

# 센서 데이터의 시각 분석 시스템\*

최문영<sup>○</sup> 김우일 윤현식 심창범 정연돈

고려대학교 컴퓨터학과

cmyell74@korea.ac.kr, wooilkim@korea.ac.kr, hyunsikyoon@korea.ac.kr,

scbeom@korea.ac.kr, ydchung@korea.ac.kr

## A Visual Analysis System for Sensor Data

Moonyoung Choi<sup>○</sup>, Wooil Kim, Hyunsik Yoon, Changbeom Shim, Yon Dohn Chung

Department of Computer Science and Engineering, Korea University

### 요 약

본 논문은 센서 데이터를 효율적으로 비교, 분석하는 시각화 시스템을 제안한다. 심층적인 분석 작업을 위해 시스템을 3가지 화면으로 나누어 상호 보완적으로 동작하도록 하였으며, 이를 통해 비정상 데이터 탐지나 시공간적 변화 양상 분석이 가능하다. 또한 실제 데이터를 분석함으로써 제안하는 시각 분석 시스템의 효과성을 보였다.

### 1. 서 론

센서에서 측정되는 데이터는 센서가 설치된 지역에 대한 핵심적인 정보를 제공한다. 센서 기술이 발전함에 따라 얻을 수 있는 데이터의 양이 기하급수적으로 증가하였다. 이에 따라 센서 데이터에서 유의미한 정보를 효율적으로 발견할 수 있는 시각화 시스템에 대한 관심이 증대되었다[1]. 또한 개별 센서 데이터를 분석하여 정보를 얻는 것 뿐 아니라, 각 센서에서 측정된 데이터 간의 비교를 통해서 정보를 얻어야 할 필요성이 대두되었다. 따라서 시간 및 지역 별로 측정된 센서 데이터 간의 연관성을 효율적으로 분석할 수 있는 시스템이 필요하다[2].

본 논문은 이러한 대용량 센서 데이터 분석을 위한 시각화 시스템을 제안한다. 본 시스템은 사용자의 접근성을 고려하여 python과 d3를 이용해 웹 기반으로 구현되었다. 사용자는 본 시스템을 이용해 센서가 설치된 지역의 이상 현상 파악과 시간 별, 지역 별 변화 양상 파악, 데이터 간 공통점 파악 등의 작업을 수행할 수 있다. 또한 제안한 시스템의 성능을 확인하기 위해 하천에 잔류하는 오염 의심 물질을 비정기적으로 수집한 실제 센서 데이터를 활용하여 사례 분석을 하였다. 제안한 시각 분석 시스템을 통해 사용자는 센서 측정 데이터의 의미를 효과적으로 파악할 수 있으며,

더 나아가 현재 센서 데이터를 측정하는 방식에 대한 문제점도 파악할 수 있다.

### 2. 설계 목표

본 논문에서 제안하는 시각 분석 시스템의 목표는 다음과 같다.

**G1.** 센서에서 측정한 데이터를 시간 흐름에 따라 시각화하여, 여러 데이터의 변화 추이를 다양한 스케일로 확인할 수 있다.

**G2.** 각 데이터를 평균, 최대, 최소, 중앙값과 같은 통계 자료와 비교하여 비정상적인 센서 데이터를 쉽게 파악할 수 있다.

**G3.** 센서 데이터를 인접한 지역 별로 비교할 수 있도록 하여, 지역 간 관련성을 파악할 수 있다.

**G4.** 특정 데이터가 비정상적으로 많이 관찰된 지역과 시간대를 파악할 수 있다.

**G5.** 사용자와 시스템의 상호작용적 탐색을 지원한다.

### 3. 분석 과제

본 논문에서 제안한 시각화 시스템을 통해 다음과 같은 분석 과제를 수행할 수 있다.

**T1.** 시간 흐름에 따른 센서 데이터의 총체적, 개별적 변화를 관찰함으로써 전체 센서 데이터의 변화 양상을 시간별, 지역별로 파악한다.

**T2.** 센서 데이터가 비정상적으로 관측된 날짜와 지역을 파악한다.

\* 이 논문은 2018년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. NRF-2017R1A2A2A05069318).

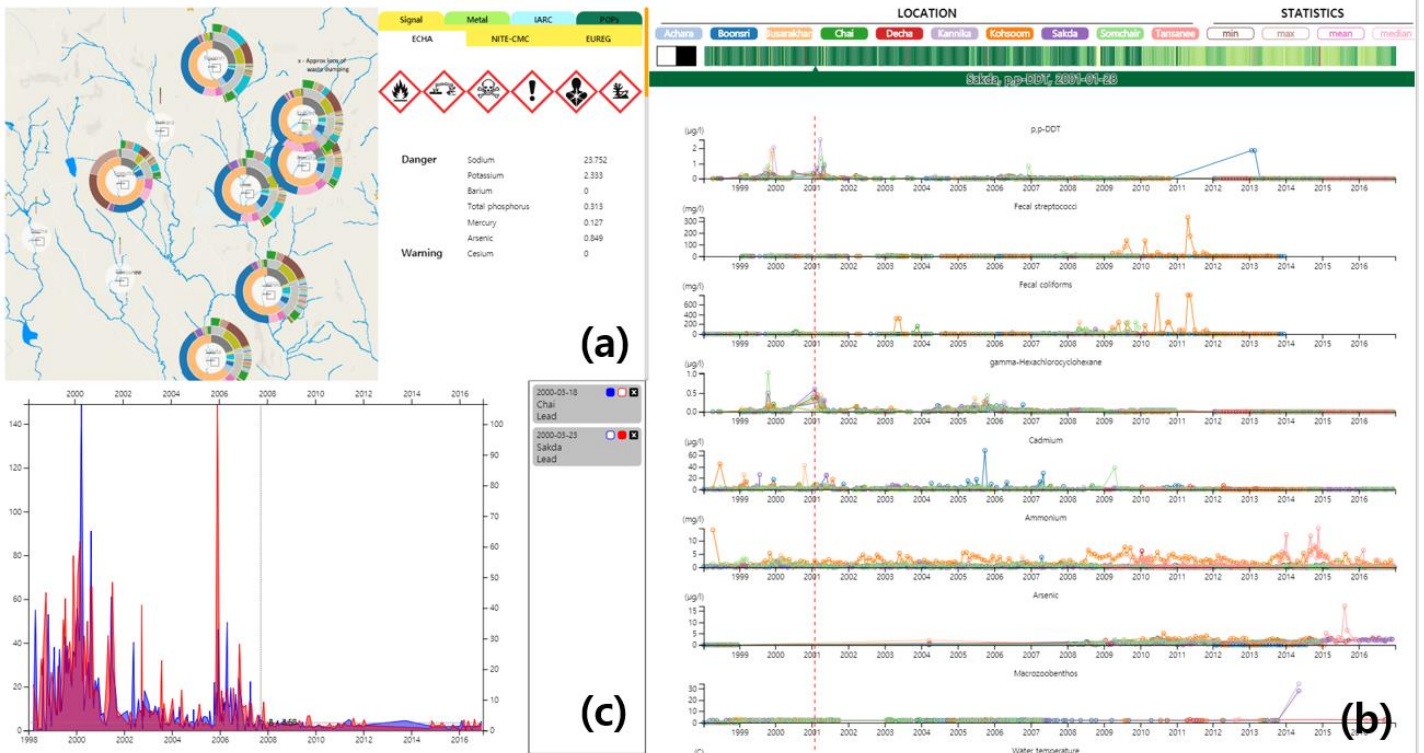


그림 1. 시각화 시스템 인터페이스: 지도 화면(a), 타임라인 화면(b), 비교 화면(c)

**T3.** 데이터 간 비교를 통해 데이터 측정 방식의 적합성 여부를 분석한다.

**T4.** 센서 데이터 간 공통점과 공통적인 변화 양상을 찾는다.

#### 4. 시스템 개요

본 논문의 시각화 시스템은 크게 3가지 화면으로 구성된다. 그림 1은 시각화 시스템 인터페이스이다.

##### 4.1 데이터 전처리

###### 1) 센서 데이터 선형 보간(Linear interpolation)

센서 데이터는 정기적 혹은 비정기적으로 측정된 한 시점의 값이다. 서로 다른 시점에 측정된 센서 데이터를 동시에 비교하기 위해 선형 보간하였다.

###### 2) 데이터 분류(Clustering)

사용자가 센서에서 측정된 데이터의 특성을 파악하고, 비슷한 특성을 가진 데이터를 쉽게 비교, 분석하기 위해 데이터를 분류한다. 본 논문에서 사용한 실제 데이터는 해당 데이터의 도메인 지식을 바탕으로 계층적 클러스터링 (Hierarchical clustering) 기법을 통해 분류하였다. 분류 기준은 5장에 자세히 서술하였다.

##### 4.2 지도 화면(Map view)

지도 화면에서는 인접한 지역 별로 센서 데이터를 비교하여 관련성을 찾아내고, 특징을 파악할 수 있다 (G3, G5). 그림 2는 시스템의 지도 화면이다. 각 지역

에서 측정한 센서 데이터가 전처리를 통해 분류된 카테고리에 따라 파이 차트로 각각 표현되어 있다. 파이 차트는 카테고리에 따라 계층 구조로 구성되어 있으며, 파이 차트 각 구성 요소 위에 커서를 올리면 중심에 해당 구성 요소에 할당된 센서 데이터와 검출량이 표시된다(G3). 지도 오른쪽의 표를 통해서 분류 기준을 다르게 적용할 수 있으며, 해당 분류에 따라 파이 차트가 재구성된다. 표에는 각 분류 별로 센서 데이터의 목록과 검출량이 제시된다(G5, A3).

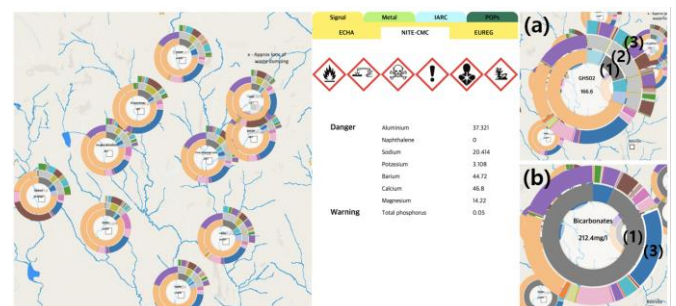


그림 2. 지도 화면: 계층적 구성의 파이 차트. (a)는 2단계로 분류(안쪽부터 차례로 대분류(1), 소분류(2), 센서 데이터(3)). (b)는 1단계로 분류(안쪽부터 차례로 대분류(1), 센서 데이터(3))

##### 4.3 타임라인 화면(Timeline view)

타임라인 화면에서는 센서 데이터 수집 시간에 따른 변화 추이를 확인할 수 있다(G1, T1). 그림 3은 시

시스템의 타임라인 화면이다. 해당 화면은 위치 버튼 바(a), 통계 값 바(b), 시간대 바(c), 그리고 그래프 표시부(d)로 구성된다. 그래프 표시부에는 검출량이 많은 데이터부터 내림차순으로 그래프가 표시된다(T1, T2). 그래프 표시부에서는 위치 버튼 바를 통해서 사용자가 원하는 지역의 센서 데이터 그래프를 띄울 수 있다. 통계 값 바를 통해 각 센서 데이터의 연도별 최솟값, 최댓값, 평균값, 중간값을 센서 데이터 그래프와 비교할 수 있다(G2, G3, G4). 시간대 바를 통해 사용자는 선택한 날짜의 센서 데이터 그래프를 볼 수 있고, 해당 날짜의 센서 데이터를 사용해 지도 화면이 갱신된다. 시간대 바 옆 버튼을 통해 실제 데이터와 보간 데이터의 구분 여부를 결정할 수 있다. 시간대 바를 확대한 (e)를 보면, 시간대 바에서 해당하는 날짜의 색깔이 빨간색에 가까울수록 평균치와 차이가 큰 비정상적인 센서 데이터가 존재한다는 뜻이고, 초록색에 가까울수록 정상 데이터에 가깝다는 것을 의미한다(G4, T2).

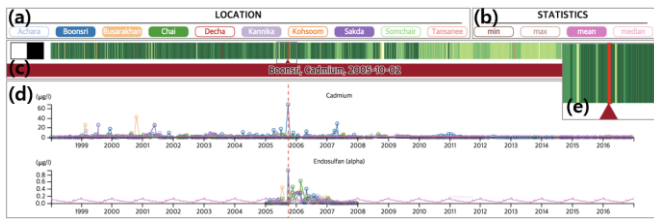


그림 3. 타임라인 화면: 그래프 표시부(d)에는 각 원소 별 꺾은선 그래프 출력(가로 축은 측정 시간, 세로 축은 센서 데이터 측정값)

#### 4.4 비교 화면(Comparison view)

비교 화면에서는 센서 데이터 그래프의 범위 변화를 통해 원하는 센서 데이터를 심층적으로 비교, 분석할 수 있다(G1). 그림 4는 시스템의 비교 화면이다, 사용자는 타임라인 화면의 그래프 표시부에서 다른 센서 데이터와 비교, 분석하고 싶은 센서 데이터를 클릭하여 비교 화면 목록(b)에 추가할 수 있다. 이 경우 측정 시간, 지역, 센서 데이터의 일련이 비교 화면 목록에 추가된다. 사용자는 추가된 일련 중 2가지를 선택

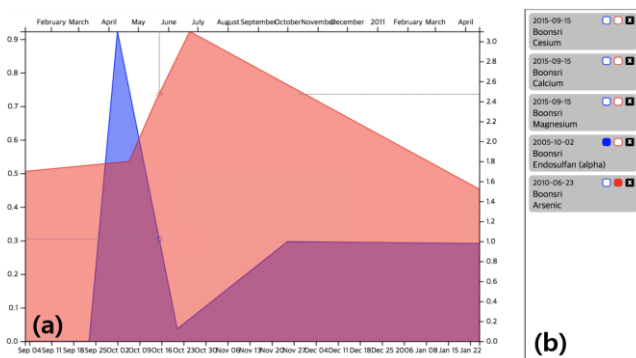


그림 4. 비교 화면

택해 그래프의 형태와 측정값을 비교할 수 있다. 그래프 간 비교화면(a)에서는 가로축인 측정 시간, 세로축인 측정값 규모를 실시간으로 변경하면서 데이터를 비교할 수 있다(T4).

#### 4.5 시각화 분석 과정(Visual analysis process)

타임라인 화면에서 분석하고자 하는 센서 데이터 측정 시간을 시간대 바에서 선택한다. 해당 측정 시간의 센서 데이터를 이용해 지도 화면이 갱신된다. 타임라인 화면의 그래프 표시부에서 측정 시간, 지역, 센서 데이터 일련을 클릭하여 비교 화면에 추가한다(G6). 그림 5는 위 분석 과정 흐름 모식도이다.



그림 5. 시각화 분석 과정 흐름 모식도

#### 5. 사례 분석(Case study)

시각화 시스템의 디자인 목표 구현, 효율적인 정보 분석 수행 여부를 판단하기 위해 하천에 잔류하는 오염 의심 물질을 비정기적으로 수집한 실제 센서 데이터[3]를 활용하여 분석 과제를 수행하였다. 해당 데이터는 다음과 같이 구성되어 있다.

1. 각 지역 지도 데이터
2. ID, 측정값, 지역, 샘플 측정 날짜, 측정 성분

지도 데이터에는 10가지의 지역이 물질을 기준으로 표시되어 있으며, Achara, Somchair, Sakda 지역, Boonsri, Kohsoom, Busarakahn, Chai, Kannika 지역이 같은 물질을 공유하고 나머지 지역은 독립적으로 존재한다. ID는 각 데이터의 고유값이다. 샘플 측정 날짜는 1998년 1월 11일부터 2016년 12월 31일까지이며, 측정 성분은 106가지이다.

센서에서 측정한 성분들의 특성과 위험도를 쉽게 파악할 수 있도록 측정된 성분을 4가지 범주: GHS, Metal, IARC, POPs로 분류하였다. GHS는 세계조화시스템(Globally Harmonized System of Classification & Labeling of Chemicals)에 따라 화학물질의 유해성, 위험성을 분류해 놓은 기준이다. Metal은 화학 원소의 특성에 따라 금속, 준금속, 비금속으로 분류해 놓은 기준이다. IARC는 국제 암 연구 기관(International Agency for Research on Cancer)이 사람에게 미치는 발암성에 따라 성분을 분류해 놓은 기준이다. POPs란 생산활동이나 폐기물의 처리 과정에서 생산되는 잔류성 유기오염물질 분류를 위한 스톡홀름 협약의 규제 기준이다.

## 5.1 사례 분석 결과

사례 1. 해당 데이터셋에서 공장 폐수와 관련된 성분이 비정상적으로 검출된 시간대와 지역을 찾고자 한다(T2). 그림 6의 (a)를 보면, 시각화 시스템 타임라인 화면의 시간대 바에서 빨간색으로 표시된 날짜 중, 구리와 철 등 Metal 범주에 속한 모든 성분들의 검출량이 급격하게 증가한 2003년 8월 15일에 공장 폐수 방출이 일어났다고 판단된다. (b)의 지도 화면에서 해당 날짜의 지역별 파이차트를 비교했을 때 같은 물질을 공유하고 있는 지역 모두 같은 현상이 관찰된다는 사실이 해당 가설을 뒷받침한다. 또한 (c)의 비교 화면에서 강의 상류에 위치한 Kohsoom에 비해 하류에 위치한 Kannika가 구리의 검출량이 적은 것을 확인할 수 있다. 그러므로 강 상류 부근부터 공장 폐수에 의한 수질 오염이 발생한다는 사실이 유추된다.

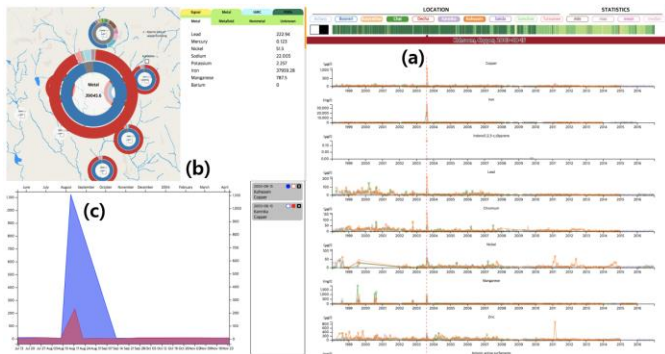


그림 6. 사례 1: 2003년 8월 15일의 이상현상을 (a)타임라인 화면, (b)지도 화면, (c)비교 화면에서 각각 확인

사례 2. 데이터 측정 기간 내 성분 검출량의 총체적, 개별적 변화를 관찰함으로써 전체 센서 데이터의 변화 양상을 지역별로 비교하려 한다(T1). 그림 7의 (b) 타임라인 화면을 보면 하천을 공유하는 Busarakhan, Chai, Kannika 지역, Sakda와 Somchair 지역(a) 각각의 성분 검출량이 비슷하고, (c) 비교 화면을 통해 분석한 결과 그래프의 추이가 거의 유사하다는 사실은 하천을 공유하고 있는 지역끼리 서로 연관성이 있다는 가설을 뒷받침한다. 그러므로 오염 성분이 물을 매개로 전달되며, 공장 폐수가 오염원이라는 사실이 합당하게 유추된다. 또한, (d)에 해당하는 Boonsri 지역은 위에서 언급한 Busarakhan, Chai, Kannika 지역과 하천을 공유함에도 불구하고 큰 연관성이 발견되지 않은 것으로 보아 공장 폐수를 방류한 지역이 Boonsri보다는 하류인 것으로 사료된다.

## 6. 결론 및 향후 연구 방향

본 논문은 실시간으로 측정되는 다량의 센서 데이터 의미 분석의 어려움을 해소하기 위한 시각화 시스템을 소개하고, 해당 시스템으로 수행할 수 있는 대표

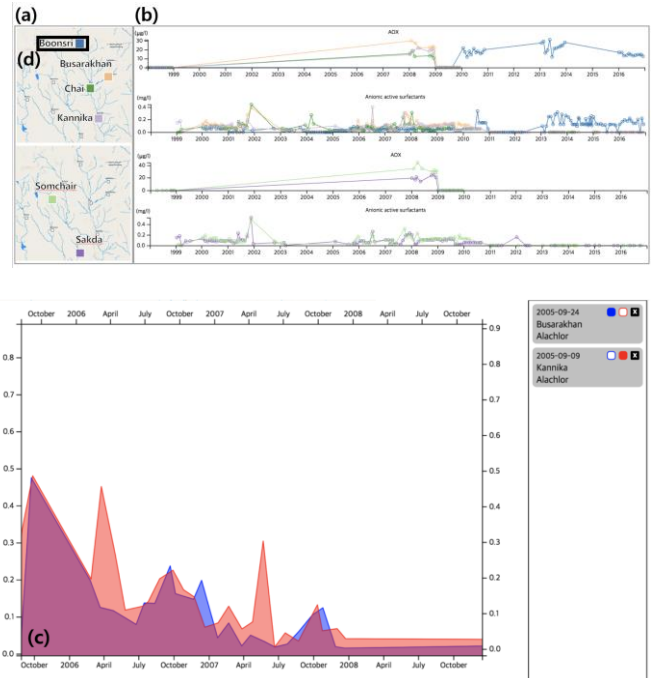


그림 7. 사례 2: (a)지역 간 연관성과 (d)지역 간 차별성을 (b)타임라인 화면, (c)비교 화면을 통해 분석

적인 분석 과제를 제안하였다. 또한 실제 수로 센서 데이터를 이용해 환경 오염 분석 과제를 효과적으로 수행하였다. 향후 연구로는 센서 데이터 간 관련성을 측정하여 그 값을 기반으로 데이터를 시각화하는 분석 시스템 연구를 수행할 계획이다.

## 7. 참고문헌

- [1] Hugo Hromic, Danh Le Phuoc, Martin Serrano, "Real time analysis of sensor data for the Internet of Things by means of clustering and event processing", IEEE International Conference on Communications(2015): p536-541
- [2] Jill A. Engel-Cox, Christopher H.Holloman, Basil W.Coutant, Raymond M.Hoff, "Qualitative and quantitative evaluation of MODIS satellite sensor data for regional and urban scale air quality", Atmospheric Environment Volume 38, Issue 16 (2004): p2495-2509
- [3] VAST challenge 2018 Mini Challenge 2(MC2) Dataset, <http://vacommunity.org/VAST+Challenge+2018+MC2>