사물인터넷융복합논문지 Vol. 8, No. 3, pp. 69-74, 2022 https://doi.org/10.20465/KIOTS.2022.8.3.069

처방적 분석 기반의 연구자 맞춤형 연구정보 서비스 설계

이정원[1], 오용선[2*]
[1]목원대학교 정보통신공학 박사과정, [2]목원대학교 정보통신공학 교수

Design of Customized Research Information Service Based on Prescriptive Analytics

Jeong-Won Lee[1], Yong-Sun Oh[2*]
[1]Ph.D. Student, Department of Information Communication Engineering, Mokwon University
[2]Professor, Department of Information Communication Engineering, Mokwon University

요 약 빅데이터 관련 분석 기법에서 처방적 분석 방법론은 적극적인 학습이 양질의 학습 데이터를 확보함으로써 수동적인 학습모델의 성능을 개선하고, 해당 시스템을 최적화하여 성능의 극대화를 통해 처리 프로세싱 과정을 다루며 판단의 근거가 되는 이유를 제시하고 있다. 그리고 범주 정보가 없는 데이터의 경우 기계가 이를 분석하여 애매한 것과 경계지점에 놓인 것들을 찾아내 수동으로 판단하게 하여 값비싼 범주 데이터를 매우 효과적으로 구축하는 방식이다. 연구자 역량을 강화하기 위하여 연구자의 연구 분야, 연구 성향, 연구 활동정보 등을 수집하여 데이터가 가진 가치를 확장하기 위해 데이터 전처리 후 실행 시점의 상황 예측하고 실행 가능한 대안 도출을 통해 상황 변동에 따른 대안 유효성 검토 등 처방적 분석을 통하여 연구자 맞춤형 연구정보 서비스를 제공한다.

주제어 : 처방적 분석, 연구자 맞춤정보, 학술정보, 객체 분류, 연구 역량

Abstract Big data related analysis techniques, the prescriptive analytics methodology improves the performance of passive learning models by ensuring that active learning secures high-quality learning data. Prescriptive analytics is a performance maximizing process by enhancing the machine learning models and optimizing systems through active learning to secure high-quality learning data. It is the best subscription value analysis that constructs the expensive category data efficiently. To expand the value of data by collecting research field, research propensity, and research activity information, customized researcher through prescriptive analysis such as predicting the situation at the time of execution after data pre-processing, deriving viable alternatives, and examining the validity of alternatives according to changes in the situation Provides research information service.

Key Words : Prescriptive analytics: Researcher personalized information: Academic information: Object classification; Research competency

1. 서론

빅데이터 분석방법은 범주 정보가 없는 데이터의 경우 기계가 이를 분석하여 애매한 것과 경계지점에 놓인 것들을 찾아내 수동으로 판단하게 함으로써, 값비싼 범주 데이터를 구축하여 최적의 방안을 도출하는 지능형 시스템에 적용 가능한 새로운 방식의 빅데이터 분석 기술이 요구되고 있다.

*교신저자 : 오용선(sysunoh@mokwon.ac.kr)
접수일 2022년 3월 29일 수정일 2022년 4월 25일 심사완료일 2022년 4월 27일
www.kci.go.kr

또한 기존의 기술적 분석, 예측적 분석을 이용하여 기하급수적으로 증가하는 데이터 프로세스 환경에서 데이터를 처리, 분석 후 개선점을 도출하고 서비스를 재편하여, 제한된 자원을 효율적으로 활용하여 최적의 대안을 찾아 “비즈니스에 도움이 되려면 무엇을 해야되는가?”에 대해 자동적으로 제시할 수 있는 기술이 확보되지 않아, 인공지능형 시스템 운영에 있어서 데이터 정확성 및 신뢰성에 많은 문제점에 대하여 해결하기 위한 방법으로 처방적 분석을 통해 최적의 대안을 제시할 수 있다.

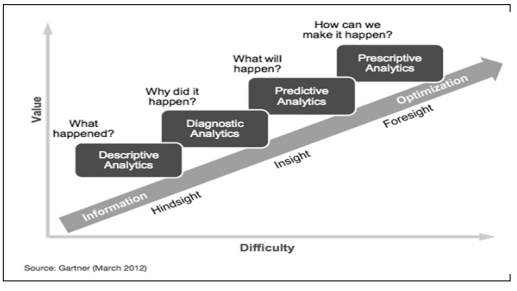
대용량 데이터에 기반한 맞춤형 서비스를 위해서는 최상의 분류 기술이 요구되며 학술논문을 대한 연구가 많이 되고 있지만 자동으로 분류하는데 어려움을 가지고 있다. 현재 분류 기반 서비스의 만족도를 높이기 위하여 분류체계를 재구성하고 분야별 데이터를 자동으로 진단하는 정보가 필요하다.

연구자 역량 강화를 위해 연구자의 연구 분야, 연구 성향, 연구 활동정보 등의 빅데이터 학술정보를 활용하여 맞춤형 연구정보 서비스가 필요하다.

2. 처방적 분석 개요

비즈니스 분석 방법 중에서 처방적 분석은 [Fig 1]와 같이 앞으로 어떻게 하면 해결되는지에 대하여 분석한다.

처방적 분석은 제한된 자원을 활용하여 최적의 대안을 제안하고 예측 결과를 토대로 의사결정 하는데 도움을 준다. 예측 분석을 통해 도출된 예측 결과를 바탕으로 최적의 의사결정을 하는데 도움을 줌으로써 사람이 수동으로 처리하는 부분이 점차 줄어들 것이다[1].



[Fig. 1] Customized research information service overview

빅데이터 개념의 세 가지 핵심 요소인 3V는 다양성(Variety), 대용량(Volume), 속도(Velocity) 이다. 다양한 유형으로 생산되는 대용량의 자원을 신속하게 처리하고 효과적인 관리를 위해 많은 분석 기반의 정보 서비스들이 분류 체계에 활용되고 있다.

기하급수적으로 증가하는 데이터 프로세싱 환경에서 대용량의 데이터 처리와 분석을 통해 개선점을 도출하고 서비스를 개선하는 순환 과정을 전문가에게만 의존하는 것이 불가능하다. 그리고 분류 성능이 보장되지 않은 상황에서 하드웨어적인 처리 방식의 도입만으로 분석 서비스를 개선하는 것도 한계가 존재한다. 만약에 성능을 보장한다면 고비용의 수작업 보다 자동화 처리가 좋은 대안이 될 것 이다. 분류되지 않은 대용량의 학술정보를 규칙적인 방법으로 분류하는 것이 우선적으로 필요하기 때문에 자동 분류 기법을 활용해야 한다.

학술 정보의 경우 분류별(공학, 의약학, 인문학 등)로 학문 분류의 특성이 달라 용어의 전문성이 다양한 요인으로 성능의 편차가 크게 나타난다.

다양한 유형으로 생산되는 대량의 정보 자원을 신속하게 처리하고 효과적으로 관리하며, 분류 되지 않은 대량의 학술 정보를 규칙적인 방법으로 분류코드를 부여하기 위해 처방적 분석 기반의 자동 분류 방식의 기술을 활용한다.

3. 맞춤형 연구정보 서비스 분석

처방적 분석 방법론은 양질의 학습 데이터를 확보함으로써 학습 모델의 성능을 개선하고, 주어진 시스템을 최적화하여 성능을 극대화하는 프로세싱 과정을 다루며, 판단의 근거가 되는 이유를 제시하고 있으며 범주 정보가 없는 데이터의 경우 기계가 이를 분석하고 애매한 것 또는 경계지점에 놓인 것 들을 찾아내 수동으로 판단함으로써 값비싼 범주 데이터를 구축하는 방식으로 맞춤형 서비스 개발을 위한 최적의 분석 방법이다.

본 연구에서 제안한 맞춤형 연구정보 서비스는 처방적 분석 기반 자동 분류된 빅데이터를 연계, 분석하여 빅데이터가 가진 가치(완전성, 활용성, 재사용성, 정확성, 편의성, 독립성, 상호운용성 등)를 확장하는 것을 의미한다. 주요 기술 분야별 연구정보 분석, 예측 및 추천 서비스는 미래의 연구 계획 및 전략을 수립하는데 중요하다[2-4].

현재 서비스하고 있는 연구 정보는 과거의 분석 정보를 제공함으로써 분석 결과는 예측 및 추천 정보가 아니므로 향후 연구 계획을 수립하는데 많은 어려움이 존재

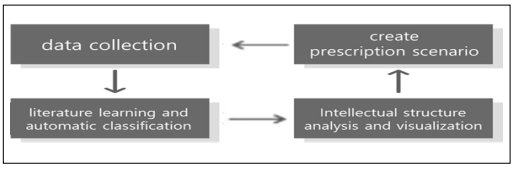
한다. 따라서 본 연구에서는 평가 요인을 내외부적 평가 요소로 구분하여 다양한 평가 요소 기반의 연구자 역량을 평가하고 최적의 맞춤형 연구정보 서비스를 제공한다.

처방적 분석은 비즈니스 분석 중에서 가장 큰 가치를 제공할 수 있는 분석방법으로 관심을 받고 있지만, 아직은 초기 단계임이 명확하며, 2015년 기준으로 가트너에 의하면 처방적 분석의 도입률은 미비한 수준이다. 하지만, 앞으로 처방적 분석에 대한 수요는 점진적으로 높아질 것으로 예상되고 있으며, 점차적으로 데이터양의 폭등, 데이터 처리의 고속화, 분석 알고리즘의 진보는 처방적 분석 기반의 의사결정에 대한 요구가 있을 것이다[5-7].

빅데이터 환경에서 대량의 학술정보를 학습하는 과정에서 자질 특성을 분석하고 제거할 때 소요되는 시간과 컴퓨팅 자원 문제가 발생하며 자질 선정기법도 한계가 존재한다. 이러한 문제점을 해결하기 위하여 본 연구에서는 데이터가 실시간으로 추가되는 환경에서 모든 데이터를 학습하지 않고 추가된 데이터만 학습 후 초기의 학습 결과와 통합하여 처리하는 학습방법이다.

대용량의 학술정보를 학습할 때 소요시간과 컴퓨팅 자원 문제를 해결하기 위해 자질 축소 기법에 의존하지 않고 부분적인 자질 추가 시에 변경요소만 추가 반영할 수 있는 범용적인 분류기법을 활용한다.

본 연구에서 제안하는 처방적 분석 기반 자동 분류 플랫폼은 [Fig. 2]와 같이 크게 4개의 컴포넌트로 이루어져 있다[8].



[Fig. 2] Components of PA-based Automatic Classification Platform

먼저 대량의 학술 논문을 데이터를 수집하고 저장하고 학습 및 자동 분류 단계에서는 수집된 대용량 문헌을 자동 분류기를 이용하여 학습 및 분류를 시행한다. 다음 단계의 분류 결과는 주제 범주 간의 확률적인 유사도 형태로 제공된다. 이와 같이 확률 데이터를 활용하면 시스템이 모호하게 분류한 학술정보 데이터를 수집할 수 있으며 학습을 통해 학술정보 데이터의 품질이 개선된 시스템으로 확장 발전시킬 수 있다. 시각화 단계에서는 자동 분류 시 생성된 오류 정보를 활용하여 학문 분야의 전역적 네트워크를 생성할 수 있다. 마지막 단계는 시나리오 생성부로써 실험 결과를 처방적 분석 서비스에서 제공할 수 있는 분류 체계 시나리오를 생성하는 단계이다. 두 번째 단계에서 자동 분류와 세 번째 단계에서 시각화는 상호작용하며 반복적으로 계산을 수행한다. 지속적인 운영을 위하여 자연스럽게 순환되도록 피드백 과정을 수행한다.

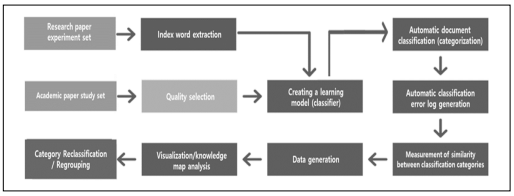
본 연구에서는 자질값 투표 분류기(Feature Value Voting Classifier: FVC)를 사용하여 대량의 학술 정보를 분류하는데 사용되는 기법이다. FVC를 이용하면 데이터의 집합을 최소 단위로 나누어 학습한 결과와 다른 학습 결과를 통합하여 빠르게 학습이 가능한 기법으로 개별 자질의 주제 범주와 유사도 벡터정보를 생성하고 다수결 방식으로 많이 득표한 주제범주를 선택하는 투표 방식 분류기이다.

또한, 프로세스는 간결하며 메모리는 적게 할당 되므로 빠르게 학술 문헌을 분류하는데 적합한 기법이다. 여러 개의 작은 단위의 학습 정보를 학습한 결과를 재 생성하는 증분 학습 알고리즘은 최적의 결과를 위한 처방적 분석 기반의 모델이며 대량의 학술 정보를 계층적으로 쌓고 빠르게 학습하기 위한 증분 학습 방식이다. 증분 학습의 장점은 원하는 학습 데이터셋을 선정 및 통합하여 학습 결과를 원하는 방식으로 생성할 수 있기 때문에 대용량 학술 정보를 학습하는데 적합한 모델이다. 현재 주제 분류 체계가 실제 데이터의 특성을 반영하지 못하고 일관성이 없는 경우 데이터 현황을 검수하여 분류 체계를 재설정할 필요가 있다[9-11].

따라서, 처방적 분석은 데이터의 품질 개선 과정을 실행할 수 있고 분류 체계를 재설정하기 위해 주제 범주 간의 연관성 측정과 의사결정에 적합한 분석 결과를 산출하는데 활용한다.

연구자 맞춤형 연구정보 서비스를 위한 학술정보의 자동 분류에서 발생한 오분류의 빈도 정보를 활용하여 학문별 주요범주 간 연관성을 측정하는 것과 기존 분류 기법들과 차별성이 있다. 요소 간의 연관 빈도행렬에 대해 측정 과정을 자동 분류의 실패로 결정된 오류 빈도를 주제 간의 유사도 측정과정으로 대체하고 학문분야의 지적 구조 분석 데이터로 활용한다. 데이터를 분류하고 분석 결과를 시각화하여 효율적으로 동작하는 처방적 분석 기반의 자동 분류 기술이다. 산출 값을 모든 주제 범주에 대하여 유사도 벡터 형태로 가지는 확률 모델을 기반으로 하는 FVC 분류 기법이다.

본 연구에서 제안하는 처방적 분석 기반 자동 분류 기법의 전체 처리 과정은 [Fig. 3]과 같다[12].



[Fig. 3] Prescriptive analysis-based automatic classification platform components

4. 처방적 분석의 연구정보 서비스

빅데이터 분석 기술의 발전은 축적된 데이터를 분석하여 파악하는 묘사 분석(descriptive analytics)과 진단 분석(diagnostic analytics), 미래를 추정하는 예측 분석(predictive analytics)을 거쳐 최적의 기법을 통해 선택 가능한 시나리오를 시스템이 제시하는 처방적 분석(prescriptive analytics)에까지 이르렀다. 처방적 분석 기반의 방법 및 개념은 앞으로 다양한 서비스에 활용되며 지속적으로 발전하고 있다[13].

앞서 자동 분류에서 오류 로그를 활용하는 방법은 학문 영역의 전체적인 구조 파악에 유용하게 활용되며 학문 분야 간의 연관정도를 보여주는 지적구조를 고려한 분류체계를 재설정하는 근거가 될 수 있다.

기계 학습과 계량정보 분석을 융합하여 시나리오를 생성함으로써 처방적 분석 시스템으로 발전시키는 방안은 기존의 분류 서비스를 의사결정 지원 서비스로 재설계하여 활용할 수 있다. 분석 서비스 제공을 위해 범주를 재설정하는 과정으로 분류 체계를 재설정함으로써 지속적으로 최종 서비스의 성능을 향상 시킬 수 있다.

연구정보에서 키워드를 추출하고 관리하는 키워드 분석 기술은 대용량 문헌 분류의 정확도를 높여줄 뿐만 아니라 연관된 관심 연구 분야 탐색, 전문가 검색 등과 같은 서비스에 매우 핵심적인 기술이다. 연구정보를 관리하기 위해 키워드를 정확하고 자동으로 정의, 관리할 수 있다면 관련 서비스의 품질 제고가 가능하여, 관심 연구자 동향 탐색, 학술 정보 객체 기반 검색, 맞춤형 추천 서비스를 제공 할 수 있다.

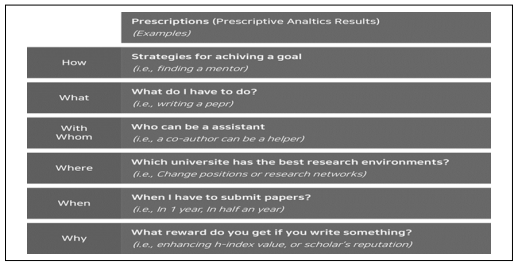
객체화를 통해 단어 객체, 본문의 단어들을 파싱하고 정지어 처리와 토큰 생성 처리 이후, 출현 빈도에 따라 연관도 높은 키워드를 추출하고 이를 독립개체로 저장하여 개체 연관 서비스에 활용된다.

기계학습, 딥러닝 및 자연어처리 등 최신기술을 활용하여 빅데이터 플랫폼 분석DB에 등록되어 있지 않은 연구 키워드 및 개체명 등의 주요 자질에 대해서도 필요 시 추출하여 활용한다.

연구정보의 메타데이터와 내용 분석을 통하여 객체를 추출하는 데이터 객체화 기술은 논문의 제목, 저자, 기관, 초록 등과 같이 개체 속성으로 존재하는 객체를 정확히 추출하고 적절히 처리하는 기법은 맞춤형 연구정보 서비스에 있어서 필수적인 전제 기술이다. 논문의 메타데이터를 활용, 언어적 분석을 통해 객체화하고 구조 정보를 추출, 데이터베이스 스키마를 구성한다[14-15].

전문적인 분석 및 서비스를 위해 요약테이블로부터 필드 정보를 선택한 후 데이터 전처리 과정이 필요하다. 요약테이블이 생성된 이후 다양한 행렬(발생행렬, 유사도 행렬, 동시발생행렬)을 생성해서 확인할 수 있다. 요약테이블과 행렬 값으로부터 차트를 만들어서 볼 수 있고 행렬 값에 대해서는 요약통계량을 만들 수 있다. 행렬을 생성한 후 클러스터링으로 군집의 계층구조를 확인할 수 있다. 마지막으로 시각화 결과로는 FDP(Force Directed Placement), 패스파인더 네트워크(PathFinder Network) 등 상호관계를 구조적으로 시각화할 수 있다.

처방적 분석 개념을 적용하여 연구자가 역량 강화를 위해서 적용된 [Fig. 4]와 같이 5W1H를 기반으로 최적화하여 연구자 맞춤형 연구정보 서비스를 제공한다[16].



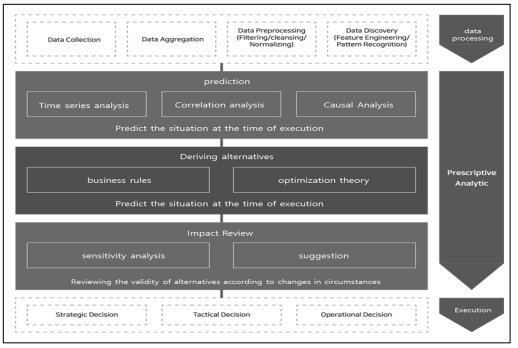
[Fig. 4] Data processing process based on prescriptive analytics 5W1H Based Prescriptive Analysis Template

처방적 분석을 통해 연구자와 관련된 이벤트가 발생할 때마다 주어진 목표와의 갭을 계산하여 새로운 분석 결과와 모니터링 할지를 결정할 필요가 있다. 그렇지만 현재까지 연구자의 역량을 강화하기 위한 처방적 분석에 적용되지 않고 있다. 그러므로 복합 이벤트 처리 방법론과 함께, 인지 과학, 행동 분석 등의 연구가 필요하다.

특정한 기술 분야에 대한 연구자 분석, 예측, 추천은

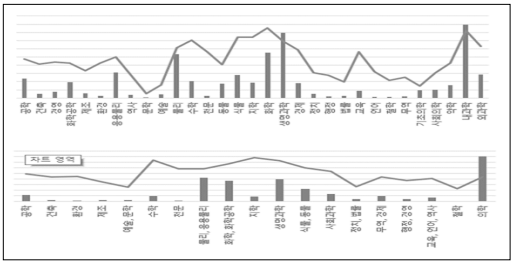
www.kci.go.kr

미래 연구 계획과 전략 수립에 중요한 부분이다. 연구자의 기관, 학술지, 관심 학술분야, 키워드 등의 연구정보를 활용하여 다양한 평가 요소를 기반으로 연구자의 역량을 평가하고 이를 기반으로 적합한 연구자 맞춤형 연구정보 서비스의 구성도는 [Fig. 5]과 같다.



[Fig. 5] Data processing process based on prescriptive analytics

대량의 학술논문을 FVC로 학습데이터와 시험데이터(8:2)를 분류하고 교차검증(K-Fold cross validation)을 통해 주제 범주 간 성능 값을 측정한다.



[Fig. 6] Data processing process based on prescriptive analytics

주제분야 공학(전산, 건축, 환경, 문학, 예술), 자연과학(수학, 물리, 천문, 화학, 생명과학), 사회과학(정치, 법률, 행정, 교육, 경제), 의학(기초의학, 내외과학, 사회의학, 약학)의 F1점수가 주제분야 내에서도 많은 편차의 성능을 보였지만 처방적 분석 기반의 분류체계를 범주 간 유사도를 고려하여 학문분야 재조정을 통해 그룹화 후 성능 값을 측정한 결과 주제분야별 F1점수 성능의 편차가 다소 감소하는 결과를 보였다[Fig 6]. 주제 분류체계는 전문가를 통해 학문적 지적 구조를 판단하여 규칙을 생성하고 성능 목표 달성을 위해 단계적으로 분류 체계 조정이 필요해 보인다.

5. 결론

처방적 분석 기반의 연구자 맞춤형 연구정보 서비스는 연구자 역량을 강화하기 위하여, 연구자의 연구 분야, 연구 성향, 환경특성, 활동정보 등을 실시간으로 수집하고, 데이터가 가진 가치(완전성, 정확성, 재사용성, 독립성, 편의성, 활용성, 상호운용성 등)를 확장하는 데이터 전처리 후 실행 시점의 상황 예측과 실행 가능한 대안 도출, 상황 변동에 따른 대안 유효성 검토 등 처방적 분석을 통하여 연구자 맞춤형 연구정보(5W1H : What, With whom, Where, When, Why, How) 서비스를 제공한다.

REFERENCES

[1] Douglas Laney et al., “Gartner Predicts 2013: Information Innovation”, Gartner Inc., 2013.

[2] Frazzetto, Davide, et al., “Prescriptive Analytics: A Survey of Emerging Trends and Technologies.” The VLDB Journal, Vol.28, No.4, pp.575-595, 2019.

[3] WN. Sadat Mosavi and M. Filipe Santos, “How Prescriptive Analytics Influences Decision Making in Precision Medicine,” Procedia Computer Science, Vol.177, pp.528-533, 2020.

[4] Santiago, A. M., and Smith, R. J., "What can "Big data" methods offer human services research on organizations and communities?" Human Service Organizations: Management, Leadership & Governance, Vol.43, No.4, pp.344-356, 2019.

[5] Riabacke, Mona; Danielson, Mats; Ekenberg, Love., “State-of-the-Art Prescriptive Criteria Weight Elicitation”. Advances in Decision Sciences, Vol.2012, pp.1-24, 2012.

[6] Lepenioti, K., Bousdekis, A., Apostolou, D., and Mentzas, G., “Prescriptive analytics: Lite-rature review and research challenges”. International Journal of Information Management, Vol.50, pp.57-70, 2020.

[7] Frazzetto, Davide Nielsen, Thomas Dyhre Pedersen, Torben Bach Šikšnys, Laurynas., “Prescriptive analytics: A survey of emerging trends and technologies”. The VLDB Journal, Vol.28, No.4, pp.575-595, 2019.

[8] Steenstrup, K.; Sallam, R. L.; Eriksen, L.; Jacobson, S. F., “Industrial Analytics Revolutionizes Big Data in the Digital Business”. Gartner Research, 2014.

[9] Hupfeld, D., Maccioni, R., Sesemann, R., Ravazzolo, D., “Fleet asset capacity analysis and revenue management optimization using advanced prescriptive analytics”. J. Revenue Pricing Manag, Vol.15, No.6, pp.516-522, 2016.

www.kci.go.kr

[10] Soltanpoor, R., Sellis, T., Prescriptive analytics for big data. In: Databases Theory and Applications-27th Australasian Database Conference, pp.245-256, 2016.

[11] Wu, P.J., Yang, C.K., The green fleet optimization model for a low-carbon economy: A prescriptive analytics. ICASI, pp.107-110, 2017.

[12] Bill Vorhies, "Prescriptive versus Predictive Analytics-A Distinction without a Difference?". Predictive Analytics Times, 2014.

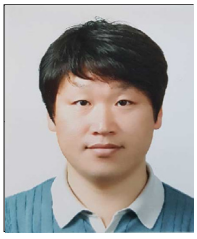
[13] Prescriptive analytics. Wikipedia[Internet], https://en.wikipedia.org/wiki/Prescriptive_analytics.

[14] L.-J. Lee, & S.-J. Kim, A Study on the Information Use Behaviors of Researchers in the Field of Business Administration for Improving Information Services. Journal of the Korean BIBLIA Society for Library and Information Science, Vol.26, No.1, pp.279-302, 2015.

[15] Jeong-Hwan. Kim, Jay-Hoon Kim and Jae-Young Hwang, "A Study on Information Users' Needs and Information Seeking Behavior of Doctoral Researchers in Digital Age." Journal of Korean Library and Information Society, Vol.42, No.3, pp.189-208, 2011.

[16] Lestari, F., Herman, T., & Sujana, A., Using The 5W1H Method in Writing Important Information with Google Forms in Elementary Schools. The 3rd International Conference on Elementary Education, Vol.3, No.1, pp.207-211, 2021.

이 정 원(Jeong-Won Lee) [정회원]



■ 2020년 3월 ~ 현재 : 목원대학교 정보통신공학 박사과정

〈관심분야〉
머신러닝, 자연어처리

오 용 선(Yong-Sun Oh) [정회원]

■ 1983년 2월 : 연세대학교 공과대학 전자공학과(공학사)
■ 1985년 2월 : 연세대학교 대학원 전자공학과(공학석사)
■ 1992년 2월 : 연세대학교 대학원 전자공학과(공학박사)
■ 1984년 3월 ~ 1986년 7월 : 삼성전자(주) 시스템개발실 연구원
■ 1987년 1월 ~ 1988년 2월 : 3J TECH. INC. 선임연구원
■ 1998년 9월 ~ 1999년 8월 : 한국해양대학교 객원교수
■ 1988년 3월 ~ 현재 : 목원대 정보통신융합공학부 교수, 한국해양정보통신학회, 한국통신학회, 대한전자공학회 IEEE 정회원.

〈관심분야〉
디지털 커뮤니케이션 시스템, 정보이론, 멀티미디어 콘텐츠

www.kci.go.kr