사물인터넷융복합논문지 Vol. 10, No. 1, pp. 21-27, 2024 https://doi.org/10.20465/KIOTS.2024.10.1.021

웹 크롤링을 통한 개인 맞춤형 정보제공 애플리케이션

김주현[1], 최정은[1], 신우경[1], 박민준[1], 김태국[2*]
[1]국립부경대학교 컴퓨터·인공지능공학부 학생
[2]국립부경대학교 컴퓨터·인공지능공학부 교수

Information-providing Application Based on Web Crawling

Ju-Hyeon Kim[1], Jeong-Eun Choi[1], U-Gyeong Shin[1], Min-Jun Piao[1], Tae-Kook Kim[2*]
[1]Student, School of Computer and Artificial Intelligence Engineering, Pukyong National University
[2]Professor, School of Computer and Artificial Intelligence Engineering, Pukyong National University

요 약 본 논문에서는 필터링(Filtering)과 웹 크롤링(Web Crawling) 기술을 이용하여 개인 맞춤형 실시간 정보제공 애플리케이션을 구현하였다. 구현한 애플리케이션은 사용자가 설정한 키워드를 웹페이지 내에서 사용자가 선택한 키워드를 기준으로 Jsoup 라이브러리를 통해 웹 크롤링을 수행하고, MySQL 데이터베이스에 저장한다. 저장한 데이터는 Flutter를 이용해 구현한 애플리케이션으로 사용자에게 제공한다. 또한 FCM(Firebase Cloud Messaging)을 이용하여 모바일 푸시 알람을 제공한다. 이를 통해 사용자는 원하는 정보를 빠르고 효율적으로 얻을 수 있다. 또한 빅데이터가 생성되는 사물인터넷(Internet of things)에도 적용하여 사용자에게 필요한 정보만 제공할 수 있을 것으로 기대한다.

주제어 : 필터링, 크롤링, 빅데이터, 애플리케이션, 사물인터넷

Abstract This paper presents the implementation of a personalized real-time information-providing application utilizing filtering and web crawling technologies. The implemented application performs web crawling based on the user-set keywords within web pages, using the Jsoup library as a basis for the selected keywords. The crawled data is then stored in a MySQL database. The stored data is presented to the user through an application implemented using Flutter. Additionally, mobile push notifications are provided using Firebase Cloud Messaging (FCM). Through these methods, users can efficiently obtain the desired information quickly. Furthermore, there is an expectation that this approach can be applied to the Internet of Things (IoT) where big data is generated, allowing users to receive only the information they need.

Key Words : Filtering, Crawling, Big data, Application, Internet of things (IoT)

1. 서론

정보량의 폭증으로 인해 정보 수집의 효율성이 강조되고 있다. 특히, 공공기관의 공지사항은 중요한 정보를 담고 있어 사용자들에게 필수적이다. 대다수의 공공기관은 홈페이지를 통해 다양한 정보를 세분화하여 제공하고 있지만, 이로 인해 사용자들은 원하는 정보를 얻기 위해 여러 페이지를 찾아다니는 불편함을 겪고 있다.

현재 웹 크롤링 기술을 활용하여 모든 공지사항을 업데이트할 때마다 알림을 제공하는 연구는 진행되고 있다. 크롤링(Crawling)이란 HTML 페이지를 가져와서, HTML/CSS 등을 파싱(Parsing)하고, 필요한 데이터만

이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(RS-2023-00242528).
*교신저자 : 김태국(king@pknu.ac.kr)
접수일 2023년 12월 29일 수정일 2024년 01월 29일 심사완료일 2024년 02월 04일

추출하는 기법이다[1-5]. 그러나 이러한 기술을 이용한 연구들은 수많은 알림 중에서 원하는 정보를 찾아보아야 하는 불편함이 존재한다.

이에 따라 본 논문에서는 실시간 웹 크롤링과 필터링을 통한 사용자 맞춤형 알림 서비스에 대한 연구를 수행했다. 이 서비스는 사용자가 원하는 정보의 키워드를 기반으로 공공기관의 여러 홈페이지에서 공지사항을 실시간으로 필터링하여 알림을 제공한다. 이 알림은 Firebase Cloud Messaging(FCM)[6,7]을 활용한 모바일 푸시 알람을 통해 사용자에게 실시간으로 제공되어, 정보를 효과적으로 얻을 수 있도록 돕는다. 또한, 알림을 받은 공지사항은 어플 내에서, 다시 확인이 가능하도록 구성되어 사용자에게 편의를 제공한다. 이를 통해 필요한 정보를 빠르게 효율적으로 얻을 수 있는 서비스를 제공하고자 한다.

2. 관련 연구

다양한 목적으로 웹 크롤링을 이용한 연구가 이루어졌다.

곽신의 연구에서는 크롤링하여 얻은 상품 리뷰 데이터를 텍스트 마이닝 기법을 사용하는 연구를 진행하였다. 상품을 검색할때 사용자가 선택한 구매 기준과 키워드를 기준으로 크롤링을 진행한다. 이를 통해 얻은 데이터들을 형태소 분석을 실행하여 각 리뷰에 대해 긍정과 부정 점수를 산출해 개인화된 맞춤형 상품 추천 서비스가 진행된다[8].

김종근의 연구에서는 웹 페이지를 모니터링하고 알람 서비스를 제안하였다. 웹 크롤링은 Beautiful Soup 라이브러리를 이용하여 데이터를 수집하였고, 이를 통해 얻은 데이터를 MySQL로 관리한다. 그리고 FCM을 이용하여 모바일 알람 서비스를 제공한다[9].

이용희의 연구에서는 데이터 수집의 효율을 높인 빅데이터의 효율적 활용을 위한 웹 크롤링 프로그램을 제안하였다. 웹에서 데이터 수집을 하는 여러 방식이 있다. 그 중, 오래전부터 사용해오던 R이나 Python을 이용하여 URL에 대한 분석을 하여 문서구조형식을 이용하는 방식을 새로 제안한 웹 크롤링 알고리즘과 비교한다. 새로 제안한 PABC 크롤링 아키텍처는 기본 아키텍처보다 보안 성능과 수집 성능 모두 보완하여 제시한 아키텍처이다. PABC는 측정지표로 보았을때도 정확성, 속도, 보안 관리 면에서 우수함을 증명한다[10].

3. 제안한 애플리케이션 설계

본 논문에서 웹 크롤링 기반의 개인 맞춤형 정보 제공 애플리케이션을 설계하고 구현하였다.

3.1 요구사항 명세서

‘웹 크롤링 기반의 개인 맞춤형 정보 제공 앱(Dingle)’에 대한 연구는 사용자의 요구사항에 맞게 소프트웨어 생명 주기에 따라 구현된다. 〈Table 1〉은 앱의 중요 기능을 중요도 5로 설정하고 부가적인 기능을 중요도 1로 설정한 요구사항 명세서(SRS, Software Requirements Specification)이다[11,12].

웹 크롤링(Web Crawling)의 결과는 Home Page에 나타나며 제목, 내용, 사진을 보여준다. 공지사항에 사진이 없는 경우에는 기본 사진으로 대체된다. 사용자의 편의 즉 앱의 UX를 위해 제목과 내용은 각각 앞에서 부터 13글자, 17글자만 나타낸다. 해당 공지사항을 누르면 링크를 통해 공지사항 웹 페이지 화면으로 이동한다.

〈Table 1〉 Software Requirements Specification

Screen Name	Requirement Name	Requirement Description	Priority
SignUp Page	Sign-up for Kakao	-Proceed with Kakao sign-up. -Obtain consent for the collection of personal information during the Kakao sign-up process.	2
Login Page	Login for Kakao	-Proceed with Kakao login.	2
Home Page	Notice Board	-Display filtered notices based on selected keywords. · Title · Content · Image *In case of notices without images, substitute with a default image.	5
Custom Page	Custom homepage and keyword	-Set preferences for receiving notifications, choosing both websites and categories. -Both websites and categories can be selected with the possibility of duplication.	1
FCM (Firebase Cloud Messaging) Notification	FCM Notification	Notify about newly updated notices via FCM (Firebase Cloud Messaging).	4

3.2 개발 환경

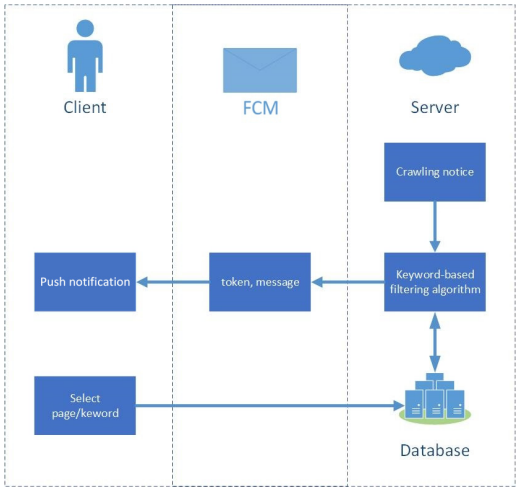
본 연구에는 크게 웹 페이지에서 공지사항의 내용을 크롤링할 환경, 이를 저장할 환경, 그리고 해당 어플리케이션에 웹 크롤링의 결과를 표시할 환경이 필요하다.

〈Table 2〉는 본 연구에서 사용된 시스템 개발 환경에 대한 표이다. Java언어를 이용한 Spring Boot 환경에서 웹 크롤링을 진행하고 데이터는 MySQL을 이용하여 관리한다. 서버 인프라의 경우 AWS(Amazon Web Services)[13.14] 클라우드를 활용하여 구축하여 배포한다. Figma[15]를 이용하여 UI를 디자인하였고 결과는 Dart 언어를 사용한 Flutter[16]로 어플리케이션을 구현하였다.

〈Table 2〉 System Development Environment

Category		Content
System Development Environment	Integrated Development Environment (IDE)	IntelliJ(Spring Boot), Android Studio(Flutter), MySQL, Figma(UI/UX), AWS
	Programming Language	Java, Flutter

3.3 서비스 흐름도



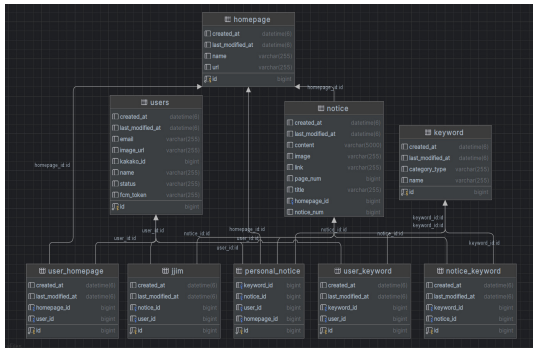
[Fig. 1] Real-time web crawling-based personalized information provision app service flowchart

'실시간 웹 크롤링을 기반 개인 맞춤형 정보제공 앱(Dingle)'은 사용자가 설정한 키워드 및 웹 페이지를 바탕으로 실시간 푸시 알림을 제공한다. [Fig. 1]은 Dingle 서비스의 흐름도이다. 사용자는 원하는 정보를 제공받기 위한 웹 페이지와 키워드를 선택한다. 서버는 특정 시간을 주기로 실시간 크롤링을 통해 새롭게 추가된 정보를 추출한다. 해당 정보가 사용자가 설정한 웹페이지에 추가된 정보이거나, 사용자가 설정한 키워드가 포함되는 경우, 토큰 값을 기반으로 어플리케이션 푸시 알림 메시지를 통해 정보를 제공한다.

3.4 데이터베이스 구조 객체-관계 모델

사용자가 설정한 정보를 기반으로 실시간 크롤링 후 정보를 리스트와 알림을 통해 제공하기 위해 MySQL 데이터베이스를 통해 데이터를 저장하고 관리한다. [Fig. 2]는 테이블 간의 관계를 나타낸 MySQL 데이터베이스의 다이어그램이다. 총 9개의 테이블로 구성되며, 연관관계를 맺고 있다.

사용자 정보는 users, 공지사항 목록은 notice, 키워드 목록은 keyword, 홈페이지 목록은 homepage 테이블을 통해 관리된다. 이 때, 사용자가 설정한 키워드는 user_keyword, 사용자가 설정한 웹페이지는 user_homepage 테이블을 통해 따로 관리된다. User_keyword 테이블의 경우 user_id와 keyword_id를 외래키(Foreign key)로 가져 users 테이블과 keyword 테이블을 참조할 수 있도록 한다. 이는 테이블 간 다대다 관계를 일대다 관계로 풀어주는 연관 테이블 역할을 한다. 서로 연관된 정보들의 경우 외래키를 통해 참조가 가능하도록 구성하였다.



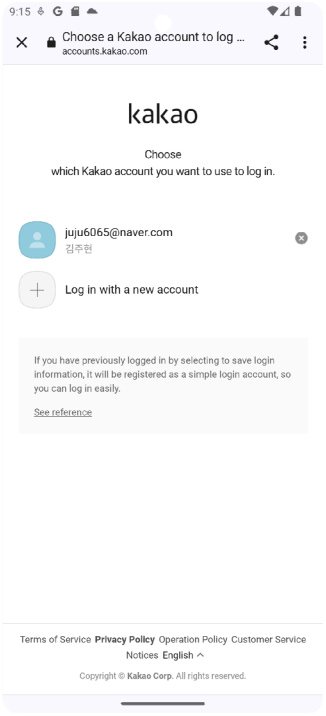
[Fig. 2] Database Diagram

4. 어플리케이션 구현

4.1 소셜 로그인

웹 크롤링 기반의 개인 맞춤형 정보 제공 앱(Dingle)에서는 OAuth를 이용한 Kakao[17] 로그인 방식을 사

용한다. OAuth는 다양한 플랫폼 환경에 권한을 부여하는 산업 프로토콜이다. OAuth 환경을 이용하면 별도의 정보 수집 없이 플랫폼에 연결된 사용자의 프로필을 이용하여 회원 가입과 로그인 과정을 단축시킬 수 있다. [Fig. 3] 은 OAuth를 이용한 회원가입 및 로그인 중 Kakao 로그인을 진행한 화면이다. [Fig. 3]과 같이 클라이언트 환경에서 Kakao 로그인을 진행하여 얻은 Accesstoken을 서버 환경에 전달하면 JWT(Json Web Token)을 반환한다. 이 JWT을 이용하여 사용자를 식별한다.



[Fig. 3] Kakao Login

4.2 실시간 크롤링과 필터링 된 리스트

4.2.1 키워드 커스텀

4.1 과정을 통해 식별된 사용자는 제공받고자 하는 홈페이지 화면과 키워드를 선택한다. 본 연구에서는 공공기관 중 국립부경대학교[18]를 대상으로 진행한다. 따라서 선택의 홈페이지 화면의 범위는 학교의 단과대 홈페이지와 학과 홈페이지이다. [Fig. 4]는 각각 홈페이지와 키워드 선택 화면이다. 홈페이지와 키워드는 중복 선택이 가능하다.

[Fig. 4] Homepage and Keyword selection screen

4.2.2 실시간 크롤링과 필터링 된 리스트

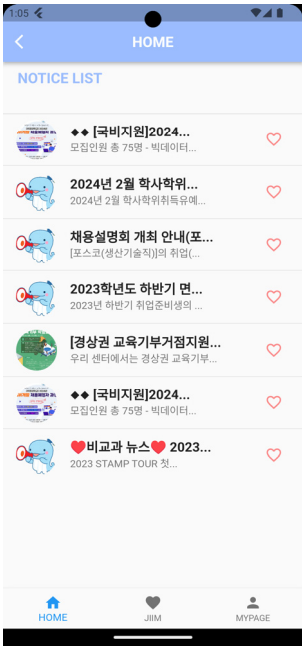
Java의 Jsoup 라이브러리를 통해 크롤링이 이루어진다. 홈페이지의 공지사항 목록 페이지에 연결 후, Connection 객체를 생성하고, 헤더, 사용자 에이전트, 메서드 등을 설정 후, 목록 페이지에서 각 공지사항 요소를 추출한다. 이 때, 데이터베이스에서 가장 최신 공지사항 번호를 조회하여 새롭게 올라온 공지사항만을 처리하도록 한다. select 함수를 통해 공지사항 번호, 제목, 내용, 링크, 이미지 정보를 추출한다. 공지사항 내용의 경우 30자만 가져오며, 이미지는 없는 경우 기본 이미지로 대체된다.

실시간으로 크롤링이 이루어지도록 하기 위해 스프링 프레임워크에서 제공하는 '@Scheduled' 어노테이션을 활용하였다. fixedDelay 속성을 통해 크롤링 주기를 3600000ms로 설정하여 1시간 간격으로 크롤링이 이루어지도록 하였다. 해당 어노테이션 사용을 위해 스프링 애플리케이션에 '@EnableScheduling'을 추가하여 스케줄링 활성화를 명시하였다.

필터링의 경우, 키워드 목록을 바탕으로 제목과 내용에서 키워드를 추출하여 해당 키워드를 설정한 사용자에게 개인화된 알림을 생성한다. 해당 홈페이지를 설정한 사용자에게도 개인화된 알림을 생성한다. 이 과정은 모두 데이터베이스에 저장된 정보를 기반으로 이루어진다.

[Fig. 5]은 위의 과정으로 크롤링과 필터링된 정보는 Flutter에 리스트화 되어서 정렬된 홈 화면이다. 정렬된 데이터 정보에는 이미지, 제목, 내용이 있다. 그리고 하트모양 이모지를 누름으로써 '좋아요' 기능이 가능하다.

제목과 내용은 요구사항 명세서에 기술한 것과 같이 어플리케이션의 UX를 위해 각각 13자, 17자까지만 제공한다.

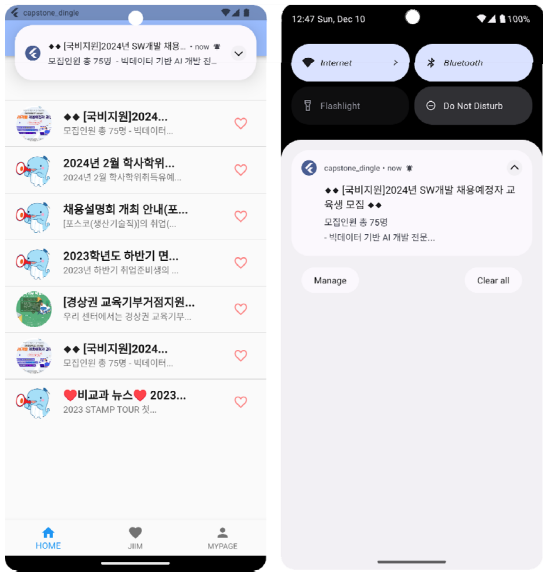


[Fig. 5] Homepage screen

4.3 푸시 알림

푸시 알림은 FCM을 통해 구현된다. FCM은 메시지를 안정적으로 무료 전송할 수 있는 크로스 플랫폼 메시징 솔루션이다. 클라우드 메시징 서버가 클라이언트와 서버 사이에 위치하여 실시간으로 사용자들에게 메시지를 전송한다.

'Dingle' 앱은 Flutter(Frontend)와 SpringBoot (Backend) 사이에 FCM을 두어 알림 전송을 구현하였다. Firebase 프로젝트 생성 후, 클라이언트와 서버에서 각각 FCM 설정이 이루어진다. 클라이언트에서, FCM 초기화 및 토큰 생성하여 로그인 시, 서버로 FCM 토큰 값을 넘겨준다. 이 때, FCM 토큰 값은 푸시 알림을 전송할 기기를 식별하기 위해 사용된다. 서버에서는 실시간 크롤링 후 사용자 맞춤형 개별 알림을 생성하여 클라이언트로부터 받은 FCM 토큰을 기반으로 알림을 전송한다. 클라이언트에서는 수신된 알림을 처리하여 [Fig. 6]과 같이 시각적으로 제공한다.



[Fig. 6] Personal Notice Push Notification

5. 결론

실시간 웹 크롤링, 필터링, 푸시 알림 기능을 통해 사용자 맞춤형 정보 제공 앱 'Dingle'을 제안하고 구현하였다. 해당 앱은 사용자가 설정한 키워드와 웹페이지를 기반으로 실시간 웹 크롤링을 통해 얻은 정보들을 푸시 알림을 통해 제공한다.

본 연구는 국립부경대학교를 대상으로 앱을 구현하였으나, 특정 지역이나 계층 등으로 확장함으로 다양하게 활용될 수 있을 것으로 기대한다. 사용자들은 가치 있는 정보를 제공받음으로 시간과 비용을 절약할 수 있다. 향후 사업화 시, 카카오 알림톡을 통한 알림 제공으로 알림에 대한 접근성을 높일 수 있을 것으로 기대된다. 현재는 제공되는 키워드 목록 중 선택하는 방식이지만, 추후에 사용자가 직접 키워드를 추가하는 방향으로 구현한다면 사용자 맞춤형에 더욱 가까워질 것으로 기대된다.

REFERENCES

[1] C.W.Na, B.W.On, "A proposal on a proactive crawling approach with analysis of state-of-the-art web crawling algorithms," *Korean Society for Internet Information,* Vol.20, No.3, pp.43-59, 2019.

[2] J.H.Kim, E.J.Kim, "WCTT: Web Crawling System based on HTML Document Formalization," *The Korea Institute*

of Information and Commucation Engineering, Vol.26, No.4, pp.495-502, 2021.

[3] Y.J.Lee, "Estimation of maximum object size satisfying mean response time constraint in web service environment," *Journal of Internet of Things and Convergence*, Vol.9, No.3, pp.1-6, 2023.

[4] B.H.Lee, "HTML specification and semantics analysis of korean news sites," *Journal of Digital Contents Society*, Vol.18, No.5, pp.949-956, 2017.

[5] H.I.Lee, J.H.Cha, "A Study on the Recognition of Korea Armed Forces Nursing Academy and Military Nursing Officers through Web Crawling and Text Mining," *Journal of the Korea Academia-Industrial cooperation Society*, Vol.24, No.5, pp.381-388, 2023.

[6] Firebase, Firebase Cloud Messaging[Internet], https://firebase.google.com/docs/cloud-messaging.

[7] A.S.Oh, "Design and Implementation of Platform for Monitoring of Notification System in Firebase Message," *Journal of Information & Communication Convergence Engineering*, Vol.19, No.1, pp.16-21, 2021.

[8] S.Kwak, "Product Recommendation System Based on User Purchase Criteria and Product Reviews," Paichai university, Master's thesis, 2021.

[9] J.K.Kim, K.H.Sim, Y.S.Lee, Y.H.Lim, "Development of real-time monitoring web BBS and the alerts service using mobile web," *Journal of Digital Contents Society*, Vol.13, No.1, pp.1-11, 2012.

[10] Y.H.Lee, "Implementation Of Web Crawling Program For Efficient Use Of Big Data," *Jounal of The Korea Society of Information Technology Policy & Management*, Vol.12, No.5, pp.1983-1989, 2020.

[11] H.W.Park, J.H.Choi, S.G.Hwang, M.S.Park, S.U.Noh,, K.H.Chung, K.H.Choi, "Development of verification tool for the soundness of software requirements specification using clustering techniques," *The Journal of Korean Institute of Next Generation Computing*, Vol.17, No6, pp.54-65., 2021.

[12] Y.K.Kim, "A Study on the Completeness Measurement of Requirements Specifications for Software Completion Appraisal," *Journal of Software Assessment and Valuation*, Vol.19, No.1, pp.11-18, 2023.

[13] T.K.Kim, "Spatial Crowdedness Measurement System using IoT and Amazon Web Services," *Journal of Internet of Things and Convergence*, Vol.9, No.4, pp.15-20, 2023.

[14] Amazon, Amazon Web Services[Internet], https://aws.amazon.com.

[15] Flutter, Flutter - Build apps for any screen[Internet], https://www.figma.com.

[16] Figma, Figma: The Collaborative Interface Design Tool[Internet], https://flutter.dev.

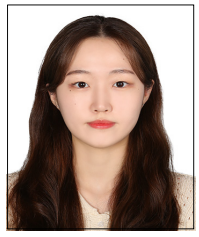
[17] Kakao, Kakao developers[Internet], https://developers.kakao.com.

김 주 현(Ju-Hyeon Kim) [준회원]

■ 2020년 3월 ~ 현재 : 국립부경대학교 컴퓨터·인공지능공학부

〈관심분야〉
빅데이터(Big Data), 데이터 분석(Data Analysis), 인공지능(AI), IT컨설팅(IT Consulting)

최 정 은(Jeong-Eun Choi) [준회원]



■ 2020년 3월 ~ 현재 : 국립부경대학교 컴퓨터·인공지능공학부

〈관심분야〉
백엔드(Backend), 데이터베이스(Database), 빅데이터(Big Data)

신 우 경(U-Gyeon Shin) [준회원]

■ 2019년 3월 ~ 현재 : 국립부경대학교 컴퓨터·인공지능공학부

〈관심분야〉
백엔드(Backend), 데이터베이스(Database), 빅데이터(Big Data)

박 민 준(Min-Jun Piao) [준회원]

■ 2020년 3월 ~ 현재 : 국립부경대학교 컴퓨터·인공지능공학부

〈관심분야〉
빅데이터(Big Data)

김 태 국(Tae-Kook Kim) [종신회원]

■ 2004년 8월 : 고려대학교 전기전자전파공학부(공학사)
■ 2006년 8월 : 고려대학교 메카트로닉스학과(공학석사)
■ 2014년 8월 : 고려대학교 모바일솔루션학과(공학박사)
■ 2016년 3월 ~ 2022년 2월 : 동명대학교 AI학부 교수
■ 2022년 3월 ~ 현재 : 국립부경대학교 컴퓨터·인공지능공학부 교수

〈관심분야〉
사물인터넷(IoT), 콘텐츠 전송 네트워크(CDN), 이동성, 인공지능(AI), 빅데이터(Big Data), 모바일 서비스