The Journal of The Institute of Internet, Broadcasting and Communication (IIBC)
Vol. 14, No. 6, pp.229-234, Dec. 31, 2014. pISSN 2289-0238, eISSN 2289-0246

http://dx.doi.org/10.7236/JIIBC.2014.14.6.229
JIIBC 2014-6-33

스마트폰 센싱에서 메타데이터의 구조적 유사도를 고려한 클러스터링 기법

A Clustering Scheme Considering the Structural Similarity of Metadata in Smartphone Sensing System

민 홍*, 허준영**

Hong Min*, Junyoung Heo**

요 약 다수의 저가 센서 노드를 통해 주변의 환경 정보를 수집하는 센서 네트워크와 스마트폰에 탑재되어 있는 다양한 종료의 센서들을 연동함으로써 사용자의 상태에 따라 주위 환경과 반응하는 응용들이 개발되고 있다. 이런 응용에서 수집된 데이터의 공유를 위해 센싱 데이터와 의미정보를 저장하는 XML 형태의 메타데이터를 함께 저장할 필요가 있다. 메타데이터는 시스템 설계자의 필요에 따라 확장되고 변형되는데 거리 기반의 클러스터링 기법을 사용할 경우 서로 다른 형태의 메타데이터가 혼재하게 되어 데이터 수집의 효율성이 떨어지는 문제가 발생한다. 본 논문에서는 효율적인 데이터 수집을 위해 클러스터를 구성할 때 각 노드의 메타데이터의 구조적 유사도를 반영함으로써 클러스터 구성에 필요한 시간을 줄이고, 구성원 간 메타데이터 유사도를 향상시키는 기법을 제안한다.

Abstract As association between sensor networks that collect environmental information by using numberous sensor nodes and smartphones that are equipped with various sensors, many applications understanding users' context have been developed to interact users and their environments. Collected data should be stored with XML formatted metadata containing semantic information to share the collected data. In case of distance based clustering schemes, the efficiency of data collection decreases because metadata files are extended and changed as the purpose of each system developer. In this paper, we proposed a clustering scheme considering the structural similarity of metadata to reduce clustering construction time and improve the similarity of metadata among member nodes in a cluster.

Key Words : Clustering, Metadata, Structural Similarity, Smartphone Sensing

Ⅰ. 서 론

최근 스마트폰은 휴대용 통신기기로써의 역할뿐만 아니라 탑재되어 있는 다양한 종류의 센서와 주변 센서들과의 연동을 통해 사용자의 상태와 주변 환경을 모니터링 할 수 있다[1]. 다수의 값싼 기기에 의존했던 기존의 무선 센서 네트워크와[9]는 달리 성능이 뛰어나고 사용자에 의해 원하는 목적지로 이동이 가능한 스마트폰을 활용하여 데이터를 수집하는 스마트폰 센싱 시스템은 앞으로 사용자의 사회 활동과 연관되어 많은 응용 분야에서 사

*정회원, 호서대학교, 컴퓨터정보공학부
**정회원, 한성대학교, 컴퓨터공학과
접수일자: 2014년 9월 29일, 수정일자: 2014년 10월 29일
게재확정일자: 2014년 12월 12일

Received: 29 September, 2014 / Revised: 29 October, 2014
Accepted: 12 December, 201
**Corresponding Author: jyheo@hansung.ac.kr
Dept. of Computer Engineering, Hansung University, Korea

용될 것으로 전망된다[2,10]. 예를 들어, 스마트폰 사용자들의 위치 정보를 수집하여 트래픽 모니터링을 하거나, 생체 센서를 통해 개인이나 그룹의 헬스케어, 환경 모니터링이 가능하다.

스마트폰 센싱을 통해 수집된 측정값들은 장치의 특성에 따라서 데이터의 형식이 다르고 저장하는 방식에 따라 데이터양의 차이가 크다. 또한 데이터 수집에 많은 비용을 지불하고 있지만 각 연구 집단마다 표준화 되지 않은 방식으로 데이터를 저장하고 있어 이를 공유하기가 힘들다. 연구자 상호간 데이터 교류를 지원하고 순수 데이터에서 의미 있는 정보를 효과적으로 추출하기 위해서 메타데이터의 정보를 함께 저장하는 온톨로지(Ontology) 기반 시멘틱 웹 (Semantic web) 기술들이 연구되고 있다[3,4,5].

본 논문에서는 센서 측정값과 메타데이터를 동시에 저장하고 유사한 메타데이터를 갖는 노드들끼리 클러스터를 구성하여 데이터 수집의 효율성을 높인 사전연구인 TF-IDF (Term Frequency and Inverse Document Frequency) 기반 유사성 분석 기법[6]을 기반으로 클러스터 구성에 필요한 시간을 줄이기 위한 구조적 메타데이터 유사성 비교 기법을 적용하고 실험을 통해 성능을 검증하였다.

본 논문의 구성은 다음과 같다. 2장에서는 센싱 데이터의 의미정보를 표현하는 방법에 대해 설명하고, 3장에서는 메타데이터의 유사도에 기반을 둔 클러스터링 구성 기법을 제안한다. 4장에서는 제안한 기법의 실험 및 결과에 대해 서술하고, 5장에서는 결론을 맺는다.

II. 센싱 데이터의 의미정보 표현

기존의 센서 네트워크에서와 같이 스마트폰 센싱에서도 데이터를 수집하기 위해서는 많은 비용과 노력이 필요하다. 특정 장소의 온도와 습도를 측정한 결과인 그림 1에서 보는 바와 같이 순수한 센싱 데이터는 시간의 연속성을 가지고 있으며 부가적인 정보 없이 이를 해석하는 것이 어렵다[7]. 이러한 문제로 연구자들 간에 센싱 데이터를 공유하는 것이 불가능하기 때문에 센싱 된 데이터의 의미 정보를 함께 저장하는 연구가 진행 되었다.

```
1  1  45.93  27.97  0
2  1  45.9   27.95  0
3  1  45.9   27.96  0
4  1  45.93  27.95  0
5  1  45.93  27.97  0
```

그림 1. 수집된 센싱 데이터의 예
Fig. 1. An example of collected sensing data

센싱된 데이터의 의미 정보를 저장하기 위해서 확장성을 고려한 XML(eXtensible Markup Language) 형식에 기반을 둔 RDF (Resource Description Framework)과 OWL (Ontology Web Language) 등의 온톨로지 기술 기법들이 활용된다. 그림 2는 센서 데이터에 대한 메타데이터 기술 방법의 예를 보여준다. 두 예제 모두 온도 센서로부터 측정된 데이터와 측정된 장비, 값의 단위 등과 같은 의미 정보를 함께 저장하고 있다. 이를 통해서 순수 데이터만으로는 알 수 없는 정보들을 추출 할 수 있게 된다. 그러나 두 연구 기관에서 사용하는 메타데이터의 양식이 다르고 시간이 지남에 따라 버전이 갱신되기 때문에 표준화된 양식을 지원하는 것이 어렵다.

```
<hasSensor>
  <HOBO_S-TMA-M002_TemperatureSensor
  rdf:ID="HOBO_S-TMA-M002_TemperatureSensor_001">
    <hasOffset>
      <Offset rdf:ID="Offset_10feetbelow">
        <x rdf:datatype="http://www.w3.org/2001/XMLSchema#float">
        -10.0</x>
        <z rdf:datatype="http://www.w3.org/2001/XMLSchema#float">
        0.0</z>
        <y rdf:datatype="http://www.w3.org/2001/XMLSchema#float">
        0.0</y>
      </Offset>
    </hasOffset>
  </HOBO_S-TMA-M002_TemperatureSensor>
</hasSensor>
```

(a) Coastal Environmental Sensing Networks

```
<Quantity rdf:ID="Temperature">
  <hasUomIdentifier>
    <rdf:Description
    rdf:about="http://sweet.jpl.nasa.gov/ontology/units.owl#degreeC">
      <hasDoubleValue
      rdf:datatype="http://www.w3.org/2001/XMLSchema#double">
      -7.3
      </hasDoubleValue>
    </rdf:Description>
  </hasUomIdentifier>
</Quantity>
```

(b) Swiss Experiment platform

그림 2. 의미정보를 포함한 메타데이터의 예
Fig. 2. Examples of semantic metadata

Ⅲ. 메타데이터 유사도 기반 클러스터링

1. 클러스터링 기법

스마트폰 센싱에서 각 노드는 참여도와 적극성에 따라 참여기반 노드와 기회기반 노드로 구분할 수 있다. 참여기반 노드(Participatory node)는 실험에 참여하는 과정에서 금전적 보상과 같은 특혜를 제공받기 때문에 적극적으로 데이터를 수집하며 메타데이터 및 실험 파라미터 갱신과 같은 관리자의 지시를 잘 수행한다. 기회기반 노드 (Opportunistic node)은 전송된 메시지나 영상 등과 같이 의도되지 않은 기회를 통해 센싱 데이터를 제공하기 때문에 관리자의 지시를 따르지 않는다. 이러한 구조적인 특성 때문에 참여기반 노드를 클러스터 헤드로 선정하고 기회 기반 노드를 멤버 노드로 클러스터링하는 기법이 데이터 수집에 효율적이다.

그림 3은 거리기반 클러스터링 구성 기법과 제안된 메타데이터 클러스터링 기법과의 차이를 보여준다. 기회기반 노드를 사이에서 A, B, C 버전의 메타데이터가 혼재되어 있는 상황을 가정했을 때, 기존 기법은 거리가 가까운 노드들로 클러스터를 구성하지만 제안 기법에서는 유사한 메타데이터 파일을 가지고 있는 노드들로 클러스터를 구성하게 된다.

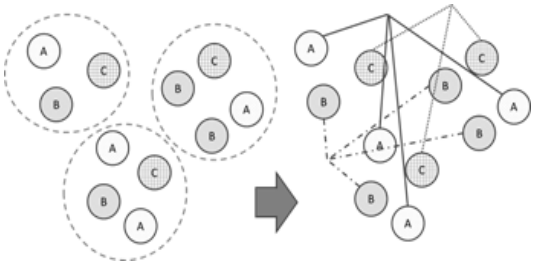

그림 3. 거리기반과 메타데이터 유사도 기반 클러스터링 비교
Fig. 3. Compare metadata similarity based clustering with distance based clustering

2. 사전 연구

사전 연구인 TF-IDF 기반 메타데이터 클러스터링은 그림 4와 같은 과정을 통해 메타데이터의 유사도를 분석한다[6].

클러스터 구성 시 태그 선정과 유사도 비교 과정에서 $O(N^3)$의 복잡도를 갖기 때문에 문서의 크기와 메타데이터의 수가 증가함에 따라 클러스터 구성에 많은 시간이 소모되고 대부분의 메타데이터 문서에서 비슷한 용어(term)을 사용하기 때문에 클러스터링 결과의 정확성이 떨어지는 문제가 발생한다.

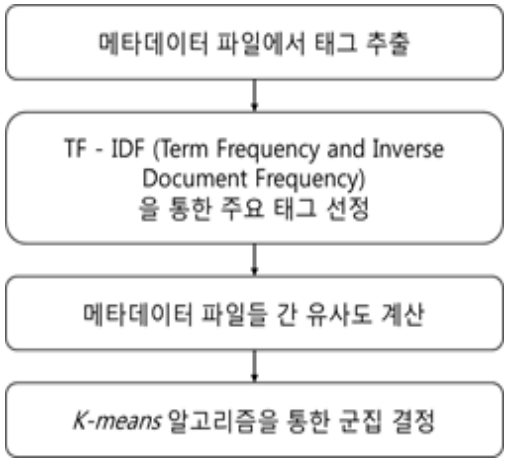

그림 4. TF-IDF 기반 클러스터 구성 기법
Fig. 4. TF-IDF based cluster construction scheme

3. 구조적 유사도를 고려한 클러스터링 기법

본 논문에서는 온톨로지에 기반을 둔 메타데이터 파일의 구조적인 특성을 활용하여 파일간 유사도를 도출함으로써 TF-IDF 기법에 비해 비교 대상의 크기를 줄여 클러스터 구성에 필요한 시간을 줄일 수 있는 기법을 제안한다.

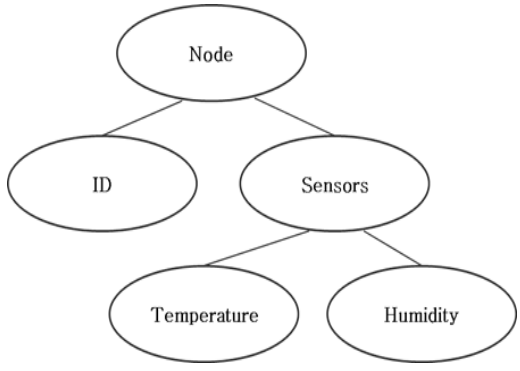

그림 5. 메타데이터 파일의 태그 트리
Fig. 5. A tree of tags in the metadata file

XML 문서에서 사용한 용어뿐만 아니라 대한 구조적 유사도를 측정하기 위해서 SCVM (Structure and Content Vector Model) 기법[8]을 활용하였다. SCVM에서는 XML 문서를 태그에 따라 트리 구조로 변환하고 루트에서부터 모든 패스를 원소로 하는 벡터를 생성한다.

그림 5에서와 같이 메타데이터 파일의 태그로 트리를 생성한 경우 구조적 특성을 나타내는 벡터는 다음과 같이 정의할 수 있다.

$$Vstruct_i = \begin{bmatrix} Node/ID \\ Node/Sensors/Temperature \\ Node/Sensors/Humidity \end{bmatrix}$$

두 메타데이터 파일들 (d_i, d_j) 사이의 유사성을 분석하기 위해서 수식 (1)을 사용한다. 즉 두 문서에서 생성한 패스들을 각각 비교하여 동일한 패스의 수를 계산하고 이를 전체 가능한 조합의 수로 나누어 준다. 따라서 두 파일사이의 유사도(S_{ij})는 동일한 패스의 수가 많을수록 높은 값을 갖는다.

$$S_{ij} = \frac{Vstruct_i^t \bullet Vstruct_j}{\| Vstruct_i \| \bullet \| Vstruct_i \|}, \qquad (1)$$

$$a_{ij} = \begin{cases} 1, \text{if } a_j \in Vstruct_i \\ 0, otherwise \end{cases}$$

마지막으로 계산된 유사도 값에 따라 높은 상관관계를 보이는 메타데이터 파일들끼리 클러스터로 묶어주는 작업이 필요하다. 이를 위해 사전에 정의된 K개의 클러스터를 구성하기 위해서 각 클러스터의 처음 멤버가 되는 시드 노드(Seed Node)를 선정한다. 시드 노드는 소수의 이상치(outlier)를 제외하고 서로간의 메타데이터 유사성이 가장 낮은 K개를 선택한다. 시드 노드를 제외한 나머지 노드들은 각 클러스터의 시드 노드들과의 유사성 비교를 통해서 가장 유사도가 높은 클러스터로 배정된다. 이와 같이 클러스터가 확장해가는 과정에서 클러스터의 구심점(centroid)인 센터 노드가 매번 변경되는데 수식 (2)에서와 같이 각 멤버 노드들 간에 메타데이터 유사도의 오차(E_i^j)를 계산하여 오차를 최소화하는 노드를 새로운 센터 노드로 선정한다. 오차는 클러스터 내의 모든 메타데이터 파일(D_k)을 대상으로 후보 노드($\vec{\mu}(D_k^i)$)와 다른 노드들과($\vec{\chi}$)의 메타데이터 유사도 차이를 제곱하여 합산한다.

$$E_k^i = \sum_{x \in D_k} |\vec{\chi} - \vec{\mu}(D_k^i)|^2 \qquad (2)$$

Ⅳ. 실험 및 결과

본 절에서는 시뮬레이션을 통해 TF-IDF 기법과 제안 기법 사이의 성능을 비교하였다. 실험의 단순화와 클러스터링 기법만의 성능 비교를 극대화하기 위해 네트워크 통신 오류로 인한 오버헤드는 무시했다.

그림 6는 각 기법들 간의 클러스터 구성 완료 시간을 비교한 것이다. 거리기반 방식의 경우 클러스터 구성 시 주변에 위치한 메타데이터에 대한 유사성 분석이 필요하지 않고, 주변 노드들에게 클러스터의 가입 여부를 확인하고 가입되지 않은 노드들을 멤버 노드로 편입하기 때문에 클러스터의 개수에 따라 클러스터 구성 시간에 큰 변화가 없다. 제안 기법의 경우 거리기반 기법에 비해 메타데이터 유사성 분석 과정에 의해 더 많은 부하가 발생하지만 TF-IDF 보다는 부하가 줄어드는 것을 알 수 있다. 이는 유사성 비교 과정에서 제안 기법이 TF-IDF 보다 투입되는 정보의 양이 작기 때문이다.

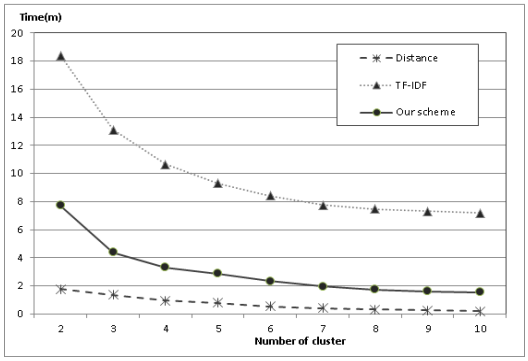

그림 6. 클러스터 구성 시간 비교
Fig. 6. Comparing the cluster construction time

그림 7은 제안 기법과 TF-IDF사의 클러스터 내의 멤버 노드들 사이의 메타데이터 유사도를 비교한 결과이다. 클러스터링 완료 후에 각 멤버 노드들 사이의 메타데이터가 얼마나 일치하는지를 분석한 결과도 제안 기법의 정확도가 높은 것을 알 수 있다. TF-IDF 기법의 경우 유사도 분석 과정의 부하를 줄이기 위해서 상위 20%의 태그만을 추출하게 되는데, 이 때문에 다른 구조를 가지고 있지만 유사한 태그를 사용한 메타데이터들 사이에서 유사성이 높게 나타나는 문제가 발생한다. 제안 기법에서는 사용하는 태그의 내용뿐만 아니라 메타데이터의 구조적 특징을 반영하여 유사도를 계산하기 때문에 정확도가 높다.

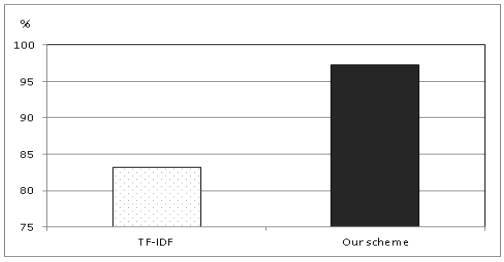

그림 7. 멤버 노드 간 메타데이터 유사도 비교
Fig. 7. Comparing the metadata similarity among member nodes

V. 결 론

스마트폰은 여러 사용자들은 네트워크를 통해 연결시켜주는 역할 뿐만 아니라[11] 탑재된 센서들을 통해 센싱 데이터를 수집하고 데이터 공유를 위해 의미 정보를 포함한 메타데이터를 함께 저장하는 연구들이 진행되고 있다[12]. 본 논문에서는 메타데이터의 유사성이 높은 노드들끼리 클러스터를 구성함으로써 중복적인 데이터를 줄여 효율적인 자료 수집이 이루어 질 수 있는 기법을 제안했다. 제안 기법이 기존의 TF-IDF 기법에 비해 투입되는 정보의 양을 줄여 클러스터 구성 과정에서의 부하를 줄이고 구조적 특성을 반영하여 메타데이터 유사도 측정의 정확도를 높일 수 있음을 실험을 통해 확인하였다.

References

[1] A. T. Campbell, "From Smart to Cognitive Phones," IEEE Pervasive Computing, Vol. 11, No.3, pp.7~11, 2012.

[2] W. Z. Khan, Y. Xiang, M. Y. Aalsalem, and Q. Arshad, "Mobile Phone Sensing Systems: A Survey," IEEE Communications Surveys & Tutorials, Vol.15, No.1, pp.402~427, 2013.

[3] M. Compton et al., "The SSN ontology of the W3C semantic sensor network incubator group," Web semantics, Vol. 17, pp.25~32, 2012.

[4] R. Bendadouche et al., "Extension of the Semantic Sensor Network Ontology for Wireless Sensor Networks", The 11th International Semantic Web Conference, pp.49~64, 2012.

[5] J. Calbimonte et al., "Semantic Sensor Data Search in a Large-Scale Federated Sensor Network", International Workshop on Semantic Sensor Networks, pp.23~38, 2011.

[6] H. Min, and J. Heo, "Document Clustering Scheme for Large-scale Smart Phone Sensing," The Journal of The Institute of Internet, Broadcasting and Communication(JIIBC), Vol. 14, No. 1, pp. 253~258, 2014.

[7] J. Calbimonte et al., "Deriving Semantic Sensor Metadata from Raw Measurements", The 5th International Workshop on Semantic Sensor Networks, pp.33~48, 2012.

[8] L. Zhang et al., "Structure and Content Similarity for Clustering XML Documents," Web-Age Information Management, Vol. 6185, pp.116~124, 2010.

[9] M. Ko et. al., "An Integrated Processing Method for Image and Sensing Data Based on Location in Mobile Sensor Networks," The Journal of The Institute of Webcasting, Internet Television and Telecommunication, Vol. 8, No. 5, pp.65~71, 2008

[10] H. Park et. al., " A Study for Context-Awareness based on Multi-Sensor in the Smart-Clothing," The Journal of The Institute of Webcasting, Internet Television and Telecommunication, Vol. 13, No. 3, pp.71~78, 2013

[11] H. Hwang, and X. Lee, "A Study of the Factors influencing User Acceptance of Social Shopping based on Social Network Service," Journal of the Korea Academia-Industrial cooperation Society, Vol. 15 No. 1, pp.61~71, 2014.

[12] J. Chang, "Efficient Retrieval of Short Opinion Documents Using Learning to Rank," The Journal of The Institute of Internet, Broadcasting and Communication(JIIBC), Vol. 13 No. 4, pp.117~126, 2013.

※ 본 연구는 한성대학교 교내학술연구비 지원 과제임.

저자 소개

민 홍(정회원)

- 2004년 : 한동대학교 전산과학 졸업 (학사).
- 2011년 : 서울대학교 컴퓨터공학부 졸업(박사).
- 2013년 ~ 현재 : 호서대학교 컴퓨터 정보공학부 조교수.

<주관심분야 : 운영체제, 무선 센서 네트워크, 스마트폰 센싱, 임베디드 시스템, 결함허용 시스템>

허 준 영(정회원)

- 1998년 : 서울대학교 컴퓨터공학과 졸업(학사).
- 2009년 : 서울대학교 컴퓨터공학부 졸업(박사).
- 2009년 ~ 현재 : 한성대학교 컴퓨터 공학과 조교수.

<주관심분야 : 운영체제, 무선 센서 네트워크, 임베디드 시스템, 기계학습>