Journal of Convergence for Information Technology
Vol. 10. No. 12, pp. 9-14, 2020
e-ISSN 2586-4440
DOI : https://doi.org/10.22156/CS4SMB.2020.10.12.009

NTIS 시스템에서 딥러닝과 형태소 분석 기반의 대화형 검색 서비스 설계 및 구현

이종원[1], 김태현[2], 최광남[3*]
[1]한국과학기술정보연구원 박사 후 연구원, [2]한국과학기술정보연구원 선임연구원, [3]한국과학기술정보연구원 책임연구원

Design and Implementation of Interactive Search Service based on Deep Learning and Morpheme Analysis in NTIS System

Jong-Won Lee[1], Tae-Hyun Kim[2], Kwang-Nam Choi[3*]
[1]Postdoctoral Researcher, Korea Institute of Science and Technology Information
[2]Research Engineer, Korea Institute of Science and Technology Information
[3]Senior Research Engineer, Korea Institute of Science and Technology Information

요 약 현재 NTIS(National Technology Information Service)는 인공지능 기술을 기반으로 대화형 검색 서비스를 구축하고 있다. 이용자의 검색 의도를 파악하고 과제정보를 제공하기 위해 딥러닝 모델과 형태소 분석기를 기반으로 대화형 검색 서비스를 구축한다. 딥러닝 모델은 NTIS와 대화형 검색 서비스를 활용할 때 적재되는 로그 데이터를 기반으로 학습을 진행하고 이용자의 검색 의도를 파악한다. 그리고 단계별 검색을 통해 과제정보를 제공한다. 검색 의도 파악은 예외처리를 용이하게 해주며 단계별 검색은 통합검색보다 쉽고 빠르게 원하는 정보를 얻을 수 있도록 한다. 향후연구로는 인공지능 기술이 접목된 성장형 대화형 검색 서비스로써 이용자에게 제공하는 정보의 범위를 확대해야 한다.

주제어 : 대화형 검색 서비스, 딥러닝, 인공지능, 지능형 서비스, 형태소 분석기

Abstract Currently, NTIS (National Technology Information Service) is building an interactive search service based on artificial intelligence technology. In order to understand users' search intentions and provide R&D information, an interactive search service is built based on deep learning models and morpheme analyzers. The deep learning model learns based on the log data loaded when using NTIS and interactive search services and understands the user's search intention. And it provides task information through step-by-step search. Understanding the search intent makes exception handling easier, and step-by-step search makes it easier and faster to obtain the desired information than integrated search. For future research, it is necessary to expand the range of information provided to users.

Key Words : Comunication Search Service, Deep Learning, AI, Intelligence Service, Morpheme Analyze

1. 서론

2020년 9월 기준으로 NTIS는 84.1만건의 과제정보를 포함하여 성과정보 등 약 740만 건의 R&D 정보를

*This research was supported by Korea Institute of Science and Technology Information(KISTI).
*Corresponding Author : Kwang-Nam Choi(knchoi@kisti.re.kr)
Received October 16, 2020 Revised November 30, 2020
Accepted December 20, 2020 Published December 28, 2020

관리하고 있다. 이용자들이 가장 많이 검색하는 정보는 과제정보이다. 과제정보를 검색할 때에는 단어나 문장으로 검색을 진행하게 된다. 한 개의 단어만을 활용하여 검색할 때에는 해당 단어가 포함된 과제정보를 모두 보여주면 되지만 단어의 개수가 늘어나거나 문장 형태로 검색을 진행하게 되면 단어의 위치에 따라 의미가 달라질 수 있다[1]. 또한 이용자의 질의가 검색을 위한 것인지도 파악할 수 있어야 한다. 이러한 이유로 이용자의 검색 의도를 파악하고 단계별 검색이 가능한 대화형 검색 서비스의 필요성이 대두되고 있다.

본 논문에서는 이용자의 검색 의도가 검색을 위한 것인지 파악하고 질의 내용과 관련된 과제정보를 제공하는 대화형 검색 서비스를 제안한다[2-4]. 이용자의 검색 의도를 파악하기 위해서 NTIS를 활용할 때 적재되는 로그 데이터를 딥러닝 모델이 학습하고 판단하도록 한다. 그리고 형태소 분석기가 질의 내용을 분석하고 이와 관련된 과제정보를 제공한다[5-7].

제안하는 시스템은 이용자들의 검색 의도를 파악하고 단계별로 검색 범위를 좁혀나가는 환경을 제공하는 것이다. 이로 인해 제안하는 시스템은 다른 대화형 검색 서비스들에 비해 관련없는 질의 내용을 용이하게 처리할 수 있으며 현재 NTIS에서 제공하는 통합검색 서비스 보다 빠르고 쉽게 검색이 가능하다. 이러한 장점들을 바탕으로 제안하는 시스템이 지능형 NTIS 서비스로써 자리매김할 것으로 기대한다.

2. 관련연구

2.1 NTIS 통합검색 서비스의 문제점

현재 NTIS에서 제공하는 통합검색 서비스는 이용자가 선택한 네비게이션 항목들을 기준으로 과제정보를 검색하는 방식이다. 검색 범위를 좁히지 않는다면 이용자가 원하는 과제정보를 얻기가 어렵다. 이러한 문제점을 해결하기 위해 단계별 검색이 가능하도록 기능들을 추가하였지만 방식이 복잡한 문제점이 있다. 보다 쉽고 빠르게 검색할 수 있는 환경을 제공하기 위해 대부분의 기업이나 공공기관에서는 대화형 검색 서비스를 제공하고 있다.

2.2 태스크 기반 대화형 검색 서비스

일반적으로 활용되고 있는 대화형 검색 서비스들은 대부분이 태스크를 기반으로 한다. 특정 질의에 대해 설정한 응답 문구나 선택지를 제공하는 방식이며 이는 변수가 적은 환경에서 주로 활용된다. 미리 설정해놓은 규칙이나 시나리오 상에서 동작하기 때문에 이용자의 검색 의도를 파악할 수 없고 단계별 검색이 불가능한 문제점이 있다[8-11].

2.3 데이터 기반 대화형 검색 서비스

데이터 기반 대화형 검색 서비스는 자연어 처리 기술이나 인공지능 기술을 접목시켜서 대화형 검색 서비스의 성능을 향상시킨 형태이다. 대표적인 시스템으로는 IBM의 Watson Assistant가 있다. 딥러닝 모델을 활용하여 특정 환경에서 생성되거나 사용되고 있는 데이터를 학습하고 정보를 제공해주는 구조이다. 데이터가 대규모로 적재되어있는 상황에서 활용하기 적합하지만 이용자들마다 질의 내용과 패턴이 다르기 때문에 딥러닝 모델만을 활용한다면 신뢰성을 보장할 수 없는 문제점이 있다[12-15].

3. 본론

3.1 시스템 설계

이용자의 검색 의도를 파악하고 단계별 검색이 가능하도록 시스템을 구축하기 위해서는 다음과 같은 사항들이 요구된다. 첫째, 이용자들이 검색을 진행할 때 입력하는 자연어를 처리하기 위해서 용어사전이 요구된다. 둘째, 기존의 대화형 검색 서비스들은 미리 지정한 질의나 범주 내에서만 응답이 가능하였다. 그리고 불용어를 등록하면 해당 불용어가 포함된 질의는 모두 예외처리를 진행하기 때문에 이용자가 원활하게 서비스를 이용할 수 없는 경우가 발생하였다. 이러한 이유들로 인해 기존의 대화형 검색 서비스들은 이용자들의 다양한 질의나 목적에 대응하는 것이 어려웠었다. 이를 해결하기 위해 제안하는 시스템은 이용자들이 NTIS를 활용했을 때 적재된 로그 데이터를 딥러닝 모델로 학습한다. 그리고 학습 결과를 바탕으로 이용자들의 검색 의도를 파악할 수 있도록 한다. 이용자들이 입력한 질의가 과제정보와 관련이 있는지 파악하는 과정은 정확도가 보장되어야 한다. 이를 위해 딥러닝 모델이 지속적으로 성장할 수 있는 환경을 구축해야 한다. 셋째, 단계별 검색이 시작되면 형태소 분석기로 질의 내용을 분석하고 이와 관련된 과제정보를 제공한다. 검색 방식은

이용자의 선택에 따라 검색 범위를 좁혀나간다. 제안하는 시스템은 총 5개의 서버로 구성하였으며 Fig. 1은 시스템의 소프트웨어 구성도이다.

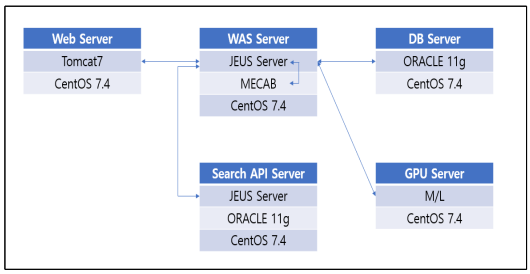

Fig. 1. Configuration of Software

질의 내용은 단어나 문장으로 구성된다. 이를 분석하기 위해서는 전처리 단계에서 문장을 단어 단위로 분할해야 한다. 그리고 분할된 단어들을 딥러닝 모델이 학습하여 과제정보와 관련된 단어인지를 판별하여야 한다. 이를 위해 필요한 처리 과정을 Fig. 2로 나타낸 것이고 Fig. 3은 딥러닝 모델의 학습과정을 나타낸 것이다.

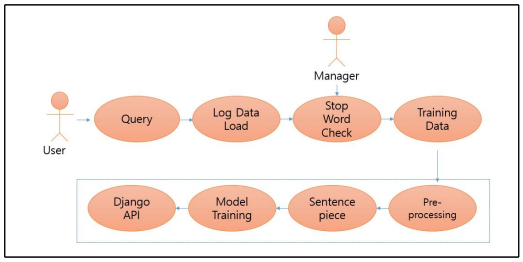

Fig. 2. Dataflow of Irrelevant Query

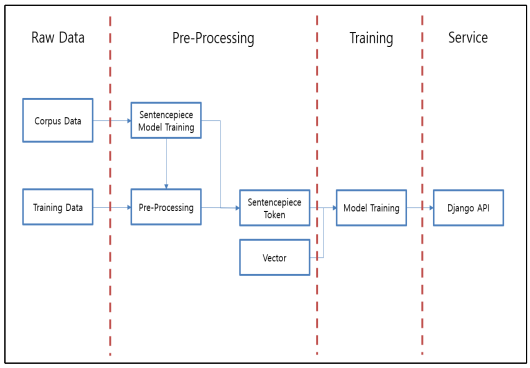

Fig. 3. Dataflow of BiLSTM Model

3.2 시스템 구현

NTIS를 이용할 때 적재된 로그 데이터들을 정제하여 딥러닝 모델이 학습할 수 있도록 전처리를 진행한다. 전처리를 진행하게 될 로그 데이터는 2018년 1월부터 2019년 8월까지 적재된 데이터를 사용하였다. 그리고 전처리는 다음과 같은 순서로 진행하였다.

로그 데이터 수집 35,207,975건, 실제 과제정보를 검색한 로그 데이터 1,610,169건, NaN 제거 후 1,383,657건, 중복 제거 후 364,343건, 샘플링 500건, 일반적인 대화 또는 관련없는 문장 생성 200건, 학습 데이터 구성 및 라벨링 700건 등 전처리 및 학습 데이터 정제를 진행하였다. Fig. 4는 전처리를 진행하여 정제된 데이터들을 나타낸다.

	text	label
202	Ultra Low CTE(1ppm/℃) Fabric 개발	1
4	축종별 전문경영체 기술수준 파악 및 모델설정 연구	1
647	다음 대통령 누가될지 알려줘	0
628	네가 반장이라고	0
626	부디 오래오래 행복을 누리소서	0
612	전래동화 인기순위가 어때	0
136	자동차용 연료전지 가격 저감을 위한 비백금	1
14	생물유전자원의 접근 및 이익공유(ABS) 국제레짐에 대비한 연구분야 대응방안 기획연구	1
624	맛있는거 좀 사줄래?	0
66	기후변화 대응 신소재 합성물 자원 창출	1

Fig. 4. Refined Dataset

딥러닝 모델을 구축할 시 학습 데이터와 검증 데이터의 비율은 7 : 3에서 9 : 1로 설정하는 것이 대부분이다. 본 논문에서는 75 : 25 비율의 학습 데이터와 검증 데이터를 활용하였다. 그리고 BiLSTM 모델을 활용하여 인공신경망 구조를 구축하였다.

전처리된 데이터들은 Django 프레임워크에서 개발한 딥러닝 모델이 학습하게 되고 이용자의 질의에 대한 답변을 제공한다.

대화형 검색 서비스를 이용할 때 과제정보 검색과 관련없는 질의가 입력되는 경우에는 딥러닝 모델의 학습에 영향을 주지 않도록 관련없는 질의를 처리할 수 있는 기능이 필요하다. Fig. 5는 관련없는 질의 예문을 입력하여 제안하는 시스템이 예외 처리를 진행할 수 있도록 설정하는 화면이고 Fig. 6은 딥러닝 모델이 검색 의도를 파악하는 과정을 나타낸 것이다.

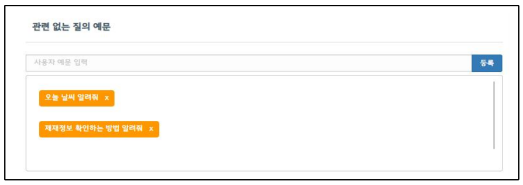

Fig. 5. Screen of Admin (Setting Irrelevant Query)

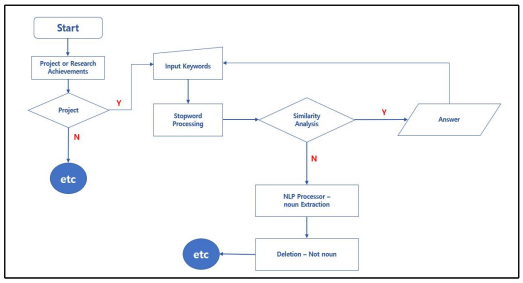

Fig. 6. Flow of Irrelevant Query Processing

Fig. 5와 같이 관련없는 질의 예문을 입력하면 딥러닝 모델이 이를 단어단위로 나누고 학습한다. 그리고 Fig. 6과 같이 이용자의 질의와 비교하여 검색 의도를 파악하게 된다. 관련없는 질의로 판단되면 Fig. 7과 같이 정해진 문구로 응답하는 화면이다.

Fig. 7. Response Screen of Irrelevant Query

기존의 대화형 검색 서비스들은 불용어로 문장을 등록한다면 해당 문장에 대해서만 예외처리를 진행하게 된다. 이에 반해 제안하는 시스템은 문장을 분석하여 단어 단위로 학습하고 과제정보와 단어들의 연관성을 분석한다. '오늘'과 '알려줘'라는 단어는 과제정보와 관련이 없기 때문에 예외처리가 진행되며 '날씨'라는 단어는 과제정보와 관련이 있기 때문에 예외처리를 진행하지 않는다. Fig. 8은 검색 의도 파악이 완료된 후 형태소 분석기가 동작되는 흐름을 나타낸 것이다.

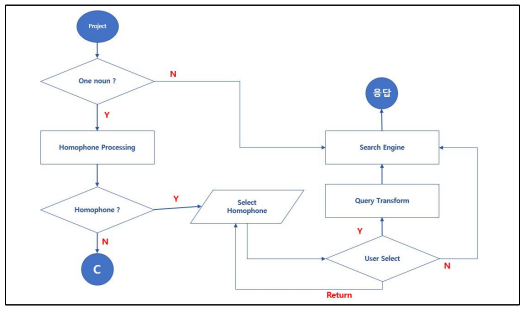

Fig. 8. Flow of Morpheme Analyzer Processing

형태소 분석기는 문장을 단어단위로 나누고 동음이의어 처리를 진행한다. 그리고 이용자의 선택에 따라 검색 엔진이 검색 범위를 단계별로 좁힌 뒤 과제정보를 제공한다. Fig. 9는 NTIS에서 대화형 검색 서비스를 활용하여 'LSTM'을 검색하고 추가 질의를 선택했을 때의 화면이다.

Fig. 9. Answer Screen of User's Query 1

Fig. 10은 추가 질의 → 기준년도 → 2020년 → 190건의 과제정보 → 과제관리기관 → 한국에너지기술평가원 → 3건의 과제정보를 보여주는 화면이다. 이용자는 도출된 3개의 과제정보에서 원하는 정보를 찾았는지에 대한 피드백을 할 수 있다. 그리고 대화이력을 딥러닝 모델이 지속적으로 학습하여 검색 의도 파악의 정확도를 향상시킨다.

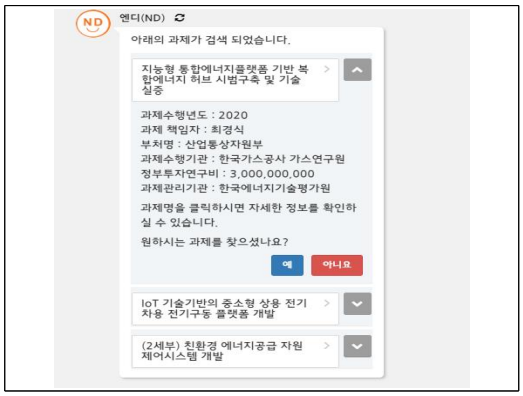

Fig. 10. Answer Screen of User's Query 2

4. 고찰

일반적인 대화형 검색 서비스들은 관리자가 정해놓은 규칙에 해당하는 답변만을 제공한다. 이는 규칙 기반의 시스템으로써 이용자의 검색 의도를 파악할 수 없고 검색과 관련된 규칙을 지속적으로 관리해야 하는 문제점이 존재한다. 다른 방식의 대화형 검색 서비스들은 특정 이용자만을 위한 시스템으로써 활용되고 있기 때문에 다양한 이용자들이 활용하기에는 적합하지 않다.

제안하는 시스템은 NTIS를 활용하는 이용자들의 로그 데이터를 학습하여 이용자들의 검색 의도를 파악한다. 다양한 이용자들이 활용하는 서비스이기 때문에 예외처리를 관리자가 직접 관리하는 것은 비효율적이다. 이를 위해 예외처리로 등록한 문장을 단어단위로 나누고 이를 딥러닝 모델이 학습한다. 그리고 단어들이 과제정보와 관련이 있는지를 판별하여 예외처리를 진행한다. 예외처리가 완료되면 단계별 검색을 진행하여 이용자들의 선택에 따라 검색 범위를 좁혀나간 뒤 과제정보를 제공한다. 그리고 NTIS를 활용할 때 생성되는 로그 데이터를 딥러닝 모델이 지속적으로 학습하여 검색 의도 파악에 대한 정확도를 향상시키는 구조이다. 검색 의도 파악이 가능하기 때문에 다른 대화형 검색 서비스들에 비해 다양한 상황에 대처가 가능하고 개인이 아닌 다수의 이용자들이 사용할 수 있다는 장점이 있다. 이는 제안하는 시스템의 우수성을 나타낸다.

5. 결론

현재 NTIS는 84.1만건의 국가연구개발사업과 관련된 과제정보를 관리하고 있고 15만명 이상의 연구자들이 과제정보를 검색하고 있다. 이러한 상황에서 통합검색보다 쉽고 빠르게 과제정보를 제공하기 위해서 데이터 기반 형태의 대화형 검색 서비스를 구축하였다.

제안하는 시스템은 딥러닝 모델과 형태소 분석기로 구성하였다. 딥러닝 모델은 NTIS를 이용하는 이용자들의 로그 데이터를 기반으로 학습을 진행하고 이용자의 질의에 대해서 예외처리 여부를 판별한다. 질의 내용이 문장일 경우 단어단위로 나누고 해당 단어들이 과제정보와 연관이 있는지 판단하게 된다. 검색 의도 파악 과정이 완료되면 단계별 검색과 이용자의 선택을 통해서 검색 범위를 좁혀나간 뒤 과제정보를 제공하게 된다.

기존의 대화형 검색 서비스들은 간단한 정보를 자동으로 제공해주는 것이 주목적이었다. 이에 반해 제안하는 시스템은 대규모로 관리되고 있는 과제정보를 대상 데이터로 다루고 수많은 이용자들에게 과제정보를 제공할 수 있도록 구축하였다. 이는 데이터 기반 형태의 대화형 검색 서비스로써 NTIS를 활용하는 모든 이용자들이 통합검색에 비해 쉽고 빠르게 과제정보를 얻을 수 있도록 도움을 줄 수 있는 검색 시스템임을 시사한다.

REFERENCES

[1] J. T. Kim & H. G. Lee & H. S. Kim. (2020). Effective Generative Chatbot Model Trainable with a Small Dialogue Corpus. *Journal of Korean Institute of Information Scientists and Engineers, 46(3),* 246-252.
DOI : 10.5626/JOK.2019.46.3.246

[2] D. A. Park. (2017). A Study on Conversational Public Administration Service of the Chatbot Based on Artificial Intelligence. *Journal of Korea Mutimedia Society, 20(8),* 1347-1356
DOI : I410-ECN-0101-2018-004-001287355

[3] M. J. Kang. (2018). A Study of Chatbot Personality based on the Purposes of Chatbot. *Journal of the Korea Contents Association, 18(5),* 319-329.
DOI : I410-ECN-0101-2018-310-002251103

[4] J. J. Kim & H. J. Jo. (2019). Development of Conversational News Chatbot System Based on User Intent Analysis. *Journal of Digital Contents Society, 20(5),* 963-972.
DOI : 10.9728/dcs.2019.20.5.963

[5] M. C. Sung. (2020). Pre-Service Primary English Teachers' AI Chatbots. *Journal of Language Research, 56(1),* 97-115.
DOI : 10.9728/dcs.2019.20.2.241

[6] S. H. Choi & J. Y. Kim & J. H. Song & S. M. Jung & S. J. Hong. (2019). Labor Law Consulting System With IBM Watson Chatbot. *Journal of Digital Contents Society, 20(2),* 241-249.
DOI : 10.9728/dcs.2019.20.2.241

[7] J. W. Kim & H. I. Jo & B. G. Lee. (2019). The Study on the Factors Influencing on the Behavioral Intention of Chatbot Service for the Financial Sector - Focusing on the UTAUT Model. *Journal of Digital Contents Society, 20(1),* 41-50.
DOI : 10.9728/dcs.2019.20.1.41

[8] X. F. Wang & H. C. Kim. (2018). Text Categorization with Improved Deep Learning Methods. *Journal of Information and Communication Convergence Engineering, 16(2),* 106-113.
DOI : 10.6109/jicce.2018.16.2.106

[9] D. H. Seo & J. S. Lyu & E. J. Choi & S. H. Cho & D. K. Kim. (2018). Web based Customer Power Demand Variation Estimation System using LSTM. *Journal of the Korea Institute of Information and Communication Engineering, 22(4),* 587-594.
DOI : 10.6109/jkiice.2018.22.4.587

[10] J. W. Lee & H. Y. Kim & H. K. Jung. (2020). Deep Learning Module Optimization based on Sequential Data Prediction. *ASM Science Journal, 13(1),* 82-91.

[11] Y. H. Kim & Y. K. Hwang & T. G. Kang & K. M. Jung. (2016). LSTM Language Model Based Korean Sentence Generation. *The Journal of Korean Institute of Communications and Information Sciences, 41(5),* 592-601.
DOI : 10.7840/kics.2016.41.5.592

[12] I. T. Joo & S. H. Choi. (2018). Stock Prediction Model based on Bidirectional LSTM Recurrent Neural Network. *Journal of Korea institute of information, electronics, and communication technology, 11(2),* 204-208.
DOI : 10.17661/jkiiect.2018.11.2.204

[13] H. I. Kim & J. Y. Lee. (2020). Prediction of Urban Flood Extent by LSTM Model and Logistic Regression. *Journal of the Korean Society of Civil Engineers, 40(3),* 273-283.
DOI : 10.12652/Ksce.2020.40.3.0273

[14] T. H. Min & H. J. Shin & J. S. Lee. (2019). Korean Spatial Information Extraction using Bi-LSTM-CRF Ensemble Model. *The Journal of the Korea Contents Association, 19(11),* 278-287.
DOI : 10.5392/JKCA.2019.19.11.278

[15] H. Y. Yu & Y. J. Ko. (2017). Expansion of Word Representation for Named Entity Recognition Based on Bidirectional LSTM CRFs. *Journal of Korean Institute of Information Scientists and Engineers, 44(3),* 306-313.
DOI : 10.5626/JOK.2017.44.3.306

이 종 원(Jong-Won Lee) [정회원]

· 2016년 2월 : 배재대학교 컴퓨터공학(공학석사)
· 2019년 2월 : 배재대학교 컴퓨터공학(공학박사)
· 2020년 6월 ~ 현재 : 한국과학기술정보연구원 박사 후 연구원/ NTIS
· 관심분야 : 빅데이터, 오픈사이언스, 인공지능
· E-Mail : jwon1991@kisti.re.kr

김 태 현(Tae-Hyun Kim) [정회원]

· 2001년 2월 : 충남대학교 컴퓨터과학과(이학석사)
· 2001년 3월 ~ 2001년 11월 : ㈜엔퀘스트테크놀러지 연구원
· 2002년 3월 ~ 2004년 2월 : 한국전자통신연구원 연구원
· 2004년 3월 ~ 현재 : 한국과학기술정보연구원 선임연구원 / NTIS 개발팀장
· 관심분야 : 정보검색, 정보분석, 전문용어사전구축, 소프트웨어공학
· E-Mail : heemang@kisti.re.kr

최 광 남(Kwang-Nam Choi) [정회원]

· 1994년 2월 : 충남대학교 컴퓨터공학과(공학석사)
· 2017년 2월 : 배재대학교 컴퓨터공학과(공학박사)
· 1994년 7월 ~ 현재 : 한국과학기술정보연구원 책임연구원 / NTIS 센터장
· 관심분야 : 정보검색, 정보분석, 빅데이터
· E-Mail : knchoi@kisti.re.kr