The Journal of The Institute of Internet, Broadcasting and Communication (IIBC)
Vol. 16, No. 6, pp.225-230, Dec. 31, 2016. pISSN 2289-0238, eISSN 2289-0246

https://doi.org/10.7236/JIIBC.2016.16.6.225
JIIBC 2016-6-29

신뢰성 빅데이터 플렛폼의 연구

Study of Trust Bigdata Platform

김정준*, 곽광진**, 이돈희***, 이용수****

Jeong-Joon Kim*, Kwang-Jin Kwak, Don-Hee Lee***, Yong-Soo Lee****

요 약 최근 네트워크와 인터넷의 발전으로 웹상에 대용량의 데이터가 생겨났으며, 이를 처리하기 위해 빅데이터 기술이라는 패러다임이 생겨났다. 빅데이터 기술은 기존의 정형 데이터뿐만 아니라 소셜 데이터 등 다양한 비정형 데이터를 이용해 다각적이고 정확한 분석을 목표로 연구되고 있다. 그러나 소셜 데이터는 전문성과 객관성을 가지고 있다고 보기는 힘들고 정보의 조작 및 은폐, 왜곡 등의 문제성이 제기되고 있다. 따라서, 본 논문에서는 신뢰성 빅데이터 플랫폼에 대하여 제안하며, 세부 관리자와 모듈에 대하여 설명한다. 본 논문에서 제안하는 신뢰성 빅데이터 플랫폼은 데이터 정제 관리자, 데이터 분석 관리자, 상호 신뢰 관리자, 시각화 관리자, 검색 관리자로 구성되어진다.

Abstract Recently, Web has arisen large amount of data that to the development of the network and the Internet. In order to process it appeared that Big Data technology. Big Data technologies have been studied aiming a multifaceted and accurate analysis using existing regular data and a variety of data social data. But social data does not have the expertise and objectivity. And such manipulation and concealment and distortion of information have been raised troubling. Thus, this paper proposes for trust big data platform and will be described in detail. The big data platform proposed in this paper consists of data refiner, Data Analyzer, co-truster, visualizer, searcher, etc.

Key Words : Big Data, System Architecture, Trust System, Social Data, Analytics

Ⅰ. 서 론

1990년 이후 네트워크와 인터넷의 발전을 통해 데이터가 네트워크를 통해 보급되게 되었으며. 2000년 이후 Web의 패러다임의 변화로 다양한 미디어 보급과 편리해진 웹을 통해 많은 데이터가 웹으로 이동하였다. 웹상의 데이터가 급증함에 따라 검색 회사들을 필두로 대용량 데이터에 대한 연구가 활발해지게 되었다. 특히 2000년대 중반 구글의 Big Table과 아마존의 Dynamo는 병렬처리를 위한 모델링을 제시함에 따라 이 연구에 패러다임을 제시하였다, 한편, 오픈 소스 진영에서는 Apache의 Hadoop을 기반으로 한 많은 오픈 소스가 개발 되었다. 이에 따라 기업들에서도 Hadoop을 기반으로 대형 병렬처리 플랫폼에 다양한 저장소, 검색, 분석, 시각화 등 다양한 소프트웨어를 접목시킨 빅데이터 플랫폼 사용하게 되었으며, 이를 통해 페이스북, 트위터, 인스타그램 등 소셜 네트워크 서비스가 발전하게 되었다. 그리고 이들 기업에서는 소셜 네트워크 분석을 통해 또 다른 새로운 가

*정회원, 한국산업기술대학교 컴퓨터공학과
**준회원, 한국산업기술대학교 컴퓨터공학과
***정회원, SK C&C
****정회원, 여주대학교 컴퓨터정보과
접수일자 : 2016년 9월 8일, 수정완료 : 2016년 10월 8일
게재확정일자 : 2016년 12월 9일

Received: 8 September, 2016 / Revised: 8 October, 2016 / Accepted: 9 December, 2016
**Corresponding Author: plastic4185@gmail.com
Dept. of Computer Science, Korea Polytechnic University, Korea

치를 창출하고 있다[4].

그러나, 소셜 데이터는 전문성과 객관성이 의심됨으로 맹신할 수는 없으며, 정치, 문화, 가치관에 따라 여론이 형성되며, 여론 조작을 하는 경우도 종종 발견된다. 따라서 본 논문에서는 신뢰성 있는 빅데이터 플랫폼을 제안하고자 한다.

II. 관련 연구

1. 빅데이터 분석

빅데이터 분석[5,6]은 다음 그림과 같이 5단계로 진행되어진다.

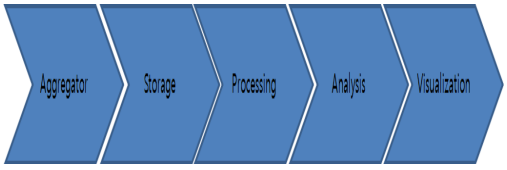

그림 1. 빅데이터 분석 체계
Fig. 1. Bifdata Analysis Process

먼저 필요한 정보 또는 관련성 있는 정보를 모으고, 이를 저장소에 저장한 후 이를 가공하여 분석 가능한 형태로 변환한다. 가공된 정보를 바탕으로 분석을 수행하고 시각화를 통해 사용자에게 필요한 정보를 제공한다.

단순한 과정이지만 거대한 데이터를 위와 같은 흐름으로 진행하기 위해서는 고도의 기술이 요구된다. 수집 단계에서는 필요한 정보와 불필요한 정보를 판단할 수 있는 능력이 요구되고, 저장단계에서는 거대한 데이터를 저장할 수 있는 공간과 이를 빠르게 찾을 수 있는 인프라와 기술이 요구된다. 가공단계에서는 요구사항에 맞추어 분석할 데이터를 요청하고 이를 분석 가능한 형태로 변환해주어야 한다. 분석단계에서는 가공된 데이터를 토대로 예측, 결정, 통계 등 다양한 결과를 제시할 수 있어야 하며, 시각화 단계에서는 사용자가 쉽게 알아볼 수 있도록 하여야 한다.

2. Hadoop ECO System

하둡은 아파치 재단의 프로젝트로 대량의 데이터를 처리할 수 있는 소프트웨어 프레임워크이며, 아파치 재단의 하둡 서브프로젝트들을 통해 다양한 빅데이터 비즈니스 도구들을 제공하고 있다. 하둡 프레임워크와 수집, 저장, 가공, 분석, 시각화를 도와주는 하둡 서브프로젝트들을 통해 하둡 에코 시스템[2]을 구축할 수 있다.

다음 그림은 하둡 에코 시스템의 구조이다.

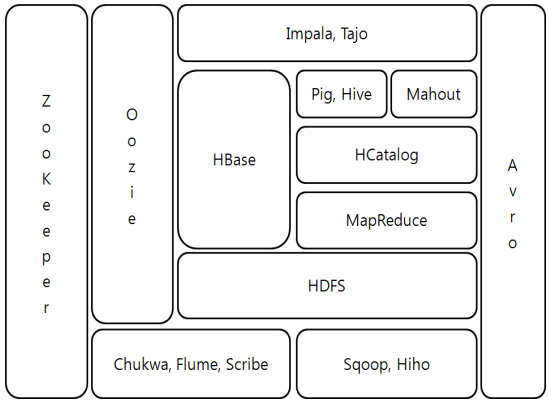

그림 2. 하둡 에코 시스템
Fig. 2. Hadoop Eco System

ZooKeeper는 분산 관리자로 장비 트래픽의 로드밸런싱과 Available을 보장하여 주며, Oozie는 워크플로우 관리자로 하둡안의 작업들의 관리를 담당한다. HBase는 HDFS 기반의 컬럼 기반 데이터베이스이며, Pig와 Hive는 SQL과 유사한 형태를 보이지만 HDFS에 저장되어진 데이터를 컨트롤한다는 점이 다르다. Mahout은 분석툴로 분류, 추천, 필터링, 군집화, 회귀분석, 패턴 마이닝 등 다양한 분석기능을 제공하며, HCatalog는 데이터 관리자로, 특정 모듈에서 생성된 데이터를 다른 모듈에서 사용할 수 있도록 한다. Avro는 원격 프로시져 콜과 데이터 직렬화를 지원하며, Chukwa와 Flume은 생성된 데이터를 HDFS에 안정적으로 저장할 수 있도록 한다. Scribe는 Chukwa와 Flume과 다르게 HDFS가 아닌 메인서버로 데이터 전송하며, JNI(Java Native Interface)를 사용하여야 HDFS에 저장이 가능하다. Sqoop은 HDFS, RDBMS, DW 등 다양한 저장소에 데이터를 전송하는 역할을 하며, Impala와 Tajo는 RDBMS와 같이 SQL을 통해 데이터를 조회, 저장하는 역할을 한다.

3. Bigdata Platform Design and Implementation Model

Bigdata Platform Design and Implementation Model[3]은 하둡 에코 시스템과 NoSQL, 각종 Open

Source들을 이용하여 병렬로 데이터를 처리할 수 있도록 확장 구성한 시스템이다.

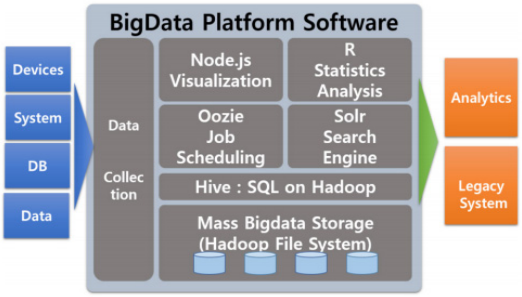

그림 3. 빅데이터 플랫폼 소프트웨어 아키텍처
Fig. 3. Bigdata platform S/W architecture

그림 3에서 보는 바와 같이 제시한 소프트웨어 아키텍처는 각종 장치와 데이터 베이스 등에서 발생한 정보를 수집하여 하둡 에코 시스템과 오픈 소스 등을 이용하여 분석 및 정책 시스템에 반영한다.

이 아키텍처에에서는 빅데이터 스토리지로 하둡의 HDFS를 사용하였고 하둡의 Hive를 이용하여 HDFS와 다른 시스템을 연결한다. 오픈 소스의 솔라(Solr), R, Node.js 는 각각 검색, 분석, 시각화를 담당하고 하둡의 우지는 각 모듈의 스케쥴링을 담당한다. 이 이외에도 수집 모듈에서는 하둡의 Flume, Sqoop을 사용하고 분석에서는 Mahout이 사용된다. 스트리밍 데이터 처리에는 Storm, S4, Spark 등을 사용하였다.

4. The Stratosphere Platform for Big Data Analytics

Stratosphere platform[1]은 매우 큰 규모의 분석 응용 프로그램으로서 병렬화 및 최적화 반복적 프로그래밍을 지원하는 확장 플랫폼이다. 데이터 웨어 하우징, 정보 추출 및 통합, 데이터 정리, 그래프 분석 및 통계 분석 어플리케이션을 포함하고 있다.

그림 4는 Stratosphere Platform의 아키텍처이며, 아래부터 저장소, 클라우드 플랫폼, 잡 관리자, 병렬 프로그램 관리자, 스크립트 관리자로 구성된다.

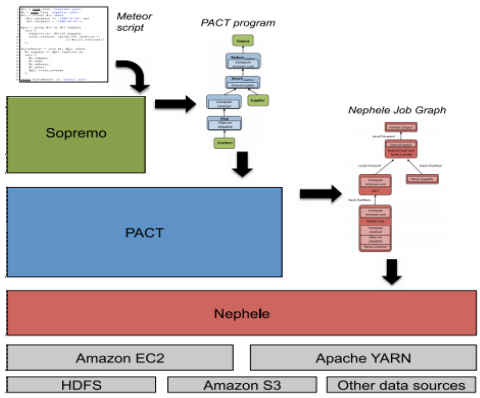

그림 4. Stratoshphere Platform 시스템 아키텍처
Fig. 4. Stratosphere Platform System Architecture

그림 4에서 보는 바와 같이 HDFS, Amazon S3, 기타 데이터 베이스를 포함한 데이터 웨어하우스들을 Amazon EC2, Apache YARN과 같은 클라우드 플랫폼으로 연동되고, Nephele라는 잡 플래너를 만들어 병렬처리의 우선 순위를 선정한다. PACT(PArallelization ConTract)는 병렬처리를 위한 과정으로 MapReduce를 기반으로 처리된다. Sopremo는 요청된 스크립트를 바탕으로 데이터를 분석하여 중복 제거 및 데이터 정렬 등의 과정을 통해 데이터를 결과를 도출한다.

Stratosphere Platform은 Meteor 스크립트라는 입력과 출력 값을 지정하는 스크립트를 작성하여 찾을 데이터와 결과에 대한 명세서를 작성하여 Sopremo에 전달한다.

그림 5. Sopremo
Fig. 5. Sopremo

그림 5의 Sopermo는 Meteor 스크립트에서 요구한 결과에 맞추어 데이터를 요청하고 중복 제거와 문장에서 주요 내용을 추출해서 결론을 작성한다.

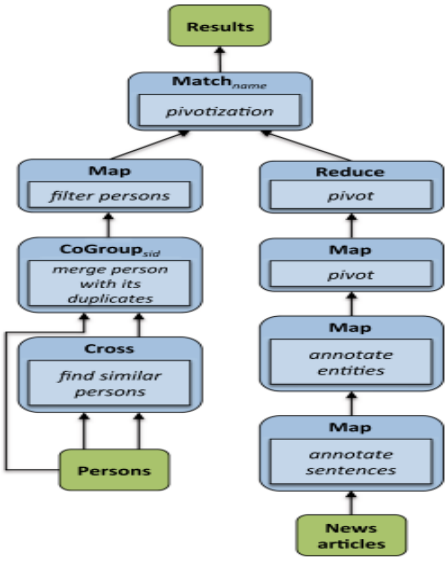

그림 6. PACT
Fig. 6. PACT

그림 6의 PACT는 Map과 Reduce 단계와 상호 그룹 연산 등을 통해 필요한 데이터들을 정제한다.

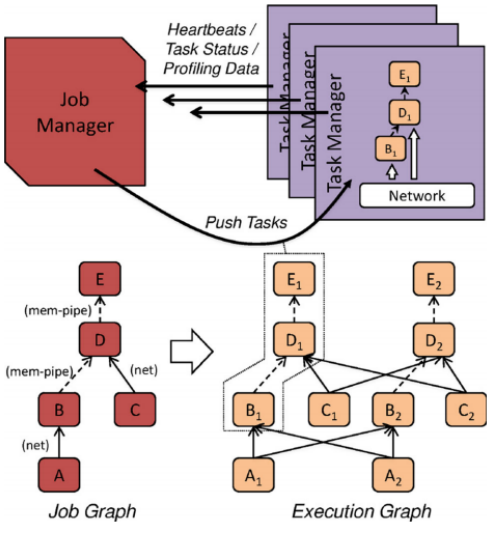

그림 7. Nephele
Fig. 7. Nephele

그림 7의 Job 관리자인 Nephele는 작업관리자로부터 실행할 작업들을 전달 받고 이를 그래프로 변경한 후 이 그래프를 토대로 실행 그래프를 생성하고 이를 다시 작업관리자에 반영시킴으로써 병렬작업의 효율성을 높였다.

Ⅲ. 시스템 설계

관련 연구에서는 제시되었던 빅데이터 플랫폼에 대하여 설명하였다. 기존 연구들에서는 다양한 매체를 통해 정보를 수집하여 이를 분석하기 위한 과정을 제시하였다. 그러나 수집한 데이터에 대한 검증이 부족하다. 본 논문에서 제안하는 시스템은 이러한 부분은 보충하고자 한다. 본 장에서는 신뢰성 빅데이터 플랫폼의 설계에 대해 설명한다.

1. System Architecture

본 논문에서 제안하는 신뢰성 빅데이터 플랫폼은 다음 그림 8과 같은 시스템 아키텍처로 되어 있다.

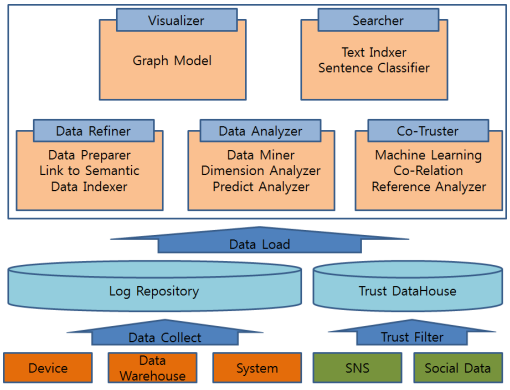

그림 8. 신뢰성 빅데이터 플랫폼 시스템 아키텍처
Fig. 8. Trust BigData Platform System Architecture

그림 8에서 보는 바와 같이 신뢰성 빅데이터 플랫폼 시스템 아키텍처는 데이터 수집에서 정제, 분석, 시각화에 이르는 빅데이터 분석 절차를 아우른다.

신뢰성 빅데이터 플랫폼의 데이터 수집은 크게 두가지로 볼 수 있다. 분석 목적을 위한 표본 데이터로 장치, 데이터 웨어하우스, 시스템에서 발생한 로그를 기반으로 하는 데이터와 이 데이터 분석에 신뢰도를 얻기 위해 트위터, 페이스북과 같은 소셜네트워크 데이터와 블로그, 신문기사, 커뮤니티 글 등을 이용한다. 이 과정에서 장치, 데이터 웨어하우스, 시스템의 로그는 인위적으로 발생한 데이터가 아니므로 신뢰할 수 있는 데이터로 볼 수 있지만 소셜 네트워크와 소셜 데이터는 주관적인 정보이므로 이에 대한 필터가 필요하다. 따라서, 두 종류의 데이터는 분리해서 처리할 필요성이 있다.

데이터 분석 및 처리에는 데이터 정제 관리자, 데이터 분석 관리자, 상호 신뢰 관리자, 시각화 관리자, 검색 관리자로 구성되어진다.

데이터 정제 관리자는 다음 그림 9와 같다.

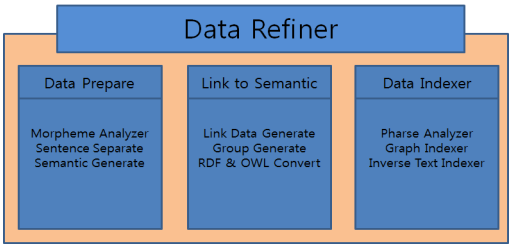

그림 9. 데이터 정제 관리자
Fig. 9. Data Refiner

그림 9에서 보는 바와 같이 데이터 정제 관리자는 데이터 전처리 모듈, 시맨틱 링크 모듈과 데이터 인덱서 모듈로 구성되어진다. 데이터 전처리 모듈은 형태소 관리 및 문장 분리, 시맨틱 생성 역할을 한다. 시맨틱 링크 모듈은 시맨틱 데이터의 연결생성 및 그룹생성, 시맨틱 데이터 연결을 위한 RDF와 OWL 포맷의 데이터 변환을 담당하며, 데이터 인덱서 모듈은 생성된 시맨틱 데이터를 그래프형 인덱스와 역텍스트 인덱스로 구성하여 빠르게 데이터를 찾을 수 있도록 도와준다.

데이터 분석 관리자는 다음 그림 10과 같다.

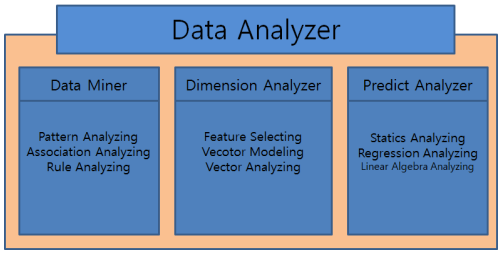

그림 10. 데이터 분석 관리자
Fig. 10. Data Analyzer

그림 10에서 보는 바와 같이 데이터 분석 관리자는 데이터 마이너 모듈, 관계 분석 모듈과 예측 분석 모듈로 구성되어진다. 데이터 마이너 모듈은 패턴 분석, 연결 분석, 규칙 분석을 수행한다. 차원 분석 모듈은 차원에 표현할 속성을 정의 및 벡터의 표현 및 그 분석을 수행하며, 예측 분석 모듈은 선형 분석, 회귀 분석, 선형 대수 분석 등을 통해 미래 예측을 수행한다.

상호 신뢰 관리자는 다음 그림 11과 같다.

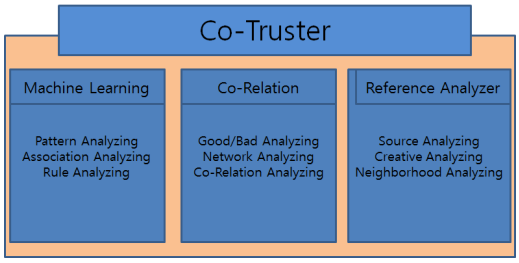

그림 11. 상호 신뢰 관리자
Fig. 11. Co-Truster

그림 11에서 보는 바와 같이 상호 신뢰 관리자는 기계 학습 모듈, 상호 분석 모듈과 참조 분석 모듈로 구성되어진다. 기계 학습 모듈은 일반화와 표현의 두 가지 기능을 수행함으로 개체에 대한 패턴과 규칙을 분석한다. 상호 관계 모듈은 객체에 대한 긍정과 부정 분석과 객체 간의 관계성을 분석하여 이에 대한 상호 연관성을 분석하며, 참조 분석 모듈은 객체의 생성처, 저작자와 저작자의 이웃들 간의 관계를 분석한다.

시각화 관리자와 검색 관리자는 그림 12와 같다.

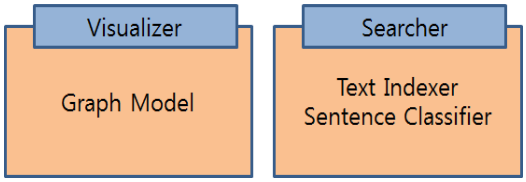

그림 12. 시각화 관리자와 검색 관리자
Fig. 12. Visualizer & Searcher

그림 12에서 보는 바와 같이 시각화 관리자는 그래프 모델을 이용해 시각화를 지원하고 검색 관리자는 텍스트 인덱스 모듈과 문장 분류 모듈을 통해 효율적인 검색을 지원한다.

V. 결 론

네트워크와 웹의 발전으로 많은 양의 데이터가 인터넷 환경으로 이주하였고, 많은 양의 데이터를 처리하기 위한 기술로 빅데이터가 주목되었다. 빅데이터 기술을 바탕으로 생성된 서비스(Social Data)들은 다시 빅데이터와 접목되면서 새로운 부가가치를 만들고 있다. 그러

나 소셜 데이터는 전문성과 객관성을 신뢰할 수 없음으로 이를 무분별하게 적용할 수는 없을 것이다. 따라서, 본 논문에서는 빅데이터 플랫폼에 소셜 데이터의 출처, 소셜 네트워크 내의 인맥관계 등을 분석하여 성향을 파악하는게 중요하다고 볼 수 있다.

본 논문에서는 이에 대한 시스템 설계를 제안하였으며 이를 구현하여 향후 성능 및 신뢰성에 대한 검증이 필요하다. 또한, 전체 시스템을 구현하기 위해서 오픈 소스에 대한 조사와 검증이 필요하다.

References

[1] Alexandrov, A., Bergmann, R., Ewen, S., and Naumann, F., “The Stratoshpare Platform Big Data Analytics.” Journal of VLDB, 2014, Vol.23, No.6, pp.939-964.

[2] Landset, S., Khoshgoftaar, T. M., Richter, A. N., and Hasanin, T., “A Survey of Open Source Tools for Machine Learning with Big Data in The Haddop Ecosystem. Journal of Big Data, 2015, Vol.2, No.24.
DOI: https://doi.org/10.1186/s40537-015-0032-1

[3] Noh, K., S.,, and Lee, D., S., “Bigdata Platform Design and Implementation Model,” Jounal. of Indian Jounal of Science and Technology, 2015. Vol.8, No.18,
DOI: https://doi.org/10.17485/ijst/2015/v8i18/75864.

[4] Shin, J., S., “SNS using Big Data Utilization Research,” Journal of IIBC, 2012, Vol.12, No.6, pp.257-265.
DOI: https://doi.org/10.7236/jiwit.2012.12.6.267

[5] Zakir, J., Seymour, T., and Berg, K., “Big Data Analytics,” Jounal of Issues in Information Systems, 2015, Vol.16, No.2, pp.81-90.

[6] Borthakur, D., and T., The Hadoop Distributed File System: Architecture and Design. Hadoop Project Website, 2007.

저자 소개

김 정 준(정회원)

• Jeong Joon Kim received his BS and MS in Computer Science at Konkuk University in 2003 and 2005, respectively. In 2010, he received his PhD in at Konkuk University. He is currently a professor at the department of Computer Science at Korea Polytechnic University. His research interests include Database Systems, BigData, Semantic Web, Geographic Information Systems (GIS) and Ubiquitous Sensor Network (USN), etc.

곽 광 진(준회원)

• Kwang Jin Kwak received his MS in Computer Science at Konkuk University in 2010 and 2015. His research interests include Database Systems, BigData, Document Clustering, Geographic Information Systems (GIS) and Data Mining, etc.

이 돈 희(정회원)

• Don Hee Lee received his BS in Computer Science at Kangweon University in 1987 and 1990, and MS in Computer Science at Yonsei University in 2002 and 2004. In 2016, he received his PhD in at Konkuk University. Currently He is working in SK. His research interests include Database Systems, Ubiquitous Sensor Network (USN), Informaton System audit, PMO, etc.

이 용 수(정회원)

• Yong-soo Lee received his MS in Computer Science at Konkuk University in 1989. In 2015, he received his PhD in Information & Control Engineering at Kwangwoon University. He is currently a professor at the Department of Computer Information at Yeoju Institute of Technology. He is the Member of the Korea Institute of Internet, Broadcasting & Communication (IIBC). His research interests include Database Systems, Data Mining, BigData, Wireless Sensor Networks and Ubiquitous Sensor Network (USN), etc.