http://dx.doi.org/10.7236/JIIBC.2014.14.1.253
JIIBC 2014-1-32

대규모 스마트폰 센싱을 위한 문서 클러스터링 기법

Document Clustering Scheme for Large-scale Smart Phone Sensing

민 홍*, 허준영**

Hong Min*, Junyoung Heo**

요 약 스마트폰에 탑재된 센서들을 사용하여 사회 조직에서 발생하는 다양한 현상들을 모니터링하는 스마트폰 센싱 분야에서 대규모 데이터 처리 및 품질 향상과 수집된 정보를 공유하기 위해 시멘틱 데이터를 관리하는 것은 중요한 이슈 중에 하나이다. 본 논문에서는 이러한 대규모 시멘틱 데이터 관리 구조에서 서버의 부하를 줄이기 위한 문서 클러스터링 기법을 제안한다. 제안된 클러스터링 기법은 헤드 노드와 멤버노드를 갖는 하이브리드 백엔드 구조에서 서버단의 부하 감소를 위해 유사한 메타데이터를 갖는 노드들로 클러스터를 구성한다. 시뮬레이션을 통해 제안 기법이 기존의 거리기반 클러스터링 기법에 비해 서버부하를 줄일 수 있다는 것을 검증 하였다.

Abstract In smartphone sensing which monitors various social phenomena of the individuals by using embedded sensors, managing metadata is one of the important issue to process large-scale data, improve the data quality, and share collected data. In this paper, we proposed a document clustering scheme for the large-scale metadata management architecture which is designed as a hybrid back-end consisting of a cluster head and member nodes to reduce the server-side overhead. we also verified that the proposed scheme is more efficient than the distance based clustering scheme in terms of the server-side overhead through simulation results.

Key Words : Smartphone sensing, Document clustering, Hybrid back-end architecture, Semantic management, Data quality

Ⅰ. 서 론

무선 센서 네트워크는 수많은 노드들이 특정 지역에 배포되어 자의적으로 네트워크를 구성하고, 탑재된 센서를 활용하여 환경 데이터를 수집하는 기반을 제공한다. 최소 비용으로 대규모의 센서 네트워크를 구성하기 위해서 센서 노드들은 낮은 성능의 프로세서, 작은 용량의 메모리와 저장 공간이 탑재된다. 또한 배터리를 통해 전원을 공급 받는 등 제한된 자원 내에서 동작해야 하는 제약을 받는다[1]. 이러한 센서 네트워크 기술은 스마트폰에 내장된 센서들을 바탕으로 환경 데이터뿐만 아니라 사회 조직 내에서 발생하는 다양한 현상들을 모니터링 할 수 있는 스마트폰 센싱 연구 분야로 확대되고 있다[2]. 스마트폰 센싱은 고성능의 프로세서와 대용량의 저장 공간이 탑재되고 안정적이고 빠른 속도 무선 통신환경을 제공한다. 하지만 개인정보를 포함한 센싱 데이터의 민감성 때문에 기존 센서 네트워크와는 다른 구조적 접근이 필요하다.

*정회원, 호서대학교, 모바일시스템공학
**정회원, 한성대학교, 컴퓨터공학과
접수일자 : 2013년 11월 6일, 수정완료 : 2014년 1월 7일
게재확정일자 : 2014년 2월 7일

Received: 6 November, 2013 / Revised: 7 January, 2014
Accepted: 7 February, 2014
*Corresponding Author: jyheo@hansung.ac.kr
Dept. of Computer Engineering, Hansung University

스마트폰 센싱에서 데이터 수집 방법은 크게 두 가지 형태로 구분할 수 있다. 참여기반 접근 방법 (Participatory approach)은 실험에 참여할 대상자들에게 특혜를 제공하면서 적극적인 참여를 권장하여 데이터를 수집하는 방법이다. 기회기반 접근 방법 (Opportunistic approach)은 의도되지 않은 기회를 통해 센싱데이터를 획득하는 방식으로 공유된 사진을 통해 위치 정보를 획득할 수 있는 것과 같은 방법으로 데이터를 수집하는 방법이다. 본 논문에서는 이러한 스마트폰 센싱의 특성을 고려하여 수집된 데이터를 공유할 수 있는 계층적 백엔드 구조 (Hierarchical back-end architecture)[3]를 설계한 사전 연구를 바탕으로 서버의 부하를 줄임으로써 시간과 비용을 줄일 수 있는 문서 클러스터링 기법을 제안한다.

계층적 백엔드 구조에서는 수집된 센싱 데이터뿐만 아니라 RDF/S (Resource Description Framework/ Schema) 기반 메타데이터 파일을 관리하는 기능도 포함하고 있기 때문에 데이터의 수집 목적, 방법, 수단 등과 같은 의미 정보 (Semantic information)를 함께 저장할 수 있다. 시간의 흐름에 따라 메타데이터의 형식과 내용이 변경되게 되는데 여러 종류의 문서가 혼재할 경우 서버에서 수행하는 의미 추론 과정 (Reasoning/ Knowledge generation)의 부하가 급격하게 증가한다. 본 논문에서는 이러한 문제를 해결하기 위해서 스마트폰들로 구성된 클러스터 생성 시 유사 문서 단위로 클러스터링 구성 및 메타데이터 병합을 진행함으로써 스마트폰 내에서의 작업 처리량을 줄일 수 있을 뿐만 아니라 서버단에서도 입력받을 메타데이터 파일의 수를 줄임으로써 처리 속도를 높일 수 있는 기법을 제안한다.

본 논문의 구성은 다음과 같다. 2장에서는 제안 기법이 전제하는 시스템의 구조를 설명하고, 3장에서는 제안한 문서 클러스터링 기법에 대해 서술한다. 4장에서는 제안한 방법의 실험 및 결과에 대해 살펴보고, 5장에서는 결론을 맺는다.

II. 계층적 백엔드 구조

1. 시스템 개요

기존 연구[3]에서는 스마트폰을 활용하여 대규모 센싱 및 메타데이터 처리를 위해 그림 1과 같은 계층적 백엔드 구조를 제안하였다.

앞에서 언급한 데이터 수집 방법에 따라 참여기반 노드에게는 클러스터 헤드의 역할을부여하고 기회기반 노드들에게는 멤버 노드의 역할을 수행할 수 있도록 클러스터를 구성한다. 클러스터 내의 데이터는 클러스터 헤드에서 병합되어 프락시로 전송된다. 프락시는 근거리 무선 통신을 제공하는 액세스 포인트를 대상으로 하며 스마트폰과 클라우드 서버 사이의 가교 역할을 수행한다. 배터리를 통해 전원을 공급받는 클러스터의 헤드의 경우 잦은 무선 통신과 많은 양의 연산을 수행 할 수 없다. 또한 스마트폰에서 클라우드 서버로 직접 자료를 업로드 할 경우 무선 통신 환경과 사용자의 상황에 따라 업로드 지연 시간이 길어지는 문제가 발생한다. 클라우드 서버에서는 수집된 센싱 데이터를 저장하고 메타데이터의 의미를 추론하여 확장 및 표준화된 메타데이터 파일을 생성하여 관리한다.

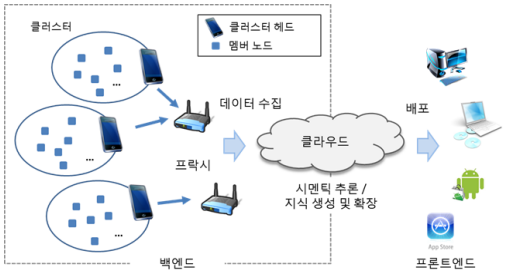

그림 1. 스마트폰 센싱을 위한 계층적 백엔드 구조
Fig. 1. Hierarchical back-end architecture for smartphone sensing

2. 메타데이터를 통한 시멘틱 관리

메타데이터는 센싱 정보를 공유하기 위한 데이터의 근원을 제시해 주며 시멘틱을 추상화하는 중요한 정보를 포함 한다. 즉 제 3자가 센싱 데이터만으로는 본래 연구의 취지를 알 수 없고, 데이터 해석에 있어서 오류를 범할 가능성이 높다. 따라서 메타데이터를 함께 저장하여 센싱데이터의 의미를 보존할 뿐만 아니라 고차원의 데이터 추론이 가능하도록 한다. 이를 위해 [3]에서는 메타데이터 관리 시스템은 메타데이터의 정의, 시멘틱 표현, 유사성에 기반을 둔 시멘틱 병합 및 확장을 지원하도록 설계하였다. 일반적으로 스마트폰 센싱에 필요한 메타데이터의 경우 크게 시간의 흐름에 따라 내용이 변하지 않는 정적 메타데이터와 내용이 변하는 동적 메타데이터로 구분할 수 있다. 이러한 메타데이터들은 그림 2에서 보는 것과 같이 관리된다.

클러스터 헤드는 멤버 노드들로부터 전송받은 센싱 데이터와 메타데이터를 바탕으로 장치 관련 메타데이터 정보를 통합하고 이를 클라우드 서버로 전송한다. 클라우드 서버는 클러스터 헤드로부터 전송 받은 장치 관련 메타데이터 정보와 연구자들의 실험 관련 메타데이터 정보를 저장하고 이를 분석하여 표준화된 메타데이터 파일을 주기적으로 클러스터 헤드에게 전송한다.

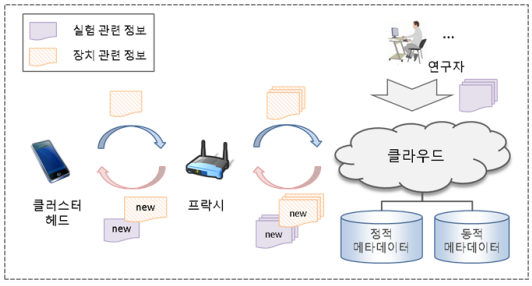

그림 2. 시멘틱 메타데이터 관리
Fig. 2. Semantic metadata management

메타데이터 정보 관리를 위해서 사용된 RDF 표준은 실세계에서 사용하는 문장의 구조와 의미를 저장하기 위해서 고안된 것으로 XML (Extensible Markup Language) 기반 문서 형식을 따른다. 문장의 주어부, 서술부, 목적부를 서브젝트(subject), 서술(predicate), 오브젝트(object)로 구성된 트리플(triple)로 저장하기 때문에 서브젝트와 오브젝트 사이의 의미 관계를 파악할 수 있다.[4] 그림 3은 RDF 표준에 따라 작성된 XML 파일의 일부를 보여주는 것으로 "Book1은 홍길동을 저자로 한다." 라는 의미를 서브젝트 (Book1), 서술 (author), 오브젝트 (Gildong Hong) 트리플로 표현한다.

```
<rdf:Description rdf:about="Book1">
    <uni:author>Gildong Hong<uni:author>
</rdf:Description>
```

그림 3. RDF 트리플 예제
Fig. 3. RDF triple example

Ⅲ. 문서 클러스터링 기법

1. 제안 시스템 개요

문서 클러스터링은 데이터 마이닝 (data mining) 분야에서 주로 연구되고 있으며 수집한 문서들을 요약하고 유사도가 높은 문서들을 종류별로 분류하는 작업을 의미한다[5]. 본 논문에서는 이러한 문서 클러스터링 기법을 활용하여 클러스터를 구성함으로써 계층적 백엔드 구조에서 메타데이터를 관리하는 클러스터 헤드와 클라우드 서버의 부하를 줄이는 기법을 제안한다.

기존의 무선 센서 네트워크에서도 센싱 데이터 수집에 필요한 부하를 줄이기 위해서 문서 클러스터링 기법을 적용한 연구들이 진행되었다. S. Manisekaran은 클러스터링 구성 시에 유사한 데이터를 수집하고 있는 노드들로 클러스터를 구성하여 싱크 노드(sink node)로 전송되는 메시지의 개수를 줄이는 기법을 제안했다[6]. K. Khanna은 파티클군집최적화(PSO: Particle Swarm Optimization) 기법을 적용하여 수집된 데이터를 2차원 공간상에 맵핑하고 임의의 지점들에서 동시적으로 클러스터를 확장해가는 기법을 서버에 적용함으로써 저장해야 할 데이터의 양을 줄이는 기법을 제안했다[7]. 그러나 앞서 언급한 연구들은 센서 노드들의 통신 거리 제약 때문에 지역적인 범위에서 중복된 데이터의 제거는 가능하지만 네트워크 전반에서 유사 데이터를 바탕으로 클러스터를 구성할 수 없다는 단점이 있다.

본 논문에서는 기존 무선 센서 네트워크에서 사용하는 거리기반 기법의 한계를 극복하고 메타데이터 관리에 필요한 부하를 줄이기 위해서 클러스터를 구성하는 단계에서 메타데이터 파일을 분류하고 유사한 메타데이터를 가지고 있는 스마트폰들로 클러스터를 구성함으로써 메타데이터 관리에 필요한 시간과 비용을 줄인다. 그림 3은 기존의 거리기반 클러스터링 구성 기법과 제안된 문서 클러스터링 기법과의 차이를 보여준다. 그림 4과 같이 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.

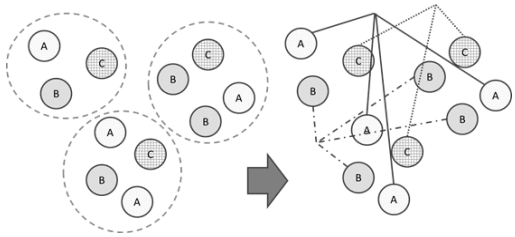

그림 4. 거리기반 클러스터링과 문서 클러스터링 기법 비교
Fig. 4. Comparing distance based clustering to document clustering

2. 클러스터 구성 과정

RDF/S 기반 메타데이터 파일이 혼재되어 있는 상황하에서 서버 부하를 줄일 수 있는 효율적인 클러스터는 그림 5와 같은 과정을 통해 구성된다.

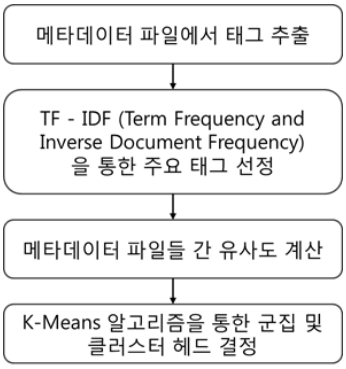

그림 5. 클러스터 구성 과정
Fig. 5. Cluster composition process

먼저 메타데이터 파일에서 중요한 의미를 가지고 있는 태그들만 파일에서 추출하여 벡터(Vector)에 저장한다. 벡터에 저장된 태그들은 각각의 빈도수에 따라 중요도를 평가받게 되는데 TF-IDF (Term Frequency and Inverse Document Frequency) 기법을 사용한다. 본 논문에서는 구현의 단순화를 위해서 다음과 같은 수식 (1)을 통해 TF-IDF 값을 도출하였다.

$$TF-IDF = tf \times idf, \quad (1)$$
$$tf = \frac{\text{특정 태그가 문서에서 발견된 횟수}}{\text{해당 문서에서 전체 태그의 개수}},$$
$$idf = \frac{\text{특정 태그가 발견된 문서의 개수}}{\text{전체 문서의 개수}}$$

각각의 태그별로 계산된 TF-IDF는 내림차순으로 정렬하여 N개의 태그들에 대해서만 유사도 계산 단계에 활용함으로써 계산에 필요한 부하를 줄인다. 유사도는 메타데이터 파일별로 선정된 태그를 비교하여 두 파일 간에 몇 개의 태그가 일치하는지 여부로 평가한다. 수식 (2)는 메타데이터 A, B 파일 간에 선정된 태그 벡터 (V_A, V_B) 들을 바탕으로 유사도를 계산하는 방법을 보여준다.

$$S_{AB} = \sum_{i=1}^{N} C(a_i, b_i) \ , a_i \in V_A, b_i \in V_B \quad (2)$$
$$C(a_i, b_i) = \begin{cases} 1, & b_i \in V_A, a_i \in V_B \\ 0, & b_i \notin V_A \text{ or } a_i \notin V_B \end{cases}$$

마지막으로 계산된 유사도 값에 따라 높은 상관관계를 보이는 메타데이터 파일들끼리 묶어서 K개의 클러스터를 구성하고 수식 (3)에서와 같이 오차(E_k)를 계산한다. 오차는 해당 클러스터 내의 모든 메타데이터 파일을 대상 (D_k) 으로 클러스터 헤드 후보자가 저장하고 있는 메타데이터 파일에서 각 태그들의 TF-IDF 값을 저장한 벡터 ($\vec{\mu}(D_k)$)와 다른 멤버 노드의 TF-IDF 벡터 ($\vec{x}$) 사이의 차를 제곱하여 계산한다. 모든 메타데이터 파일들에 대해서 오차를 계산한 후에 이를 최소화 할 수 있는 메타데이터 파일을 클러스터 내의 표준 문서로 지정하고, 해당 파일을 저장하고 있는 노드를 클러스터 헤드로 선정한다.

$$E_k = \sum_{x \in D_k} |\vec{x} - \vec{\mu}(D_k)|^2 \quad (3)$$

Ⅳ. 실험 및 결과

본 절에서는 시뮬레이션을 통해 제안 기법과 기존 거리기반의 클러스터링 기법과의 성능을 비교하였다. 시뮬레이션에 사용된 K-means 클러스터링 알고리즘은 Kunwar가 제안한 오픈 소스 라이브러리[9]를 활용하였으며 실험의 단순화를 위해서 네트워크 통신에 대한 오버헤드는 시뮬레이션 대상에 포함하지 않았다. 표 1은 실험에 사용한 파라미터들에 대한 설명과 설정 값을 보여준다.

표 1. 실험 파라미터
Table 1. Simulation Parameters

파라미터 목록		설정 값
K	클러스터의 개수	1 ~ 10
T	전체 노드의 개수	300
N	선정된 태그의 수	상위 20%

그림 6은 기존의 거리기반 클러스터링 (Distance)과 제안된 문서 클러스터링 (Doc) 기법사이의 클러스터 헤드 선출 시간을 비교한 것이다.

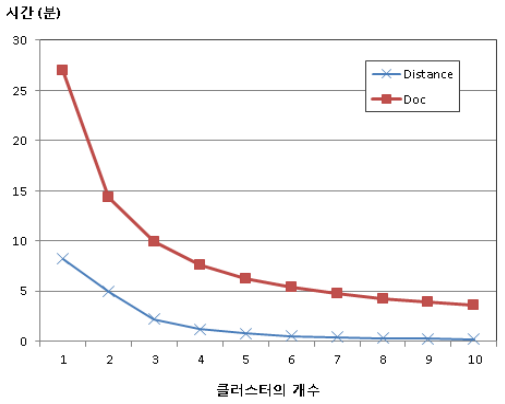

그림 6. 클러스터 헤드 선출 시간 비교
Fig. 6. Comparison of cluster head selection time

기존 기법의 경우 클러스터 내에서 모든 노드들 사이에 통신 거리를 최소화 할 수 있는 노드를 클러스터 헤드로 선정한다. 제안 기법에서는 각 노드들이 가지고 있는 메타데이터 파일에서 추출된 태그들의 유사도를 문서사이에 측정하고, 중요도를 반영하여 오차가 가장 작은 노드를 클러스터 헤드로 선정한다. 클러스터 헤드 선출에 있어서 제안 기법은 문서들 사이에 유사도를 측정하는 과정에서 기존 거리기반 클러스터링보다 많은 시간을 소비한다. 그러나 제안 기법의 경우 클러스터의 개수가 커질수록 비교 대상의 수가 줄어들기 때문에 두 기법간의 시간 차이도 줄어든다.

그림 7은 클러스터 선출 이후에 클러스터 내의 메타데이터 파일을 통합하여 표준화된 문서를 생성하는데 필요한 시간을 비교한 것이다.

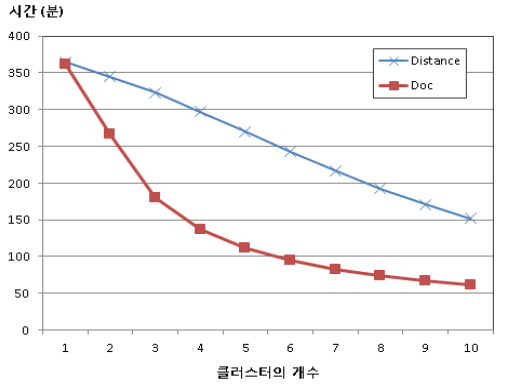

그림 7. 메타데이터파일 통합 시간 비교
Fig. 7. Comparison of metadata merge time

기존의 거리기반 클러스터링에서는 서로 다른 형식의 메타데이터 파일이 클러스터 내에 혼재되어 있어 클러스터의 개수가 증가해도 (클러스터 내의 멤버 노드의 수가 감소해도) 통합시간이 크게 단축되지 않는다. 그러나 제안 기법에서는 메타데이터 파일의 유사도에 따라 클러스터를 구성했기 때문에 클러스터의 개수가 증가함에 따라 통합 시간이 급격하게 줄어드는 것을 알 수 있다.

그림 8은 앞서 살펴본 클러스터 헤드 선출 시간과 메타데이터 통합 시간을 합산하여 클러스터링을 구성하는 전체 작업시간을 비교한 결과를 보여준다. 제안 기법의 경우 초기에 클러스터 헤드를 선출하는 과정에서 기존 거리기반 클러스터링에 비해 시간을 더 소비하지만 메타데이터 파일 통합과정에서 기존 기법에 비해 많은 시간을 단축할 수 있기 때문에 효율적이다. 또한 클러스터의 수가 증가함에 따라 기존 기법과의 수행 시간 차이가 커지기 때문에 대규모 센싱에 더 적합하다는 것을 알 수 있다.

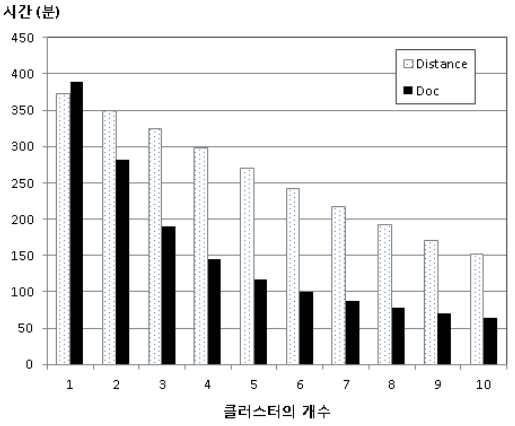

그림 8. 전체 작업시간 비교
Fig. 8. Comparison of total processing time

V. 결 론

스마트폰은 단순한 통신 기기로써의 역할뿐만 아니라 내장되어 있는 센서들을 바탕으로 사회 조직 내에서 발생하는 다양한 현상들을 모니터링 할 수 있는 응용 플랫폼으로써의 역할을 수행할 수 있다[10]. 본 논문에서는 스마트폰을 활용하여 사회 현상을 모니터링 하는 시스템 구축 시 수집된 정보를 연구자들 간에 공유할 수 있도록 메타데이터를 관리하는 시스템에서 스마트폰간의 효율

적인 클러스터 구성에 대한 기법을 제안했다. 제안 기법에서는 스마트폰에 저장된 메타데이터 파일의 유사도를 바탕으로 클러스터를 구성하기 때문에 스마트폰 사용자 사이의 거리로 클러스터를 구성하는 기존의 기법에 비해 대규모 클러스터 운용에 있어서 효율성을 보이며 시뮬레이션을 통해 이를 검증하였다.

References

[1] S. Iyengar, and R. Brooks, “Distributed Sensor Networks: Sensor Networking and Applications,” CRC Press, 2012.

[2] A. Madan, M. Cebrian, S. Moturu, K. Farrahi, and A. Pentland, "Sensing the 'Health State' of a Community," IEEE Pervasive Computing, Vol. 11, No. 4, 2012.

[3] H. Min, and P. Scheuermann, "A Hierarchical Back-end Architecture for Smartphone Sensing," Proc. of ACM RACS, 2012.

[4] G. Antoniou, P. Groth, F. Harmelen and R. Hoekstra, "A Semantic Web Primer," The MIT Press, 2012.

[5] C. Aggarwal, and C. Zhai, "A Survey of Text Clustering Algorithms," Springer, 2012.

[6] S. Manisekaran, R. Venkatesan, and G. Deivanai, “Mobile Adaptive Distributed Clustering Algorithm for Wireless Sensor Networks,” International Journal of Computer Applications, Vol. 20, No. 7, 2011.

[7] K. Khanna, and M, Yadav, “An Improved Swarm Based Approach for Efficient Document Clustering,” International Journal of Computer Trends and Technology, Vol. 4, No. 6, 2013.

[8] V. Singh, N. Tiwari, and S. Garg, "Document Clustering using K-means, Heuristic K-means and Fuzzy C-means," Proc. of IEEE CICN, 2011.

[9] K-means documents clustering library, http://www.codeproject.com/Articles/439890/Text-Documents-Clustering-using-K-Means-Algorithm

[10] M. Lee, K. Kim, K. Lee, "Design and Implementation of BAN System using ECG Sensor based on Smartphone," The Journal of The Institute of Internet, Broadcasting and Communication, Vol. 10, No. 4, 2010.

민 홍(정회원)

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

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

※ 이 논문은 2013년도 호서대학교의 재원으로 학술연구비 지원을 받아 수행된 연구임 (2013-0074)