키워드 중심 학술정보서비스 개선 연구*

- NDSL 추천 및 분류를 중심으로 -

An Improvement study in Keyword-centralized academic information service

- Based on Recommendation and Classification in NDSL -

김 선 겸(Sun-Kyum Kim)† ** · 김 완 종(Wan-Jong Kim)† ***
이 태 석(Tae-Seok Lee)† · 배 수 영(Su-Yeong Bae)†

〈목 차〉

초 록

최근 정보의 폭발적인 증가로 인해 사용자에게 적합한 정보를 제공하기 위한 정보의 필터링이 매우 중요시 되고 있다. 한국과학기술정보연구원에서 운영하고 있는 학술정보서비스인 NDSL은 방대한 자료를 보유함에도 불구하고 사용자들은 검색 외에 자료 획득이 쉽지가 않다. 본 논문은 사용자에게 적합한 정보를 제공하기 위하여 키워드 특성을 활용한 서비스인 PIN(Profiling service In NDSL)을 제안한다. PIN은 키워드만을 가지고 검색하는 것이 아닌 사용자 본인 및 유사 사용자가 등록한 관심 키워드, 동시이용 키워드, 검색 키워드로 분석된 워드 클라우드를 제공하고 이를 통하여 사용자에게 맞춤형 논문, 보고서, 특허, 동향의 콘텐츠를 추천한다. 또한 콘텐츠를 보다 쉽게 접근하기 위하여 중복분류가 가능한 학술연구분류체계 기반 분류를 제공한다. 이를 검증하기 위해 NDSL의 축적된 2016년도의 국내논문의 데이터를 기반으로 분류별로 키워드를 추출하고 이를 통해 매칭 기반의 분류 모델을 만든 후 트레이닝 및 테스트를 거쳐 결과를 도출한다.

키워드: 추천, 분류, 키워드, 워드 클라우드, 학술연구분류체계, NDSL

ABSTRACT

Nowadays, due to an explosive increase in information, information filtering is very important to provide proper information for users. Users hardly obtain scholarly information from a huge amount of information in NDSL of KISTI, except for simple search. In this paper, we propose the service, PIN to solve this problem. Pin provides the word cloud including analyzed users' and others' interesting, co-occurence, and searched keywords, rather than the existing word cloud simply consisting of all keywords and so offers user-customized papers, reports, patents, and trends. In addition, PIN gives the paper classification in NDSL according to keyword matching based classification with the overlapping classification enabled-academic classification system for better search and access to solve this problem. In this paper, Keywords are extracted according to the classification from papers published in Korean journals in 2016 to design classification model and we verify this model.

Keywords: Recommendation, Classification, Keyword, Word cloud, the academic classification system, NDSL

* 본 연구는 2018년도 한국과학기술정보연구원(KISTI) 주요사업 과제로 수행한 것입니다.
** 한국과학기술정보연구원 박사후연구원(skyum@kisti.re.kr) (제1저자)
*** 한국과학기술정보연구원(wjkim@kisti.re.kr) (교신저자)
† 한국과학기술정보연구원{skyum, wjkim, tsyi, sybae}@kisti.re.kr
•논문접수: 2018년 11월 20일 •최초심사: 2018년 11월 27일 •게재확정: 2018년 12월 18일
•한국도서관정보학회지 49(4), 265-294, 2018. [http://dx.doi.org/10.16981/kliss.49.201812.265]

Ⅰ. 서론

인터넷의 발달로 인해 정보가 폭발적으로 증가함에 따라 정보를 이용하는 사람의 수가 증가하였고 단순 검색을 통한 정보의 획득이나 의미있는 정보를 판단하기가 매우 어려워졌다. 특히 부정확한 정보로 인해 신뢰도가 높은 정보를 찾기 위해 많은 시간을 투자하는 것이 강제되고 있는 실정이다.

따라서 사용자에게 적합한 정보를 제공하기 위한 정보의 필터링인 추천 (이락규 외 2011, 2)과 분류 (정영미 2005, 12-14)가 매우 중요시 되고 있다. 추천은 정보나 서비스가 개인 또는 특정집단의 요구에 맞게 적용된 개인화된 정보를 제공하는 형태를 말하며 이러한 기능을 하는 시스템을 추천 시스템 또는 추천 서비스라고 한다 (이락규 외 2011, 2). 추천 시스템은 사용자가 관심있어할 만한 정보를 선택해주며, 그렇지 않은 복잡하거나 불필요한 정보에 대해서도 걸러 주는 기능도 한다. 초기 추천 서비스는 인터넷 쇼핑몰이나 뉴스와 같은 취향을 분석하여 높은 선호도가 예상되는 아이템을 추천하였다면 최근에는 영화, TV, 음악 등 다양한 영역에서 추천 서비스들이 활용되고 있다 (Smeaton and Callan 2005, 3).

문서의 분류는 유사한 내용을 가진 문서들을 모아 집단화하는 작업이다 (정영미 2005, 12-14). 특히 연구자 입장에서 문서의 분류는 보다 신속하게 정보를 획득하고 연구의 효율을 높이는 데에 매우 중요한 작업이며, 분류화된 문서로 인해 직관적이고 쉽게 접근이 가능하다 (박창호, 염성숙, 이정모 2000, 2).

KISTI(한국과학기술정보연구원)의 NDSL(국가과학기술정보센터)은 국내 최대의 과학기술정보 포털 사이트로 제공자 입장에서 국내 과학기술 연구자들의 연구 분야는 중요한 이슈로 받아들이고 있다 (이태석 외 2012, 2). 하지만 방대한 자료를 보유함에도 불구하고 사용자들은 검색 외에 자료 획득이 쉽지가 않으며, 자료의 분류에 있어서도 도서관 DDC 기반의 분류를 하고 있으나 분류명칭의 모호함으로 인해 접근하기 쉽지 않고 최신 트렌드를 반영하기 어렵다. 또한 KDC나 LC 분류를 적용시키에도 도서 기반의 분류이기 때문에 학술정보를 분류함에 있어서 애매한 부분이 존재한다.

키워드는 문서의 내용을 정확하게 함축하여 나타내는 주요한 요소로서, 추출과 검색 등 다양한 정보연구에 활용되며 NDSL은 키워드 기반으로 자료를 검색을 하고 이로 인해 원하는 자료에 접근이 가능하다.

본 논문은 이러한 문제를 해결하고 사용자에게 적합한 정보를 제공 및 접근을 위해 이러한 키워드 특성을 활용하여 기존 NDSL을 개선한 추천 및 분류 서비스인 PIN (Profiling service In NDSL)을 제안한다.

PIN은 전체 키워드만을 가지고 구성된 것이 아닌 사용자 본인 및 유사 사용자가 등록한 관심 키워드, 동시이용 키워드, 검색 키워드를 통해 수집과 분석된 워드 클라우드를 제공하고

이를 통하여 사용자에게 맞춤형 논문, 보고서, 특허, 동향의 콘텐츠를 추천한다. 뿐만 아니라 사용자의 이력과 히스토리를 한눈에 파악함으로써 자신의 기록이 추천에 어떻게 영향을 주는지 확인이 가능하다.

또한 PIN은 콘텐츠를 보다 쉽게 접근하기 위하여 사용자에게 학술연구서비스 및 대학의 학과 분류에서 사용하는 한국연구재단의 학술연구분류체계를 바탕으로 한 논문 분류를 제공한다. 이를 검증하기 위해 기계학습을 통한 키워드 매칭 (McCallumzy and Nigamy 1999, 2-3) 기반으로 해당 논문의 키워드가 속한 분류체계에 따라 논문을 분류하고 테스트를 하였다. 키워드 매칭은 정보의 양이 너무 적은 경우에 적용되기 어려운 단점 (김은경, 최진오 2005, 1)이 있지만 많은 양의 데이터를 활용할 수 있는 NDSL의 데이터를 이용하기에 적합하다고 할 수 있다. NDSL의 2016년도의 국내논문의 데이터를 기반으로 저자 키워드를 추출하고 규칙 기반(Rule based)의 분류 모델을 만든 후 테스트셋을 이용하여 학술연구분류체계의 분류별로 얼마나 정확히 분류되는지 결과를 도출하였다.

본 논문의 구성은 다음과 같다. 2장은 관련 연구, 3장은 이 연구를 위한 배경을 설명한다. 4장은 제안하는 추천 시스템 및 분류에 대해 기술하고 5장의 결론으로 마무리를 한다.

Ⅱ. 관련 연구

추천 시스템은 크게 내용 기반 필터링 (Pazzani and Billsus 2007, 1-2), 협업 필터링 (Goldberg et al. 2001, 1), 이 기법을 합한 혼합형 (Burke 2002, 1-2)으로 나뉘게 되며 각 추천 시스템이 사용하는 기법들을 <표 1>에 정리를 하였다. 유사성의 대상에 따라 내용기반, 협업, 혼합형으로 나뉘고 과거 선호도의 이용여부에 따라 메모리 기반과 모델 기반으로 나뉜다. 메모리 기반은 다시 사용자 중심인지 아이템 중심인지에 따라 사용자 중심과 아이템 중심으로 나뉜다 (손지은 외 2015, 8-13; 여운동 외 2010, 4-6; Das et al. 2007, 1-2; Adomavicius and Tuzhilin 2005, 1).

내용 기반 필터링 기법은 정보 검색이나 정보 필터링 연구로부터 발전한 것으로, 아이템과 아이템 또는 아이템과 사용자간의 유사도를 측정하고 그 결과를 순위화하여 추천해준다 (Wu and Chen 2000, 1-2). 이 방식은 영화, 음악, 도서 및 텍스트 기반의 뉴스나 인터넷 기사 등을 추천하는데 널리 쓰이고 있다. 하지만 텍스트와 무관한 주제에 대해서는 적용을 할 수 없고 새로 추가된 아이템인 경우 평가가 없으므로 추천이 불가능하며, 취향이나 선호도를 반영하지 못하기 때문에 다양성을 보장할 수 없는 단점이 있다 (Balabanović and Shoham 1997, 1-2).

협업 필터링은 추천 시스템 중에서 현재까지 가장 우수한 성능 보이고 가장 널리 알려진 기법이다. 협업 필터링은 사용자의 선호도를 수집하여 데이터베이스를 구축하고, 유사한 취

〈표 1〉 추천 시스템 분류

필터링	기법	
내용 기반	- TF-IDF	
협업 필터링	메모리 기반	- Nearest neighborhood - Graph theory
	모델 기반	- Bayesian classifiers - Clustering - Neural networks - Dimensionality reduction - Latent semantic - Regression-based - MDP-based - Decision tree
혼합형	- Linear combination - Voting schemes - Incorporating one component	

향의 특정 사용자들의 데이터를 찾아 이들이 선호하는 아이템을 사용자에게 추천한다. 내용 기반 필터링은 사용자의 아이템의 정보만 의존하여 선호도를 예측하는 반면, 협업 필터링은 유사 사용자가 아이템에 대해 평가한 정보를 사용하여 선호도를 예측하는 것이 차이점이다 (여운동 외 2010, 4-6). 유사 취향의 사용자들의 데이터를 이용하기 때문에 다양한 아이템을 추천받을 수 있다. 이러한 협업 필터링은 메모리 기반 필터링과 모델 기반 필터링으로 나뉜다 (Breese, Heckerman, and Kadie 1998, 2-5).

메모리 기반 필터링은 앞서 설명한 방식으로 유사 사용자의 선호도를 이용하여 추천해주는 방식이다. 여기서 선호도를 계산하는 기준이 사용자인지 아이템인지에 따라 사용자 기반 또는 아이템 기반 필터링으로 나뉘어진다. 모델 기반 필터링은 군집화, 분류, 예측의 단계에서 기계학습 또는 데이터마이닝 기법을 활용하는 것이며, 베이시안, 뉴럴 네트워크, 차원 축소, 회귀분석, 결정트리 등이 있다.

하지만 협업 필터링은 새로운 사용자가 시스템에 가입 하였을 때 사용자의 정보가 없어서 이웃을 찾을 수 없게 되는 문제와, 신규 아이템에 대해 평가가 없기 때문에 발생하는 문제가 존재한다. 또한 기계학습을 하는 만큼 상당히 많은 계산이 필요하다.

내용 기반 접근방식과 협업 필터링은 각각 장·단점을 가지고 있으며 이러한 문제를 해결하기 위해 최근에는 각 방식의 장점을 높이면서 단점은 보완하는 혼합형 필터링에 대한 연구가 이루어 지고 있다 (Burke 2002, 1-2). 이러한 혼합형 필터링은 다양한 정보를 활용하며 이 과정에서 내용 기반 및 협력 필터링의 방법들을 결합한다. 하지만 각 필터링으로부터 획득할 수 있는 정보는 다르기 때문에, 혼합형 추천 시스템을 설계하기 위해서는 명확하게 목적을 정의하고, 이에 따른 적합한 방식과 데이터를 이용해야 한다.

학술연구분류체계 (학술연구분류체계, https://www.nrf.re.kr/biz/doc/class/view?menu

_no=323)는 한국연구재단이 학술연구지원사업을 효율적으로 하기 위하여 <표 2>와 같이 학술연구분야를 정리해 놓은 것이다. 이 학술연구분류체계를 실제로 학술연구서비스뿐만 아니라 대학의 학과 구분도 이 체계에 준해서 구분을 하고 있으며, 한국대학교육협의회에서 사용하는 대학입학정보 포탈의 모집단위별 분류 체계도 여기 기준을 따르는 구조로 되어있다. 인문학, 사회과학, 자연과학, 공학, 의약학, 농수해양학, 예술체육학의 7개 대분류와 위의 분류 중 2가지 이상이 섞여있는 복합학을 포함하여 총 8가지의 대분류로 분류되며, 대분류(8)-중분류(152)-소분류(1,551)-세분류(2,468)의 4개 단계로 수직적으로 구성되어 있다. 이 체계의 개선을 위한 연구는 크게 세차례 있었으며, 현재 사용되고 있는 분류는 1999년 11월 수행된 연구결의 내용과 큰 차이가 없으며, 큰 개편없이 사용되고 있다.

〈표 2〉 학술연구분류체계

대분류	중분류	소분류	세분류
인문학	23	167	298
사회과학	22	269	479
자연과학	13	135	371
공학	28	310	457
의약학	39	409	648
농수해양학	7	64	132
예술체육학	12	104	61
복합학	8	93	22
합계	152	1,551	2,468

제안하는 PIN은 기존의 단순 검색 서비스에서 다수의 적합한 정보를 제공하기 위해 추천과 분류를 제공한다. 추천 서비스는 사용자 본인과 유사 사용자의 관심 키워드 및 검색 키워드로 구성되는 혼합형 필터링을 이용한 워드 클라우드 기반 추천 서비스를, 분류는 학술연구분류체계를 이용하여 기계학습 기반 비교 마이닝 기법인 키워드 매칭을 활용한 분류를 제공한다.

Ⅲ. 배경

1. NDSL

National Digital Science Library (NDSL, 국가과학기술전자도서관)은 국내 연구자들에게 필요한 해외 학술저널 및 프로시딩 서비스에 대한 대책마련과 전자저널의 급속한 확산에 따른 디지털 콘텐츠 확충과 서비스 개발이 요구되는 시대적 요청에 따라 지난 2001년부터

서비스를 개시 (이주현, 이응봉, 김환민 2006, 3; NDSL Homepage, http://www.ndsl.kr) 하였고, 국내 학계, 연구계, 산업계의 모든 연구자들을 위한 해외 학술저널 및 프로시딩 포털로서 2018년 11월 현재 약 9천만건의 논문과 3천 7백만건의 특허, 31만건의 보고서와, 5만건의 동향, 38만건의 저널/프로시딩을 비롯하여 86만건의 연구자 데이터, 35만건의 연구기관 데이터를 보유하고 있으며 연구자 중심의 서비스로 거듭나기 위해 추천을 비롯한 서비스 개선에 매진하고 있다. <그림 1>은 현재 NDSL의 메인 페이지이며 검색을 비롯하여 논문, 보고서, 특허, 동향 등의 콘텐츠를 제공한다.

메인 페이지는 화면 중간의 검색과 메인 화면과 우측의 NDSL의 콘텐츠를 표기하는 표로 구성되어 있다. 사용자는 화면 상단의 검색바를 통해 약 1억건의 콘텐츠를 검색하여 자료를 획득할 수 있다. 메인 화면은 과학기술동향과, 공지사항과 KISTI에서 출판하는 과학도서 '과학향기'를 빠르게 접근을 가능하며, 이슈&NDSL은 현재 최신의 이슈를 정리하여 보여준다. 오른쪽 상단에 로그인 버튼이 있으며 로그인 시에 원문 복사 신청이 가능하다. 로그인 버튼 옆에 MList를 통해 내계정, About, 사이트맵, 고객센터, 공지사항, 이슈&NDSL, 과학향기, iCON, DB현황에 접근이 가능하다. 내계정은 로그인, About은 NDSL 소개, iCON은 KISTI에서 운영하는 정보서비스 동향지식 포럼으로의 링크이다.

<그림 1> NDSL 메인 페이지

2. 사용자 프로파일 기반 서비스

사용자 프로파일 기반 서비스는 사용자의 정보를 프로파일에 저장하고 관리하며 사용자의 프로파일을 이용하여 개인의 성향을 분석하고 추론한다 (김광영, 곽승진 2011, 4). 사용자

프로파일 수집 방법에는 명시적 방법과 암시적 방법이 있다. 명시적 방법은 관심 정보나 평가 등급과 같이 피드백 정보를 요청하는 방법이며, 명시적 방법은 온라인 문서를 읽는 시간과 같은 사용자의 행동을 관찰하는 방법이다 (Speretta and Gauch 2005). 현재 검색 시스템 대부분은 사용자 프로파일을 이용하여 추천을 제공해주고 있으며, 이를 위해 퍼지 개념 네트워크, 온톨로지, 커뮤니티, 태그 카테고리, 사용자 브라우징 데이터, 폭소노미, 소셜관계 분석 등을 활용하고 있다 (김광영, 곽승진 2011, 4-5). 특히 유사한 학술 검색 서비스인 DBPia (DBPia, http://www.dbpia.co.kr) 의 경우 사용자의 다운로드 기록을 이용하여 추천을 하고 있으며, 신규 발행기관/저널, 검색결과, 저자, 논문피인용, 맞춤 추천 등을 알림 서비스로 제공한다. 네이버 학술정보 (네이버 학술정보, https://academic.naver.com) 의 경우 사용자 프로파일 서비스를 지원하지 않는 대신에 축적된 학술 데이터를 분석하여 분야별/키워드별 연구 트렌드를 제공한다.

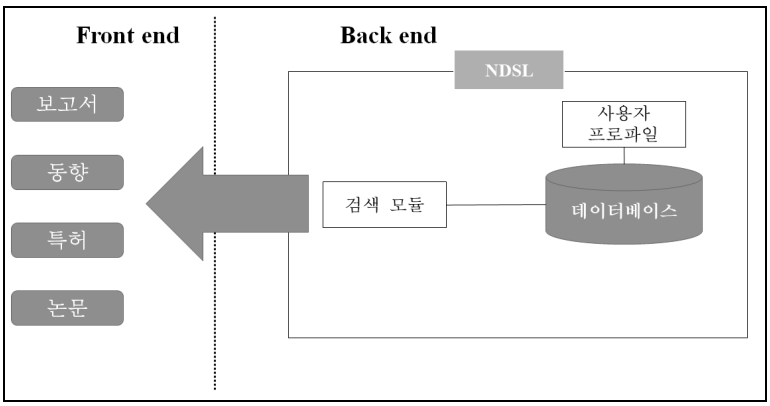
3. 워드 클라우드 서비스

워드 클라우드는 대표적인 텍스트 시각화 기법중 하나로 빈도에 따라 문자의 크기를 결정함으로써 해당 문자를 강조하는 기법이다. 직관적으로 빈도를 알 수 있는 장점이 있으며 이로 인해 키워드 분석시에 사용되어 지고 있다. SNS 키워드, 텍스트 마이닝, 연구 프로젝트 보고서, 교과서, 뉴스 데이터를 분석한 연구 등 다양한 분야의 연구에 사용되고 있다 (정덕영, 이준석, 박상성 2016, 3; 김남규, 이동훈, 최호창 2017, 8-9). DBPia (DBPia, http://www.dbpia.co.kr) 는 단순 추천이 아닌 분류별로 키워드 빈도가 높은 순으로 워드 클라우드를 지원한다.

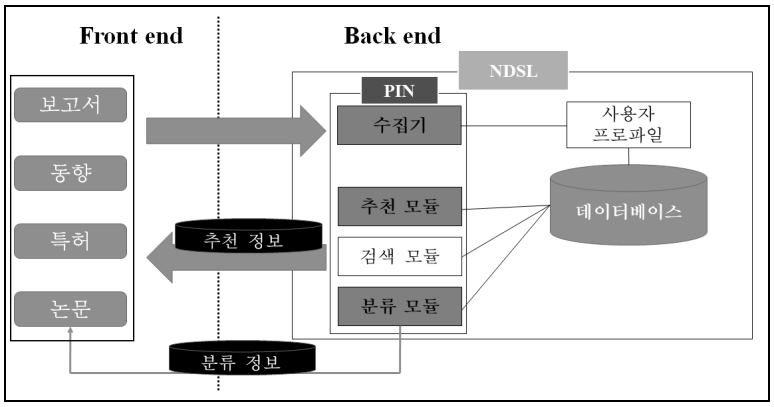
Ⅳ. PIN 서비스 모델

검색 서비스의 사용자 프로파일 이용은 필수가 되었으며, 이를 적용하여 PIN은 기존 NDSL에 사용자 프로파일 기반의 워드 클라우드를 활용한 추천을 제공하며, 논문 분류를 통해 정보의 접근성을 높인다. 서비스 <그림 2>의 (a)와 (b)는 기존 서비스의 모델과 제안하는 PIN 서비스 모델을 보여준다. 기존 모델의 경우, 사용자는 데이터베이스에 연결된 검색 모듈을 통해 논문, 보고서, 특허, 동향 등의 학술 정보 컨텐츠를 확인할 수 있었다. PIN은 기존 모델과 비교하여 데이터를 수집하는 수집기, 추천을 위해 데이터를 가공하는 추천기, 분류를 위한 분류기 모듈을 포함한다. 사용자는 검색을 통해 콘텐츠 데이터를 이용하게 되면 수집기로 데이터를 기록하고 로그인/비로그인 사용자에 따라 데이터를 수집 및 가공하게 되며 다시 추천과 논문의 분류 형태로 사용자에게 제공된다. 추천의 경우, 로그인 사용자에게는 로그

인 사용자가 자주 이용하는 키워드에 따라 콘텐츠를 추천하고 비로그인은 모든 사용자가 이용한 기록을 이용한다. 논문 분류의 경우 로그인/비로그인과 상관없이 데이터베이스에 주기적으로 마이그레이션된 논문 데이터를 갱신하여 분류기로 분류된 데이터를 사용자에게 제공한다.



(a) 기존 서비스 모델



(b) PIN 서비스 모델

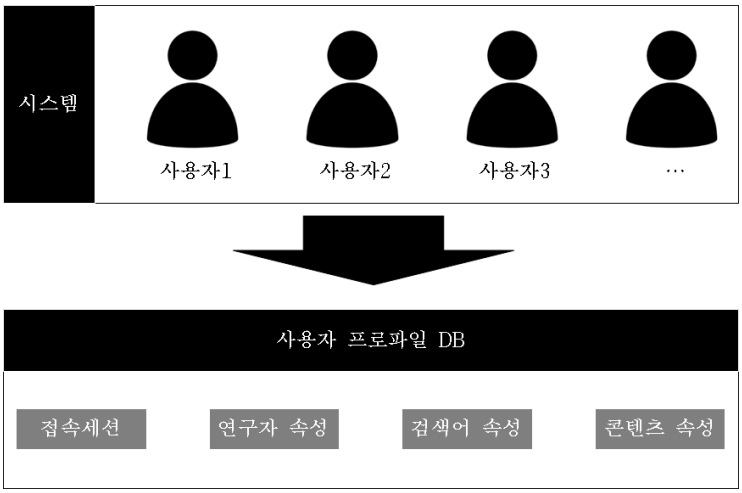
<그림 2> 기존 및 PIN 서비스 모델

1. 추천 서비스

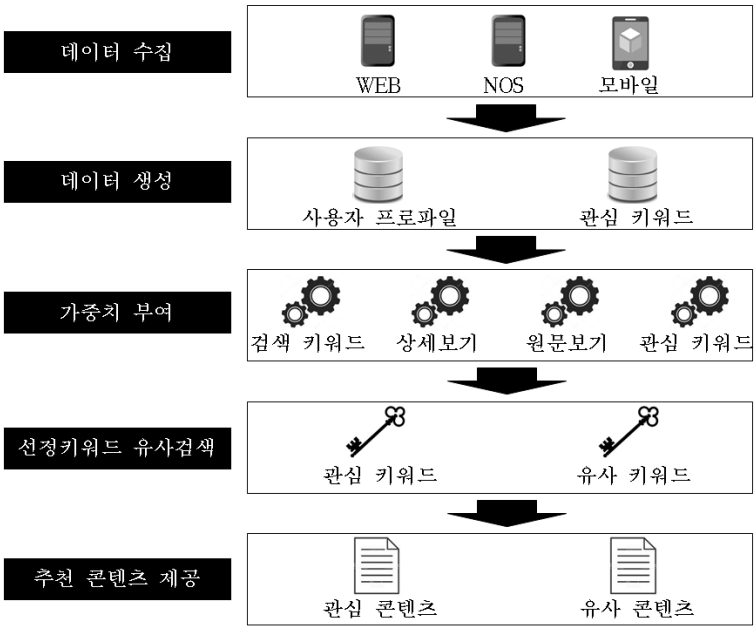
가. DB 및 처리 프로세스

NDSL에서 기존 단순 검색위주의 서비스에서 벗어나 추천 서비스의 필요성을 느끼고 사용자 프로파일 기반의 추천 서비스를 제안한다. 해당 서비스를 위하여 기존에 축적된 사용자 로그 데이터를 정리하기 위해 사용자 데이터베이스를 개선하고 앞으로 추천에 데이터를 활용하기 위해 '사용자 프로파일 DB'를 만든다. <그림 3>은 NDSL에서의 수집을 위한 사용자

프로파일 DB 속성이며, NDSL은 이를 통해 사용자들의 접속 세션, 연구자 속성, 검색어 속성, 콘텐츠 속성을 수집한다. 2018년 11월 현재 총 건수 약 9백만건이며, 일평균 16만건이 축적된다. 이를 DB를 토대로 사용자 맞춤형 콘텐츠 제공기반마련의 토대가 되어 처리 프로세스를 거쳐 사용자에게 추천 콘텐츠를 제공한다.



〈그림 3〉 사용자 프로파일 DB 속성



〈그림 4〉 처리 프로세스

<그림 4>은 추천 서비스의 처리 프로세스를 보여준다. WEB과 NOS (NDSL Open Service, API로 NDSL에 접근하는 서비스), 모바일로부터 사용자 및 키워드의 데이터를 수집하고 추천 서비스를 위해 사용자 프로파일 및 관심 키워드를 분석하여 데이터를 생성한다. 이 생성된 데이터에서 검색 키워드, 상세보기, 원문보기, 관심 키워드를 추출하여 동시발생빈도 룰 (Yutaka and Ishizuka 2004, 1-2)을 적용하여 데이터를 분석하고, 중요도에 따라 가중치를 부여하며, 가중치가 적용된 선정 키워드인 관심 키워드와 유사한 타 사용자의 관심 키워드 및 동시이용 키워드, 검색 키워드 등의 유사 키워드를 활용하여 관심 콘텐츠와 유사 콘텐츠를 추천한다.

나. 비로그인 및 로그인 추천

PIN의 추천은 크게 로그인 추천과 비로그인 추천이 있으며, 로그인 추천시 PIN 전용 페이지에 접근이 가능하다. 메인 페이지와 PIN 전용 페이지는 사용자들의 최신 검색 키워드와 사용자 관심 키워드를 적용한 워드 클라우드 기반의 추천을 보여주며, 논문 클릭시에 나타나는 '원문보기'는 로그인과 상관없이 해당 논문의 원문을 보여준다. <표 3>은 제안하는 PIN의 비로그인/로그인별 메인 페이지, PIN 전용 페이지, 원문보기에서의 추천 서비스를 정리한 것이다. 여기서 NDSL에서의 콘텐츠란 논문, 특허, 보고서, 동향을 말한다.

<표 3> PIN이 제공하는 추천 서비스

	비로그인	로그인
메인 페이지	- 최신 콘텐츠 - 인기 콘텐츠	- 개인화된 추천 콘텐츠 - 개인화된 추천 콘텐츠와 함께 이용한 콘텐츠
PIN 전용 페이지	이용 불가	- 개인화된 추천 콘텐츠 - 개인화된 추천 콘텐츠와 함께 이용한 콘텐츠 - 나의 콘텐츠 이용률 - 접속 이력 - 검색키워드 히스토리 - 상세보기 히스토리 - 원문보기 히스토리
원문보기	- 개인화된 추천 콘텐츠 - 동시이용 콘텐츠 - 저자의 다른 논문 - 참고문헌 - 이 논문과 함께 출간된 논문	

비로그인 추천은 모든 사용자에게 공통으로 최신 및 인기 콘텐츠를 추천하는 것이며, 메인 페이지 중앙에 위치하였기 때문에 쉽게 접근이 가능하다. <그림 5>는 비로그인시에 표시되는 메인 페이지를 보여준다. 비로그인시의 워드 클라우드는 기존에 NDSL을 사용했던 축적된 로그로부터 전체 사용자들이 가장 많이 검색한 키워드를 기반으로 구성된다. 비로그인시의

워드 클라우드에서 기본적인 키워드 색은 파란색이며, 색이 진할수록, 크기가 클수록 검색 빈도수가 높은 키워드이다. 키워드가 부족하거나 트렌드를 반영하기 위해 관리자 키워드를 입력할 수 있으며 이 키워드는 노란색으로 표시된다. 현재 <그림 5>에서는 '메르스', '바지락', '동치미', '재조합 단백질 생산' 순으로 랭크되어 화면에 출력되었고 '4차 산업혁명'과 '인공지능'은 관리자 키워드로 표시되었다. 이러한 워드 클라우드를 바탕으로 빈도 기반의 최신 콘텐츠와 인기 콘텐츠를 워드 클라우드 하단에 추천해준다. 사용자가 검색에 이용한 질의어에서 추출한 최신 키워드는 워드 클라우드 우측 상단의 '실시간 검색어' 영역에서 확인할 수 있다. 워드 클라우드 상단의 논문, 특허, 보고서, 동향 탭을 선택하면 워드 클라우드 및 최신과 인기 콘텐츠가 재구성된다. 탭을 선택하지 않아도 실시간 결과를 반영하여 30분단위로 워드 클라우드 및 해당 콘텐츠가 갱신된다.

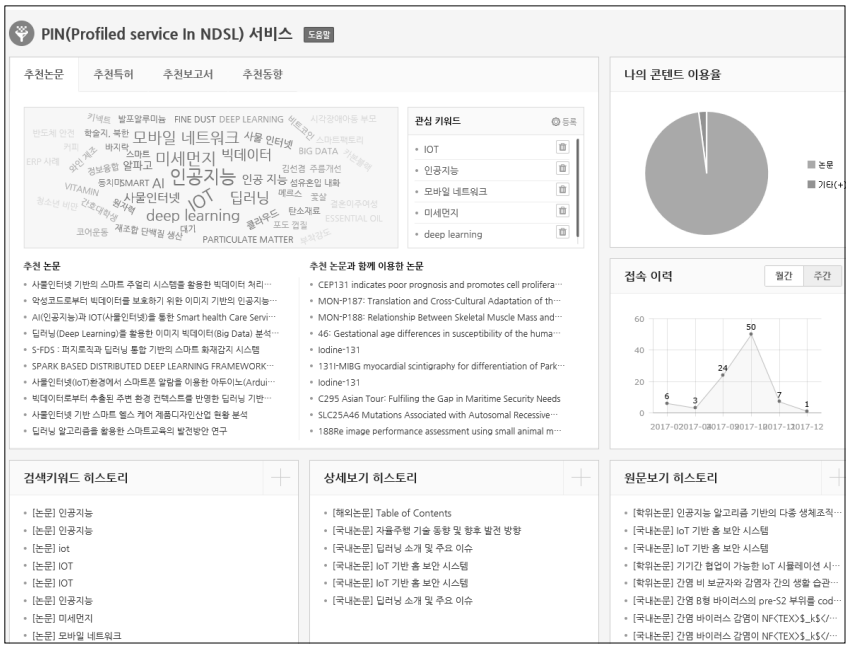
<그림 5> 비로그인 - 메인 페이지

<그림 6>은 로그인시의 메인 페이지를 보여준다. 메인 페이지에서는 사용자가 검색하고 등록하는 키워드에 따른 추천 콘텐츠를 제공하고 있으며, 비로그인시의 실시간 검색어 영역이 사용자가 직접 등록할 수 있는 '관심 키워드' 영역으로 바뀐다. 로그인시의 워드 클라우드는 로그인 사용자의 관심 키워드, 동시이용 키워드, 검색 키워드와 전체 사용자의 동시이용 키워드, 전체 검색 키워드로 구성된다. 워드 클라우드에 관심 키워드를 적용시키려면 사용자는 관심 키워드를 최소 3개 이상 등록을 해야 한다. <그림 6>에서 예시로 관심 키워드에 'IOT', '인공지능', '모바일 네트워크', '미세먼지', 'deep learning'을 등록하였으며, 워드 클라

우드에서는 빨간색으로 표시된다. 특히 사용자 관심 키워드가 워드 클라우드 구성 키워드 중 가장 높은 가중치가 적용되고 각 키워드도 검색빈도에 따라 가중치 값이 적용이 되기 때문에 '인공지능', 'IOT', '미세먼지', '모바일 네트워크', 'deep learning' 순으로 랭크값이 정해지며 워드 클라우드에 표시된다. 해당 관심 키워드와 관련한 동시이용 키워드로 추출된 '딥러닝', '알파고', '사물인터넷', 'FINE DUST', '클라우드', '빅데이터' 등이 구성 키워드 중 두 번째 큰 가중치 값이 적용이 되어 초록색으로 표시된다. 가중치 값이 사용자가 등록하는 키워드 중 가장 작은 사용자 검색 키워드인 '메르스', '바지락', '동치미', '재조합 단백질 생산'과 비로그인시 화면에 나타났던 키워드들이 가장 작은 크기로 파란색으로 표시된다. 앞서 입력한 사용자가 등록한 관심 키워드로 검색되는 논문들을 워드 클라우드 하단의 '추천 논문'으로 나타내며 다른 사용자들이 사용하였던 동시에 이용한 동시이용 키워드를 통해 획득된 논문을 '추천 논문과 함께 이용한 논문'으로 보여진다. 워드 클라우드 상단에 있는 탭을 클릭하면 추천 논문, 추천특허, 추천보고서, 추천동향을 추천해주며, 우측 상단에 있는 추천 및 전체 탭으로 비로그인의 추천과 로그인의 추천을 자유롭게 이동할 수 있다.

〈그림 6〉 로그인 - 메인 페이지

<그림 7>는 PIN 전용 페이지를 보여준다. PIN 전용 페이지는 메인 페이지와는 달리 로그인 사용자만 이용할 수 있으므로 비로그인 추천 탭으로는 진입하지 못한다. PIN 전용 페이지는 추천뿐만 아니라 사용자가 NDSL을 어떻게 이용하고 있는지 확인할 수 있다. 콘텐츠 이용률은 사용자가 논문, 특허, 보고서, 특허, 동향 뿐만 아니라 과학향기, 이슈&NDSL 등의 NDSL



〈그림 7〉 로그인 - PIN 전용 페이지

에서 활용되는 모든 콘텐츠들 중 어떠한 콘텐츠를 가장 많이 이용해왔는지 알 수 있으며, 접속 이력은 언제 얼마만큼 동안이나 NDSL을 사용하였는지 월별, 주별로 확인할 수 있다. 그 외에 검색키워드, 상세보기, 원문보기의 히스토리를 제공함으로써 사용자는 본인이 어떠한 키워드를 검색하고 어떠한 콘텐츠를 상세보기 및 원문보기를 하였는지 파악이 가능하며, 이로인해 자신의 기록이 추천에 어떻게 영향을 주는지 알 수 있다.

마지막으로 해당 콘텐츠를 클릭할 때 표시되는 원문보기는 로그인과 상관없이 사용자가 보고자 하는 콘텐츠의 원문을 PDF로 보여주며, 화면 우측에 이와 관련된 '이 논문과 함께 이용한 콘텐츠', '연관 콘텐츠', '저자의 다른 콘텐츠', '참고 문헌', '이 논문과 함께 출판된 콘텐츠'를 보여준다. 단, 원문보기의 콘텐츠는 데이터의 특성상 논문과 보고서만 해당되며 특허와 동향은 제외한다. '이 논문과 함께 이용한 콘텐츠'는 다른 사용자가 해당 논문을 보고 다른 콘텐츠를 봤을 때의 동시발생에 따른 추천이며, '연관 콘텐츠'는 해당 콘텐츠가 보유한 키워드로 검색시에 획득한 콘텐츠이다. '저자의 다른 콘텐츠'는 원문의 저자가 쓴 다른 콘텐츠를 추천하며, '참고 문헌'은 해당 콘텐츠의 참고문헌 리스트, '이 논문과 함께 출판된 콘텐츠'는 해당 논문이 출판될 때 동시에 출판된 콘텐츠들의 리스트를 보여준다. 원문보기 클릭할 때마다 처리가 되는 것이 아닌 데이터베이스에 관련 데이터가 기록되어 있으며, 해당 추천 데이터가 없을 때에는 제공하지 않는다.

다. 워드 클라우드

앞서 서술한대로 비로그인과 로그인 모두 워드 클라우드 기반의 추천을 하고 있다. 본 장은 이 워드 클라우드가 어떻게 구성되고 갱신이 되는지 알아본다. <표 4>은 이 워드 클라우드를 구성하는 키워드를 보여준다. 워드 클라우드는 총 50개의 키워드로 이루어져 있으며, 기존의 단순히 전체 검색 키워드만으로 구성되는 것이 아닌, 사용자의 관심 키워드, 동시이용 키워드, 사용자 검색 키워드, 전체 사용자 검색 키워드로 이루어진다. 관심 키워드는 사용자가 등록한 키워드이며, 동시이용 키워드는 사용자 또는 전체 사용자가 동일한 관심 키워드와 함께 검색한 키워드를 말하며 관심 키워드는 최소 3개, 동시이용 키워드는 관심 키워드별 최대 5개까지 저장된다. 동시이용 키워드의 유사도의 랭크는 동시발생빈도 행렬 (Yutaka and Ishizuka 2004)로 구하며 <표 5>는 <그림 6>에서 입력하였던 사용자 관심 키워드를 가지고 동시이용 키워드를 구한 동시발생빈도 행렬의 예이다.

키워드는 'IOT', '인공지능', '사물인터넷', '딥러닝'이며 행렬 안의 값은 동시에 발생한 빈도를 의미한다. 'IOT'의 경우 '사물인터넷'이 5번으로 가장 높았고 '인공지능'은 '딥러닝'이 7번으로 가장 높았기 때문에 'IOT'와 '인공지능'의 동시이용 키워드는 각각 '사물인터넷'과 '딥러닝'이 된다. 사용자의 관심 키워드와 실시간으로 새로 입력되는 키워드를 이용하여 동시발생빈도의 값을 DB에 기록하고 이를 워드 클라우드가 갱신될 때마다 구성에 포함된다.

〈표 4〉 워드 클라우드 구성 키워드

구분	크기	대상
관심 키워드(x)	사용자당 3개 이상	- 사용자가 등록한 키워드
동시이용 키워드(y)	관심 키워드 별 최대 5개	- WEB과 모바일 검색 키워드 - NOS 검색키워드 제외
사용자 검색 키워드(z)	전체	- WEB과 모바일 검색 키워드 - NOS 검색키워드 제외 - 불용키워드 제외
전체 사용자 검색 키워드(k)	전체	- WEB과 모바일 검색 키워드 - NOS 검색 키워드 제외 - 불용키워드 제외

〈표 5〉 동시발생빈도 행렬

	IOT	AI	사물인터넷	딥러닝
IOT	4	1	5	1
AI	1	2	2	7
사물인터넷	5	2	3	2
딥러닝	1	7	2	3

검색 키워드는 검색한 키워드 전체를 말한다. 동시이용 키워드와 검색 키워드는 모두 불용 키워드 및 NOS 검색 키워드를 제외하고 WEB과 모바일에서 검색된 키워드를 바탕으로 수집된다. 이 키워드를 바탕으로 비로그인시에는 전체 사용자 검색어로만 구성된 워드 클라우드 기반으로 추천을 하고 로그인시에는 관심 키워드, 동시이용 키워드, 사용자 검색 키워드, 전체 사용자 검색 키워드를 이용한다. 식 1은 로그인시의 사용자 i의 워드 클라우드 구성 집합 U_i을 나타낸다.

$$U_i = \left\{\{x_i \cdot w_1\}, \left\{(y_{ii} + \sum_{j=1}^{n} y_{ij}) \cdot w_2\right\}, \{z_i \cdot w_3\}, \{k \cdot w_4\}\right\} \quad (1)$$

단, $w_1 = 5000$, $w_2 = 1000$, $w_3 = 100$, $w_4 = 1$

로그인시에는 사용자 i의 워드 클라우드 U_i는 사용자 i 본인의 모든 관심 키워드 x_i와 사용자 검색 키워드 z_i 및 사용자의 관심 키워드에 대한 동시이용 키워드 y_{ii}를 추출하고, 사용자와 동일한 관심 키워드를 가지는 유사 사용자의 동시이용 키워드 y_{ij}와 전체 사용자 검색 키워드 k로 이루어진다. 사용자 전체 검색어 k는 사용자 전체 검색 관련이므로 식별자를 붙이지 않았다. 키워드별 가중치 $w_1 \sim w_4$가 부여되며 관심 키워드 5,000점, 동시이용 키워드 1,000점, 사용자 검색 키워드 100점, 전체 사용자 검색 키워드 1점의 가중치를 부여 받아 랭크를 정하고 랭크에 따라 워드 클라우드에 출력한다.

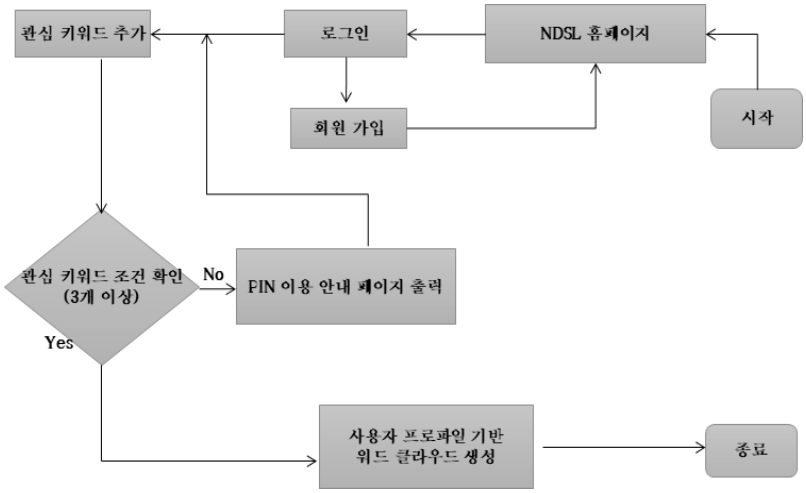
<표 6>은 이 워드 클라우드에서 키워드 선정 방법을 나타낸다. 워드 클라우드는 매 5일치마다 데이터를 정리 한다. 앞서 언급한대로 비로그인시 워드 클라우드는 관리자 등록 키워드와 전체 사용자 검색 키워드로 구성되고, 로그인시의 워드 클라우드는 관심 키워드, 동시이용 키워드, 사용자 검색 키워드, 전체 사용자 검색 키워드로 이루어진다. 동시이용 키워드는 신규 검색어는 등록시마다, 기존 키워드는 1시간마다 갱신한다. 실시간 키워드는 실시간 단위로 조회하며 워드 클라우드는 30분 단위로 갱신하며 비로그인시의 관리자 키워드와 로그인시의 사용자 키워드는 실시간으로 갱신한다.

〈표 6〉 키워드 선정 방법

구분	크기	대상	배치	가중치 점수측정
워드 클라우드 키워드 (비로그인)	D-5 일	- 관리자 등록 키워드 - 전체 사용자 검색 키워드	- 30분 단위로 데이터 갱신 (전체 사용자 키워드) - 관리자 등록 키워드는 실시간 수집	- 관리자 키워드는 직접 입력 - 전체 사용자 키워드 1점
워드 클라우드 키워드 (로그인)		- 관심 키워드(최소 3개 이상) - 동시이용 키워드(관심키워드별 최대 5개) - 사용자 검색 키워드 - 전체 사용자 검색 키워드	- 30분 단위로 데이터 갱신 (전체 사용자 키워드) - 사용자 키워드는 실시간 수집(관심, 동시이용, 검색 키워드)	- 관심 키워드 5,000점 - 동시이용 키워드 1,000점 - 사용자 검색 키워드 100점 - 전체 사용자 검색 키워드 1점

라. 추천 서비스 전개 절차

PIN은 워드 클라우드 기반의 비로그인/로그인 추천을 제공하며, 사용자들은 비로그인시에는 접근이 쉽지만 로그인시에 맞춤형 추천 서비스를 이용하기 위해서는 절차를 따른다. <그림 8>은 해당 서비스 전개 절차를 보여준다. 비로그인시에 절차는 메인 페이지에서 바로 추천 정보를 확인할 수 있기 때문에 제외시킨다.



<그림 8> 서비스 전개 절차

사용자는 먼저 NDSL 홈페이지 들어가 로그인시의 추천을 이용하기 위해 로그인을 한다. 만약 로그인을 위한 ID가 없다면 가입절차에 따른 ID를 생성한다. 관심 키워드가 없다면 추가한다. 3개 이상의 관심 키워드가 존재하지 않는다면 워드 클라우드도 비로그인시의 워드 클라우드와 동일하게 출력한다. 조건이 충족이 된다면 관심 키워드와 그전에 이용하였던 검색 키워드와 관심 키워드를 통해 검색하였던 동시이용 키워드 및 검색 키워드를 기반으로 워드 클라우드를 생성한다. 이로써 사용자는 메인 페이지 또는 PIN 전용 페이지에서 추천 서비스를 이용할 수 있게 된다.

2. 논문 분류

가. 학술연구분류체계

제안하는 논문 분류는 한국연구재단 학술연구분류체계를 따른다. 단, NDSL에 적용시킨 중분류의 경우, DBPia (DBPia, http://www.dbpia.co.kr)와 네이버 학술정보 (네이버 학술정보, https://academic.naver.com)를 참고하였으나, 키워드의 특성과 논문 데이터를 고려하여 다소 변경시켰다. 소분류 이하는 고려하지 않았다. <표 7>은 NDSL에서 적용시킨 분류이다.

〈표 7〉 NDSL에 적용된 한국연구재단 기반 분류

대분류

인문학	사회과학	자연과학	공학	의약학	농수해양학	예술체육학	복합학

중분류 - 인문학

인문학일반	역사학	철학	종교학/신학	언어학	문학	한국어문학	중국어문학
일본어문학	영어문학	프랑스어 문학	독일어문학	러시아어 문학	서양고전어 문학		

중분류 - 사회과학

사회과학일반	경제학	경영학	관광/무역학	교육학	군사학	민속학	법학
사회/ 사회복지학	신문방송학	정치외교학	지리/ 지역학	행정학			

중분류 - 자연과학

자연과학 일반	물리학	생물학	식물학	생활과학	수학	지구과학/ 천문학	통계학
화학							

중분류 - 공학

공학일반	건축공학	기계공학	산업공학	원자력공학	자원공학	재료공학	금속공학
전자전기/제어계측공학	전자/ 정보통신	컴퓨터공학	조선/ 해양공학	토목공학	환경공학	화학/ 생물공학	생명공학
항공우주 공학	교통/도로/ 철도공학	안전공학					

중분류 - 의약학

간호학	수의학	약학	의학	내과학	보건학	정신건강의학	물리치료학
치의학	한의학						

중분류 - 농수해양학

농학	해양학	수산학	식품과학	임학	조경학	축산학	해상운송학

중분류 - 예술체육

예술체육일반	건축	디자인	미술	미용	연극	영화	음악
의상	체육						

중분류 - 복합학

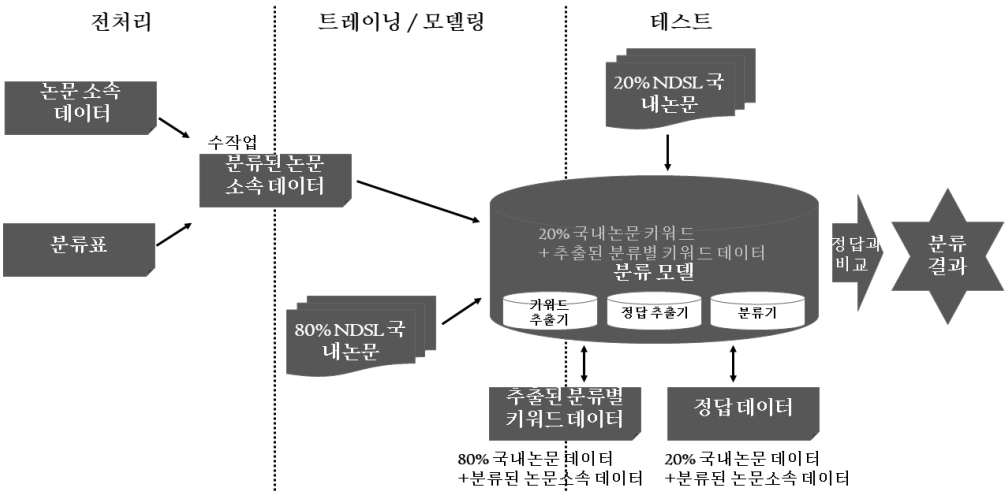
복합학일반	감성과학	과학기술학	문헌정보학	인지과학

인문학은 인문학일반부터 서양고전어문학까지 14개의 중분류로, 사회과학은 사회과학일반부터 행정학까지 13개의 중분류로 분류한다. 자연과학은 자연과학일반부터 화학까지 9개, 공학은 공학일반부터 안전공학가지 19개, 의약학은 간호학부터 한의학까지 10개로 분류한다. 농수해양학은 농학부터 해상운송학까지 8개, 예술체육학은 예술체육일반부터 체육까지 10개, 복합학은 복합학일반부터 인지과학까지 5개로 분류한다. 이렇게 분류된 8개의 대분류와 88개의 중분류를 제안하는 모델에 적용시킨다. PIN에 적용시에 워드 클라우드 자리 위치하며 워드 클라우드와 함께 선택 탭으로 원하는 화면 방식을 선택할 수 있게 한다.

나. 분류 방법

제안하는 분류 모델은 학술분류체계 기반 논문이 가지고 있는 저자 키워드를 이용하여 논문을 분류하며, 논문 분류시 기계학습에 따른 키워드 매칭 모델을 이용하여 논문을 분류시킨다. 전처리, 모델링, 테스트의 세단계로 나뉜다. <그림 9>은 이 모델의 처리과정을 나타내며 이는 데이터 마이닝시에 기계학습 및 테스트와 동일하다. 이해를 돕기 위해 <표 8>은 <그림 9>에 처리된 데이터의 내용 및 사용단계를 보여준다.

전처리 단계에서는 논문-학회/논문지의 번호로 이루어진 '논문 소속 데이터'와 앞에 언급한 학술연구분야분류체계가 저장된 '분류표'를 매칭시켜 어떠한 학회/논문지가 어떠한 분류에 속하는지 분류된 '분류된 논문 소속 데이터'를 만든다. 입력된 전체 NDSL 논문 데이터 중에서 80%를 모델링에, 20%를 테스트용으로 삼는다. 모델링 단계에서 모델링용 데이터는 분류된 논문 소속 데이터와 함께 키워드 추출기, 정답 추출기, 분류기가 포함된 분류 모델을 만든다. 이 때 키워드 추출기는 모델링용 데이터를 이용, 기계학습을 통해 분류 모델의 핵심이 되는 '추출된 분류별 키워드 데이터'를 만든다. 이 때 검증을 위하여 데스트용 데이터를



〈 그림 9〉 분류 처리 단계

입력하여 정답 추출기를 통해 테스트용 데이터의 정답을 추출한다. 최종적으로 추출된 분류별 키워드 데이터와 테스트용 데이터, 정답 데이터를 분류 모델에 입력하여 분류기를 통해 분류된 결과를 정답과 비교하여 정분류, 오분류, 미분류를 구분하고 중복분류를 출력한다. 지속적으로 기계학습을 통해 트레이닝이 되기 때문에 보다 정확한 결과를 보여주게 되며 사용자에게는 <표 7>에 기반한 분류를 제공한다. 제안하는 논문 분류 모델은 기존의 분류들과 다른 저자 키워드 분류이며, 이를 통해 중복 분류가 가능한 새로운 논문의 분류를 보여주고 사용자에게 키워드에 따른 논문에 접근이 용이하도록 한다. 본 논문은 이를 구현하기 위해 기존 논문 데이터를 가지고 실험을 통해 검증해 보았다.

〈표 8〉 단계별 사용 데이터

데이터명	내용	사용단계
논문 소속 데이터	소속번호, 소속명	전처리
분류표	학술연구분야분류체계	전처리
분류된 논문 소속 데이터	중분류, 소속번호	전처리, 모델링
논문 데이터1	80% 논문	모델링
분류별 추출된 키워드 데이터	분류별 키워드	모델링, 테스트
논문 데이터2	20% 논문	모델링, 테스트
정답 데이터	20% 논문 정답	테스트

다. 분류 결과

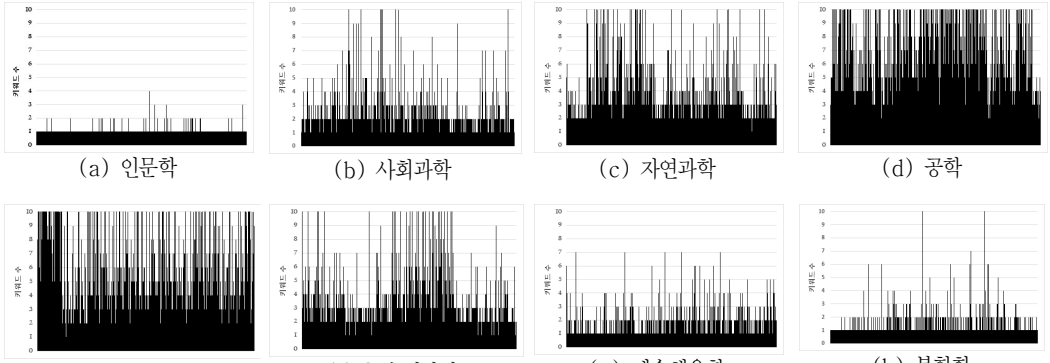
1) 입력 데이터

테스트를 위해 NDSL이 보유한 2016년 국내논문을 이용하였다. 키워드 매칭을 기반으로 동작하므로 앞서 적용시킨 학술연구분류체계를 이용하여 분류별 키워드 데이터의 빈도를 분석하였다. <표 9>는 키워드의 종류와 중복을 포함하는 키워드의 수를 나타내며, <그림 10>은 학술연구분류체계에 따른 2016년 국내논문 키워드 빈도이다.

<그림 10>에서 (a)부터 (h)까지 각 분류의 키워드 빈도를 나타내고 있으며, 가로축은 키워드 종류, 세로축은 해당 키워드의 수를 보여준다. 10개 이상을 보유한 키워드가 존재하였으나, 분류별로 키워드 수를 정확하게 비교하기 위해 임의로 10개 이하만 표시된다. 그림에서 10으로 표시된 키워드들은 모두 10이상이 되었으며 실제로는 모든 키워드를 포함하여 실험을 하였다. <그림 10>과 <표 9>에서 보는 것처럼 키워드 종류에 비해 중복을 나타내는 전체 키워드 수가 많을수록 고른 빈도를 나타내는 경향이 있으며 이것은 키워드 매칭이 더 좋은 성능을 보일 수 있다는 것을 의미한다. 과학기술분야에 특화된 NDSL이기 때문에 많은 수의 데이터를 보유한 공학이 가장 고른 키워드 빈도를 나타내고 있으며 반대로 인문학이 가

장 낮다. 실제 키워드 매칭을 통한 분류를 할 시에 빈도가 가장 고른 공학과 의약학이 가장 높은 성능을, 상대적으로 고른편에 속하는 사회과학, 자연과학, 농수해양학이 그 다음으로 높은 성능을 보여줄 것으로 예측할 수 있으며, 빈도가 고르지 않은 인문학, 예술체육학과 복합학은 성능이 낮을 것으로 예상하였다.

분류를 위해 사용한 데이터는 NDSL이 보유한 2016년 국내논문 데이터 38,612편를 이용하였다. 80%의 논문인 30,890편은 모델링에 사용하고, 20%의 논문인 7,722편은 테스트에 활용하였다. 약 20만개의 키워드 데이터를 포함하고 있으며, 700여개의 학회/논문지 소속으로 되어 있다. 테스트의 분배를 편차를 줄이기 위해 논문 데이터 셋을 셔플하여 20개의 셋으로 만들고 평균값을 내어 측정하였다.



<그림 10> NDSL이 보유한 2016년 국내논문 키워드 빈도

<표 9> NDSL이 보유한 2016년 국내논문 키워드 빈도

	키워드 종류	전체 키워드 수
인문학	611	652
사회과학	13,050	14,429
자연과학	26,227	30,451
공학	97,001	115,024
의약학	22,651	30,702
농수해양학	15,902	19,183
예술체육학	5,598	6,198
복합학	2,573	2,902

2) 분류측정

예측된 값과 정답 데이터와 비교하여 정분류, 오분류, 미분류와 중복분류로 구분하였다. 하단의 <표 10>은 예측을 통한 분류측정방법이다. 정답 데이터는 논문 편당 1개로 구분되어

있으며 해당 논문이 속하는 본래의 학회/논문지가 속하는 분류가 정답이 된다. 해당 논문의 저자키워드 중 1개라도 정답과 일치하면 정분류, 예측된 값이 정답과 다르면 오분류, 예측한 답이 없으면 미분류로 분류하며, 그 외에 타분류에 속하는 결과 모두 중복분류로 포함시켰다.

〈표 10〉 분류측정

예측	분류
성공	정분류
타분류	오분류, 중복분류
X	미분류

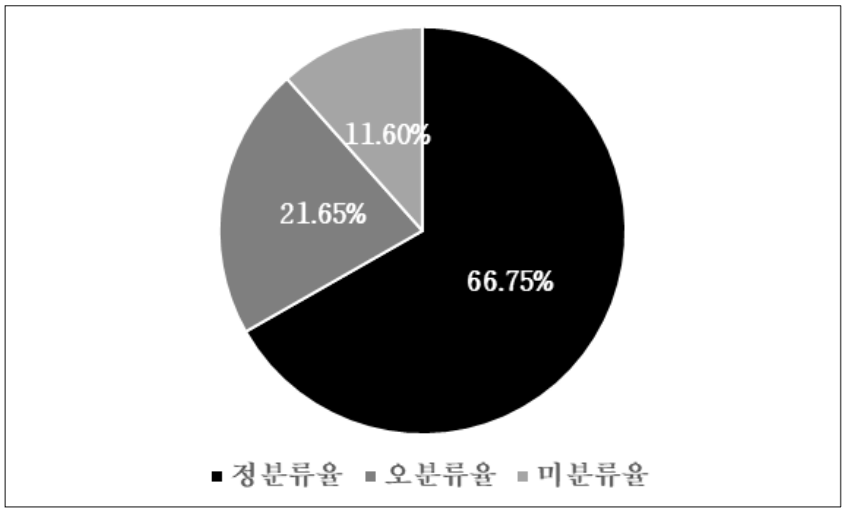
3) 실험 결과

<표 11>은 전체 정분류 횟수를 보여주고 이를 토대로 <그림 11>은 정분류, 오분류, 미분류의 전체 분류율을 보여준다. 테스트 데이터 7,722편의 논문중 정분류율은 5,154.9개의 66.75%이고, 오분류율은 1,672.15개의 21.65%, 미분류율은 894.95개의 11.6%라는 결과를 얻을 수 있었다. 중복분류는 53,672.35번으로 테스트 논문 편수의 약 6.95배에 달하는 수치이며, 산술적으로는 논문 편당 평균적으로 약 7개의 분류로 중복분류가 되었다. 다소 높지 않은 정분류율, 다소 낮지 않은 오분율과 미분류율을 보여주는데 이것은 저자 키워드 자체의 특성이 필터링 되지 않은 자연어이고 테스트를 위한 학술분류체계가 한 분야 이상의 분류를 허용하지 않기 때문이다. 하지만 이를 감안해서라도 자연어의 특성임에도 충분히 높은 정분류을 보여주었으며, 실제로 기존 학술분류체계가 명확한 답은 아니기 때문에 학회/논문지에 종속적인 단분류에서 벗어난 다양한 분야로의 중복분류가 가능한 분류를 제공하고자 한다.

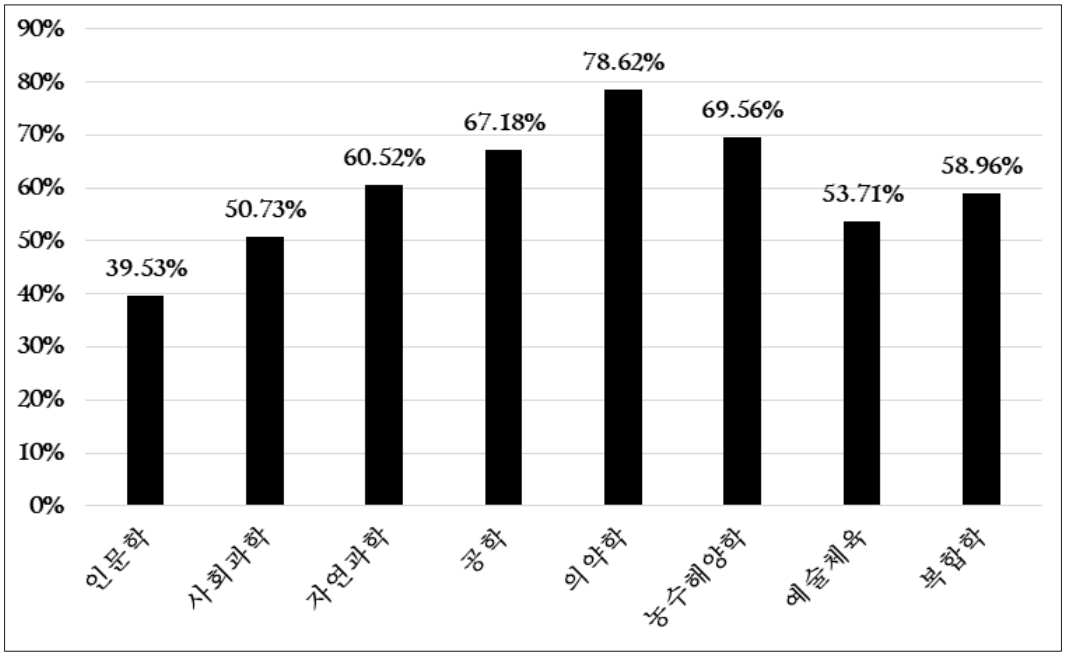
<그림 12>는 대분류별 정분류율을 보여준다. 의약학이 78.62%로 가장 높고 인문학이 39.53%로 가장 낮았다. 이는 앞서 키워드 빈도에서 확인되었던 키워드 종류에 비해 전체 키워드 수가 많을수록, 키워드 빈도가 기복없이 균일할수록 성능이 높다는 것을 보여준다. 공학의 키워드 빈도가 더 좋음에도 의약의 성능이 높다는 것은 공학에 비해 파생어가 많지 않고 전문적인 단어 위주로 일관성이 높은 키워드를 구성한다는 사실을 알 수 있으며 이것은 해당 분류의 특성이 반영되었다고 할 수 있다. 반면에 인문학은 키워드의 종류와 수가 적을 뿐만 아니라 빈도도 균일하지 않고 고저차가 심하므로 성능이 높지 않다는 것을 확인할 수 있었다.

〈표 11〉 전체 정분류 횟수

전체	정분류	오분류	미분류	중복분류
7,722	4,662.5	1,808.25	1,251.25	48,009.2



〈그림 11〉 전체 분류율



〈그림 12〉 대분류별 정분류율

보다 자세하게 확인하기 위해 중분류별 정분류율도 수치화하였다. <표 11>는 중분류별 정분류율이다. 인문학에서의 정분류율은 52.74%로 언어학이 가장 높으며, 다음은 49.59%의 철학이 가장 높지만 인문학 전체 평균 정분류율은 40%도 넘지 못하였다. 사회과학의 경우

높은 순으로 경영학이 56.46%, 교육학이 52.20%이며, 데이터가 없었던 민속학을 제외하면 정치외교학이 4.75%로 가장 낮음을 알 수 있었다. 자연과학은 생물학이 77.66%로 가장 높은 성능을 보여주고 있으며 물리학이 37.83%로 가장 낮았다. 공학의 경우 76.43%로 전자/정보통신이 가장 높으며 항공우주공학이 35.95%로 가장 낮았다. 의학은 간호학이 95.97%로 전체 분류중에서도 가장 높은 성능이었으며, 수의학은 61.46%로 가장 낮지만 무려 62%에 육박하는 성능을 보여주어 의약학이 전체 대분류중에 가장 높은 성능을 보였다. 농수해양학의 경우 식품과학이 84.3%로 가장 높았으며, 해상운송학이 43.92%로 가장 낮았다. 예술체육은 70.1%로 체육분야가 가장 높았으며 분류가 없는 연극과 영화 분야를 제외하고 미술이 0%로 가장 낮은 성능을 보여주었다. 마지막으로 복합학에서는 문헌정보학과 감성과학이 각각 60.47%와 15.4%로 가장 높은 성능과 가장 낮은 성능을 보여주었다. 이러한 결과는 NDSL에서 해당 분야에서 데이터가 많을수록 높은 성능을 보여준다고 할 수 있겠지만 키워드 종류에 대한 전체 키워드 수에 따라 비례한다고 할 수 있다. 의약학은 높은 키워드의 일관성을 가지기 때문에 중분류의 정분류율이 높았으며, 이로인해 대분류 성능이 높아지기 때문에 분야의 키워드가 가지는 특성이 매우 중요하다는 것을 알 수 있다.

중복분류는 제안하는 모델에 가장 큰 장점으로 분류 횟수와 상관없이 하나의 논문은 키워드에 따라 여러 분야로 중복분류가 될 수 있다. <표 13>은 중복분류가 가장 많이 되는 상위 13개의 분류를 보여준다. 가장 높은 중복분류율을 가진 분류는 5.23%의 공학일반이었다. 공학은 상위 8개의 분류 중에서 네번째의 의학을 제외하고 모든 분류의 대분류가 공학으로 매우 높은 중복분류율를 보였다. 치의학을 제외하고 자연과학분야의 생물학과 농수해양학의 농학이 12번째로 상위권으로 랭크되었으며, 표에는 표시되지 않았지만 경영학이 사회과학에서 가장 높은 중복분류율로 15번째에 랭크되었다. 반면에 인문학, 예술체육학은 중복분류에서 찾아보기가 쉽지 않았다. NDSL의 특성상 공학 분야의 데이터가 가장 많기도 하지만 그만큼 공학이 기반이 되어 타분야와 융합연구가 활발하다는 것을 보여주었다.

〈표 12〉 분류별 정분류율

	인문학일반	역사학	철학	종교학/신학	언어학	문학	한국어문학
정답	5.7	1.65	2.25	0	11.8	1.85	0
예측	0.95	0	0.9	0	4.7	0.7	0
정분류율	18.60%	0.00%	35.42%	–	40.54%	34.08%	–

	일본어문학	중국어문학	영어문학	프랑스어문학	독일어문학	러시아어문학	서양고전어문학
정답	0	0	0	0	0	0	0
예측	0	0	0	0	0	0	0
정분류율	–	–	–	–	–	–	–

(a) 인문학 분류율

	사회과학일반	경제학	경영학	관광/무역학	교육학	군사학	민속학
정답	59.2	17.05	145.6	9.3	81.55	18.1	0
예측	27.35	7.2	82.05	3.65	42.65	1.1	0
정분류율	46.09%	41.71%	56.46%	38.33%	52.20%	5.95%	-

	법학	사회/사회복지학	신문방송학	정치외교학	지리/지역학	행정학
정답	11.2	30.35	8.2	0	72.7	9.25
예측	5.6	9	0.35	0	30.65	2.9
정분류율	48.12%	29.89%	4.48%	-	42.18%	32.35%

(b) 사회과학 분류율

	자연과학일반	물리학	생물학	식물학	생활과학
정답	23.1	80.85	330.5	78.7	26.6
예측	11.45	24.8	233.05	44.35	9.3
정분류율	49.25%	30.76%	70.49%	56.34%	36.04%

	수학	지구과학/천문학	통계학	화학
정답	156.9	249.9	43.8	153.55
예측	57.45	138.85	19.45	71
정분류율	36.48%	55.54%	44.56%	48.29%

(c) 자연과학 분류율

	공학일반	건축공학	기계공학	산업공학	원자력공학	자원공학	재료공학
정답	423.9	112.85	310.55	59.05	46.25	85.15	124.85
예측	258.4	58.8	182.15	17.8	20.5	46.45	79.9
정분류율	60.95%	52.06%	58.74%	30.19%	44.15%	54.06%	64.09%

	금속공학	전기전자/제어계측공학	전자/정보통신	컴퓨터공학	조선/해양공학	토목공학
정답	69.65	492	592.2	505	75.5	270.35
예측	47.55	298.3	423.9	325.9	38.1	176.55
정분류율	68.70%	60.70%	71.58%	64.56%	50.63%	65.19%

	환경공학	화학/생물공학	생명공학	항공우주공학	교통/도로/철도공학	안전공학
정답	88.35	234.6	132.4	61.95	45.8	44.85
예측	46.55	156.5	77.3	17.55	18.4	15.15
정분류율	52.81%	66.67%	58.53%	28.49%	40.95%	33.93%

(d) 공학 분류율

	간호학	수의학	약학	의학	내과학
정답	142.55	28.95	114.25	490.7	59.1
예측	134.6	14.2	67.8	357.6	37.95
정분류율	94.42%	49.98%	59.68%	73.17%	66.01%

	보건학	정신건강의학	물리치료학	치의학	한의학
정답	61.25	19.5	38.15	176.05	164
예측	34	11.6	23.15	124.3	116.65
정분류율	55.45%	59.22%	60.87%	70.57%	71.19%

(e) 의약학 분류율

	농학	해양학	수산학	식품과학	임학	조경학	축산학	해상운송학
정답	156.5	53.4	85.2	216.9	85.2	45.85	32.65	16.4
예측	103.5	26.9	44.4	169.85	46.65	21.55	19.45	5.45
정분류율	66.21%	50.49%	52.68%	78.09%	54.88%	48.55%	59.28%	36.09%

(f) 농수해양학 분류율

	예술체육일반	건축	디자인	미술	미용
정답	12.65	30.15	27.6	0.3	9.05
예측	5.95	11.9	10.55	0	4.4
정분류율	44.39%	39.88%	36.20%	0.00%	48.93%

	연극	영화	음악	의상	체육
정답	0	0	13.9	89.75	9.3
예측	0	0	5.85	50.1	5.8
정분류율	-	-	42.54%	55.84%	58.88%

(g) 예술체육학 분류율

	복합학일반	감성과학	과학기술학	문헌정보학	인지과학
정답	54.85	8.4	9.45	60.4	4.7
예측	29.9	1.3	4.15	36.45	1.35
정분류율	53.98%	15.40%	43.35%	60.47%	31.49%

(d) 복합학 분류율

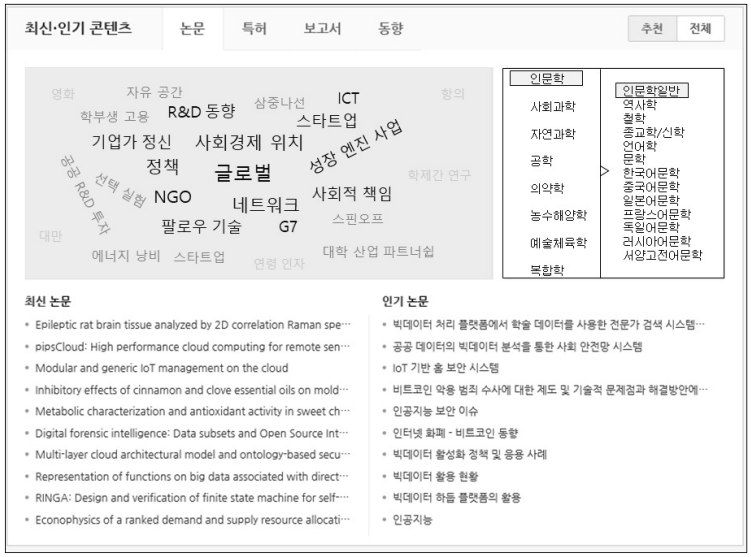
라. 분류 정보 제공

분류된 논문 정보를 활용하기 위해 PIN 서비스의 워드 클라우드를 이용한다. PIN 내부의 분류 모듈은 중복분류가 가능한 학술정보분류체계에 따라 분류시킨 후, 저자 키워드를 추출

〈표 13〉 중복분류율

순위	분류명	분류횟수	분류율
1	공학일반	2537.15	5.28%
2	전자/정보통신	2272.05	4.73%
3	컴퓨터공학	2143.55	4.46%
4	의약학	1724.25	3.59%
5	전기전자/제어계측공학	1624.65	3.38%
6	화학/생물공학	1426.1	2.97%
7	토목공학	1402.25	2.92%
8	지구과학/천문학	1397.25	2.91%
9	식품과학	1392.35	2.90%
10	생물학	1371.55	2.86%
11	치의학	1167.45	2.43%
12	농학	1166.3	2.43%
13	기계공학	1291.9	2.41%

하고, 이를 검색빈도순으로 워드 클라우드를 구성한다. 이 워드 클라우드는 대분류와 중분류별로 세분화하여 구성되며, 사용자는 이 대분류와 중분류를 선택함으로써 해당 분류에 따른 워드 클라우드가 선택이 되며, 키워드가 중복분류 되어 구성되기 때문에 학술지/학회/저널의 종속적이지 않은 다양한 분야로의 분류를 보여준다. 워드 클라우드에서 보여지는 키워드를 선택하면 해당 키워드를 저자 키워드로 보유한 논문을 추천해 준다. <그림 13>는 PIN의 논



〈그림 13〉 분류 정보 제공

문 분류 정보 제공 화면이다. 비로그인시에 실시간 검색어와 로그인시에 관심 키워드 선택 화면에서 분류를 보여준다. 대분류의 인문학, 중분류의 인문학일반을 선택하여 워드 클라우드를 출력한다. 현재 인문학-인문학일반에서의 가장 빈도수가 높은 저자 키워드는 글로벌, 정책, 사회경제 위치 순이며, 해당 키워드를 클릭시 이 키워드를 보유한 논문을 추천해준다.

Ⅴ. 결론

본 논문은 한국과학기술정보연구원에서 제공하는 국내 최대의 과학기술정보 포털 사이트인 NDSL에서 사용자들에게 검색뿐 아니라 방대한 자료에 쉽게 접근하기 위해서 사용자들의 프로파일을 분석, 맞춤형 추천 및 분류 서비스인 PIN을 제안하였다. 추천 서비스의 경우, 비로그인과 로그인시에는 관심 및 유사 키워드를 이용하여 워드 클라우드를 구성하고 이를 통하여 비로그인시에는 최신 및 인기 콘텐츠, 로그인시에는 연관 및 동시이용한 콘텐츠를 제공하였다. 원문보기는 로그인과 상관없이 공통적으로 원문보기를 통해 보여지는 콘텐츠와 관련된 다양한 콘텐츠를 제공하며, PIN 전용 페이지를 통해 사용자의 이력과 히스토리를 한눈에 파악이 가능하다.

분류의 경우 기존의 학술정보서비스에서 활용한 학회/저널의 종속적인 단분류 또는 소수의 중복분류의 문제를 해결하고 보다 쉬운 접근과 최신 트렌드를 반영 및 연구자 중심 서비스의 일환으로 NDSL에서 키워드 매칭을 이용하여 학술분류체계 기반 분류 모델을 제안하였다. 이를 검증하기 위해 2016년 NDSL 논문 데이터를 이용하였다. 정분류는 66.75%, 오분류 21.65%, 미분류는 11.6%, 그 중에서 의학이 78.62%, 인문학이 39.53%로 가장 높은 성능과 낮은 성능을 보여주었고, 중복분류는 전체 7,722편의 논문중에 53,672.35번의 중복분류로 편당 약 7개의 분류로 분류되어 학회/논문지에 종속적인 단분류에서 벗어나 다양한 분야로의 중복분류가 가능하였으며 실제로 분류 모델의 활용 가능성을 보여주었다. 이러한 PIN 서비스는 사용자에게 적합한 정보를 제공 및 접근을 할 수 있도록 도와주게 된다.

향후에는 인공지능 알고리즘을 도입하여 모델 기반 필터링을 통해 보다 정확한 추천과 오분류을 줄이고, 논문뿐만 아니라 NDSL이 보유한 보고서, 특허, 동향을 분류하여 사용자에게 보다 더 접근이 쉬운 서비스 목표로 하는 연구를 할 예정이다.

참고문헌

김광영, 곽승진. 2011. 주제분류 기반의 개인화 검색시스템에 관한 연구.『한국문헌정보학회지』, 45(4): 77-102.

김남규 외. 2017. 텍스트 분석 기술 및 활용 동향.『한국통신학회논문지』, 42(2): 471-492

김은경, 최진오. 2005. 효율적인 키워드 검색을 지원하는 학습자료의 구조화 방법 연구,『한국해양정보통신학회』, 9(1): 1063-1066.

네이버 학술정보, http://academic.naver.com/

박창호, 염성숙, 이정모. 2000. 사용자 중심의 홈페이지 분류체계가 분류 검색에 미치는 효과.『인지과학회지』, 11(1): 47-65.

손지은 외. 2015. 추천 시스템 기법 연구동향 분석.『대한산업공학회지』, 41(2): 185-208.

여운동 외. 2010. 연구논문 추천시스템의 전자도서관 적용방안.『한국콘텐츠학회논문지』, 10(11): 10-19.

이락규 외. 2011. 모바일 환경에서 콘텐츠 추천 시스템 설계 및 구현.『한국콘텐츠학회』, 11(12): 40-51.

이주현, 이응봉, 김환민. 2006. NDSL 웹사이트 분석 및 서비스 품질평가.『정보관리연구』, 37(4): 69-91.

이태석 외. 2012. 용어 자동분류를 사용한 검색어 범주화의 분석적 고찰.『정보처리학회지』, 19(2): 133-138.

정덕영, 이준석, 박상성. 2016. 워드 클라우드를 이용한 기술트렌드 분석.『한국지능시스템학회』, 26(1): 17-18.

정영미. 2005.『정보검색연구』. 서울: 구미무역 출판부.

학술연구분류체계, <https://www.nrf.re.kr/biz/doc/class/view?menu_no=323>

Adomavicius, Gediminas, and Alexander Tuzhilin. 2005. "Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions". *IEEE Transactions on Knowledge and Data Engineering,* 17(6): 734-749.

Balabanović, Marko, and Yoav Shoham. 1997. "Fab : content-based, collaborative recommendation". *Communications of the ACM,* 40(3): 66-72.

Breese, John S., David Heckerman, and Carl Kadie. 1998. "Empirical analysis of predictive algorithms for collaborative filtering". *Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence,* 43-52.

Burke, R. 2002. "Hybrid recommender systems : Survey and experiments". *User modeling and user-adapted interaction,* 12(4): 331-370.

Das, Abhinandan S. et al. 2007. "Google news personalization: scalable online collaborative filtering". *Proceedings of International Conference on World Wide*

Web, 271-280.

DBPia, http://www.dbpia.co.kr

Goldberg, Ken. et al. 2001. "Eigentaste : A constant time collaborative filtering algorithm". *Information Retrieval,* 4(2): 133-151.

Matsuo, Yutaka, and Mitsuru Ishizuka, M. 2004. "Keyword extraction from a single document using word co-occurrence statistical information". *International Journal on Artificial Intelligence Tools,* 13(1): 157-169.

McCallum, Andrew, and Kamal Nigam. 1999. "Text classification by bootstrapping with keywords". *EM and shrinkage.*

NDSL Homepage, http://www.ndsl.kr

Pazzani, Michael J., and Daniel Billsus. 2007. "Content-based recommendation systems". *Proceedings of the adaptive web,* 325-341.

Smeaton, Alan F., and Jamie Callan. 2005. "Personalisation and recommender systems in digital libraries". *International Journal on Digital Libraries.* 57(4): 299-308.

Wu, Yi-Hung, and Arbee LP Chen. 2000. "Index structures of user profiles for efficient web page filtering services". *Proceedings of International Conference on Distributed Computing Systems,* 644-644.

국한문 참고문헌의 영문 표기

(English translation / Romanization of reference originally written in Korean)

Kim, Kwang-Young and S. Kwak. 2011. "A study on personalized search system based on subject classification." K*orean Society for Library and Information Science,* 45(4): 77-102

Kim, Namgyu. et al. 2017."Investigations on Techniques and Applications of Text Analytics." *Korean Institute of Intelligent Systems,* 42(2): 471-492.

Kim, Eun-Kyung and J. Choi, 2005. "A study on structuring method of study data supporting efficient keyword search." *Maritime information and communication sciences*, 9(1): 1063-1066.

Lee, Nak-Gyu. et al. 2011. "Design and implementation of a contents recommendation system in mobile enviornment." *Korea Contents association,* 11(1): 40-51.

Naver Academic, http://academic.naver.com/

Park, Chang-Ho, S. Youm, and J. Lee. 2000. "The effect of user-centered

categorization system of homepages on directory Search." *Cognitive science,* 11(1): 47-65.

Son, Jieun. et al. 2015. "Review and Analysis of Recommender Systems." *Journal of the Korean Institute of Industrial Engineers,* 41(2): 185-208.

Yeo, Woon-Dong. et al. 2010. "Application of Research Paper Recommender System to Digital Library." *Korea contents association*, 10(11): 10-19.

Lee, Ju-Hyun, E. Lee, and H. Kim. 2006."Analysis and service quaility evaluation on NDSL website." *Journal of information management,* 37(4): 69-91.

Lee, Tae-Suk. et al. 2012."An Analytic Study on the Categorization of Query through Automatic Term Classification." *KIPS Transactions:PartD,* 19(2): 133-138.

Jeong, Duckyoung, J. Lee and S. Park. 2016. "A technology trend analysis using world cloud." *Proceedings of KIIS Spring conference,* 26(1): 17-18.

Jung, Young-Mi. 1993. *Information retrieval*, Seoul: Kumibooks,

The academic classification system. <https://www.nrf.re.kr/biz/doc/class/view?menu_no=323>