다.

본 논문에서는 스트림 큐브 기법[1]과 기존 단순 베이지안 분류기(Naive Basian Classifier)[2]를 변형하여 다차원 분석 질의 및 이상치 탐지 기능을 제공하는 지능적인 센서 데이터 처리 시스템의 구현 방법을 제안한다.

본 논문의 구성은 다음과 같다. 2장에서는 관련연구로 스트림 데이터 처리를 위한 관련 연구들에 살펴보고, 3장에서는 본 논문에서 제안한 센서 데이터 처리 시스템의 구조를 설명하고 4장에서는 제안한 시스템의 결과를 설명하며 6장에서는 결론 및 향후 연구를 기술한다.

II. 관련 연구

센서 데이터와 같은 스트림 데이터는 연속적이며 복잡하고 한시적인 접근만이 가능한 특징을 갖는다. 또한 메모리 사용에 제한적이며 시간에 따른 순서를 가지기 때문에 기존 DBMS에서 관리 되는 데이터셋과는 달리 랜덤한 접근이 불가능하다[3]. 따라서 스트림 데이터의 효과적인 처리를 위해서는 새로운 데이터 구조, 기술 그리고 알고리즘을 이용한 데이터 스트림 처리 시스템(Data stream Management System)이 필요하다. 기존 데이터 스트림을 효율적으로 처리하기 위한 다양한 프로젝트가 존재하며 대표적인 프로젝트의 기능은 다음과 같다.

- STREAM[3]: 스탠포드 대학에서 개발한 STREAM은 연속적으로 입력되는 데이터 스트림과 저장되어 있는 데이터 집합에 대한 연속 질의(Continuous Query: CQ)를 기존 DBMS의 SQL을 확장하여 지원한다.
- AURORA[4]: 하나의 질의를 다수의 오퍼레이터로 나누고 효과적인 질의 스케줄링을 하여 대용량 스트림 데이터의 실시간 처리를 지원한다.
- TelegraphCQ[5]: 연속 질의를 사용하여 매우 유동적이고 대용량의 데이터 스트림으로부터 원하는 정보를 효과적으로 얻기 위해 오퍼레이터에 대한 스케줄링을 제공하며, 오퍼레이터간의 통신을 제공하는 기능을 제공한다.
- COUGAR[6]: 센서 네트워크용 분산 데이터 처리 시스템으로 네트워크 변화에 동적으로 적응 가능하며 높은 유연성과 확장성을 제공한다.
- NiagaraCQ[7]: 인터넷 환경에서 연속 질의를 효과적으로 처리하기 위해 유사한 구조를 갖는 연속 질의를 그룹화하고 공통적으로 필요한 연산을 한 번만 수행함으로써 질의 처리 성능 향상과 CPU 및 메모리 자원 사용을 최소화할 수 있는 기능적 특성을 가진다.
- OpenCQ[8]: 인터넷 상의 분산되어 존재하는 이질 정

보들의 변화를 효과적으로 모니터링 할 수 있다.

이상에서 언급한 프로젝트들은 주요 처리 대상이 되는 데이터가 온도, 강수량, 습도 등의 간단한 센서 데이터를 기준으로 한다는 점과 요약정보를 저장하여 스트림 데이터를 관리한다는 공통점이 있다. 하지만 본 논문에서 제안하는 센서 데이터 처리 시스템은 USN 환경에서의 실시간 모니터링 응용에서 다차원 분석 질의 처리기능과 학습 기반의 분류기를 통한 이상치 탐색 기능을 제공하는 보다 지능적인 센서 데이터 처리 시스템이라는 차별성을 갖는다.

스트림 데이터의 효율적인 처리를 위해 필요한 관련 요소 기술은 스트림 데이터를 저장하기 위한 무한한 공간을 확보할 수 없어 근사적인 답을 얻기 위한 압축 기술, 스트림 데이터의 다양한 분석 수행을 위한 스트림 데이터 마이닝 기술 및 다차원 분석 질의 처리 기술이 필요하며 대표적인 관련연구는 〈표 1〉과 같다.

표 1. DSMS를 위한 관련 기술
Table 1. Related technologies for DSMS

기술	설명
압축 기술	▸ 임의 추출[9,10,11] 임의의 표본을 추출하여 저장 관리
	▸ 슬라이딩 윈도우[12,13,14] 임의로 표본을 추출하는 것 대신에 최근 스트림 데이터를 주기적으로 갱신하여 저장 관리
	▸ 히스토그램, 웨이블릿[15,16,17,18] 스트림 데이터의 빈도 분포를 이용한 대략적인 요약 정보 저장 관리
마이닝 기술	▸ 분류[19,20,21,22] 스트림 데이터에 대한 실시간 의사 결정 지원
	▸ 군집화[23,24,25,26] 스트림 데이터의 유사성을 기반으로 실시간 그룹화 및 이상치 결정 지원
	▸ 패턴 분석[27,28,29] 스트림 데이터에 대한 빈발 패턴 및 이상치 결정 지원
질의 처리	▸ 데이터 큐브[30,31,32] OLAP 데이터 모델인 큐브 생성을 통한 다차원 분석 질의 처리

III. 센서 데이터 처리 시스템

1.1 센서 데이터 처리 시스템 구조

센서 데이터 처리 시스템은 이종의 센서들로 구성된 다양한 유비쿼터스 네트워크와 이를 활용하여 사용자에게 서비스를 제공하기 위한 다양한 목적의 시스템 사이에서 동작한다. 〈그림 1〉은 본 연구에서 개발한 센서 데이터 처리 시스템의 전체 구조를 보인다.

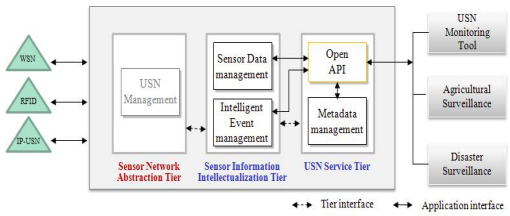


그림 1. 센서 데이터 처리 시스템 구조
Fig. 1. Sensor data management system

센서 데이터 처리 시스템은 센서 네트워크 추상화 계층(Sensor Network Abstraction Tier), 센서 정보 지능화 계층(Sensor Information Intellectualization Tier)과 USN 서비스 계층(USN Service Tier)으로 구분된다. 가장 하위 계층은 센서 네트워크 추상화 계층으로 다양한 복수의 이기종 네트워크 공통인터페이스 기능을 제공하고 다양한 센서네트워크의 상태에 대한 지속적 모니터링과 제어 기능을 제공한다. 본 논문에서 개발한 센서 데이터 처리 시스템은 센서 네트워크 추상화 계층에 USN 관리 컴포넌트(USN Management component)를 통해 센서 정보를 수집한다. 중간 계층인 센서 정보 지능화 계층은 센서 정보를 지능화하는 계층이다. 이는 USN 인프라로부터 제공되는 센서정보의 실시간 관리 기능과 센서정보에 대한 다양한 질의 처리 기능 및 능동적인 이상치 탐지 기능을 제공하는 역할을 한다. 마지막으로 센서 데이터 처리 시스템의 최상위 계층인 USN 서비스 통합 계층은 USN을 기반으로 하는 다양한 응용 서비스를 지원하기 위한 역할을 담당한다. 본 논문에서는 중간 계층인 센서 정보 지능화 계층의 센서 데이터 관리기와 지능적인 이벤트 관리기 위주로 설명한다.

1.2 센서 데이터 관리기 구조

센서 데이터 관리기는 센싱 데이터를 기반으로 사용자에게 다차원 분석 질의를 지원하는 기능을 수행한다. 〈그림 2〉는 센서 데이터 관리기의 구조를 보인다.

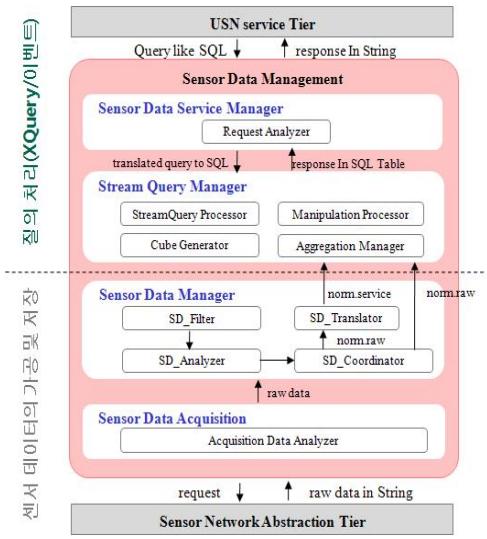


그림 2. 센서 데이터 관리 컴포넌트 구조
Fig. 2. Sensor data management component

센서 데이터 관리기는 크게 4개의 모듈로 구성되며 각 모듈의 기능은 다음과 같다.

- Stream Data Service Manager: USN service Tier로부터 들어오는 다양한 형태의 질의를 분석하여 해당 질의를 Stream Query Manger에게 전달하는 역할을 한다.
- Stream Query Manager: Stream Service Manager를 통해 요청받은 다차원 분석 질의 처리 및 Sensor Data Manager를 통해 들어오는 정형화된 센서 데이터를 관리하는 역할을 한다.
- Sensor Data Manager: Sensor Data Acquisition을 통해 수집된 특정 센서 데이터에 대해 유효성 검사, 단위 필드정보 인출, 정규화 된 레코드 생성 및 가공된 데이터를 생성하여 대량의 스트림 데이터를 걸러주는 역할을 한다.
- Sensor Data Acquisition: 다양한 종류의 센서 데이터(WSN, RFID, IP-USN) 처리 및 실제 센서 데이터를 수집하는 역할을 한다.

1.2.1 센서 데이터 복합 질의 처리

센서 데이터에서의 복합 질의는 수집된 센서 데이터에 대한 다차원적인 시각으로 요약된 통계 정보에 대한 질의를 말한다. 이러한 다차원 데이터 분석은 기존의 데이터베이스 및 데이터 웨어하우스 시스템에서의 OLAP 기술이 가장 대표적이다. OLAP은 데이터 집합을 다차원 배열 관점으로 바라보고 데이터를 가시화하거나 요약된 통계를 생성해주는 기술이다. 하지만 OLAP 기술을 이용하여 USN 환경에서 발생하는 방대한 센서 데이터 즉, 스트림 데이터에 대한 다차원 분석을 처리하기에는 무리가 있다. USN에서 발생하는 센서 데이터의 방대한 양과 지속성으로 인한 메모리 공간의 한계와 다차원 질의 결과를 위한 많은 조인 연산으로 인한 응답속도 지연 때문이다.

본 논문에서는 부분 데이터 큐브를 이용하여 다차원 분석 시 문제가 되는 메모리 공간 및 응답 속도 문제를 개선한다.

1.2.2 다차원 데이터 모델(데이터 큐브)

다차원 분석 질의 처리를 위한 데이터 모델은 OLAP을 위한 데이터 모델인 데이터 큐브를 이용한다. 이 모델은 데이터를 데이터 큐브 형태로 본다. 데이터 큐브는 데이터가 여러 차원으로 모델링되고 보이도록 해주며, 차원(dimension)과 사실(fact)로 정의된다. 일반적인 용어로, 차원은 한 조직이 그것에 대하여 기록하기를 원하는 시각이나 개체이며 사실은 수치측도(numerical measures)를 갖는다. 데이터 웨어하우징 분야에서는 이러한 데이터 큐브를 큐보이드(cuboid)라고 부른다. 그리고 주어진 차원들의 집합에서 큐보이드의 격자(lattice)를 구축할 수 있는데, 각 큐보이드는 다른 단계의 요약, 즉, group by로 데이터를 보여준다. 이 큐보이드의 격자를 데이터 큐브라고 부른다. 각 큐보이드는 일반적으로 개념계층을 갖는다. 개념계층은 하위개념들의 집합으로부터 상위의 보다 일반적인 개념들로의 사상의 연속을 정의한다.

실제 대전 서구 3개동(괴정동, 가장동, 용문동)을 화재 감시 지역(zone)으로 설정한 <그림 3>을 참고하여 차원 지역에 대한 개념계층을 설명한다.

그림 3. 타겟 지역(대전 서구 지역)
Fig. 3. Target zone(seo-gu, Daejeon)

지역에 대한 거리이름(street_name) 값은 ○○길, ○○

길, ○○길 등을 포함한다. 하지만, 각 거리이름(street_number)은 그것이 속해 있는 ○○동(dong)으로 사상될 수 있다. 예를 들어, 괴정길은 괴정동으로 새들길은 가장동으로 사상될 수 있다. 이러한 사상들은 하위개념들의 집합(예:street_name)을 보다 일반적인 상위개념(예:dong)으로 사상하면서 지역차원에 대한 개념계층을 형성한다. 시설물(facility)에 대한 개념계층은 대형 시설물(F_large)과 그 내부에 위치하는 소형 시설물(F_small)로 개념계층을 형성한다. 즉, 소형 시설물은 해당하는 대형 시설물로 사상이 가능하다. 마지막으로, 차원 센서타입(type)의 개념계층은 센서의 카테고리(category)와 그에 해당하는 센서의 종류(sensor)로 개념계층을 형성한다. 예를 들어 온도센서(sensor)는 환경센서(category)로 사상되며 배터리센서는 계측센서 카테고리로 사상된다. 〈표 2〉는 본 연구진에서 제안하는 차원과 각 차원에 대한 개념계층을 나타낸다.

표 2. 화재감시 응용을 위한 차원 및 개념계층
Table 2. Dimensions and Hierarchies

Dimension	Zone	Facility	Type
Hierarchy	dong (Z1) street_number (Z2)	F_large (F1) F_small (F2)	category (T1) sensor (T2)

지금까지 설명한 데이터 웨어하우스를 위한 차원과 사실값은 일반적으로 관계형 DBMS를 이용하여 표현이 가능하며 널리 쓰이는 개념적 스키마 구조는 스타스키마, 눈송이 스키마, 사실성군이 존재한다.

본 논문에서 제안하는 센서 데이터 처리 시스템에서는 스타 스키마 구조를 사용하여 다차원 분석을 위한 통합된 데이터 집합 저장소를 구축한다.

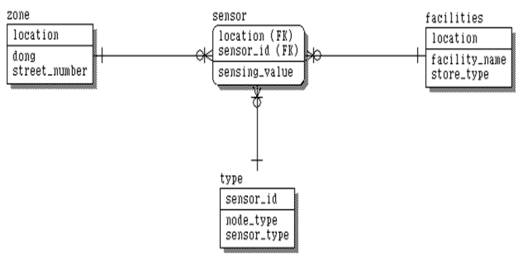


그림 4. 논리적 데이터 모델
Fig. 4. A logical data model

〈그림 4〉는 화재감시응용에서의 스타 스키마 구조를 따르는 화재 감시 응용을 위한 논리적 데이터 모델이다. 〈그림 4〉에서 보듯이 화재감시응용에서는 지역, 시설물, 센서타입 차원에 대한 테이블과 센싱값을 위한 테이블의 관계를 정의한다. 센서로부터 수집되는 센싱값들을 이용하여 다차원 분석을 위한 데이터 큐브를 생성한다. 이때 차원에 따른 데이터 큐보이드의 수가 정해지며 이는 메모리 사용량 및 응답시간에 지대한 영향을 주는 중요 요인이 된다. 차원의 수(n)에 대한 큐보이드 경우의 수는 2n 으로 차원의 수가 많아질수록 생성해야 하는 큐보이들의 수가 급격하게 증가하게 된다. 더욱이 각 차원에서의 개념계층까지 고려한다면 큐보이드들의 경우의 수는 $\prod_{i=1}^{n}(L_i+1)$ 와 같다. 여기에서 Li 는 차원 i의 개념계층의 수를 의미한다. 마지막으로, 각 차원의 카디널러티(cardinality)까지 고려한다면 그 수는 굉장히 크게 된다. 따라서 이러한 문제를 해결하기 위해 전체 데이터 큐브를 구성하는 대신 핵심 데이터 큐보이드만으로 이루어진 부분 데이터 큐브를 생성하는 방법을 사용한다. 이를 통해 스트림 데이터에 대한 다차원 분석 시 문제가 되는 메모리 사용량 및 응답시간을 효율적으로 해결한다.

1.2.3 부분 데이터 큐브

화재감시응용을 위한 차원은 〈표 2〉에서 보듯이 3개이며 각각 2개의 개념계층으로 정의한다. 차원에 대한 카디널리티를 고려하지 않은 전체 큐보이드의 수는 3x3x3 = 27개 이다. 스트림 데이터를 처리해야 하는 27개 이상의 데이터 큐보이드를 저장 관리하는 것은 성능상 비효율적이다. 따라서 본 연구진은 핵심 큐보이드만을 이용하여 부분 데이터 큐브를 생성한다. 이를 위해 o-layer와 m-layer라는 두 개의 레이어를 정의한다. o-layer는 핵심 데이터 큐브 중 가장 상위 데이터 큐보이드를 의미하며 m-layer는 가장 하위 데이터 큐보이드를 의미한다. 따라서 두 레이어를 정의하고 그 사이에 있는 데이터 큐보이들만 관리를 하게 되면 메모리 사용량 및 응답속도 문제에 대한 해결이 가능하다. 하지만 전체 큐보이드를 구성하지 않기 때문에 구성되지 않은 큐보이드들을 처리하지 못하는 문제가 있다. 따라서 두 개의 레이어는 해당 응용에서의 전문가가 충분히 고려하여 정의한다. 이와 같은 부분 데이터 큐브에 더하여 성능 향상을 이끌어내기 위해서 부분 데이터큐브에 속한 큐보이들 중에서 조금 더 빈번하게 발생하는 큐보이드만을 인기경로라고 지정하여 인기경로에 속한 큐보이드들을 실시간 유지관리 한다. 인기경로 또한 해당응용에 전문가의 의견을 통하여 정하거나 통계적인 분석으로 얻을 수 있다. 〈그림 5〉는 두 개의 레이어를 이용하여 구성한 부분 데이터 큐브를 보인다.

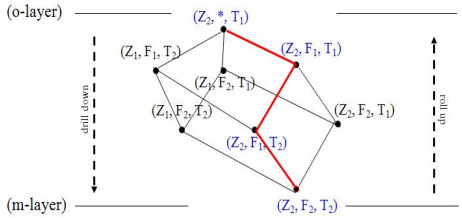


그림 5. 부분 데이터 큐브
Fig. 5. A partial data cube

〈그림 5〉에서 빨간선으로 연결된 큐보이드들이 바로 인기 경로에 속한 데이터 큐보이드들이다. 인기 경로에 속한 큐보이드의 셀 값에는 요약된 센싱정보가 실시간으로 업데이트 되어 관리된다. 이로 인해서 인기경로에 속한 큐보이드들의 셀 값을 요청하는 질의에 대해서는 즉각적인 응답을 할 수 있게 된다. 또한 인기경로에 속하지 않은 큐보이드들에 대한 질의도 응답이 가능하며 그 원리는 각 차원 개념계층 간의 사상 능력이다.

1.2.4 경동 시간 윈도우

경동 시간 윈도우(tilted time window) 프레임은 시간 차원을 요약하기위한 효과적인 방법이다. 기존의 OLAP처럼 요약 데이터를 추가 저장하는 방식은 제한되어 있는 메모리상에서 데이터를 처리해야 하는 데이터 스트림에는 적합하지 않기 때문에 기존 연구에서는 데이터 스트림의 특성을 기반으로 데이터 큐브를 재구성하였다. 데이터 분석가들은 오래전 데이터보다는 최근의 데이터변화에 더욱 관심이 많기 때문에 최근의 데이터일수록 정밀하게 저장되어지길 원하게 된다. 따라서 시간의 정밀도에 차이를 두어 최근 데이터일수록 정밀하게 저장되도록 경동 시간 윈도우 프레임을 이용한다. 〈그림 6〉에서와 같은 경동 시간 구조는 15분은 1쿼터, 4쿼터는 1시간, 24시간은 1일로 구성 되어져 있다. 매 분마다 데이터가 들어오게 될 경우, 경동 시간 구조에서는 15분이 되는 시점에 데이터를 집계하여 쿼터에 저장하고, 분 데이터는 다시 처음부터 저장하게 되기 때문에 기존의 분단위로 요약 데이터를 추가하는 방식보다 상당한 메모리 절감 효과가 있을 수 있다. 하지만, 이런 저장 방식으로 인하여 과거 데이터에 대해서는 세세한 질의를 할 수 없게 된다.

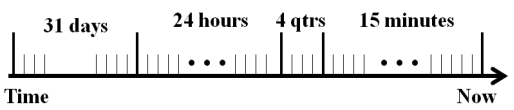


그림 6 경동 시간 윈도우 예
Fig. 6. An example of tilted time window

1.3 지능적 이벤트 관리기 구조

지능적 이벤트 관리기(Intelligent Event Management)는 센서 데이터 관리기를 통한 복합질의 응답 기능에 더하여 주기적으로 추가되는 각 센서 값에 대한 의사결정을 도와주는 응답 결과를 지능적으로 제공하는 역할을 한다. 예를 들어 화재감시 응용에서 주기적으로 수집되는 센서 값에 대한 화재발생 위험도를 실시간으로 제공한다면 화재 예방효과를 제공할 수 있게 된다. 화재발생 위험도는 복합 질의를 통해 얻을 수 없는 고차원적으로 가공된 정보이며 이를 효율적으로 처리하기 위한 방법이 필요하다. 본 논문에서는 스트림 데이터에 대한 실시간 의사결정을 지원하기 위해 학습기반의 모델을 구축하고 주기적으로 추가되는 센서 값의 화재발생 위험도를 결정하는 방법을 제안한다. 〈그림 7〉은 지능적 이벤트 관리기의 구조를 보인다.

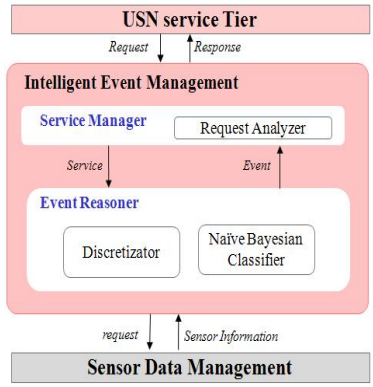


그림 7. 지능적 이벤트 관리기 구조
Fig. 7. An example of Intelligent Event Management

지능적 이벤트 관리기는 크게 두개의 모듈로 구성되며 각 모듈의 기능은 다음과 같다.

- Service Manager: USN service Tier로부터 들어오는 다양한 타겟 응용을 분석하여 해당 응용을 Event Reasoner에게 전달하는 역할을 한다.
- Event Reasoner: Service Manager를 통해 전달받은 응용에 대해 학습기반의 모델을 생성하며 이를 위해 센서 데이터 관리기에서 기 수집된 센서 요약정보를 활용한다. 또한, 모델 생성을 위한 센서 요약정보에 대한 이산화를 수행하는 역할을 한다.

스트림 데이터의 실시간 의사결정 지원을 위해 학습기반의 분류 기법인 단순 베이지안 분류기를 이용하여 주기적으로 추가되는 센서값의 화재발생 위험도를 실시간으로 결정한다. 즉, 제안 방법은 정상범위에 있는 센서값에 대한 모델을 구축

하고 화재발생 위험이 높은 비정상범위의 데이터를 실시간으로 모니터링 하여 USN 응용에 비정상 이벤트에 대한 능동적인 정보제공을 가능하게 한다. 데이터 마이닝 분야에서는 이와 같은 비정상 이벤트를 이상치(Outlier, Anomaly)라고 하며 이를 탐색하는 기법을 이상치 탐지 기법이라 한다.

본 논문에서는 실제 센서에 대한 클래스라벨이 있는 훈련 데이터 셋의 획득이 현실적으로 불가능하여 가상의 클래스 라벨 속성을 갖는 훈련 데이터 셋을 이용한 지도학습 기반의 단순 베이지안 분류기를 이용하여 이상치를 탐지한다. 단순 베이지안 분류기를 통해 이상치를 검색하는 방법은 비교적 간단하다. 하지만 USN환경에서의 속성 집합은 센서종류에 따른 센싱값들로 이루어졌는데 이러한 센싱값들은 범주형 속성값이 아닌 연속형 속성값을 갖는다. 따라서 단순 베이지안 분류기의 활용을 위해서는 이러한 연속형 속성값들을 범주형 속성값으로 변환시키는 작업이 필요하다. 이는 연속형 속성값의 범위가 매우 방대하기 때문에 몇 개의 연속형 속성값을 하나의 범주형 속성값으로 변환하여 연속형 속성값의 범위를 축소하는 과정을 의미한다. 이 과정을 이산화 과정이라 한다. 하지만 센서값은 연속적으로 생산되는 스트림 데이터이기 때문에 이산화 과정이 쉽지 않다. 동등-폭, 동등-빈도, 불순도 측정, 확률 분포 함수와 같은 스트림 데이터가 아닌 데이터에 대한 연속형 속성값을 범주형 속성값으로 변환시키는 방법이 존재하지만 스트림 데이터의 특성을 반영할 수 없어 그대로 적용이 불가능하다. 따라서 본 논문에서는 동등-폭 방법과 동등-빈도 방법을 순차적으로 적용한 방법을 이용하여 센서값에 대한 이산화를 시킨다.

표 3. 샘플 훈련 데이터
Table 3. sample training data

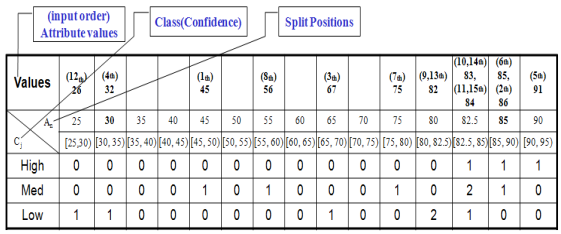
input order	temp	confidence
1	45	Medium
2	86	High
3	67	Low
4	32	Low
5	91	High
6	85	Medium
7	75	Medium
8	56	Medium
9	82	Low
10	83	Medium
11	84	Medium
12	26	Low
13	82	Low
14	83	High
15	84	Low

온도센서 값에 대한 이산화 과정을 예로 들어 설명한다. 〈표 3〉은 온도센서 값 속성에 대한 샘플 훈련데이터이다. 이외에도 습도센서 값, 연기센서 값이 있지만 설명을 위해 온도센서 값의 일부 레코드만을 이용한다.

〈표 3〉에서 보듯이 샘플 레코드들의 온도의 범위는 26부터 91까지이다. 이러한 연속형 데이터는 단순 베이지안 분류기를 통한 분류를 어렵게 한다. 따라서 다음과 같은 방법으로 이산화를 진행한다.

스텝 1: 온도범위를 임의로 지정한다.
(MIN(30) ~ MAX(85))
스텝 2: 분할개수를 임의로 지정한다. (k = 11)
스텝 3: 스텝 1과 2를 통해 간격을 구한다.
step = (MAX - MIN)/k
스텝 4: 분할지점과 분할지점에 속한 레코드의 개수를 벡터값으로 저장
스텝 5: 분할지점의 splitting을 위한 threshold 지정 (전체 값의 33%)
스텝 6: 실제 훈련데이터 셋을 이용하여 이산화 진행
스텝 7: 값이 MIN보다 작으면 MIN- step 으로 새 분할지점 을 추가
스텝 8: 값이 MAX보다 크면 MAX + step 으로 새 분할지점 추가
스텝 9: 해당 분할지점에 속한 레코드 개수가 정의한 threshold를 넘어서면 해당 분할지점을 splitting

〈표 3〉에 대하여 스텝 1~9을 진행하면 〈그림 8〉과 같은 클래스 라벨과 분할지점 매트릭스를 얻을 수 있다.



Values	(12a) 26	(4a) 32			(1a) 45		(8a) 56		(3a) 67		(7a) 75	(9,13a) 82	(10,14a) 83, (11,15a) 84	(6a) 85, (2a) 86	(5a) 91
A_2 / C_j	25	30	35	40	45	50	55	60	65	70	75	80	82.5	85	90
	[25,30)	[30, 35)	[35, 40)	[40, 45)	[45, 50)	[50, 55)	[55, 60)	[60, 65)	[65, 70)	[70, 75)	[75, 80)	[80, 82.5)	[82.5, 85)	[85, 90)	[90, 95)
High	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1
Med	0	0	0	0	1	0	1	0	0	0	1	0	2	1	0
Low	1	1	0	0	0	0	0	0	1	0	0	2	1	0	0

그림 8 분할지점 및 클래스라벨로 이루어진 매트릭스
Fig. 8. Matrix composed of break point and class label

다음 단계는 전 단계에서 생성된 매트릭스에 대한 동등-빈도를 적용하는 것이다. 연속형 속성 값의 범위를 줄이기 위한 방법이며 결과는 〈그림 9〉와 같다. 〈그림 8〉에서 연속형 속성값의 범위는 25 ~ 90사이에서 15개의 범위가 되는 반면 〈그림 9〉는 25~55, 60~80, 82.5, 85~90 이렇게 4개로 줄어든다. 이렇게 연속형 속성 값의 범위를 축소시키면 단순 베이지안 분류기 적용 시 효과적으로 계산 복잡도를 낮추게 된다. 이와 같은 방법으로 각 연속형 속성에 대한 이산화를 진행하면 단순 베이지안 분류기를 적용하여 우리가 알고자 하는 새로운 센싱 값에 따른 클래스라벨을 획득할 수 있다.

values	26	32			45		56		67		75	82	83,84	85,86	91
A_{ij} / C_j	25	30	35	40	45	50	55	60	65	70	75	80	82.5	85	90
	[25,30)	[30, 35)	[35, 40)	[40, 45)	[45, 50)	[50, 55)	[55, 60)	[60, 65)	[65, 70)	[70, 75)	[75, 80)	[80, 82.5)	[82.5, 85)	[85, 90)	[90, 95)
High	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1
Med	0	0	0	0	1	0	1	0	0	0	1	0	2	1	0
Low	1	1	0	0	0	0	0	0	1	0	0	2	1	0	0

그림 9. 최종 이산화 결과
Fig. 9. Final discretization result

IV. 실험 결과

이번 장에서는 USN 기반의 화재 감시 응용을 위한 제안 시스템의 원시 데이터의 증가에 따른 다차원 분석질의 응답 속도를 이용하여 제안 시스템의 성능을 평가한다. 또한 다차원 분석 질의 결과 및 이상치 탐지 결과 화면을 보인다.

제안 시스템의 성능 평가를 위한 실험 환경은 다음 〈표 4〉와 같다.

표 4. 실험 환경
Table 4. Experimental environment

OS	Windows XP SP3
CPU	Core 2 Duo E6750
RAM	4G
개발언어	java jdk 1.6.0_23
DB	MS-SQL server 2008
개발환경	Eclipse Helios R1
센서 데이터	D3L3C5T2000k (D: Dimension, L: Level, C: Cardinality, T: Total)

실험을 위해 인기 경로에 속한 질의를 다음과 같이 정의한다.

질의 1 : OO 거리에 있는 모든 시설물의 배터리 센서들의 평균값은?

질의 2 : OO 거리 OO 상가의 배터리 센서 평균값은?

질의 3 : OO 거리 OO 상가의 온도 센서값의 최대값은?

질의 4 : OO 거리 OO상가 중 점포 별 온도 센서값의 최대값은?

〈그림 10〉은 원시 데이터의 크기에 따른 질의 1~4의 평균응답시간을 나타낸다. 실험을 통해 인기경로를 이용한 부분 데이터 큐브(popular-path)의 응답시간과 전체 데이터 큐브(full-cubing)의 응답 시간을 비교 한다.

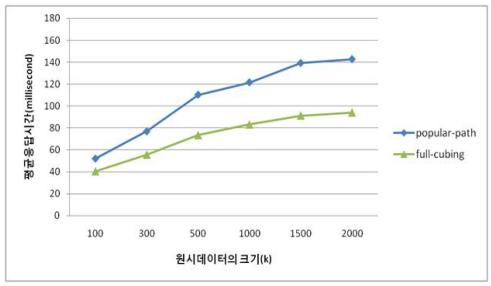


그림 10 원시 데이터 크기 vs. 평균응답시간
Fig. 10. Size vs. Average response time

〈그림 10〉에서 제안 시스템의 인기경로를 이용한 부분 데이터 큐브 기법의 평균응답시간이 약 20~30% 줄어드는 것을 확인할 수 있다.

〈그림 11〉은 인기경로에 속한 질의 1에 대한 예시를 나타낸다.

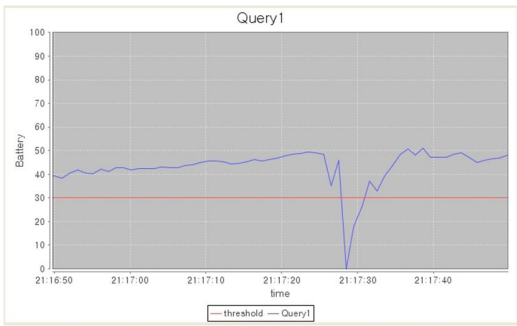


그림 11 질의 1에 대한 결과 화면
Fig. 11. a result of query 1

〈그림 12〉는 USN 기반의 화재 감시 응용에서의 이상치 탐지 결과 화면이다.

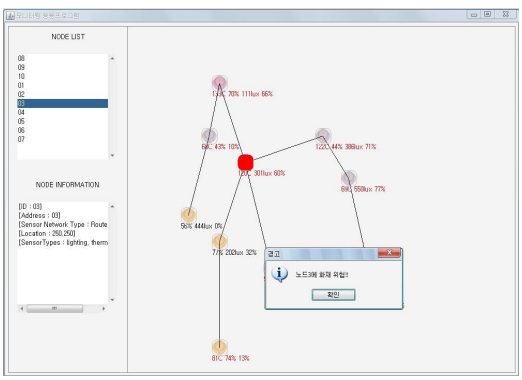


그림 12. 화재 위험 탐지 예시
Fig. 12. a example of fire detection

V. 결론

본 논문에서는 USN 기반의 화재 감시 응용을 위한 센서 데이터 처리 시스템을 개발하였다. 제안한 센서 데이터 처리 시스템은 다차원 분석 질의 처리 기능과 지능적 이벤트 관리 기능을 효과적으로 제공한다. 연속적이면서 대용량의 스트림 데이터에 대한 부분 데이터 큐브만을 생성하여 다차원 분석 질의 시 응답속도 문제를 효과적으로 해결하였다. 또한 단순 베이지안 분류기를 통해 실시간으로 수집되는 개별 센서 데이터에 대한 의사 결정을 지원한다. 이는 센서 데이터에 대한 특징을 고려한 데이터 크기 축소, 데이터 정제 및 변환을 통해 가능하다.

본 논문에서는 센서 데이터 처리 시스템의 실효성을 검증하기 위해, 화재 감시 응용의 스트림 데이터를 이용하여 실험 결과를 보였다. 향후에는 향상된 부분 데이터 큐브 생성 방법 및 이상치 탐지 방법을 제안할 예정이며 제안 시스템과의 세밀한 비교 평가를 진행할 것이다.

참고문헌

〔1〕 Jiawei, H., et al., "Stream Cube: An Architecture for Multi-Dimensional Analysis of Data Streams," Distrib. Parallel Databases, 18(2), pp. 173-197, 2005.

〔2〕 P. Domingos and M. Pazzani, "Beyond Independence: Conditions for the Optimality of The Simple Bayesian Classifier," In Proc. Int. Conf. Machine Learning, pp. 105-112, 1996.

〔3〕 A. Arasu, et al., "Stream: The Stanford Data Stream Management System," in IEEE Data Engineering Bulletin, 4(1), 2003.

〔4〕 C. Don, U, et al.,"Monitoring Streams: A New Class of Data Management Applications," In Proc. Int. Conf. Very Large DataBase, pp.215-226, 2002.

〔5〕 C. Sirish, et al., "TelegraphCQ: Continuous Dataflow Processing," In Proc. Int. Conf. Innovative Data Systems Research, pp.11-18, 2003.

〔6〕 Cougar Project, http://www.cs.cornell.edu/boom/2003sp/ProjectArch/CougarSM/index.php

〔7〕 C. Jianjun, et al.,"NiagaraCQ: A Scalable Continuous Query System for Internet Databases," SIGMOD Rec., pp. 379-390, 2000.

〔8〕 Calton Pu, Ling Liu, "Update Monitoring: The CQ Project," In Proc. Int. Conf. Worldwide Computing and Its Applications, Tuskuba, Japan, Lecture Notes in Computer Science, pp.396-411, 1998.

〔9〕 R. Philipp, sch, and L. Wolfgang, "Sample synopses for approximate answering of group-by queries," In Proc. Int. Conf. Extending Database Technology, ACM, pp.403-414, 2009.

〔10〕 K. Henning, hler, Z. Xiaofang, S. Shazia, S. Yanfeng, and T. Kerry, "Sampling dirty data for matching attributes," In Proc. Int. Conf. Management of data, Indianapolis, Indiana, USA, ACM, pp.63-74, 2010.

〔11〕 Alfredo Cuzzocrea , Sharma Chakravarthy, "Event-based Lossy Compression for Effective and Efficient OLAP over Data Streams," Data & Knowledge Engineering, v.69 n.7, p.678-708, July, 2010.

〔12〕 Yufei Tao , Dimitris Papadias, "Maintaining Sliding Window Skylines on Data Streams," IEEE Trans. on Knowledge and Data Engineering, 18(3), pp.377-391, March, 2006.

〔13〕 Abhirup Chakraborty , Ajit Singh, "A Disk-based, Adaptive Approach to Memory-limited Computation of Windowed Stream Joins," In Proc. Int. Database and expert systems applications: Part I, Aug. 30-Sep. 03, Bilbao, Spain, 2010.

〔14〕 Hua-Fu Li , Suh-Yin Lee, "Mining Frequent Itemsets over Data Streams using Efficient Window Sliding Techniques," Expert Systems with Applications: An Int. Journal, 36(2), pp.1466-1477, March, 2009.

〔15〕 W. Hai, and C.S. Kenneth, "Histograms b ased on the Minimum Description Length Principle," The VLDB Journal, pp. 419-442, 2008.

〔16〕 L. Xin, and G. Jihong, "A New Approach to Building Histogram for Selectivity Estimation in Query Processing Optimization," Comput. Math. Appl., pp. 1037-1047, 2009.

〔17〕 H. Ming-Jyh, C. Ming-Syan, and S.Y. Philip, "Integrating DCT and DWT for Approximating Cube Streams," In Proc. Int. Conf. Information and knowledge management, pp.179-189, 2005.

〔18〕 Xiao-Bo Fan , Ting-Ting Xie , Cui-Ping Li , Hong Chen, "MRST: An Efficient Monitoring Technology of Summarization on Stream Data," Journal of Computer Science and Technology, 22(2), pp.190-196, March, 2007.

〔19〕 Hanady Abdulsalam , David B. Skillicorn , Patrick Martin, "Classifying Evolving Data Streams Using Dynamic Streaming Random Forests," In Proc. Int. Conf. Database and Expert Systems Applications, September 01-05, Turin, Italy, 2008.

〔20〕 Wei Qu , Yang Zhang , Junping Zhu , Qiang Qiu, "Mining Multi-label Concept-Drifting Data Streams Using Dynamic Classifier Ensemble," In Proc. Int. Conf. Machine Learning: Advances in Machine Learning, November 02-04, Nanjing, China,2009.

〔21〕 Qu Wei , Zhang Yang , Zhu Junping , Wang Yong, "Mining Multi-label Concept-drifting Data Streams Using Ensemble Classifiers," In Proc. Int. Conf. Fuzzy systems and knowledge discovery, Tianjin, China, August 14-16, 2009.

〔22〕 Mohammad M. Masud , et al., "Classification and Novel Class Detection of Data Streams in a Dynamic Feature Space," In Proc. Int. Conf. Machine learning and knowledge discovery in databases: Part II, Barcelona, Spain, September 20-24, 2010.

〔23〕 Panagiotis Antonellis , Christos Makris , Nikos Tsirakis, "Algorithms for Clustering Clickstream Data," Information Processing Letters, 109(8), pp.381-385, March, 2009.

〔24〕 Li Wan , et al., "Density-based Clustering of Data Streams at Multiple Resolutions," ACM Trans. on Knowledge Discovery from Data), 3(3), pp.1-28, July 2009.

〔25〕 Li Tu , Yixin Chen, "Stream Data Clustering based on Grid Density and Attraction," ACM Trans. on Knowledge Discovery from Data, 3(3), pp.1-27, July 2009.

〔26〕 Maria Kontaki , Apostolos N. Papadopoulos , Yannis Manolopoulos, "Continuous Trend-Based Clustering in Data Streams," In Proc. Int. Conf. Data Warehousing and Knowledge Discovery, Turin, Italy, September 02-05, 2008.

〔27〕 Bai-En Shie , Vincent S. Tseng , Philip S. Yu, "Online Mining of Temporal Maximal Utility Itemsets from Data Streams," In Proc. Int. Conf. Applied Computing, Sierre, Switzerland, March 22-26, 2010.

〔28〕 T. Syed Khairuzzaman, A. Chowdhury Farhan, J. Byeong-Soo, and L. Young-Koo, "Efficient Frequent Pattern Mining over Data Strea ms," In Proc. Int. Conf. Information and knowledge management, pp.1447-1448, California, USA, October 26-30, 2008.

〔29〕 Yoshiaki Yasumura, Naho Kitani, Kuniaki Uehara, "Quick Adaptation to Changing Concepts by Sensitive Detection," In Proc. Int. Conf. Industrial, engineering, and other applications of applied intelligent systems, Kyoto, Japan, June 26-29, 2007.

〔30〕 Alfredo Cuzzocrea , Sharma Chakravarthy, "Event-based Lossy Compression for Effective and Efficient OLAP over Data Streams," Data & Knowledge Engineering, 69(7), pp.678-708, July, 2010.

〔31〕 Alfredo Cuzzocrea, "CAMS: OLAPing Multidimensional Data Streams Efficiently," In Proc. Int. Conf. Data Warehousing and Knowledge Discovery, Linz, Austria, August 31-September 02, 2009.

〔32〕 S ébastien Nedjar , Alain Casali , Rosine Cicchetti , Lotfi Lakhal, "Emerging Cubes: Borders, Size Estimations and Lossless Reductions," Information Systems, 34(6), pp.536-550, September, 2009.

저 자 소 개

박 원 익

2004: 충남대학교 컴퓨터과학과 이학사.

2006: 충남대학교 컴퓨터공학과 공학석사.

2007~현재: 충남대학교 컴퓨터공학과 박사 과정.

관심분야: 지능형 추천 알고리즘, 모바일 정보시스템, 데이터 마이닝, OLAP

Email : wonik78@cnu.ac.kr

김 영 국

1985: 서울대학교 계산통계학과 이학사.

1987: 서울대학교 계산통계학과 이학석사.

1995: 버지니아대학교 컴퓨터과학과 공학박사.

1996~현 재: 충남대학교 컴퓨터공학과 교수.

관심분야: 실시간데이터베이스, 모바일정보시스템, 전자상거래시스템

Email : ykim@cnu.ac.kr