

# 센서 네트워크의 데이터 스트림 마이닝을 위한 온톨로지 기반의 전처리 기법

정재은

영남대학교 컴퓨터공학과  
(j2jung@gmail.com, j2jung@intelligent.pe.kr)

.....

다양한 센서의 개발과 센서 네트워크 구축으로 인해 특정 공간의 환경 데이터를 수집할 수 있다. 보다 유용한 정보 및 지식의 발견을 위하여 데이터 마이닝(Data mining) 기법이 활용되는 연구들이 소개되었다. 본 연구에서는 이와 같은 데이터 마이닝 기법의 효율성 증대를 위하여 센서 네트워크로부터의 데이터 스트림의 전처리 과정(Preprocessing)을 수행하고자 한다. 제안하는 센서 스트림 데이터의 전처리 과정은 i) 세션 확인(Session identification)과 ii) 오류 검증(Error detection) 문제를 해결하고자 한다. 특히, 이를 위해 각 센서 장비로부터 수집되는 데이터의 의미(Semantics)를 표현하고 있는 온톨로지(Ontology)를 적용한다. 본 연구 결과의 성능 평가를 위하여 센서 네트워크 테스트 환경을 교내에 설치하였으며 30여일 동안 수집된 데이터를 이용하여 시뮬레이션을 실행하였다.

.....

논문접수일 : 2009년 08월 20일    논문수정일 : 2009년 09월 05일    게재확정일 : 2009년 09월 14일    교신저자 : 정재은

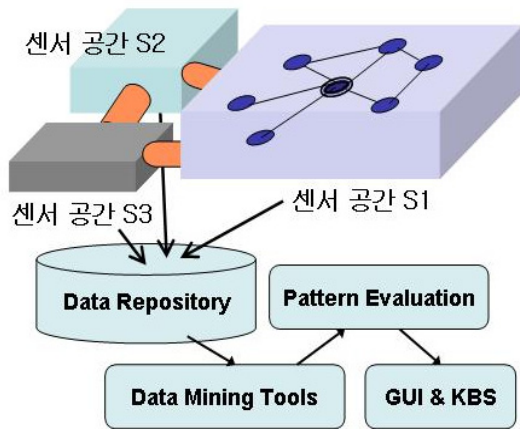
## 1. 서론

유비쿼터스(Ubiquitous) 컴퓨팅은 “언제 어디서나” 사용자들의 상황(Context)를 정확하게 인식함으로써 그들에게 시기적절한 정보 및 서비스를 제공할 수 있는 지능형 기술 및 시스템을 일컫는다. 이 개념은 80년대 후반에 Mark Weiser에 의해 소개된 이후, 다양한 기술 개발이 진행되어오고 있으며 우리 사회와 삶에 많은 변화를 있다(Weiser et al., 1999). 이와 같은 유비쿼터스 컴퓨팅 기술들 중에 하나로서 센서 네트워크(Sensor network)가 많은 관심을 받아왔으며 다양한 종류의 센서 장비들이 개발되었다(Culler et al., 2004). 이것은 주어진 특

정 공간으로부터 다양한 환경 정보를 수집할 수 있음을 의미하여, 다시 말해, 해당 센서 공간 내의 사용자들에게 더욱 효과적인 정보의 제공 및 추천이 가능하게 되었음을 제시한다(Jung, 2009). 특히, 이를 통해 특정 센서 환경이 사용자들에 대한 충분한 데이터를 수집하고 있지 않은 초기 단계에서도 효과적인 서비스 제공을 가능케 한다.

본 연구에서는 <그림 1>과 같이 각 센서 공간들 간의 의미적 이형질성(Semantic heterogeneity) 문제를 해결하기 위하여 온톨로지(Ontology)를 활용한다. 온톨로지 기반 센서 네트워크는 일반적으로 센서 노드들 간의 무선 통신이 가능한 특수한 형태의 센서 네트워크에 온톨로지를 적용하여 센서 네

\* 이 논문은 2009년 산학협동재단 지원으로 수행된 연구임.



<그림 1> 이형질적 센서 공간의 데이터 수집 및 데이터마이닝 프로세스

트위크들 간의 의미적 이질성을 해결한 개방형 네트워크 환경을 말하며, 주로 환경 모니터링 및 제어 (Environmental monitoring and controlling)를 위해 사용되어왔다. 특히, 온톨로지 기반 센서 네트워크는 교통 상황 정보, 지리 정보, RFID 정보 등을 수집 및 공유할 수 있으며, 기존의 인터넷이나 웹 (WWW) 기반 시스템들과의 연동이 가능하다는 점에서 더욱 각광 받고 있다.

최근까지 이와 같은 온톨로지 기반 센서 네트워크 연구들은 다음과 같이 크게 두 가지로 정리된다.

다양한 센서 네트워크들로부터 수집되는 데이터의 융합 및 통합(Estrin et al., 2002; Lim et al., 2005; Jabeur et al., 2009)

삶의 질 향상을 위한 센서 데이터의 활용한 지능형 서비스 제공(Eagle and Pentland, 2006; Sheth, 2009)

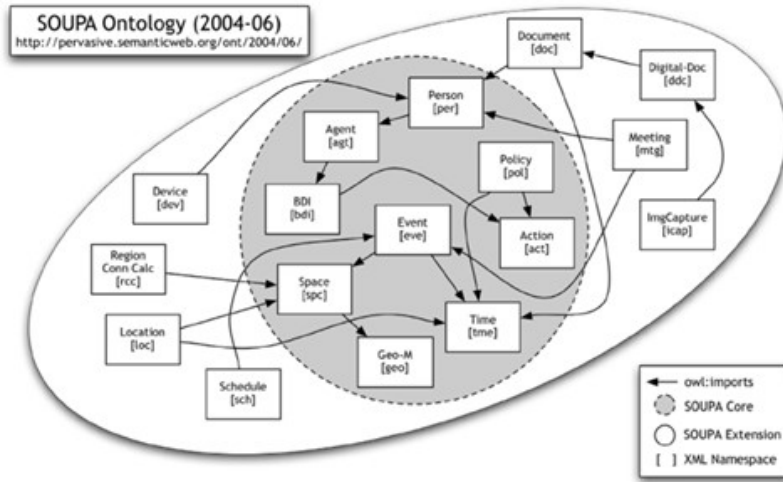
본 연구의 목적은 이와 같은 이질적인(Heterogeneous) 센서 네트워크 환경들 간의 정보 공유를 통하여 정보 융합 컴퓨팅(Information fusion computing) 플랫폼을 설계하고 개발하고자 한다. 센서 네트워크의 각 센서 노드(Node)들을 통해 수집되는 센서 신호(Sensor signal)는 온톨로지를 통해 정

확한 의미를 추론할 수 있으며, 사용자들에게 보다 효과적인 상황 인식 기반의 서비스(Context-aware service)를 제공할 수 있게 된다. 제안하는 플랫폼에서의 지능형 서비스 제공은 크게 두 단계로 이루어져 있다.

- 1) 첫 번째 단계는 **센서 공간들간의 정보 공유**이다. 센서 노드들을 통해 수집된 데이터는 해당 센서 공간의 게이트웨이(Gateway)를 통해서로 공유된다. <그림 1>과 같이 센서 공간  $S_1$ 은 특정 사용자  $U_k$ 에게 서비스를 제공하기 위해 제안하는 정보 융합 플랫폼을 통해  $S_2$ 와  $S_3$ 들이 저장하고 있는 해당 사용자에 대한 추가 정보를 전송 받을 수 있다.
- 2) 두 번째 단계는 플랫폼 내에서의 지능형 서비스 제공을 위한 적절한 데이터마이닝(Data mining) 알고리즘을 적용을 통한 유용한 패턴의 발견이다. 예를 들어, 지질 정보, 교통량(Traffic) 정보, 날씨 정보의 결합을 통한 세 가지 센서 웹 환경의 정보 융합이 될 수 있을 것이며, 충분한 데이터의 수집이 있다면 교통량(Traffic) 정보와 날씨 정보 간의 효과적인 관계를 추론해 내기 위한 통계학적 데이터 마이닝 방법론을 적용할 수 있다.

하지만, 이와 같이 이질적인 센서 공간들로부터 동시에 수집되는 스트리밍 데이터에 일반적인 데이터 마이닝 기술을 직접 적용할 수 없다. 즉, 이와 같은 온톨로지 기반 센서 네트워크 환경으로부터 수집되는 데이터 스트림의 분석을 위해서는 전처리과정(Pre-process)이 매우 중요한 프로세스로 인식되고 있기 때문이다. 이를 위해서 본 연구에서는 다음과 같이 두 가지 문제에 초점을 맞추고자 한다.

- 1) 온톨로지 기반의 시맨틱 어노테이션(Annota-



<그림 2> SOUPA 온톨로지(Chen et al., 2004)

tion)을 통해 일반적인 시계열 센서 데이터 스트림(Time-series sensor stream)의 구체적인 의미(Semantics)을 부여함으로써 센서 공간들의 비교 분석 및 유사도(Similarity) 측정이 가능하도록 한다.

- 2) 수집되는 센서 스트림이 분할(Segmentation)되어야 하며 분할된 센서 시퀀스(Sensor sequence)는 연관 관계(Association)가 있는지 판단하게 된다.

본 연구는 구성은 다음과 같다. 제 2장에서는 온톨로지 기반 센서 네트워크에 대한 배경 지식과 관련 연구들을 살펴본다. 제 3장에서는 센서 스트림 데이터의 전처리 과정을 위한 온톨로지를 이용한 센서 스트림의 분할 기법을 소개한다. 그리고 제 4장에서는 제안하는 분할 기법을 구현하고 평가하기 위한 실험 환경 및 실험 데이터를 설명하고, 기존의 방법론들과 비교하고자 한다. 마지막으로 제 5장에서는 본 연구의 결론 및 향후 연구 방향에 대하여 언급하고자 한다.

## 2. 배경 지식 및 관련 연구

본 장에서는 온톨로지 기반 센서 공간에서 수집되는 센서 스트림의 전처리 과정을 설명하기 위한 주요 배경 지식으로 i) 온톨로지 기반 센서 네트워크와 ii) 스트림 데이터 마이닝 기술들을 설명한다. 이와 더불어, 본 연구와 유사한 기존 연구들을 소개하고자 한다.

### 2.1 온톨로지와 온톨로지 기반 센서 네트워크

기본적으로 온톨로지는 컴퓨터가 처리 가능하도록 구체적으로 표현되어 지식 재사용이 가능한 메타 모델이라 정의 할 수 있다. 또한 온톨로지는 다음과 같이 표현된다.

$$O = \langle C, R, \langle_{c \times c}, \langle_{R \times R}, I \rangle \quad (1)$$

여기서  $C, R, I$ 는 각각 개념(Concept), 관계(Relation)와 인스턴스(Instance)의 집합들이며,  $\langle_{c \times c}$ ,



```

<?xml version="1.0"?>
<rdf:RDF
xmlns="http://intelligent.pe.kr/SemSensorWeb/2007/03/ContextOnto.owl#"
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:xsd="http://www.w3.org/2001/XMLSchema#"
xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
xmlns:owl="http://www.w3.org/2002/07/owl#"
xmlns:units="http://intelligent.pe.kr/SemSensorWeb/2007/03/Units.owl#"
xml:base="http://intelligent.pe.kr/SemSensorWeb/2007/03/ContextOnto.owl">
<Temperature rdf:ID="Temp8">
  <Cxt_Property_7 rdf:datatype="http://www.w3.org/2001/XMLSchema#float">
    27.0
  </Cxt_Property_7>
</Temperature>
<Time rdf:ID="t1">
  <hasTemperature rdf:resource="#Temp8"/>
</Time>
</rdf:RDF>

```

- 1) Open Geospatial Consortium (OGC) : 이와 같은 다양한 센서들로부터 수집된 데이터의 융합을 위해서 몇 가지 연구 프로젝트들이 수행되고 있다. 최근 지리정보시스템 분야의 Open Geospatial Consortium(OGC)에서는 다양한 센서들로부터 수집된 지리정보를 통합하여 일반 웹 브라우저를 통해 검색할 수 있는 시스템의 개발을 진행 중에 있다. 이를 위해 OGC에서는 지리정보의 메타데이터 표준화에 힘쓰고 있다.
- 2) SensorWare Systems : NASA 센서 웹 프로젝트의 일부분으로서 농업분야의 실시간 센서 데이터 처리에 많은 연구를 진행하고 있다.
- 3) SenseWeb : 마이크로소프트에서 수행중인 프로젝트로서 센서로부터의 데이터를 Sensor-Map이라는 지도 서비스에 메쉬업 형태로 적용되어 사용되고 있다.
- 4) GeoSensor Web : 캐나다 Calgary 대학에서 수행중인 GeoSensor Web 프로젝트는 GIS 시스템에 센서 웹 기술을 접목한 것이다. 이 시스

- 템 역시 농업 및 환경 분야에 활용되고 있다.
- 5) 52North : 공간 데이터(Spatial data)의 통합을 위한 상호운용 가능한 웹 서비스 개발을 위한 컨소시움이다.
- 6) Semantic Sensor Web : 미국의 Wright State University에서 수행하는 프로젝트로서 센서로부터 수집된 데이터에 Semantic 메타데이터를 어노테이션함으로써 공유 및 재사용이 가능하도록 하기 위한 프로젝트이다.

## 2.2 스트리밍 데이터 마이닝

스트리밍 데이터 마이닝(Streaming data mining)은 시간의 흐름에 따라 순서대로 정렬되어 수집되는 데이터로부터 동향 분석(Trend analysis)이나 주기성 분석(Periodical pattern analysis)과 같이 다양한 형태의 지식 및 패턴을 발견하기 위한 기법이다. 다시 말해, 분석하고자 하는 데이터의 특성 및 목적에 따라 적용하고자 하는 마이닝 알고리

&lt;표 1&gt; 다양한 센서를 이용한 센서 스트림 데이터 수집

Time stamp	RFID Reader	Temperature (°C)	Lights (On/Off)	Air Conditioner (Switch)	Project Screen (Pull-down)
T <sub>0</sub>	U <sub>1</sub>	27	Off	Off	False
T <sub>1</sub>	U <sub>1</sub>	28	On	On	False
T <sub>2</sub>	U <sub>1</sub> , U <sub>2</sub>	30	Off	On	True
T <sub>3</sub>	U <sub>1</sub> , U <sub>2</sub> , U <sub>3</sub>	28	Off	Off	True
T <sub>4</sub>	U <sub>1</sub> , U <sub>2</sub> , U <sub>3</sub>	27	On	On	False
T <sub>5</sub>	U <sub>2</sub> , U <sub>3</sub>	25	On	On	False
T <sub>6</sub>	U <sub>2</sub>	25	On	On	False
T <sub>7</sub>	U <sub>2</sub>	26	On	Off	False
T <sub>8</sub>	U <sub>1</sub> , U <sub>2</sub>	26	Off	Off	True

증도 달라지게 된다.

본 연구에서는 센서 스트림 데이터로부터 빈발 순차 패턴(Frequent sequential pattern)의 발견(Agrawal and Srikant, 1995)에 초점을 맞추고자 한다. 빈발 순차 패턴은 특정 데이터들의 순서를 고려하면서 자주 발생하는 패턴을 일컫는다. 지금까지는 주로 사용자의 웹 로그(Web log) 데이터 분석을 통한 개인화 서비스나 정보 검색 기능 향상에 초점을 맞추었다. 이와 같은 문제의 경우, 정확성을 향상시키기 위한 가장 중요한 과정은 스트림 데이터의 잡음 데이터 제거 및 세션화(Sessionization 또는 Session identification)와 같은 전처리 단계라 할 수 있다. 예를 들어, <표 1>에서와 같이 특정 주기로 샘플링(Sampling)하는 5가지 종류의 센서로부터 수집되는 데이터 스트림을 보인다.

### 3. 온톨로지 기반의 센서 스트림 데이터 전처리

센서 노드들로부터 수집되는 데이터는 온톨로지를 이용하여 해당 데이터가 내포하는 의미를 기술(어노테이션)한다. 이것은 센서 스트림 데이터에 명시적인 메타데이터(Explicit metadata)를 추가하는 것과 같다. 더욱 중요한 점은 센서 스트림 데이

터가 온톨로지를 통해 어노테이션됨으로써 RDF 메타데이터 스트림(Metadata stream)으로 변환된다는 것이다. 특히, 메타데이터로의 변환은 다수의 센서 공간으로부터 수집되는 센서 정보를 융합을 용이하게 한다. 이와 같은 과정을 실행하기 위하여 메타데이터 스트림은 특정 시간 간격(Time slot) 내에서의 의미 유사도(Semantic similarity, SS)를 계산하며, 시간이 지남에 따라, 의미 유사도의 분포를 조사하게 된다.

#### 3.1 시맨틱 어노테이션(Semantic annotation)

센서 공간에 설치되어 있는 센서 노드들은 공간 내에서 감지되는 모든 센서 스트림의 의미를 온톨로지 내의 개념들의 집합을 이용해 기술한다. 각 샘플링 이벤트가 발생했을 때마다 온톨로지로부터 명시적인 메타데이터를 추출함으로써 인스턴스화(Instantiation) 과정이 이루어진다. 특히, RFID 리더의 경우는, 사용자 및 객체들의 고유한 ID값을 어노테이션에 사용하게 된다. 예를 들어, <표 1>에서와 같이,  $t_1$  이벤트에서는 어노테이션을 위해 다음과 같은 RDF 형식의 메타데이터가 생성 된다.

결국에 시간이 지남에 따라 충분한 메타데이터 스트림을 수집할 수 있으며, 다른 센서 공간의 정보와의 융합을 위해 게이트웨이 센서 노드를 통해

중앙 서버에 집중된다.

### 3.2 센서 스트림 데이터 처리

일반적으로 센서 네트워크는 다양한 센서들로부터 감지된 정보를 융합하고 활용할 수 있는 환경을 말한다. 특히, 각각의 센서 네트워크로부터 수집되는 센서 데이터 스트림들은 미들웨어(Middleware)를 통해 중앙 저장소(Repository)에 저장(Aggregation)되며, 다른 센서 네트워크로부터의 데이터 스트림과의 의미 있는 관계를 발견함으로써 센서 데이터들 간의 인과 관계를 유추할 수 있도록 한다. 이와 함께, 유선 온라인상의 외부 데이터 소스(External data source)들과의 연동이 가능하여 센서 스트림 데이터를 보완할 수도 있다. 센서 스트림 데이터의 융합은 메타데이터 어노테이션의 RDF 통합(Merging)을 통해 이루어진다. 본 연구에서는 모든 센서 공간에서 동일한 온톨로지를 사용하므로(다시 말해, 동일한 네임스페이스(Name-space)를 가지므로), 별도의 온톨로지 매핑 과정이 필요하지 않는다.

이와 같은 융합된 센서 스트림 데이터의 연관 관계 발견을 위해서는 스트림 데이터 저장소(Stream repository)에서는 센서 스트림 데이터의 전처리(Pre-processing) 과정이 필요하다.

- 동기화(Synchronization) 처리 : 각 센서 공간의 센서 노드마다 상황 정보 수집을 위한 샘플링 주기가 다르므로 스트림 데이터의 Time-stamp를 참조하여 동기화 한다.
- 결측치(Missing data) 처리 : 발생한 결측치는 전후 데이터의 보간법(Interpolation)을 통해 처리한다.
- 세션화(Session identification) : 스트림 데이

터로부터 유용한 순차(Sequential) 패턴의 발견을 위해서는 센서 공간 상의 특정 상황(Contextual situation)별로 스트림 데이터가 구분되어야 한다.

### 3.3 의미 유사도 분포를 이용한 세션화

다양한 센서 스트림의 융합을 통해 데이터는 충분해졌으나, 기존의 데이터 마이닝 알고리즘을 적용하여 빈발 순차 패턴의 발견을 위해서는 스트림 데이터의 세션화가 필수적이다. 일반적인 세션화 기법(휴리스틱)은 다음과 같은 두 가지 방법이 주로 사용되고 있다.

- *Time-oriented* 휴리스틱 : 미리 지정된 단위 시간( $T_s$ )을 기준으로 세션화하는 방법
- *Frequency-oriented* 휴리스틱 : 수집되어지는 스트림 데이터의 변화량 추세에 따라 단위 시간의 값이 유동적으로 변하는 방법

하지만, 이와 같은 단순한 기법들은 센서 공간 내의 상황 정보를 적절하게 고려할 수 없다는 단점이 있다.

이와 같은 문제를 해결하기 위해서, 본 연구에서는 온톨로지를 이용한 의미 유사도의 분포(SS distribution) 정보를 적용하는 세션화 기법을 제안한다. 이 방법을 위해서는 슬라이딩 윈도우(Sliding window,  $W$ )를 사용하는데, 윈도우의 크기는  $|W| = w$ 로 표기한다. 두 시점( $t_i, t_j$ )에서의 RDF 센서 데이터를 각각  $r_i, r_j$ 라고 하며, 두 데이터 간의 의미 유사도  $\Delta$ 는 다음과 같이 구할 수 있다.

$$\Delta_{ij} = 1 - \delta(r_i, r_j) \quad (2)$$

여기서,  $\delta$ 는 두 RDF 센서 데이터 간의 Vector

Distance<sup>3)</sup>이다. 슬라이딩 윈도우 내의 RDF 센서 데이터들의 모든 가능한 조합간의 의미 유사도(SS) 들 계산하면 식 (2)는 다음과 같이 의미 유사도 행렬 (SS matrix, SSM)  $\Delta_W$ 로 확장된다.

$$\Delta = \begin{pmatrix} 0 & \Delta_{21} & \cdots & \Delta_{w1} \\ \Delta_{12} & 0 & \cdots & \Delta_{w2} \\ \cdots & \cdots & \cdots & \cdots \\ \Delta_{1w} & \Delta_{2w} & \cdots & 0 \end{pmatrix} \quad (3)$$

이 행렬의 특징은 대각요소들이 0인 대칭 행렬 (즉,  $\Delta_{ij} = \Delta_{ji}$ )이라는 것이다. 일단 의미유사도 행렬이 구해지면, 다음과 같은 통계적인 데이터를 추가적으로 구할 수 있다.

- 의미 유사도 평균(SS mean)는 슬라이딩 윈도우 내의 RDF 센서 스트림의 의미적 유사도의 산술적 평균값을 나타낸다.

$$\mu_w = \frac{2 \sum_{j=1}^{w-1} \sum_{i=j+1}^w TRIANGLE_{ij}}{w(w-1)} \quad (4)$$

- 의미 유사도 표준 편차(SS Standard Deviation)는 슬라이딩 윈도우 내의 RDF 센서 스트림의 의미적 유사도의 산술적 표준 편차값을 나타낸다.

$$\sigma_w = \sqrt{\frac{2 \sum_{j=1}^{w-1} \sum_{i=j+1}^w (\Delta - \mu_w)}{w(w-1)}} \quad (5)$$

이와 같은 값들은 시간이 지남에 따라, 슬라이딩 윈도우가 이동(Shifting) 되고, 의미 유사도 평균과 의미 유사도 표준 편차를 이용한 분포(Distribution)를 관찰할 수 있으며 세션화를 위한 휴리스틱의 설정

에 활용된다. 본 논문에서는 우선 다음과 같은 네 가지 휴리스틱을 설정하고 수집한 센서 데이터를 이용하여 세션화 실험을 실시하였다(자세한 실험 결과는 제 4장에서 소개된다).

H 1-1 : 의미 유사도 평균이 기준값( $\tau_\mu$ ) 이하 일 때, RDF 센서 스트림은 새로운 세션을 시작함을 의미한다. 본 연구에서는 실험적으로  $\tau_\mu$ 을 0.4로 설정한다.

H 1-2 : 의미 유사도 표준 편차가 기준값( $\tau_\sigma$ ) 이상 일 때, RDF 센서 스트림은 새로운 세션을 시작함을 의미한다. 본 연구에서는 실험적으로  $\tau_\sigma$ 을 0.4로 설정한다.

H 2-1 : 의미 유사도 평균이 상승에서 하강으로 변할 때, RDF 센서 스트림은 새로운 세션을 시작함을 의미한다.

H 2-2 : 의미 유사도 표준 편차가 하강에서 상승으로 변할 때, RDF 센서 스트림은 새로운 세션을 시작함을 의미한다.

예를 들어 설명하면, <표 1>에서의 센서 스트림 데이터는 슬라이딩 윈도우의 크기가 3으로 설정되었다면, <표 2>와 같은 의미 유사도 행렬을 구할 수 있다. 이 정보를 기반으로, 의미 유사도 평균과 표준편차의 시간에 따른 분포를 구해보면 <그림 4>와 같다.

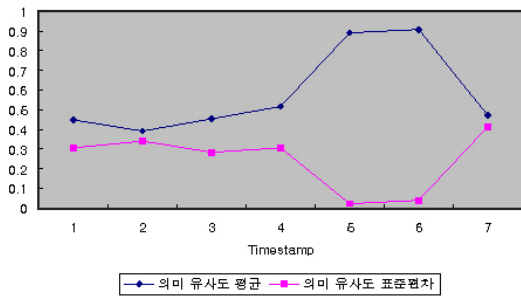
설정된 세션화 휴리스틱에 의해 세션화가 이루어져야 할 시점을 찾아보면 다음과 같이 각각의 휴리스틱에 의해 다음과 같이 3개의 세션화가 일어날 시점을 발견할 수 있다.

- H1-1에 의해 (t1, t2)사이
- H1-2에 의해 (t6, t7)사이
- H2-1에 의해 (t1, t2)와 (t6, t7)사이

3) 자세한 내용은 Jung (2007) 참조.

<표 2> 의미 유사도 행렬

	t0	t1	t2	t3	t4	t5	T6	t7	t8
t0	0	0.8	0.3						
t1	0.8	0	0.25	0.15					
t2	0.3	0.25	0	0.78	0.26				
t3		0.15	0.78	0	0.32	0.36			
t4			0.26	0.32	0	0.87	0.92		
t5				0.36	0.87	0	0.89	0.88	
t6					0.92	0.89	0	0.95	0.24
t7						0.88	0.95	0	0.23
t8							0.24	0.23	0



<그림 4> 의미 유사도 평균과 표준편차의 시간적 분포도

- H2-2에 의해 (t1, t2), (t3, t4)와 (t6, t7)사이

이와 같이 구성된 세션들을 이용하여 연관 관계를 구해보면 다음과 같은 패턴을 발견할 수 있다. 즉, 이 패턴은 다음과 같이 해석될 수 있다. 센서 공간을 회의실과 같은 여러 사람들이 토론을 하기 위한 공간이며 Project Screen이 내려오면 Light가 꺼지는 상황이 일어난다.

- Project Screen Pull-Down(True) → Light(Off)

#### 4. 실험 및 토론

실험을 위해 4개의 센서 공간(3개의 강의실과 복도에) 5가지 종류의 센서들을 설치하여 30여일(2009년 3월 9일부터 2009년 4월 6일까지)동안 센서 데이터를 수집하였다. 본 논문에서 제안하는 데이터 전처리 기법(M<sub>Onto</sub>)의 성능 비교를 위해 다음과 같은 방법들과 함께 발견되는 패턴의 타당성을 사람이 직접 평가를 실시하였다. 데이터마이닝 기법은 Agrawal and Srikant (1995)의 Sequential pattern mining 기법을 적용하였다.

- M<sub>Generic</sub> : 세션화 과정이 없다. 즉, 단순히 각각의 스트림 데이터들이 항상 유일한 상황의 의미로만 이루어져있다고 가정하는 방법이다.
- M<sub>Time</sub> : Time-oriented 세션화 방법이다. 시간이 지남에 따라 주어진 Time Slot만큼 자동으로 세션화한다.
- M<sub>Sampling</sub> : 센서 스트림 데이터의 임의 추출법(Random sampling)하는 방법이다. 대규모의 센서 스트림 데이터 처리에 대한 효율성을 비

&lt;표 3&gt; 제안하는 전처리 기법의 비교 평가(MOnto를 위한 윈도우의 크기는 5분으로 설정)

		M <sub>Generic</sub>	M <sub>Time</sub>	M <sub>Sampling</sub>	M <sub>Onto</sub>
패턴 개수		4	5	4	7
패턴 정확성	타당	1(25%)	3(60%)	2(50%)	5(71.4%)
	부당	2(50%)	2(40%)	1(25%)	2(28.6%)
	모름	1(25%)	0(0%)	1(25%)	0(0%)

&lt;표 4&gt; 슬라이딩 윈도우 크기의 변화에 따른 성능 평가

W(분)	세션 개수	발견된 패턴 개수	타당성
1	73,192,929,039	15	67.7%
2	6,853,742,836	14	68.5%
3	1,958,289,740	11	70.5%
5	368,526,283	7	71.4%
10	40,102,492	4	86.3%
30	2,510,275	3	90.5%
60	231,022	3	94%

교하고자 한다.

<표 3>과 같이, 본 연구에서 제안하는 전처리 기법(M<sub>Onto</sub>)가 다른 세 가지 기법들과 비교하여 더욱 다양한 패턴이 발견되었다. 이와 더불어, 발견된 패턴의 정확성을 비교 평가하기 위하여 실험 참가자들에게 무작위로 각 방법에 대한 타당성을 설문하였으며, M<sub>Onto</sub>가 다른 방법론들에 비해 다소 높은 타당성을 보였다.

다음으로 중요한 이슈는 센서 공간 내의 센서 개수의 증가 또는 다른 센서 공간들과의 융합에 따르는 확장성(Scalability) 문제이다. 대용량의 스트림 데이터 처리를 위한 샘플링 기법 M<sub>Sampling</sub>가 비교적 저조한 성능을 보인 것은 센서의 종류가 많지 않음으로써 샘플링의 효과를 볼 수 없었으며, M<sub>Time</sub>와 유사한 작업을 수행한 결과를 보였다. 이 문제와 관련하여 향후 추가적인 센서의 설정이 요구된다.

슬라이딩 윈도우 크기(W)의 변화에 따른 성능 평가를 위하여 우리는 <표 4>와 같이 슬라이딩 윈도우 크기를 바꿔가면서 세션 개수, 발견된 패턴의 개수 및 타당성을 측정해보았다. 윈도우 크기가 길어짐에 따라 세션 개수는 지속적으로 감소함을 볼 수 있었으며 타당성은 비교적 향상됨을 알았다. 보다 중요하게 윈도우 크기가 길어짐에 따라, 센서 공간 내의 Long-term 패턴을 발견할 수 있음을 밝혔다.

## 5. 결론 및 향후 연구

센서 네트워크 환경의 중요성이 대두되면서 다양한 센서의 개발이 이뤄지고 있다. 즉, 우리 주변의 다양한 환경 정보를 손쉽게 수집할 수 있음을 예상할 수 있다. 본 연구에서는 전통적인 데이터 마이닝 기법이 효과적으로 센서 네트워크로부터의 데이터 스트림에 적용되게 하기 위한 온톨로지 기반의 데이터 전처리 과정(Preprocessing)을 제안하고 평가하였다.

본 연구 결과의 성능 평가를 위하여 센서 네트워크 테스트 환경을 교내에 설치하였으며 30여일 동안 수집된 데이터를 이용하여 시뮬레이션을 실행하였다.

Jung(2005)을 포함한 다양한 시맨틱 어노테이션 연구들이 일반적으로 문서나 이미지와 같은 데이터에 관심이 있다면, 본 연구에서는 센서 데이터에 시맨틱 어노테이션을 시도했다는 것이 중요한

연구 성과라 할 수 있다.

향후 연구 방향으로서 크게 다음과 같은 두 가지 이슈에 초점을 맞추고 있다.

- 스트림 추론(Stream reasoning)에 대한 연구가 필요하다. 최근 Valle et al.(2008)과 같이 본 연구에서 제안하는 데이터 전처리 과정에서 발생할 수 있는 논리적 오류의 검출이 필수적이라 할 수 있다.
- 실제 환경에 적용하여 사례 연구를 실시한다. Sheth et al.(2008)에서는 날씨 정보 시스템과의 연동을 통해 실제 응용 시스템을 연구하고 있다. 본 연구의 데이터 처리 과정이 실제 시스템에 적용되어야 한다.

## 참고문헌

- Agrawal, R. and Srikant, R., "Mining Sequential Patterns", In P. S. Yu and A. S. P. Chen, editors, Proceedings of the 11th International Conference on Data Engineering, Taipei, Taiwan, 1995, 3~14.
- Chen, H., F. Perich, T. Finin, and Joshi, A., "SOUPA: Standard Ontology for Ubiquitous and Pervasive Applications", In Proceedings of the 1st Annual International Conference on Mobile and Ubiquitous Systems : Networking and Services(Mobiquitous 2004), Boston, Massachusetts, USA, August(2004), 22~26.
- Culler, D., D. Estrin, and M. Srivastava, "Overview of Sensor Networks", *Computer*, Vol.37, No.8 (2004), 41~49.
- Eagle, N. and A. Pentland, "Reality Mining: Sensing Complex Social Systems", *Personal and Ubiquitous Computing*, Vol.10, No.4(2006), 255~268.
- Eid, M., R. Liscano, and A. Saddik, "A Universal Ontology for Sensor Networks Data", In Proceedings of the 2007 IEEE International Conference on Computational Intelligence for Measurement Systems and Applications, Ostuni, Italy, June(2007), 27~29.
- Estrin, D., D. Culler, K. Pister, and G. Sukhatme, "Connecting the Physical World with Pervasive Networks", *IEEE Pervasive Computing*, Vol.1, No.1(2002), 59~69.
- Fox, M. S., "The TOVE Project Towards a Common-Sense Model of the Enterprise", In F. Belli and F. J. Radermacher, editors, Proceedings of the 5th International Conference on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems (IEA/AIE-92), Paderborn, Germany, June 9-12, 1992, Lecture Notes in Computer Science, Vol.604, 25~24.
- Jabeur, N., J. D. McCarthy, X. Xing, and P. A. Graniero, "A Knowledge-oriented Meta-framework for Integrating Sensor Network Infrastructures", *Computers and Geosciences*, Vol.35, No.4(2009), 809~819.
- Jung, J. J., "Semantic preprocessing of web request streams for web usage mining", *Journal of Universal Computer Science*, Vol.11, No.8 (2005), 1383~1396.
- Jung, J. J., "Exploiting Semantic Annotation to Supporting User Browsing on the Web", *Knowledge-Based Systems*, Vol.20, No.4(2007), 373~381.
- Jung, J. J., "Contextualized mobile recommendation service based on interactive social network discovered from mobile users", *Expert Systems with Applications*, Vol.36, No.9(2009), 11950~11956.
- Lim, H. B., Y. M. Teo, P. Mukherjee, V. T. Lam, W. F. Wong, and S. See, "Sensor Grid : Inte-

- gration of Wireless Sensor Networks and the Grid”, In Proceedings of the 2005 IEEE Conference on Local Computer Networks (LCN’05), Sydney, Australia, 2005.
- Sheth, A., C. Henson, and S. S. Sahoo, “Semantic Sensor Web”, *IEEE Internet Computing*, Vol.12, No.4(2008), 78~83.
- Sheth, A., “Citizen Sensing, Social Signals, and Enriching Human Experience”, *IEEE Internet Computing*, 2009.
- Storey, M.-A., M. Musen, J. Silva, C. Best, N. Ernst, R. Ferguson, and N. Noy, “Jambalaya : Interactive visualization to enhance ontology authoring and knowledge acquisition in Protege”, *In Proceedings of the International Workshop on Interactive Tools for Knowledge Capture (K-CAP 2001)*, 2001.
- Valle, E. D., S. Ceri, D. F. Barbieri, D. Braga, and A. Campi, “A First Step Towards Stream Reasoning”, The Large Knowledge Collider (LarKC), European Union Framework Programme 7, Deliverables, 2008.
- Weiser, M., R. Gold, and J. S. Brown, “The origins of ubiquitous computing research at PARC in the late 1980s”, *IBM Systems Journal*, Vol.38, No.4(1999), 693~696.

Abstract

## Ontology based Preprocessing Scheme for Mining Data Streams from Sensor Networks

Jason J. Jung

By a number of sensors and sensor networks, we can collect environmental information from a certain sensor space. To discover more useful information and knowledge, we want to employ data mining methodologies to sensor data stream from such sensor spaces. In this paper, we present a novel data preprocessing scheme to improve the performances of the data mining algorithms. Especially, ontologies are applied to represent meanings of the sensor data. For evaluating the proposed method, we have collected sensor streams for about 30 days, and simulated them to compare with other approaches.

**Key Words** : Ontology, Sensor networks, Knowledge discovery, Stream Mining

---

\* Department of Computer Engineering, Yeungnam University

## 저 자 소 개



정재은

저자는 현재 영남대학교 컴퓨터공학과에 조교수로 재직 중이다. 연구분야는 Description Logic과 Knowledge Engineering이다.