The Journal of The Institute of Internet, Broadcasting and Communication (IIBC)
Vol. 24, No. 3, pp.171-176, Jun. 30, 2024. pISSN 2289-0238, eISSN 2289-0246

https://doi.org/10.7236/JIIBC.2024.24.3.171
JIIBC 2024-3-24

IT 관련 논문 빅데이터를 활용한 한국과 미국의 IT 동향 분석

Analysis for IT Trends in Korea and the United States using Big Data in IT-related Papers

황승연*, 장석우**

Seung-Yeon Hwang*, Seok-Woo Jang**

요 약 IT 관련 분야는 너무나도 다양하다. 2018년 현재 4차 산업혁명으로부터 이루어진 IT 혁명은 이전과는 다른 새로운 분야들의 등장을 이루어 냈을 뿐만 아니라 과거에 이미 이슈가 되었지만 파묻혀 있던 여러 분야들의 재조명 또한 이루어 냈다. 기업이나 공공기관에서는 이러한 현 상황에 맞추어 IT 동향 파악에 큰 관심을 가지고 있기에 본 논문에서도 국내의 논문들이 제공하는 키워드를 분석하는 방법으로 IT 동향을 파악하였다. 본 논문에서는 기존에 이루어졌던 산업 동향 분석이나 경제 분석과는 달리 직접 이루어진 IT 관련 연구에 대한 석박사들의 논문들이 제공하는 키워드를 분석하는 방법에 초점을 맞추어 더 원초적이며 직접적인 IT 동향을 파악한다. 이러한 분석은 IT 업계 쪽으로 나아가려는 학생들 혹은 이러한 학생들에게 방향을 제시하려는 교육자들에게 IT 기술을 연구한 학술적인 논문들을 분석한 데이터를 토대로 앞으로의 비전을 예측하고 제시한다.

Abstract IT-related fields are very diverse. As of 2018, the IT revolution from the Fourth Industrial Revolution not only brought out the new fields that were different from the previous ones, but it also made a reexamination of various fields that had already been an issue in the past. Companies and public institutions have a great interest in understanding IT trends in this situation. Therefore, in this paper, IT trends are identified through the analyzation of keywords provided by domestic papers. Moreover, unlike previous industry trend analysis or economic analysis, this paper focuses on analyzing the keyword provided by the doctoral thesis or master's thesis about direct IT-related research, and grasps the more basic and direct IT trend. This analysis predicts and presents the vision based on the data of the analysis from the academic papers that researched in IT technology for IT related students or IT related educators.

Key Words : Big data, Crawling, IT trends analysis, Pig, R

*준회원, 안양대학교 컴퓨터공학과
**정회원, 안양대학교 소프트웨어학과
접수일자 2024년 4월 1일, 수정완료 2024년 5월 1일
게재확정일자 2024년 6월 7일

Received: 1 April, 2024 / Revised: 1 May, 2024 /
Accepted: 7 June, 2024
**Corresponding Author: swjang@anyang.ac.kr
Dept. of Software at Anyang University, Korea

Ⅰ. 서 론

4차 산업혁명의 도래로 전 세계적으로 소프트웨어 시장이 점차 확대되고 있다. 국내에서는 정부의 5개년 국정 운영 계획에 따라 1) 소프트웨어 기반 성장을 통한 국가 경쟁력 강화, 2) 소프트웨어 일자리 창출을 통한 소프트웨어 산업 활성화, 3) 격차 해소를 통한 우수 소프트웨어 인재 유치를 목표로 하고 있다. 그러나 국내 소프트웨어 시장은 해외 시장에 비해 그 규모가 여전히 부족한 상황이다. 이에 본 연구는 해외 소프트웨어 시장과 국내 시장을 비교 분석하여 국내 소프트웨어 시장의 미래 비전을 예측하고자 한다. 이 연구가 아직 전문 분야를 결정하지 못한 컴퓨터 공학도나 컴퓨터 공학을 전공하고자 하는 대학생들에게 큰 도움이 될 것으로 기대된다.

본 연구에서는 크롤링을 통해 국내 논문 데이터를 수집하고 미국 논문 데이터는 공공데이터를 이용하였다. 이러한 빅데이터를 hadoop 공간에 분산 저장하고 pig와 hive를 통해 처리한 후 R을 이용하여 해외와 한국의 IT 분야 논문의 동향을 분석 및 시각화한다.

Ⅱ. 관련 기술

1. 빅데이터

빅데이터란 통상적으로 사용되는 기존 데이터베이스 관리 도구의 능력(데이터 수집, 관리 및 처리)을 넘어서는 대량의 정형 또는 비정형 데이터의 집합으로부터 가치를 추출하고 결과를 분석하는 기술을 뜻한다[1]. 빅데이터에 대한 정의는 상당히 많지만, 이들 대부분은 3V로 알려진 대표적인 속성을 포함하는데 데이터의 크기(Volume)와 다양성(Variety), 그리고 수집&저장 및 처리&분석 속도(Velocity)가 그것이다.

2. Web Crawling

웹 크롤링이란 조직적, 자동화된 방법으로 월드 와일드 웹을 탐색하는 컴퓨터 프로그램인 웹 크롤러를 사용하여 웹 페이지 내 특정 형태의 정보를 수집하는 기술을 뜻한다. 웹 크롤러는 방문한 페이지의 복사본을 생성하는데 사용되며 이렇게 생성한 복사본에서 목적에 맞는 데이터를 추출하여 수집 가능하도록 한다. 웹 크롤링은 검색 엔진과 같은 여러 사이트에서 데이터의 최신 상태를 유지하기 위해 사용하기도 하며, 링크 체크나 HTML 코드 검증과 같은 웹 사이트의 자동 유지 관리 작업을 위해서 쓰이기도 한다. 또한, 크롤링을 스크래핑(Scrapping) 혹은 스파이더링(Spidering)이라고도 한다[2].

3. Apache Hadoop

하둡은 대량의 자료를 분산하여 저장하고 처리할 수 있는 대규모 컴퓨터 클러스터에서 동작하는 분산 응용 프로그램을 지원하는 프리웨어 자바 소프트웨어 프레임워크이다[3]. 하둡은 대용량의 빅데이터를 하둡 분산 파일 시스템(HDFS, Hadoop Distributed File System)에 저장하고, 맵 리듀스(Map Reduce)를 이용하여 처리한다.

4. Apache Pig

피그는 대표적으로 조인(Join) 연산과 같이 맵 리듀스에서 처리할 수 없는 부분들을 지원하는 대용량 데이터 셋을 다루기 위한 스크립트 언어이다. 실행 시에 내부적으로 맵 리듀스 작업으로 변경되나, 피그는 다중 값과 중첩된 형태를 보이는 좀 더 다양한 데이터 구조를 지원할 뿐만 아니라 데이터에 적용할 수 있는 변환 종류도 훨씬 다양하게 지원한다. 즉 프로그래밍하기 쉬우며 최적화할 수 있는 방법을 제공하고 나아가 사용자가 특수 목적을 위한 자신의 함수를 만들 수 있는 확장성을 제공한다.

5. R

R은 통계 소프트웨어 개발과 자료 분석에 널리 사용되고 있는 오픈소스 프로그래밍 언어 및 환경이다[4]. 통계 계산과 그래프 작업(시각화)을 지원하는 언어로 R에서 사용할 수 있는 수많은 통계 관련 패키지가 이미 통계학자들에 의해 구현되어있기 때문에 이 패키지들을 설치하는 식으로 쉽게 기능 확장이 가능하다. 시각화 또한 그래픽 관련 패키지를 설치하면 간단하게 다양한 그래프를 활용할 수 있으며 구글이나 네이버 지도를 불러오거나 이를 활용해 GIS 용도로 쓰는 것도 가능하다.

Ⅲ. 본 론

1. 데이터 수집

본 연구를 위해 웹과 공공데이터에서 제공하는 국내와 해외 논문 데이터를 이용한다. 국내 논문 데이터는 웹 크롤링을 이용하여 총 194,682개를 수집하였고 전처리를 통해 카테고리가 불분명한 데이터를 제외하고 약 51,000개의 데이터를 추출하였다. 해외 논문 데이터는

RISS에서 제공하는 해외 논문 공공데이터를 총 41,600개를 수집하였다.

	A	B
1	Title	entryDate
2	Proactive and reactive approaches for dynamic workloads	2017
3	Advances in deterministic, stochastic, and semistochastic quantum chemistry	2017
4	Isolation in cloud storage	2017
5	Capacity Region and Degrees of Freedom of Bidirectional Networks	2017
6	Human-Centric Debugging of Entity Matching	2017
7	Building evolvable distributed systems for dynamic data center environments	2017
8	Hidden Markov model-based homology search and gene prediction in NGS ERA	2017
9	Deep sequential and structural neural models of compositionality	2017
10	A framework for domain-driven development of personal health informatics technologies	2017
11	Integrating Exponential Dispersion Models to Latent Structures	2017
12	AIMOS: Automated Inferential Multi-Objective Optimization System	2017
13	Declarative Languages and Scalable Systems for Graph Analytics and Knowledge Discovery	2017
14	Testing Analogical Transfer in Pigeons (<i>Columba livia </i>) and Humans (<i>Homo sapiens</i>)	2017
15	On the Complexity of Intersection Non-Emptiness Problems	2017
16	Service-driven Secure Outsourced Computing	2017
17	Privacy Preserving Representation Learning using Deep Neural Networks	2017
18	The effect of complexity on the optimal rate of trait mutation	2017
19	On the Complexity of Market Equilibria and Revenue Maximization	2017
20	A novel framework for understanding atypical images	2017
21	<i>BioONT</i>: Improving knowledge organization and representation in the domain of biometric authentication	2017
22	A Comprehensive Method for Automating Test Collection Creation and Evaluation for Retrieval and Summarization Systems	2017
23	Simple Mechanisms and Behavioral Agents: Towards a Theory of Realistic Mechanism Design	2017
24	ArCuS: An Architecture for ASIC Cloud based Servers	2017
25	Decentralized Allocation of Tasks with Temporal and Precedence Constraints to a Team of Robots	2017
26	Nursing faculty experiences of virtual learning environments for teaching clinical reasoning	2017
27	A Real-Time Temporal Clustering Algorithm for Short Text, and its Applications	2017
28	Design and Optimization of Networks-on-Chip for Future Heterogeneous Systems-on-Chip	2017
29	Probabilistic data analysis - machine learning approach	2017
30	System and Analysis for Low Latency Video Processing using Microservices	2017
31	Comparative analysis of load balancing algorithms in cloud computing	2017
32	Power, Thermal, Reliability and Variability Management of Mobile Devices	2017
33	Automatic and Featureless <i>Sim</i>(3) Calibration of Planar Lidars to Egomotion Sensors	2017
34	Horizontal and vertical integration of bio-molecular data	2017

그림 1. riss 공공데이터
Fig. 1. RISS public data

2. 데이터 저장

본 논문에서 수집한 국내 및 해외 논문 데이터를 Apache hadoop을 이용하여 HDFS에 저장하였다. hadoop 환경은 가상 머신을 이용하여 구축하였고 수집한 데이터를 가상 머신과 연결하여 데이터를 마운트하였다.

3. 데이터 처리

데이터의 처리는 Pig와 R을 이용하였다. 본 연구에서 분석하고자 하는 것은 연도별로 어떠한 키워드를 포함하는 연구가 행해지고 있는지에 대한 분석이므로 국내 데이터와 해외 데이터를 연도별로 그룹화하고 관련된 키워드를 맵핑한다.

먼저, Pig에서 논문에 대한 데이터를 읽어 들여 발행연도, 키워드 열을 추출한다. 다음으로 조건문을 통해 연도 별 논문을 각 릴레이션에 나누어 저장한다. 연도별 릴레이션에 대해서 키워드 컬럼을 합병하는 작업을 수행한다. 이렇게 1차로 처리된 데이터를 R 프로그램으로 다시 한번 처리하게 된다. R에서는 중복되는 데이터의 reduce 과정을 거치게 되는데 정렬되고 처리하기 쉽게 추출된 키워드들의 카운트 연산을 하게 된다.

해외 논문 데이터의 경우 영어에서 사용되는 전치사와 같은 불용어를 제거하고, bigram을 사용하여 단어 카운트를 수행한다.

4. 분석

데이터 분석을 위해 전처리된 데이터들을 R의 table 형태로 워드 카운트를 진행한다. 국내 논문의 경우에는 해당 논문들이 제공하는 키워드를 카운트하고, 해외 논문은 논문의 제목을 단어로 구분하여 워드 카운트를 진행한다. 이렇게 만들어진 데이터를 상위 20위까지 추출하여 분석에 사용한다.

가. 국내 논문 키워드 분석

2000년부터 2017년까지 국내 논문에서 높은 빈도로 출현하는 키워드를 이용하여 5년 간격으로 동향 분석을 수행하였고 2016년의 분석 결과 대신 2017년의 분석 결과를 작성하였다.

2000년도의 경우에는 rsa, montgomery algorithm 등의 단어가 많이 등장하는 것으로 보아 인증 방법에 많은 관심이 있었던 것을 알 수 있다.

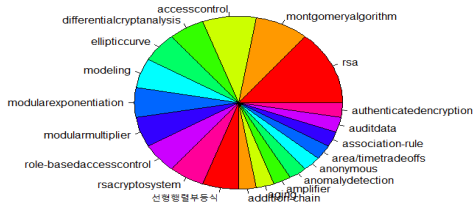

그림 2. 국내 논문의 2000년도 키워드 파이차트
Fig. 2. Keyword pie chart of domestic papers in 2000

2004년도는 2001년도에 주요 관심사였던 fuzzy 논리가 큰 비중을 차지하면서 다시 등장하게 되고, 여전히 neural network와 genetic algorithm의 키워드가 많이 등장하는 것으로 보아 인공지능에 꾸준한 관심을 보이는 것을 확인할 수 있다.

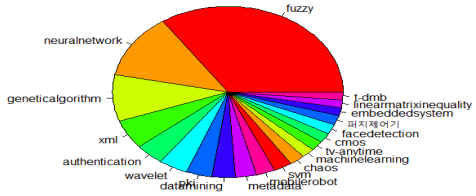

그림 3. 국내 논문의 2004년도 키워드 파이차트
Fig. 3. Keyword pie chart of domestic papers in 2004

2008년부터 2010년 까지는 계속해서 ubiquitous와 sensor network[5], ontology의 키워드가 지속적으로 등장하고, 추가적으로 authentication과 security 키워드가 등장하는 것으로 보아 IT 기기들에 대한 보안의 관심도가 증가했던 것으로 예상할 수 있다.

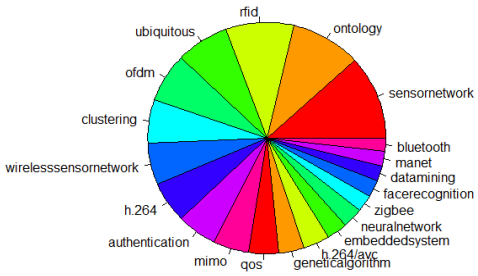

그림 4. 국내 논문의 2008년도 키워드 파이차트
Fig. 4. Keyword pie chart of domestic papers in 2008

2012년은 smart phone과 연관되는 키워드들로 lte 통신과 android 모바일 OS 키워드가 많은 비중을 차지하면서 smart phone에 발전이 이루어짐을 알 수 있다.

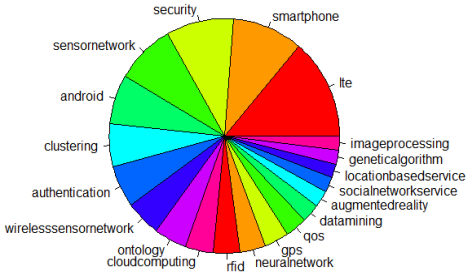

그림 5. 국내 논문의 2012년도 키워드 파이차트
Fig. 5. Keyword pie chart of domestic papers in 2012

2017년에는 2013년에 새롭게 등장한 big data[6], cloud computing, iot에 여전히 많은 연구들이 이루어지고 있고, neural network deep learning 키워드가 매우 큰 비중으로 나타나면서 인공지능에 대한 연구가 다시 이루어지고 있음을 알 수 있다.

나. 해외 논문 키워드 분석

2000년부터 2017년까지 해외 논문에서 높은 빈도로 출현하는 키워드를 이용하여 5년 간격으로 동향 분석을 수행하였고 2016년의 분석 결과 대신 2017년의 분석 결과를 작성하였다. 2012년의 분석 내용은 2008년과 큰 차이를 보이지 않아서 2013년의 분석 내용으로 작성하였다.

2000년에는 실시간 처리, 객체지향과 neural network의 키워드가 등장하게 된다. 국내 키워드보다 1년 정도 앞서서 neural network 키워드가 등장한 것이다.

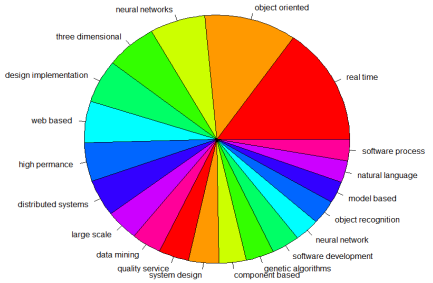

그림 6. 해외 논문의 2000년도 키워드 파이차트
Fig. 6. Keyword pie chart of international papers in 2000

2004년에는 2002년에 관심을 가졌던 hoc network와 3D에 여전히 연구가 지속되고 있고, 2003년에 높은 비중을 차지했던 data mining 키워드가 더 많은 관심을 받고 있다. 국내에서는 2002년에 data mining이라는 키워드가 적은 비중으로 처음 등장하고 해외에서는 그보다 2년 빠른 2000년도에 적은 비중으로 등장했었는데, 해외 논문에서 data mining에 큰 관심을 보이는 것을 알 수 있다.

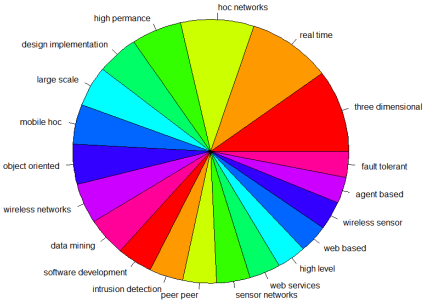

그림 7. 해외 논문의 2004년도 키워드 파이차트
Fig. 7. Keyword pie chart of international papers in 2004

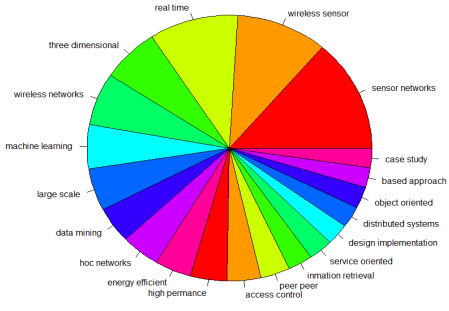

그림 8. 해외 논문의 2008년도 키워드 파이차트
Fig. 8. Keyword pie chart of international papers in 2008

2008년부터 2012년까지는 크게 다르지 않은 결과가 나타난다. 여전히 무선 네트워크에 관해 연구가 지속적으로 이루어지고, 실시간 대규모 처리들에 대해 꾸준하게 연구가 이루어진다.

2013년에는 무선과 센서 네트워크에 여전히 관심을 가지는 동시에 machine learning이라는 키워드가 높은 비중을 차지하면서 처음 등장 한다. 국내의 논문 키워드들과 비교해 보았을 때 무선 네트워크, 스마트 폰에 집중된 연구를 하고 있던 국내와는 다르게 machine learning에 관한 연구가 진행되었다.

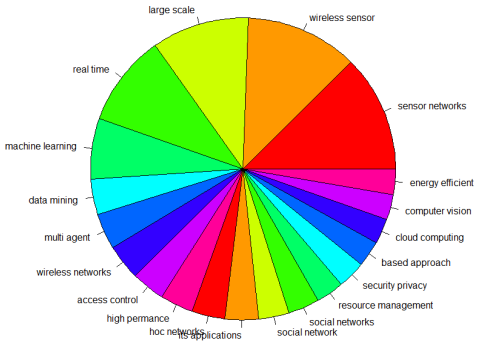

그림 9. 해외 논문의 2013년도 키워드 파이차트
Fig. 9. Keyword pie chart of international papers in 2013

2017년은 machine learning과 deep learning, big data 등의 키워드가 지속적으로 등장한다. 이는 국내 논문의 키워드 분석과 마찬가지로 인공지능, big data에 해외 또한 많은 관심을 가지고 연구하고 있다고 판단할 수 있다.

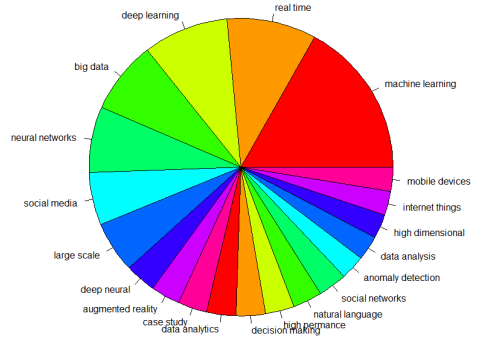

그림 10. 해외 논문의 2017년도 키워드 파이차트
Fig. 10. Keyword pie chart of international papers in 2017

Ⅳ. 실험 및 결과

표 1은 국내와 해외 연구의 동향과 지속성을 비교한 결과를 나타낸다. 국내 연구는 다양한 키워드와 트렌드를 반영하며 변화에 민감하게 반응하는 반면, 해외 연구는 핵심 기술에 대해 오랜 기간 집중적으로 연구하고 있음을 알 수 있다. 또한, 해외 연구는 기반 기술의 연구가 더 빨리 진행되는 경향이 있으며, 연구 지속성 측면에서도 장기적인 키워드가 더 많이 나타나는 특징이 있다.

표 1. 국내외 연구 동향 및 연구 지속성 비교
Table 1. Comparison of Domestic and Overseas Research Trends and Continuity

항목	내용
국내 연구 동향	- 키워드들이 매우 다양하게 나타남 - 시대의 변화에 민감하게 반응 - 기반 기술 연구가 늦음 - 알고리즘 연구 시기는 해외와 비슷 - 불규칙하게 변동하는 키워드 - 약 6년간 연구 후 감소하는 키워드들
해외 연구 동향	- 일부 핵심 기술들에 대한 키워드가 오랜 기간 지속적으로 나타남 - 오랜 기간 지속적으로 특정 연구를 진행 - 기반 기술 연구가 빠름 - 알고리즘 연구 시기는 국내와 비슷, 하지만 기반 기술 연구는 더 활발 - 일시적으로 등장했다 사라지는 키워드 - 장기간에 걸쳐 연구가 이어지는 키워드들

Ⅴ. 결 론

국내 논문들의 키워드 분석을 통해 IT 업계에서 최근 4차 산업혁명으로 인해 많은 관심을 가지게 된 IoT와 Big data, Deep learning에 대해서 향후 지속성에 대하여 IoT와 Big data의 경우 국내 논문 키워드에서 2015년부터 연구가 지속되었으므로 앞으로 약 3년 정도 연구가 더 지속될 것으로 추측하며 deep learning의 경우에는 2017년부터 많은 연구가 이루어졌으므로 앞으로 5년 정도의 연구가 지속될 것으로 추측한다.

다음으로 키워드의 지속 형태 중 꾸준히 지속되는 키워드들을 살펴보면, network 관련 연구들과 security 관련 보안 연구들, cloud computing 에 관한 연구들은 앞으로도 꾸준히 연구가 지속될 것으로 추측한다.

마지막으로 해외 논문에서는 가끔 등장하는 키워드이지만, 국내에서는 등장하지 않는 키워드를 분석하여 앞으로 국내에서 어떤 연구가 추가적으로 더 이루어져야 하는지 또한 알아볼 수 있다. natural language, case

study, decision making, data analytics 등의 키워드들이 그 키워드들에 해당된다.

References

[1] Ji. J. D. "A Study on the Development Direction of Korea Cadastral Information", Journal of Cadastre & Land InformatiX, Vol. 43, No. 1, pp. 1-22, 2013.

[2] Kim. K. J., "A study on the Status and Methods for realizing of the Right to be Forgotten Online : Focusing on the Digital Extinction Technology", Korean Journal of Law & Society, Vol. 57, pp. 95-125, 2018.
DOI: https://doi.org/10.33446/KJLS.57.4

[3] Cho. J. H., "Utilization and Prospect of Sport Big Data," The Korean Journal of Measurement and Evaluation in Physical Education and Sports Science, Vol. 14, No. 3, pp. 1-11, 2012.
DOI: https://doi.org/10.21797/ksme.2012.14.3.001

[4] Song. Y. A, "A Case Study on the Big Data Analysis Curriculum for the Efficient Use of Data," Journal of Practical Engineering Education, Vol. 12, No. 1, pp. 23-29, 2020.
DOI: https://doi.org/10.14702/JPEE.2020.023

[5] Lee. J. H, Kim. Y. Ki., "Acquisition and Utilization of Geographic Information for Participatory GIS by using Multiple Sensing Paradigms," Journal of The Korean Cadastre Information Association, Vol. 17, No. 1, pp. 127-145, 2015.

[6] Oh. C. W. "Analysis of Meaning of Social Conflict Discussion in Korea: Focusing on Key Word Network in Major Portals", Journal of Political Communication, Vol. 45, pp. 37-67, 2017.
DOI: https://doi.org/10.35731/kpca.2017..45.002

저 자 소 개

황 승 연(준회원)

• Seung-Yeon Hwang received his BS in Department of Computer Science at Korea Polytechnic University in 2019. He is currently studying MS in Department of Computer Science at Anyang University. His research interests include Database System, Big Data, Data Analysis, Machine Learning, etc.

장 석 우(정회원)

• Seok-Woo Jang received his BS and MS in computer science at Soongsil University in 1995 and 1997, respectively. In 2000, he received his PhD at Soongsil University. He is currently a professor in the department of software at Anyang University. His research interests include Artificial Intelligence (AI), Big Data, Video Processing and Block Chain.