JKIICE Journal of the Korea Institute of Information and Communication Engineering

한국정보통신학회논문지(J. Korea Inst. Inf. Commun. Eng.) Vol. 21, No. 12 : 2305~2311 Dec. 2017

XML구조를 이용한 공공 빅데이터의 선별 저장 및 시각화 방법

백봉현[1] · 하일규[2*]

A Method for Selective Storing and Visualization of Public Big Data Using XML Structure

BongHyun Back[1] · Il-Kyu Ha[2*]

[1]Argos Co. Ltd. Daegu, 41710, Korea
[2*]Department of Computer Engineering, Kyungil University, Gyeongsan, Gyeongbuk, 38428, Korea

요 약

최근들어 공공 정보화와 함께 정부기관, 지자체 및 다양한 정부산하기관에서 보유하고 있는 데이터를 공개하고 있는 추세이다. 즉, 공공기관이 업무수행의 결과물로 생성 및 수집한 다양한 전자화된 형태의 데이터를 공공데이터 포털사이트에서 개방하고 있다. 하지만 이를 사용하는 사용자는 데이터 형식의 이해와 데이터 처리 지식의 부족, 데이터에 대한 접근과 관리의 어려움, 수집 및 저장한 데이터의 이해를 위한 시각화 기술의 부족 등으로 빅데이터의 활용에 제한을 받고 있다. 따라서 본 연구에서는 다양한 공공 사이트에서 제공하는 빅데이터를 데이터셋의 URL 및 API를 사용하여 데이터 포맷에 관계없이 데이터를 수집하며, 수집된 데이터를 XML 구조를 이용하여 재가공하여 데이터베이스화하며, 데이터 융합을 통한 시각화가 가능하도록 하는 공공 빅데이터 수집, 선별 저장 및 시각화 플랫폼을 제안한다.

ABSTRACT

In recent years, there have been tries to open public data from various government agencies along with publicization of public information for the public interest. In other words, various kinds of electronic data generated and collected by the public institutions as a result of their work are opened in the public portal sites. However, users who use it are limited in their use of big data due to lack of understanding of data format, lack of data processing knowledge, difficulty in accessing and managing data, and lack of visualization data to understand collected and stored data. Therefore, in this study, we propose a big data collection, storing and visualization platform that can collect big data provided by various public sites using data set URL and API regardless of data format, re-process collected data using XML structure.

키워드 : 공공 빅데이터, 빅데이터 처리, XML, 빅데이터 시각화

Key word : Public Big Data, Big Data Processing, XML, Big Data Visualization

Received 03 October 2017, **Revised** 01 November 2017, **Accepted** 05 November 2017

* **Corresponding Author** Il-Kyu Ha(E-mail:ikha@kiu.kr, Tel:+82-53-600-5564)
Department of Computer Engineering, Kyungil University, Gyeongsan, Gyeongbuk 38428, Korea

Open Access https://doi.org/10.6109/jkiice.2017.21.12.2305 print ISSN: 2234-4772 online ISSN: 2288-4165

Ⅰ. 서 론

최근 ICT 융합산업의 발전과 함께 데이터 산업도 기업의 질적 성장, 진화된 데이터 관련 기술 및 서비스 확대, 해외 시장 진출 등을 위하여 큰 시장이 형성되면서 성장세를 보이고 있다[1,2]. 최근들어 공공 데이터의 경우 정부기관, 지자체 및 다양한 정부산하기관에서 데이터를 공개하고 있는 추세이다. 특히, 공공기관이 업무수행의 결과물로 생성 및 수집한 다양한 전자화된 형태의 데이터를 공공데이터 포털사이트에서 개방하고 있다[3,4].

공공 데이터의 종류는 과학기술, 보건의료, 사회복지, 공공행정, 재정금융 등 다양하며 API 또는 파일 형태로 지원된다. 한편으로 기업과 개인은 다양한 방법으로 빅데이터를 수집, 가공하여 다양한 목적에 활용하고자 한다. 하지만, 다음과 같은 몇 가지 문제점으로 인해 공공 빅데이터는 그 활용이 제한을 받고 있다.

첫째, 데이터 형식의 이해와 데이터 처리 지식의 문제이다. 현재 국내에서는 다양한 기관 즉, 정부, 지방자치단체, 의료분야기관, 농산물기관, 날씨기관 등 다양한 분야에서 XML, EXCEL, CSV, JSON 등 다양한 데이터 형식으로 데이터를 제공하고 있다. 그러나, 데이터 형식의 불일치로 사용자의 이해를 어렵게 하고 있다. 예를 들어 개인의 경우 평범한 EXCEL유형 이외에 XML, JSON 등의 데이터 포맷에 대한 지식과 데이터 처리에 대한 지식이 낮아 데이터 활용에 어려움이 있다.

둘째, 데이터에 대한 접근과 관리의 어려움이다. 현재 각 기관 및 지방자치단체 별로 데이터를 제공하고 있으나, 데이터 접근을 위해서는 모든 사이트별로 회원등록 및 로그인 절차가 필요하며, 수집된 데이터의 관리를 위해서는 데이터파일의 다운로드 등을 통해 자신의 컴퓨터에 저장하고 이용자 스스로 관리를 하여야 한다. 또한 데이터의 분석을 위해서는 외부의 SAS등의 통계분석 시스템과 연계해야 하며, 이에 따라 절차와 시간이 매우 소모가 된다.

셋째, 수집 및 저장한 데이터의 시각화 문제이다. 수집된 빅데이터를 적절하게 저장하고 이를 가공하여 업무에 활용하기 위해서는 다양한 형태로의 시각화하는 기술이 필요하다. 시각화된 자료는 사용자에게 보다 직관적인 정보를 제공하기 때문이다. 하지만 대부분의 일반 사용자는 빅데이터의 수집 및 저장과 함께 시각화에 대한 인식과 기술이 부족하다.

마지막으로, 대용량 데이터의 빠른 저장과 검색을 위한 데이터베이스화의 어려움이다. 사용자가 필요로 하는 대용량 데이터의 안정적인 저장과 빠른 검색을 위해서는 파일 형식이 아닌 데이터베이스가 필요하다. 일반인이 인터넷 상에서 수집한 데이터를 손쉽게 데이터베이스화하는 방법이 필요하며, 기존 RDB를 이용할 경우 비정형적인 빅데이터 처리에 어려움이 있으므로 이를 극복할 수 있는 데이터베이스화 방법이 필요하다.

이와 같이 현재의 공공 데이터의 수집과 분석 및 시각화는 정보처리 전문가가 아닌, 일반 사용자가 다루기에는 상당히 어렵고 불편하다. 다양한 출처의 데이터를 수집하고 저장하며 데이터를 재가공하여 데이터베이스하는 작업은 일반 사용자에게 매우 어려우며 데이터를 효율적으로 시각화하는 작업도 매우 어려운 실정이다.

따라서, 이와 같은 문제점을 해결하기 위해 본 연구에서는 다양한 공공 사이트에서 제공하는 빅데이터를 데이터셋의 URL 및 API를 사용하여 데이터 포맷에 관계없이 데이터를 수집하며, 수집된 데이터를 XML 구조를 이용하여 재가공하여 데이터베이스화하며, 데이터 융합을 통한 시각화가 가능하도록 하는, 공공 빅데이터 수집, 선별 저장 및 시각화 플랫폼을 제안한다.

Ⅱ. 관련 연구

현재까지 진행된 공공 또는 개인 빅데이터 플랫폼 구축 및 시각화 관련 연구들을 살펴보면 표 1과 같이 정리할 수 있다. 빅데이터 수집 및 처리에 관련된 기존 연구는 빅데이터의 구체적인 응용방법이나 응용사례가 부족하며[5-10], 구체적인 시스템의 구축방법에 대한 설명이 부족하다[11-14]. 또한 시각화와 관련된 연구[8, 10,13]들도 구체적인 시각화 방법과 시각화 응용 시스템의 제시가 미흡한 편이다.

따라서 본 연구에서는 기존의 시스템과는 다른 다음과 같은 특징을 가지는 빅데이터 수집 및 저장 그리고 시각화 플랫폼을 제안한다. 첫째, API를 이용하여 공공데이터를 자동으로 수집하고 원하는 항목을 골라 데이터베이스화할 수 있다. 둘째, XML구조를 이용하여 데이터베이스화된 데이터를 재가공하고 다른 데이터와 융합하여 다양한 정보를 얻어낼 수 있다. 셋째, 처리된

데이터를 사용자의 목적에 맞게 시각화할 수 있다.

Table. 1 Comparison of Public Big Data Processing and Visualization Studies

Study	Features	Weakness	Category
[5]	Provide an architecture for analysis and utilization	Lack of specific application methods and applications	Architecture
[6]	Provide a public big data reference model	Lack of specific application methods and applications	Reference model
[7]	Analysis of public data usage examples for spatial intelligence	Lack of adequate specific application system	Analysis
[8]	Characteristic analysis of visual representation of the app providing public data	Lack of specific applications	Analysis
[9]	A case study on big data application as National Service	Lack of adequate specific application system	Analysis
[10]	A case study on big data expression	Lack of adequate specific application system	Visualization example
[11]	Provide examples of public big data utilization for companion animals	Simplicity of purpose	Application example
[12]	Utilizing public big data to provide weather and travel information	Simplicity of purpose	Application example
[13]	Provide visual representation of public data	Restricted to visualization of voting record of parliamentary bill	Visualization example
[14]	Construction of social safety network using public big data	Simplicity of purpose	Application example

Ⅲ. 제안된 플랫폼

3.1. 제안 플랫폼 구조

본 연구에서 제안한 데이터의 수집 및 관리, 데이터베이스화, 데이터 재가공, 데이터셋의 융합과 시각화를 위한 자동화 플랫폼의 전체적인 구조는 그림 1과 같다. 제시된 플랫폼에서의 데이터 처리는 다음과 같은 절차에 의해 이루어진다.

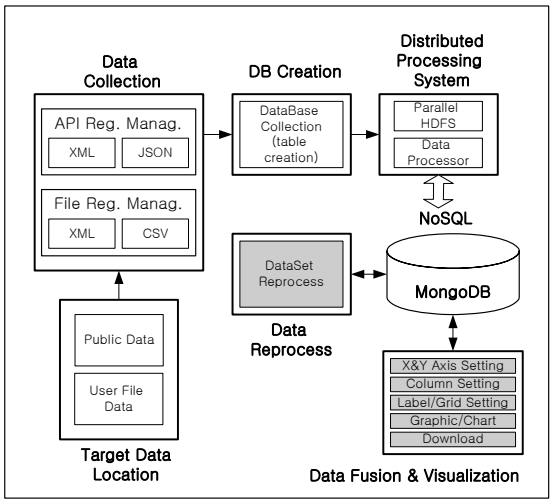

Fig. 1 Structure of the proposed platform

3.2. 데이터 수집 단계

제시된 플랫폼에서 데이터 처리의 첫 단계는 데이터 수집단계이다. 사용자가 공공 데이터를 제공하는 사이트(Target Data Location)의 URL과 Key를 이용하여 데이터를 수집(Data Collection)하는 단계이다. 이를 위하여 제안된 플랫폼에서는 사용자가 데이터의 수집과 연계를 위해 공공데이터에 대한 정보를 카테고리 별로 등록하여 데이터 리스트를 한눈에 볼 수 있도록 처리하며, URL, Key, 파일명, 제목 등의 체크를 통해 수집된 데이터의 중복성을 배제하도록 한다. 기관별로 등록된 Key는 추후의 업데이터된 데이터의 추출에 이용될 수 있다. 표 2는 XML/JSON 포맷의 데이터 제공 기관의 API등록 정보를 보여준다.

3.3. 데이터베이스 생성 단계

둘째 단계는 데이터베이스 생성(DB Creation) 단계이다. 첫째 단계에서 등록된 정보를 토대로 빅데이터 및 비정형 데이터가 저장될 수 있는 NoSQL(Mongo DB)를 활용하여 수집된 데이터(XML, JSON, EXCEL, CSV 등)의 데이터를 데이터베이스화하는 단계이다. 그림 2는 표 2의 등록된 API에 따라 수집된 XML 데이터의 예를 보여준다.

Table. 2 Required properties for API registration

Items	Function
Resource	Affiliation name as a data resource
Title	Specific title of the collected data
Specification	Specific contents of the collected data
URL	URL value of the collected data
KEY	Specific Key value of the collected data

XML	
<?xml version="1.0" encoding="UTF-8"?> <html> <head></head> <body> <rss version="2.0"> <channel> <title> KMA Mid-term Land Forcase </title> http://www.kma.go.kr/weather/forecast/mid-tem/ <description> KMA Weather Service </description>	<language> ko </language> <generator> Korea Meteorological Admin </generator> <pubdata> 2017. 09. 19. (Tue) 18:00 </pubdata> <item> <author> Korea Meteorological Admin </author> <category> Mid-term Land Forcase </category>

Fig. 2 Structure of collected XML data

수집된 XML 데이터는 데이터베이스 생성에 사용된다. 표 3은 데이터베이스 생성을 위한 요구되는 정보를 표현한 것이다. 수집할 데이터가 파일인 경우 (EXCEL, CSV 등) MongoDB 내에서 원본파일을 바이너리 형식의 chunk 단위로 데이터베이스에 저장한다. Chunk 단위로 저장된 파일 데이터를 화면에 불러온 후 데이터베이스화(테이블 생성) 한다.

Table. 3 Information required for database creation

Items	Contents
Collection name	Collection (Table) name for database creation
Description	Detailed description of the collection
Columns	Items to be database-ized among XML tags
Collection Creation	Creation of MongoDB Collection
KEY	Specific Key value of the collected data

3.4. 데이터 재처리 단계

수집된 데이터셋은 웹 사용자 인터페이스(User Interface)를 통해 사용자가 필요로 하는 데이터 항목을 선택적으로 재가공하여 문서화 또는 데이터베이스화한다. 그림 3은 수집된 XML 데이터의 구조로부터 데이터베이스를 생성하는 과정을 보여준다. 그림과 같이 API를 통해 수집된 XML의 구조로부터 필요한 항목(XML 태그)를 선별하고 이를 저장한다. 저장된 항목은 수정 또는 변경할 수 있다. 최종적으로 결정된 항목은 데이터베이스 생성을 위해 사용된다. 즉 결정된 항목으로 구성되는 테이블이 생성된다.

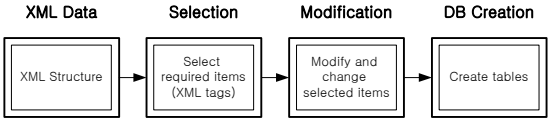

Fig. 3 The process of creating a database from an XML data

3.5. 데이터 융합 및 시각화 단계

마지막으로 분석에 필요한 복수의 데이터 셋을 결합하여 보다 가치 있는 정보를 도출하며, 결합된 정보를 시각화 하는 단계이다. 이를 위해서는 가치있는 인포그래픽을 생성하기 위해 그래프/차트/지도 등의 시각화 자료가 만들어 진다. 제안된 플랫폼에서는 데이터 시각화를 위해 6종(bar, line, area, step, pie, donut 등)의 그래프/차트를 지원하도록 한다.

3.6. 빅데이터 처리 구조

제안된 플랫폼에서는 안정적인 데이터의 저장과 대용량 빅데이터의 빠른 처리 속도를 위해 하둡(Hadoop) 기반의 분산병렬 시스템을 지원하도록 하며, NoSQL을 활용하여 데이터의 확장성과 빠른 응답을 지원하도록 한다. 그림 4는 수집된 빅데이터를 저장하고 처리하기 위한 하둡기반의 분산병렬 시스템을 보여준다. URL과 KEY값을 가지는 API를 이용하여 정보제공 기관에 접근하여 데이터를 수집하고 이를 MongoDB에 저장한다. 수집된 데이터는 종류에 따라 메인메모리 또는 하드디스크에 일시적으로 저장된다. 데이터의 처리와 가공을 거쳐 서버에 저장된다. 저장된 데이터는 재가공되어 데이터베이스를 생성하고 시각화를 위한 처리 이후에 사용자에게 제공된다. 수집된 데이터의 저장과 처리를 위해 하둡기반 분산처리시스템을 이용한다[15]. 분산처리 시스템은 안정적인 운영을 위하여 Primary server, Secondary server 그리고 Data server로 구성된다. 이는

향후 보다 다양한 빅데이터를 수집하고, 이를 활용한 다양한 데이터의 가공 및 분석에 활용될 수 있도록 안정적인 분산처리시스템으로 구축한다.

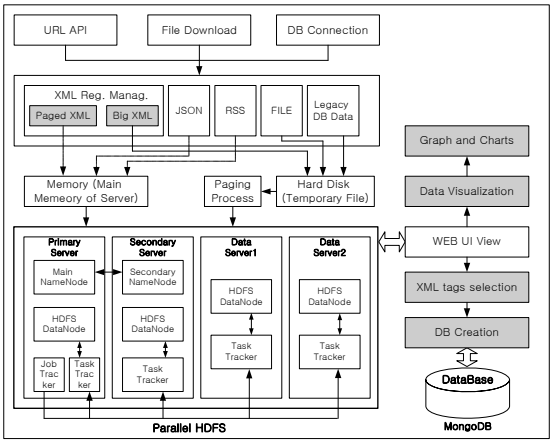

Fig. 4 Big data processing structure in the proposed platform

Ⅳ. 구 현

제시된 플랫폼을 기반으로 빅데이터를 수집하고 저장하며 이를 재가공 및 시각화하는 빅데이터 처리 시스템이 개발된다. 구현된 시스템에 사용된 빅데이터는 기상청으로부터 받은 기상 빅데이터, 건강보험심사원으로부터 받은 보건 빅데이터, 교육부에서 제공하는 교육 빅데이터 등이다. 보건 그림 5는 공공 빅데이터의 자동 수집을 위하여 수집원 기관의 URL과 API KEY를 등록하는 화면이다.

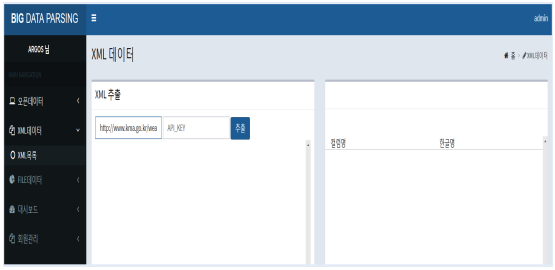

Fig. 5 Registration of URL and API Key

그림 6은 등록된 URL과 API KEY를 이용하여 추출한 XML 데이터의 예를 보여주고 있다. 예는 기상정보 저장을 위한 XML 구조이고, 도시간 날씨 비교를 위하여 XML데이터를 추출하였다.

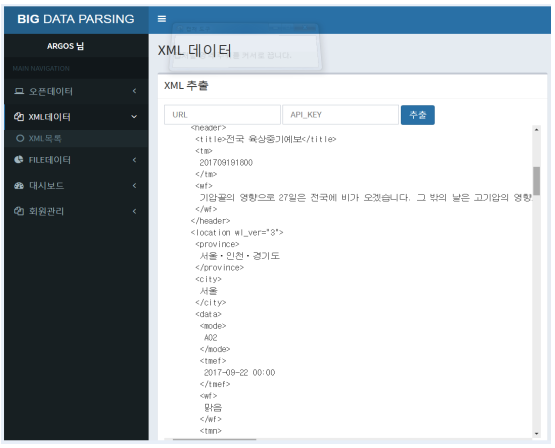

Fig. 6 Structure of extracted XML data (example of weather information)

그림 7은 추출된 XML 데이터로부터 필요한 특정 항목을 선택하여 데이터베이스를 생성하는 화면이다. 공공기관으로 제공되는 데이터는 사용자의 요구에 정확하게 일치하지 않으므로 사용자의 요구에 맞는 항목을 골라 저장할 수 있도록 데이터 재가공 기능을 부여한 것이다. 선택된 항목은 수준에 따라 표시되며 추가 또는 삭제가 가능하도록 개발된다.

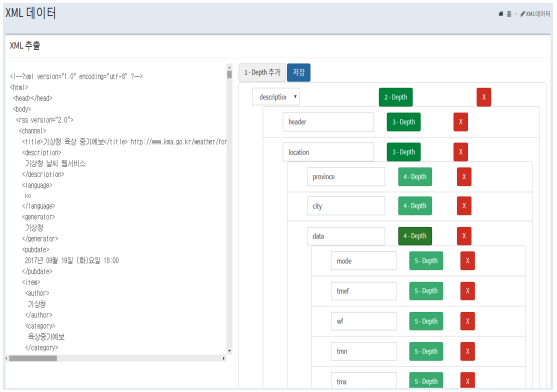

Fig. 7 Selection of specific items from the XML structure and Creation of database

그림 8은 저장된 데이터베이스 테이블의 특정한 항

목을 선택하여 이를 시각화한 예이다. 화면의 왼쪽에서 시각화를 위한 항목이 선별되고 그래프의 유형이 선택되며 이에 따라 오른쪽 화면에 시각화 자료가 표시된다. 그림에서 보여주는 예는 특정기간 동안 어느 한 도시의 최고기온과 최저기온을 비교한 것이다.

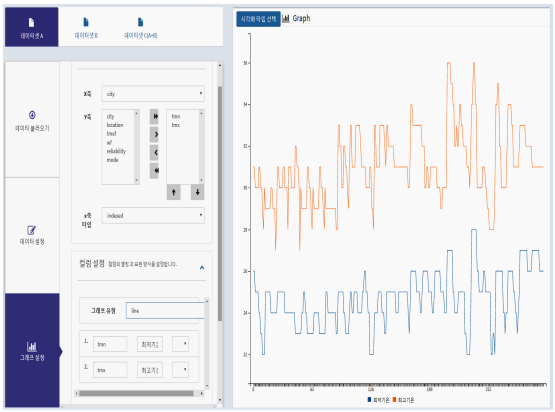

Fig. 8 Selection and visualization of specific data items

그림 9는 데이터 매칭기능을 이용하여 시각화된 데이터를 재가공한 예를 보여준다. 예는 일정기간 동안 두 개 도시의 최고 기온과 최저 기온을 비교한 것이다. 두 데이터셋의 타입 등을 매칭하여 다른 형태의 그래프로 표현한다.

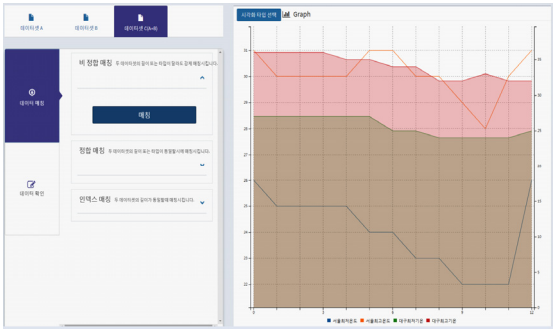

Fig. 9 Visualization using matching function

Ⅴ. 결 론

본 연구에서는 공공기관에서 제공하는 데이터를 수집하고 저장하며 이를 재가공하고 시각화하는 자동화 플랫폼을 제안하였다. 제안한 플랫폼은 다음과 같은 분야에서의 효과를 기대할 수 있을 것이다.

첫째, 공공의 이익에 기여할 수 있다. 공공데이터를 수집 및 저장하고 이를 고급 정보화하여 다양한 수요자에게 공급함으로써 공공기관의 정책수립, 업무개선, 민간산업의 기술발전, 학계의 연구력 향상 등 다양한 공공의 이익에 기여할 수 있다.

둘째, 기술우위 확보 및 정보산업 발전에 기여할 수 있다. 빅데이터 구축 및 정보제공 자동화 플랫폼 구축에 관한 기술을 보급함으로써 대외적인 기술 경쟁력을 확보할 수 있다. 최근 4차 산업혁명과 함께 요구되는 국가의 기술경쟁력 확보에 크게 기여할 것으로 판단한다.

셋째, 이익 창출에 기여할 수 있다. 향후 제안한 플랫폼을 웹기반 또는 앱기반 시스템으로 개발하여 운영하여 이익 창출을 기대할 수 있다. 또한 기술이전 등을 통해 기술을 확산함으로써 직접 또는 간접적으로 다양한 이익을 창출할 수 있다.

개발된 시스템은 변경된 데이터의 자동수집, 가공데이터의 고급화에 한계가 있다. 향후 정보제공기관의 수를 늘리고 변경된 데이터를 자동수집하고, 이를 재가공 및 융합함으로써 다양한 고급화된 시각화 정보를 제공할 수 있도록 할 계획이다.

ACKNOWLEDGMENTS

Following are results of a study on the "Leaders in INdustry-university Cooperation+(LINC+)" Project, supported by the Ministry of Education (MOE) and the National Research Foundation of Korea(NRF).

REFERENCES

[1] X. Wu, X. Zhu, D. Wu and W. Ding, “Data mining with big data,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 1, pp. 97-107, Jan. 2014.

[2] C. Cheng and C. Zhang, “Data-intensive applications, challenges, techniques and technologies: A survey on Big Data,” *Information Sciences*. vol. 275, pp. 314-347, Aug. 2014.

[3] G. Kim, S. Trimi and J. Chung, “Big-data applications in the government sector,” *Communications of the ACM*, vol. 57, no. 3, pp. 78-85, Mar. 2014.

[4] Y. Kang, K. Kim, M. Han, J. Kim, “A Study on the Business Strategies based on Big Data Analysis,” *Asia-pacific Journal of Multimedia Services Convergent with Art, Humanities, and Sociology*, vol.5, no.5, pp.145-152, 2015.

[5] S. Bang, H. Ha and C. Kim, “A Study on Big Data-based Software Architecture Design for Utilizing Public Open Data,” *Journal of Advanced Information Technology and Convergence*. vol. 13, no. 10, pp. 99-107, Oct. 2015.

[6] Y. Chol, “A Study on the Development of Public Big Data Platform Reference Model,” *Journal of Information Technology and Architecture*. vol. 12, no. 4, pp. 495-503, April 2015.

[7] M. Kim and D. Choi, “An Analysis of the Public Data for Making the Ambient Intelligent Service,” *JOURNAL OF DIGITAL CONVERGENCE*, vol. 12, no. 12, pp. 313-321, Dec. 2014.

[8] Y. Moon and J. Jung, “The visualization of application utilizing public data,” *The Treatise on The Plastic Media*. vol. 18, no. 1, pp. 63-76, Jan. 2015.

[9] K. Lee, G. Nam, J. Sim, K. Cho and W. Ryu, “Construction of Knowledge Base for The Utilization of Big Data in Public Domain,” *Communications of the Korean Institute of Information Scientists and Engineer*. vol. 30, no. 6, pp. 40-46, 2012.

[10] S. Ju, J. Jeong and G. Ryu, “Big Data Technology Trends Big Data Visualization and Public Data Visualization Examples,” *It's Smart Media*. vol. 2, no. 3, pp. 37-43, 2013.

[11] J. Lee and G. Oh, “A study of application development case built on public section big data - PET 119 IN SUWON,” *Journal of The Korea Big Data Service Society*. vol. 2, no. 1, pp. 19-24, 2015.

[12] S. Lee and S. Shin, “Design of Health Warning Model on the Basis of CRM by use of Health Big Data,” *Journal of the Korea Institute of Information and Communication Engineering*. vol. 20, no. 8, pp. 1460-1465, 2016.

[13] M. Lee and B. On, “An Example of Public Data Visualization based on the Big Data Approach,” *Information & communications magazine*. vol. 29, no. 11, pp. 36-42, 2012.

[14] S. Lee, J. Jung, G. Cha, G. Son, S. Kim and J. Kim, “Social safety net system through big data analysis of public data,” Journal of Satellite, *Information and Communications*. vol. 10, no. 4, pp. 77-82, 2015.

[15] I. Ha, B. Bak and B. Ahn, “MapReduce functions to analyze sentiment information from social big data,” *International Journal of Distributed Sensor Networks-Special issue on Advanced Big Data Management and Analytics for Ubiquitous Sensors*.vol. 2015. Jan. 2015.

백봉현(BongHyun Back)

1999.2 동국대학교 전자계산학 학사
2002.2 영남대학교 컴퓨터공학과 석사
2014.8 영남대학교 컴퓨터공학과 박사
2005~2009 株式会社 セキュアヴェイル (일본) PL
2009~2010 ㈜ 유비엔 부장
2011~현재 ㈜ 아르고스 대표이사
※관심분야 : 빅데이터분석및처리, 개인정보보호 응용 등

하일규(Il-Kyu Ha)

1992.2 영남대학교 전산공학과 학사
2003.8 영남대학교 컴퓨터공학과 박사
1992~1995 증권감독원 전산업무실 5급사무원
2002~2007, 2008~2015 영남대학교 컴퓨터공학과 강사, 객원교수, 선임연구원
2015~현재 경일대학교 컴퓨터공학과 조교수
※관심분야 : 무선센서네트워크, Body Area Networks, UAV응용, 빅데이터분석및처리 등