과학기술정보 서비스 플랫폼에서의 빅데이터 분석을 통한 개인화 추천서비스 설계*

Personal Recommendation Service Design Through Big Data Analysis on Science Technology Information Service Platform

김 도 균 (Dou-Gyun Kim)**

초 록

연구자들에게 지식을 습득하여 연구 활동에 도입하는데 걸리는 소요시간을 단축하는 것은 연구생산성 향상에 필수적인 요소라고 할 수 있다. 본 연구의 목적은 한민족과학기술자네트워크(KOSEN) 사용자들의 정보 이용 패턴을 군집화하고 그룹화 된 사용자들에게 맞는 개인화 추천서비스 알고리즘의 최적화 방안을 제안하는 것이다. 사용자들의 연구활동과 이용정보에 기반하여 적합한 서비스와 콘텐츠를 식별한 후 Spark 기반의 빅데이터 분석 기술을 적용하여 개인화 추천 알고리즘을 도출하였다. 개인화 추천 알고리즘은 사용자의 정보검색에 소요되는 시간을 절약하고 적합한 정보를 찾아내는데 도움을 줄 수 있다.

ABSTRACT

Reducing the time it takes for researchers to acquire knowledge and introduce them into research activities can be regarded as an indispensable factor in improving the productivity of research. The purpose of this research is to cluster the information usage patterns of KOSEN users and to suggest optimization method of personalized recommendation service algorithm for grouped users. Based on user research activities and usage information, after identifying appropriate services and contents, we applied a Spark based big data analysis technology to derive a personal recommendation algorithm. Individual recommendation algorithms can save time to search for user information and can help to find appropriate information.

키워드: 추천시스템, 개인화시스템, 과학기술정보, 빅데이터, 한민족과학기술자네트워크
Recommended System, Personalization System, Science and Technology Information, Big Data, Korean Scientist Network

---

* 본 연구는 2017년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 과학기술국제화 사업 연구임(2012K1A3A7A05033175).

** 한국과학기술정보연구원 선임기술원(koras@kisti.re.kr)

논문접수일자 : 2017년 12월 22일 논문심사일자 : 2017년 12월 24일 게재확정일자 : 2017년 12월 25일
한국비블리아학회지, 28(4): 501-518, 2017. 〔http://dx.doi.org/10.14699/kbiblia.2017.28.4.501〕

1. 서 론

1.1 연구의 배경과 목적

고도화 된 정보화 시대에서 학문 분야의 융·복합화가 진행되면서 온라인 협업 환경이 조성되고 연구자들의 온라인 지식 교류 활동은 지속적으로 증가하고 있다. 연구자들에게 지식을 습득하여 연구 활동에 도입하는데 걸리는 소요시간을 단축하는 것이 연구생산성 향상에 필수적인 요소라고 할 수 있다.

새로운 아이디어를 필요로 하는 연구개발에서 기하급수적으로 생산되는 방대한 자료 중에서 내게 필요한 정보를 빠른 시간 안에 습득하는 것은 연구개발의 성패를 가르는 중요한 요소라 하겠다. 특히 과학기술분야 간의 융·복합 연구가 활발한 지금 과학기술분야 간의 연구협력과 전문가 네트워크의 활용이 중요한 역할을 차지한다.

지식교류 커뮤니티 사이트는 정보의 생산과 유통의 정보 소비 활동이 일어나는 공간이다. 한국과학기술정보연구원(KISTI)에서 운영 중인 한민족과학기술자네트워크(이하 KOSEN)는 과학기술 전문가 네트워크를 통한 전문가들 간의 정보교류 활성화를 통한 정보순환 생태계를 이루고 있는 과학기술 학술정보 서비스 플랫폼이다.

정보의 홍수시대에 방대한 양의 자료 중에서 개별 연구자에게 맞는 정보를 찾기란 쉽지 않은 과정의 연속이다. 회원들의 정보 이용과 참여를 통해 운영되는 시스템의 경우, 수집된 다양한 정보 속에서 회원들이 원하는 서비스를 적절히 제공할 수 있어야 참여율 증대와 이용 활성화를 기대할 수 있다. 이를 위해서는 무엇보다 효율적인 추천 알고리즘의 도입이 중요하다.

본 연구는 KOSEN 사용자들의 정보 이용 패턴을 군집화하고 그룹화 된 사용자들에게 맞는 개인화 추천서비스 알고리즘의 최적화를 진행하였다. 개인화 추천 알고리즘은 사용자의 정보검색에 소요되는 시간을 절약하고 적합한 정보를 찾아내는데 도움을 줄 수 있다.

1.2 연구의 범위와 방법

본 연구에서는 먼저 추천시스템의 정의와 종류에 대한 선행연구에 대해서 소개하고, 과학기술 서비스 플랫폼인 KOSEN의 서비스 구성에 대해서 설명한다. 다음으로 실제적인 서비스 환경에서 빅데이터 기반의 추천 알고리즘의 발견을 위해서 기존 시스템을 이용하는 회원들의 배경 및 활동에 기반하여 회원에게 적합한 서비스 및 콘텐츠를 식별한 후 Spark 기반의 빅데이터 기반의 분석 기술을 적용하였다. 마지막으로 회원의 선호 서비스 도출 및 적합한 추천 알고리즘을 도출하였다.

2. 이론적 배경

2.1 추천 시스템

추천시스템은 콘텐츠를 이용하는데 있어서 사용자나 콘텐츠가 가지고 있는 특성 정보를 이용하여 사용자에게 필요하거나 선호되어질 만한 콘텐츠를 선별하여 제시함으로서 사용자에게 검색 편의성을 제공해주는 시스템이다. 국내

외 대형 플랫폼 화사들을 비롯한 많은 회사들이 빅데이터 기반의 머신러닝 기법들을 활용하여 사용자들에게 고차원적인 정보들을 제공하고 있다. 기존의 방법들이 사용자별 맞춤정보 제공을 위해 사용자에게 수동적인 정보 입력을 필요로 했다면, 현재는 사용자들이 관심 있는 정보를 이용하면서 발생하는 로그 형식의 데이터인 방문 페이지 체류 시간, 클릭 횟수, 연관 콘텐츠 이용 등을 분석하는 등의 자동적인 정보 수집 및 분석을 통해 사용자에게 맞춤형 정보를 추천해 줌으로써 사용자의 수고를 덜어주고 히트율을 높여주고 있다(유영석 외 2017).

최근에는 웹의 활성화에 따라 웹상에서의 개인화에 대한 연구가 활발히 진행되고 있다. 웹상에서의 개인화는 웹사이트에 접속하는 이용자의 성향과 행태별로 세분화하여, 이용자가 선호할 수 있는 적절한 정보 또는 상품을 제공함으로써 보다 적극적인 서비스를 제공하는 것을 의미한다. 이와 같은 서비스 전략은 이용자의 요구를 만족시킴으로써 해당 웹사이트에 대한 이용자의 충성도를 높여줄 뿐 아니라 타깃 마케팅(target marketing)과 일대일 마케팅(one-to-one marketing)을 가능하게 해준다는 점에서 의의가 크다(김용, 문성빈 2007).

추천시스템은 대표적으로 인구통계학적 추천(demographic recommendation), 협업 필터링(collaborative filtering), 내용 기반 필터링(contentbased filtering), 또한 각각의 방법이 가지고 있는 한계점을 보완하기 위해 두 가지 방법의 장점을 취합하는 혼합 필터링(hybrid filtering) 기법 등이 있다(정인용, 양새동, 정회경 2015).

특히 개인화된 콘텐츠 추천을 위하여 아이템 기반의 협업 필터링(collaborative filtering) 알고리즘을 사용하여 아이템 간의 유사도를 도출하고 이를 기반으로 특정 사용자에게 가장 선호도가 높을 것으로 추정되는 콘텐츠를 추천하였다(최성우, 한성희, 정병희 2014).

대부분의 추천 시스템에서는 협업 필터링 기법을 사용한다. 협업 필터링 기법을 사용한 추천 시스템은 사용자들의 선호도를 수집한 뒤 이를 기반으로 사용자들의 관심사나 유사한 취향을 예측한다. 협업 필터링 기법에서는 사용자 정보와 항목 기반으로 정보를 선별한다. 사용자 정보 기반의 협업 필터링은 사용자들의 선호도나 상품의 평가들을 수집하고 비슷한 성향을 가진 사람들을 연결한다. 그리고 이를 기반으로 사용자에게 정보를 제공한다. 항목 기반의 협업 필터링 기법은 사용자들이 평가한 상품들의 정보를 사용한다. 사용자들은 과거에 선호했던 제품들과 유사한 제품을 선호하는 경향이 있다는 점을 사용하여 유사한 제품들의 정보를 제공한다. 내용 기반의 필터링 기법은 자연 언어 처리나 정보 검색 분야에 기반을 두고 있으며 정보의 내용이나 사용자의 정보들을 비교하여 사용자에게 적합한 정보를 제공 해준다. 내용 기반의 필터링 기법에 사용되는 주요 모델은 불리언 모델과 벡터공간 모델, 확률모델 등과 같은기법을 사용하며 사용자가 과거에 사용했거나 평가했던 상품의 유사도를 측정하여 정보를 제공한다(정인용, 양새동, 정회경 2014).

KOSEN 회원에 대한 개인화 연구들은 회원들의 정보와 활동 로그 데이터를 활용하여 각 그룹별로 서비스를 추천하는 알고리즘을 도출하고 기존 시스템에 알고리즘을 적용한 연구(최가현, 황윤영, 윤정선 2017)와 적용된 시스

템에 대해서 28명을 대상으로 '전공' 항목을 추가하여 성능을 평가를 적용함으로서 추천 정확도가 향상되었음을 평가한 바 있다(박성은, 황윤영, 윤정선 2017).

본 연구에서는 Spark 기반의 빅데이터 개인화 추천서비스의 설계를 위한 기술들을 바탕으로 콘텐츠간의 유사도 분석방법과 추천 콘텐츠를 추출하고자 하는 설계에 관한 연구와 회원의 활동 로그 기반의 추천 시스템의 적용 방법을 위해서 SQL 질의 처리의 분산과 병행 최적화를 통한 시스템의 구성안에 대해서 연구하였다.

2.2 KOSEN의 특징

KOSEN은 (구)과학기술부에서 해외 한인 과학자의 인맥을 형성하고, 이를 통해 해외 정보 및 국제 협력의 기반을 다지고자 1999년 한민족 과학기술자 네트워크(KOSEN)라는 웹사이트를 구축하였다. 1999년 7월 서비스를 시작하였고 현재 한국과학기술정보연구원(KISTI)에 의해 운영되고 있으며, KOSEN은 회원의 정보활동을 통해 커뮤니티의 가치를 높이고 있다. 전문가회원이 참여해 생성하는 고급 보고서와 해외 회원 및 연구 현장 회원들이 제공하는 정선된 해외 연구 정보가 제공되며, 커뮤니티 구성원간의 전문적인 질의응답과 자료 교환이 활발히 이루어진다.

해외에 있는 과학기술자들은 이 네트워크를 통해 국내 과학기술계의 현황을 파악 할 수 있으며 여러 가지 헝태로 국내의 과학기술 발전에 기여할 수 있다. 국내의 과학기술자들은 해외 전문가들이 제공하는 정보를 통해 보다 신속하게 현지 정보를 입수할 수 있으며 국제 공동연구의 파트너를 찾는 등 네트워크를 통한 다각적인 교류가 가능하다. KOSEN을 통해 연결된 과학기술자들은 유용한 정보로 개인의 역량을 강화할 수 있으며, 결집된 역량은 국내 과학기술계로 유입시켜 국가의 과학기술 경쟁력을 제고 할 수 있다.

KOSEN의 사용자의 회원 통계를 살펴보면 총회원수 137,070명(2017.12.15.기준)으로 이중 해외 회원이 10,860명이며, 석사 학위 이상 소지자는 전체 회원의 59.1%에 이른다. 그리고 해외 회원 중 53.3%는 박사학위를 소지하고 있다. 또한 연구 개발의 핵심인력이라 할 수 있는 30~40대가 62.06%를 차지하고 있다.

현재 KOSEN에서는 전 세계 전문가들이 추천하고 분석한 고급기술동향 자료, 학회보고서, 첨단기술 보고서 등을 제공받을수 있는 '코센 리포트', 회원들끼리 과학기술 전문지식에 대해 문의하고 답변을 받을 수 있는 'What is?', 국내·외 연구실 디렉토리를 등록하는 오픈랩에는 현재 7,046개(2017.12.15 기준)의 랩 정보가 등록되어 있으며 이외에도 강의자료, 커뮤니티, 코센 전문가 제도 운영과 오프라인 모임 지원 등 다양한 서비스가 활발히 제공되고 있다.

관련된 연구로는 주로 KOSEN의 특징인 휴먼네트워크 연구는 네트워크 구성원 간 상호교류의 중요성과 회원의 참여 유도 및 융합연구의 확산 등이다.

휴먼네트워크는 특별한 목적을 달성하기 위해 소수의 전문가를 선발 하여 임무를 부여하는 폐쇄형 휴먼네트워크와 커뮤니티를 기반으로 정보를 운영하는 것은 개방형 휴먼네트워크로 구분할 수 있다. 한선화(2005)는 개방형 네

트워크는 참여자가 정보의 제공자인 동시에 수혜자가 되며, 네트워크 구성원 간 상호 교류가 매우 중요하고 다수의 커뮤니티 구성원을 대상으로 정보수집 및 가공의 임무를 개방하기 때문에, 이를 위한 관리 방법 및 조직이 갖추어져야 커뮤니티 내에서의 정보 흐름을 관리할 수 있다고 주장하였다. 개방형 커뮤니티의 성공 여부는 구성원들의 참여를 얼마나 활발하게 끌어낼 수 있느냐에 달려 있다. 따라서 커뮤니티 구성원들이 관심을 갖고 지속적으로 접속하여 참여할 수 있는 다양한 장치들이 필요하다(한선화 2005).

인터넷으로 제공하는 DB 구축 서비스의 경우 사용자들의 참여에 의해서 전적으로 운영되는 서비스는 정보의 관리가 쉽지 않으며 회원들의 참여를 유도하기도 상당히 어렵다. 정보의 제공에 따르는 인센티브나 참여 동기가 확실해야 사용자들의 참여를 이끌어낼 수 있다. 회원 간 질의응답 서비스나 토론 게시판의 경우는 회원들의 참여를 끌어내기가 상대적으로 쉽지만, 자신이 보유한 정보를 단순히 제공하는 게시판의 경우 회원들의 참여를 유도하기가 쉽지 않다(윤정선, 정혜주, 한선화 2010).

과학기술분야 전문연구는 책임자 중심의 실험실 단위로 수행되고 있고 연구 분야 또한 세분화되고 있는 추세이다. 새로운 학문의 출현과 과학기술분야의 국가 정책에 따라 연구 분야의 유동성도 큰 편이다. 요즘은 특히 융합연구와 학제 간 교류가 많이 필요한 시대라 연구실 단위의 연계가 많이 이루어지고 있다(정선양, 김기동 2008).

3. 개인화 추천서비스 설계

3.1 추천서비스의 목표

본 연구에서의 개인화 추천시스템은 첫째, 회원들의 활동 정보를 기반으로 빅데이터 기반의 추천 알고리즘을 적용하여, 회원들에게 개인 맞춤형 정보 서비스 및 콘텐츠 제공을 목적으로 한다.

둘째, 개인 회원의 선호도를 빅데이터 기반의 분석 및 추천 알고리즘을 통해 개인에게 최적화된 서비스 및 콘텐츠를 제공하고, 동일 유저 그룹의 선호 서비스를 개인에게 추천하여 서비스의 활성화를 유도하여 선제적 개인화 맞춤 서비스를 효율적인 정보의 제공한다.

3.2 개인화 추천서비스의 범위

KOSEN의 회원들의 활동 로그를 기반으로, 빅데이터 처리 기술에 기반하여 최근에 이용한 콘텐츠간의 유사도를 분석하고 추천 콘텐츠를 추출해 개인화 서비스를 제공한다. 이는 과학기술 정보서비스 내에서의 특정 서비스에 대한 편중적인 소비를 막고 가치 있는 연구성과물들의 생명주기를 연장하며 보다 많은 활용도를 높이기 위함이다. 이를 통해 각 회원에게 필요한 콘텐츠가 충분하게 전달 될 수 있도록 하여 콘텐트의 소비를 촉진하고 회원들의 배경(학위, 거주국, 직종 등) 및 활동에 기반하여 그들에게 적합한 콘텐츠를 선별하고, 필요한 시점에 제공한다.

이를 통한 회원 및 회원의 활동 로그 기반의 분석 및 추천 시스템의 적용과 관리자 페이지

내 회원 그룹별 추천 관리 기능을 통한 유연한 시스템의 구현을 그 범위로 한다. 이에 따라 선호 서비스 추천 알고리즘 개발되어 적용되며 리스트, 카드, 위젯, 이미지 서비스 등 개인화 지원을 위한 웹 인터페이스가 적용되었다.

KOSEN의 이용자는 크게 회원과 비회원으로 구분하고, 회원은 다시 보통회원, 휴면회원, 신규회원으로 구분된다. 초기 사용 정보가 거의 없는 신규회원과 휴면회원은 신규회원이 속한 유저 그룹의 서비스를 추천하여 보다 높은 활용성을 유도하고, 비회원에게는 관리자가 설정한 정보를 제공한다.

방송 콘텐츠 추천 시스템에서는 사용자의 소비이력을 바탕으로 아이템기반의 협업필터링 알고리즘을 사용하여 아이템간의 유사도를 도출하고 사용자가 선호하는 콘텐츠와 유사한 콘텐츠를 추천한다. 또한 방송 콘텐츠의 복잡한 메타데이터 구조를 분석하여 이를 사용자 선호 정보 수집에 활용하고 추천 알고리즘에 반영하여 방송 콘텐츠 추천에 적합한 시스템을 구성한다(오수영 외 2011).

3.3 추천서비스에 적용된 데이터

본 고에서의 연구개발을 위해 사용된 데이터는 2012~2016년을 기준으로 자체 생산한 전문정보 2,596건, 국내외 연계수집 자료 29,001건이며 회원 간 지식교류를 통해 축적된 정보는 23,271건(2015년 기준)이다. 이와 함께 회원 활동 웹로그 정보와 DB에 저장된 회원 정보를 활용하여 군집화에 활용하였다. 회원 정보는 학위, 직종, 거주국, 회원타입, 회원구분, 연구분야가 주요항목으로 설정되었다.

3.4 Spark 기반 빅데이터를 기반으로 한 개인화 추천서비스

최근 빅데이터를 빠르고 정확하게 고도 분석(advanced analytics)할 수 있는 기술은 새로운 정보 및 지식 발견과 직·간접적으로 연결되어 있어 기업 및 국가 경쟁력의 차별화 전략 수립에 많은 주목을 받고 있다. 고도 분석이란 대용량의 데이터에서 의사결정을 지원하기 위해 축적된 데이터 안에서 숨겨져 있는 유용한 정보를 발견하기 위한 분석 기법으로 연관성 분석, 클러스터링, 분류 등 데이터 마이닝 기법에 기반한 기술 분석(descriptive analytics) 뿐만 아니라 미래에 발생할 사건의 확률이나 경향에 대한 예측을 수행하기 위한 의사결정나무, 회귀분석, 신경망, 유전자 알고리즘, 기계학습 등 예측분석(predictive analytics)을 의미한다.

빅데이터의 고도 분석을 기반으로 재난, 재

〈표 1〉 사용이력에 따른 KOSEN 회원구분

구분	회원구분	내용
회원	보통회원	최근 180일 이내에 로그인 이력이 있는 회원
	휴면회원	180일 이내에 로그인 이력이 없는 회원
	신규회원	최근 60일 이내에 가입한 회원
비회원	비회원	미가입 회원

해, 식품안전, 테러, 범죄 등 사회현안에 합리적・선제적 대응을 강화할 수 있다. 따라서 빅데이터 분석을 필요한 정보를 수집 및 분류하고 이를 분석하여 빠른 의사결정에 활용하는 기법에 대한 연구가 활발하게 진행되고 있다. 고도 분석 지원을 위한 시스템의 진화는 하둡 플랫폼 기반, 자체 엔진을 내장한 하둡 플랫폼 기반 그리고 스파크 플랫폼 기반으로 구분할 수 있다(2016 정재화).

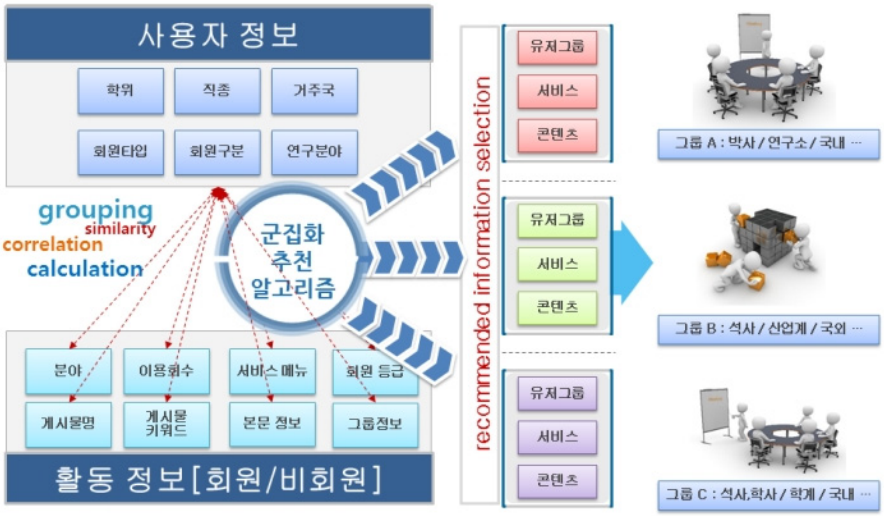
본 고에서 추천시스템에 도입한 Spark는 하둡의 맵리듀스 작업에서 성능의 병목현상으로 지적되던 디스크 I/O 비용을 최소화하고 데이터 분석 작업 에 용이한 인메모리 컴퓨팅 기반의 범용적 데이터 분산처리 시스템으로 실행속도가 빠른 것이 장점이다.

3.5 데이터 수집 및 분석 항목 도출

KOSEN의 회원정보 및 회원의 활동정보를 기반으로 데이터를 수집하여 수집된 데이터 중에서 회원의 특성을 구분하고 클러스터링 시 유저그룹의 변별력을 확연히 구분할 수 있는 항목에 대한 가중치를 부여 하고자 하고 사용자 정보에서는 학위, 직종, 거주국, 회원타입, 회원구분 및 연구 분야를 주요 항목으로 하여 데이터를 추출했다. 또한 정보서비스 시스템의 사용자 별 서비스 이용로그를 기반으로 빅데이터 기반의 저장 처리 기술을 통해 빅데이터화 하고 이를 기반으로 분석, 추천 알고리즘의 적용 및 관리자 관점에서의 사용자 추천을 위한 사용자 그룹, 사용자 연령등의 그룹핑을 통해 추천 할 수 있는 관리자 기능과 추천 기술을 적용하였다(〈그림 1〉 참조).

3.6 Apache Spark 시스템의 구성

정보서비스 시스템의 구성 레이어는 오라클 DB로부터 사용자의 활동정보, 사용자 정보, 콘



〈그림 1〉 사용자 그룹별 분석항목 수집

〈그림 2〉 KOSEN 추천서비스를 위한 Spark 시스템의 구성도

텐츠 정보 등을 추출하는 추출 레이어, 사용자 활동정보 및 사용자 정보를 수집하는 수집레이어, 수집된 데이터에서 사용자 별 군집, 콘텐츠 메타데이터 분서 및 서비스 로그 분류등의 기능을 수행하는 분석레이어와 추천 및 모니터링을 담당하는 서비스레이어등의 4개 영역으로 구성된다. 서비스 시스템은 사용자에게 추천 서비스를 제공하는 사용자 기반의 웹 서비스 페이지와 모바일 페이지로 구분되며, 리스트 서비스, 카드 서비스, 위젯서비스, 이미지 서비스의 정보 제공을 관리하는 관리자 기능으로 구분된다(〈그림 2〉 참조).

3.7 회원정보 및 이용로그 기반 클러스터링 산정

클러스터링에는 회원 이용자 정보 및 이용자 활동 정보 DB에서 Raw 데이터, 훈련 데이터(주제, 직위, 직종, 거주국 등), 회원 정보의 데이터 표준화와 특성 가중치 부여, 개별 사용자 활동 정보, 회원 특성별 서비스 메뉴 이용 합산, 군집별 서비스 메뉴 활용 빈도 합산과 정렬, 그룹별 서비스 메뉴 추천 결과를 포함한다.

4. 개인화 추천시스템 구축

4.1 추천시스템의 구성

KOSEN 정보서비스 시스템의 사용자는 크게 회원과 비회원으로 구분하고, 회원은 다시 보통회원, 휴면회원, 신규회원으로 구분하며 초기 사용정보가 거의 없는 휴면회원과 신규회원은 휴면회원과 신규회원이 속해져 있는 유저 그룹의 서비스를 추천하여 사용을 유도하고, 비회원에게는 관리자가 설정한 정보의 제공을 통해

〈그림 3〉 추천시스템의 구성

KOSEN의 서비스를 제공한다. 또한, 휴면회원과 신규회원은 시스템에서 설정되어 있는 설정 주기에 따라 자동으로 그룹의 설정을 통해 일정 시기가 지나면 개인에게 제공되는 맞춤형 정보 서비스를 확인 할 수 있도록 하는 것이 추천 시스템의 주요 기능이다(〈그림 3〉 참조).

4.2 회원 관심 서비스 및 콘텐츠 데이터 구축

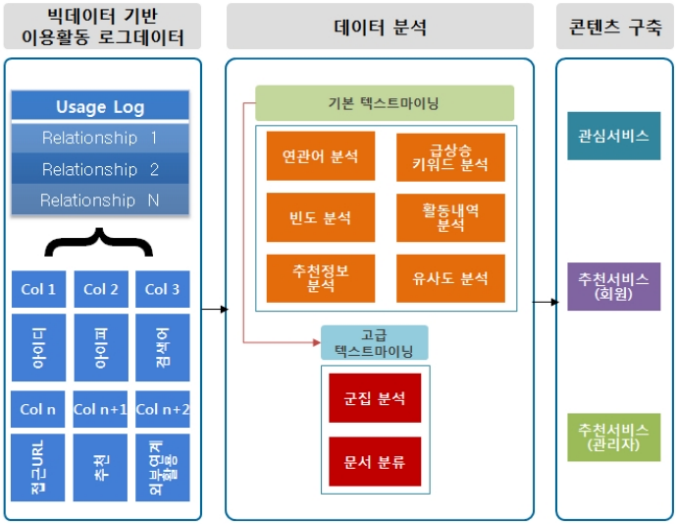
개인화 서비스 활용 증대 위하여 회원 이용자의 기본정보(직군, 학력, 전공분야, 회원등급 등)와 활동정보를 로그화 한 대량의 데이터를 기반으로 텍스트마이닝 과정을 통한 분석을 통하여 회원 개인별 관심서비스 정보를 구축하도록 하겠습니다. 또한, 분석 된 정보를 기반으로 회원 개인에게 추천할 수 있는 서비스 목록과 그에 따른 콘텐츠 정보를 구축하였다. 이는 전체 회원의 서비스 이용 활동 데이터 로그를 기반으로 대량의 데이터 분석을 거쳐 회원 이용자들의 배경 및 활동 정보를 텍스트마이닝 과정을 거쳐 빅데이터 처리기술에 기반한 개인화 서비스 제공을 위한 콘텐트를 구축하였다(〈그림 4〉 참조).

4.3 개인화 추천 알고리즘 개발

4.3.1 조건별 가중치 조정

개인화 추천 알고리즘 개발을 위한 조건별 가중치 조정은 회원정보와 서비스 이용 활동 정보를 기준으로 하였다. 일차 가중치 조정을 위한 회원정보의 레코드 수는 128,997건이며 회원정보는 〈표 2〉와 같다.

유사그룹 분류 생성을 위한 가중치 특성 적용하고 가중치는 총합계를 1로 설정하였으며, '분야〉학위=직종군〉거주국=회원등급=회원구분'의 조건별로 구분하였다. 유사그룹 가중치 적용 1과 적용 2는 〈그림 5〉, 〈그림 6〉과 같다.



〈그림 4〉 개인화 서비스 기반 데이터 구축

〈표 2〉 가중치 조정을 위한 회원정보

구 분	회원정보
전공분야	총 34분야
학 위	박사 / 석사 / 학사 / 기타
직 종	산업체 / 연구소 / 학교
거 주 국	총 86개국
회원종류	관리자 / 교정가 / 일반회원 / 전문가
회원구분	보통회원 / 신규회원 / 휴면회원

USER_SEQ	KOR_SCI_I	USER_DEC	USER_JOB	KOR_COU	USER_GRA	USER_CONDITION_TYPE
143119	생명과학	석사	학계	대한민국	일반회원	휴면회원
143120	건설/교통	박사	학계	대한민국	일반회원	휴면회원
143121	정보/통신	기타	연구소	대한민국	일반회원	휴면회원
143122	언어	기타	학계	대한민국	일반회원	휴면회원
143123	심리	박사	학계	대한민국	일반회원	보통회원
143124	화학	박사	연구소	대한민국	일반회원	보통회원
143125	언어	학사	학계	대한민국	일반회원	휴면회원
143126	보건의료	박사	연구소	대한민국	일반회원	휴면회원
143127	화학	기타	학계	캐나다	일반회원	휴면회원
143128	건설/교통	석사	연구소	대한민국	일반회원	휴면회원
143129	화학	석사	연구소	대한민국	일반회원	휴면회원
143130	생활	학사	학계	대한민국	일반회원	휴면회원
143131	환경	석사	연구소	대한민국	일반회원	휴면회원
143132	수학	석사	학계	대한민국	일반회원	휴면회원
143133	에너지/자	기타	학계	대한민국	일반회원	휴면회원
143134	기계	학사	산업계	대한민국	일반회원	휴면회원
143135	지구과학(	학사	학계	대한민국	일반회원	휴면회원
143136	전기/전자	석사	산업계	대한민국	일반회원	휴면회원
143137	생명과학	학사	학계	대한민국	일반회원	휴면회원
143138	에너지/자	학사	학계	대한민국	일반회원	휴면회원
143139	생명과학	학사	학계	대한민국	일반회원	휴면회원
143140	건설/교통	학사	학계	대한민국	일반회원	휴면회원
143141	건설/교통	기타	학계	대한민국	일반회원	휴면회원
143142	미디어/커	기타	학계	대한민국	일반회원	휴면회원
143143	지구과학(	석사	학계	대한민국	일반회원	휴면회원
143144	건설/교통	학사	산업계	대한민국	일반회원	휴면회원
143145	뇌과학	석사	학계	미국	일반회원	휴면회원
143146	건설/교통	석사	학계	미국	일반회원	휴면회원
143147	경제/경영	석사	학계	대한민국	일반회원	휴면회원
143148	건설/교통	석사	학계	미국	일반회원	휴면회원
143149	보건의료	석사	학계	미국	일반회원	휴면회원
143150	생명과학	석사	연구소	대한민국	일반회원	휴면회원

〈그림 5〉 유사그룹 가중치 적용 1

USER_SEQ	SERVICE_102000	SERVICE_103000	SERVICE_104000	SERVICE_201000	SERVICE_202000	SERVICE_204000	SERVICE_205000
105538	1	0	0	20	526	1	0
101459	0	2	1	0	2	0	0
106340	1	1	0	9	2	8	0
137609	0	0	0	1	2	0	0
23984	0	0	34	3	5	0	0
87142	0	0	4	7	214	2	0
22576	10	2	5	42	151	2	0
46687	0	1	0	57	15	2	0
128651	5	7	7	1	0	0	0
87516	7	1	7	13	4	0	0
146590	1	0	0	2	0	5	0
140976	0	0	0	10	5	0	0
2618	0	0	0	11	0	0	0
146410	0	0	0	1	0	0	0
128413	3	1	1	52	123	3	0
134699	1	0	3	1	1	0	0
87969	0	4	0	21	60	2	0
128812	4	3	0	20	17	4	0
31862	0	3	0	6	4	1	0
12131	2	0	0	1	42	0	0
95991	0	0	4	0	0	0	0
91278	1	1	2	4	0	5	0
106599	3	2	1	2	2	0	0
60412	0	0	0	0	4	0	0
85571	10	0	2	19	14	0	0
135493	0	1	0	24	11	1	0

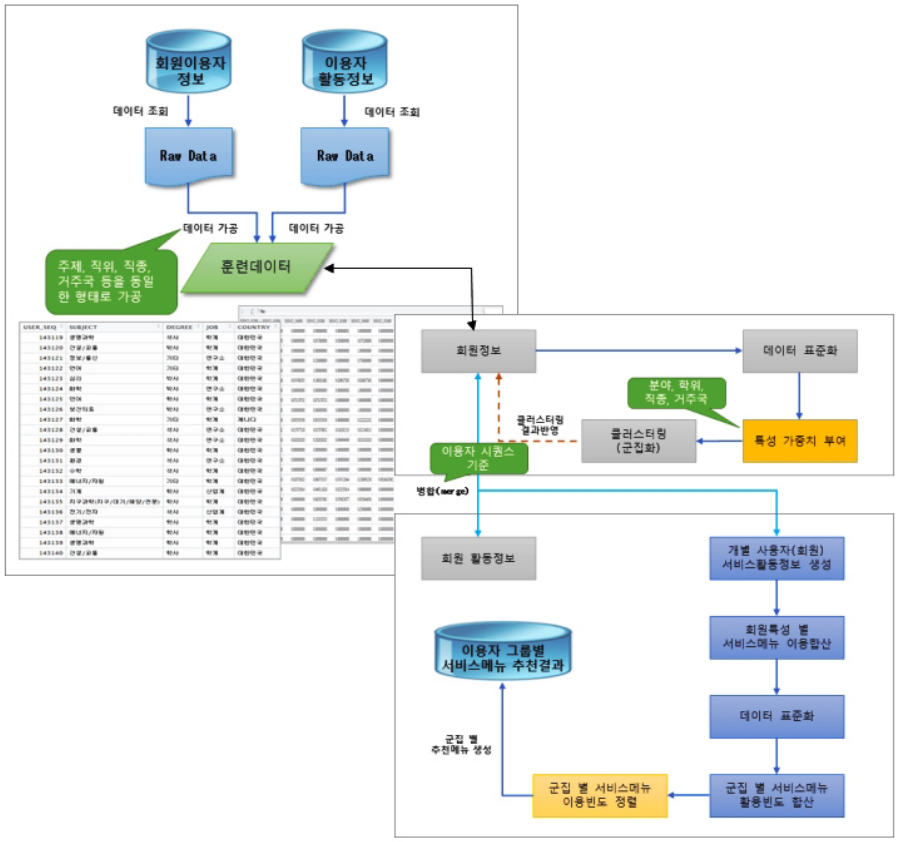
〈그림 6〉 유사그룹 가중치 적용 2

개인화 추천 알고리즘 개발을 위한 이차 조건별 가중치 조정을 위한 서비스 이용 활동 정보의 대상은 KOSEN 서비스 가입회원이 활동한 활동정보를 회원 별로 그룹핑 한 데이터이다. 서비스 이용정보는 이용자 27,265명이 이용한 활동정보를 서비스 메뉴 기준으로 분석하였다. 서비스는 리스트식 서비스와 카드식 서비스로 구분할 수 있다. 리스트식 서비스는 학회정보, What is?, 사업공고, 채용공고, 동향보고서, 코센리포트, 지식공감, 분석신청, 연구실정보이다. 카드식 서비스는 이슈토론, 글로벌뉴스, 동영상자료, 큐레이터, 행사알림, 보도자료, 해외기업, 코센웹진, 활용사례, 출장지원, 인포그래픽 등이다.

4.3.2 회원정보 및 이용로그 기반 클러스터링 산정 적용

회원 이용자 정보와 이용 활동정보를 추출하고 분야와 직위, 직종, 거주국을 동일 형태로 가공한 후 회원 정보 데이터 표준화 작업을 거쳐 특성 가중치를 부여하였다. 각 클러스터링 된 그룹의 산정 로직은 〈그림 7〉과 같다.

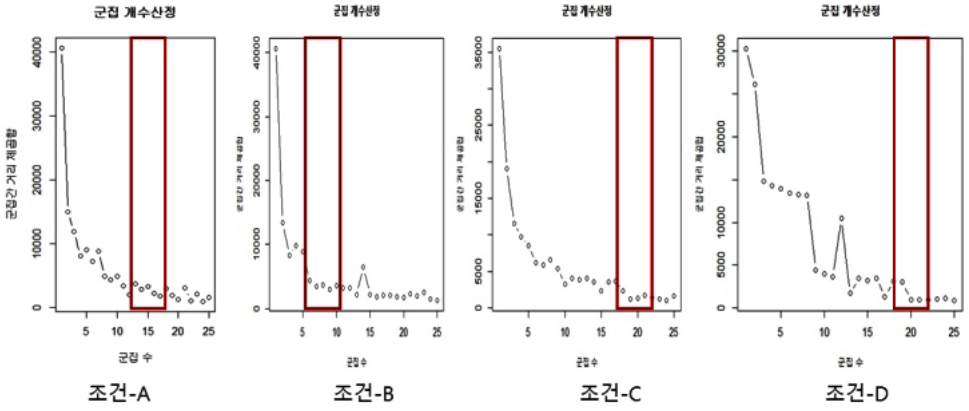
클러스터링 된 그룹의 가중치 조정 테스트 결과는 〈표 3〉과 같다.



〈그림 7〉 회원정보 및 이용로그 기반 클러스터링 산정 로직

〈표 3〉 조건별 가중치 적용 테스트

조건	클러스터수	분야	학위	직종	거주국	회원 등급	회원 타입
A	15	0.1	0.2	0.5	0.1	0.05	0.05
B	7	0.1	0.2	0.5	0.1	0.05	0.05
C	20	0.05	0.4	0.3	0.05	0.1	0.1
D	20	0.05	0.2	0.1	0.05	0.3	0.3
E	20	0.02	0.3	0.3	0.3	0.04	0.04
E-1	10	0.02	0.3	0.3	0.3	0.04	0.04
E-2	6	0.02	0.3	0.3	0.3	0.04	0.04
F-1	9	0.01	0.5	0.4	0.01	0.04	0.04
F-2	13	0.01	0.5	0.4	0.01	0.04	0.04
G-1	9	0	0.5	0.5	0	0	0
G-2	9	0	0.7	0.3	0	0	0
G-3	9	0	0.3	0.7	0	0	0



〈그림 8〉 군집 개수 산정 테스트 A, B ,C, D

군집 개수 산정 테스트 결과는 〈그림 8〉과 같다.

4.4 군집화 결과 분석

조건별로 군집 가중치 테스트를 실시하고 군집화 결과를 분석했다. 학위, 회원타입, 직종, 회원구분, 거주국에는 가중치를 부여하고 최종적으로 변별력 확보를 위해 학위와 직종에 각각 0.0485 가중치 적용하고 회원타입, 회원구분, 거주국에 0.01을 부여하여 총 1점이 되도록 가중치를 적용하였다. 이러한 군집화 분석을 통해서 9개 리스트 서비스의에 대해서 9개 그룹의 군집을 도출해내고 카드식 서비스에 대해서는 11개 메뉴에 대한 9개 그룹을 도출하였다.

회원을 특정 할 수 있는 항목을 5개(학위, 직종, 회원타입, 회원구분, 거주국)로 정의하고, 각 항목(독립변수)별로 가중치를 달리 적용하여 KOSEN 회원 그룹별 서비스 군집화 도출 및 콘텐츠 추천을 적용하였다. 군집화 된 회원

그룹별 개인화 기반의 서비스 메뉴 추천결과는 〈그림 9〉와 같다.

회원을 특정 할 수 있는 항목을 5개(학위, 직종, 회원타입, 회원구분, 거주국)로 정의하고, 각 항목(독립변수)별로 가중치를 달리 적용하여 KOSEN 회원 그룹별 서비스 군집화 도출 및 콘텐츠 추천을 적용하였다. 개인화 된 서비스 페이지의 예시는 〈그림 10〉과 같다.

		그룹 1	그룹 2	그룹 3	그룹 4	그룹 5	그룹 6	그룹 7	그룹 8	그룹 9
메뉴형식	추천순위	메뉴명	메뉴명	메뉴명	메뉴명	메뉴명	메뉴명	메뉴명	메뉴명	메뉴명
리스트	1	코센리포트	코센리포트	코센리포트	코센리포트	코센리포트	코센리포트	코센리포트	분석신청	코센리포트
리스트	2	분석신청	What is?	지식공감	분석신청	학회정보	What is?	What is?	What is?	동향보고서
리스트	3	동향보고서	동향보고서	동향보고서	What is?	What is?	지식공감	연구실정보	코센리포트	지식공감
리스트	4	지식공감	지식공감	What is?	동향보고서	지식공감	분석신청	지식공감	지식공감	분석신청
리스트	5	What is?	채용공고	학회정보	지식공감	연구실정보	동향보고서	동향보고서	학회정보	What is?
리스트	6	채용공고	분석신청	채용공고	학회정보	동향보고서	학회정보	분석신청	채용공고	학회정보
리스트	7	학회정보	학회정보	사업공고	채용공고	분석신청	채용공고	학회정보	동향보고서	채용공고
리스트	8	연구실정보	연구실정보	연구실정보	연구실정보	채용공고	연구실정보	채용공고	사업공고	연구실정보
리스트	9	사업공고	사업공고	분석신청	사업공고	사업공고	사업공고	사업공고	연구실정보	사업공고
메뉴형식	추천순위	메뉴명	메뉴명	메뉴명	메뉴명	메뉴명	메뉴명	메뉴명	메뉴명	메뉴명
카드식	1	지식큐레이터	지식큐레이터	글로벌뉴스	지식큐레이터	지식큐레이터	지식큐레이터	지식큐레이터	지식큐레이터	지식큐레이터
카드식	2	글로벌뉴스	글로벌뉴스	지식큐레이터	해외출장지원	글로벌뉴스	글로벌뉴스	글로벌뉴스	글로벌뉴스	글로벌뉴스
카드식	3	해외출장지원	동영상자료	동영상자료	글로벌뉴스	동영상자료	해외출장지원	동영상자료	해외출장지원	동영상자료
카드식	4	동영상자료	해외출장지원	해외기업네트워크	해외기업네트워크	해외출장지원	동영상자료	해외출장지원	동영상자료	해외출장지원
카드식	5	해외기업네트워크	해외기업네트워크	해외출장지원	동영상자료	해외기업네트워크	해외기업네트워크	해외기업네트워크	이슈토론	해외기업네트워크
카드식	6	인포그래픽	이슈토론	KOSEN 웹진	KOSEN 웹진	KOSEN 웹진	KOSEN 웹진	KOSEN 웹진	해외기업네트워크	KOSEN 웹진
카드식	7	KOSEN 웹진	KOSEN 웹진	인포그래픽	인포그래픽	이슈토론	이슈토론	이슈토론	KOSEN 웹진	인포그래픽
카드식	8	이슈토론	인포그래픽	이슈토론	이슈토론	인포그래픽	인포그래픽	인포그래픽	인포그래픽	이슈토론
카드식	9	행사알림	행사알림	행사알림	활용사례	행사알림	행사알림	행사알림	행사알림	행사알림
카드식	10	활용사례	활용사례	활용사례	행사알림	활용사례	활용사례	보도자료	보도자료	활용사례
카드식	11	보도자료	보도자료	보도자료	보도자료	보도자료	보도자료	활용사례	활용사례	보도자료

〈그림 9〉 군집화 된 회원그룹 별 개인화 기반의 서비스메뉴 추천결과

〈그림 10〉 개인화 된 서비스 페이지 예시

4.5 개인화가 적용된 서비스 페이지

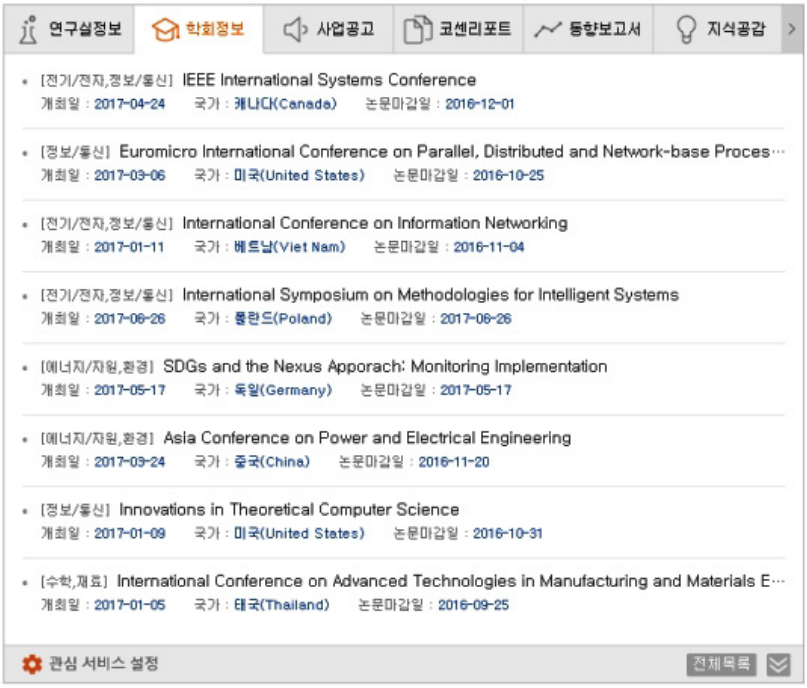
4.5.1 개인화가 적용된 리스트식 서비스

리스트 서비스는 총 9개의 서비스 메뉴로 구성하고, 로그인 후 개인화가 적용된 시점에 유저가 포함되어져 있는 유저그룹의 주 이용 서비스를 기반으로 9개 서비스 메뉴에 대해 순차적으로 제공한다. 9개의 서비스 메뉴는 연구실정보, 학회정보, 사업공고, 코센리포트, 동향보고서, 지식공감, Whatis?, 채용공고, 분석신청 등으로 구성된다.

로그인 시점에 리스트 서비스의 하단에 관심서비스 설정 기능이 활성화 되고, 활성화 이후 개인 사용자의 설정에 의해 9개 메뉴 중에서 본인이 선호하는 서비스를 추가로 설정할 수 있다. 단, 관리자가 추천을 위해 최우선 순위로 설정한 메뉴에 대해서는 변경이 불가능하며 주기적 분석에 의해 추천되는 서비스의 순서는 변경 적용된다. 개인화가 적용된 리스트식 서비스 예시는 〈그림 11〉과 같다.

4.5.2 개인화 적용된 카드식 서비스

카드 서비스는 총 11개의 서비스 메뉴로 구성하고, 로그인 후 개인화가 적용된 시점에 유저가 포함되어져 있는 유저그룹의 주 이용 서비스를 기반으로 11개 서비스 메뉴에 대해 순차적으로 제공한다. 11개의 서비스 메뉴는 글로벌뉴스, 지식큐레이터, 동영상자료, 해외출장지원, 해외기업, 코센웹진, 이슈토론, 인포그래픽, 한인행사, 활동사례, 보도자료 등으로 구성된다.



〈그림 11〉 개인화가 적용된 리스트식 서비스 예시

〈그림 12〉 개인화가 적용된 카드식 서비스 예시

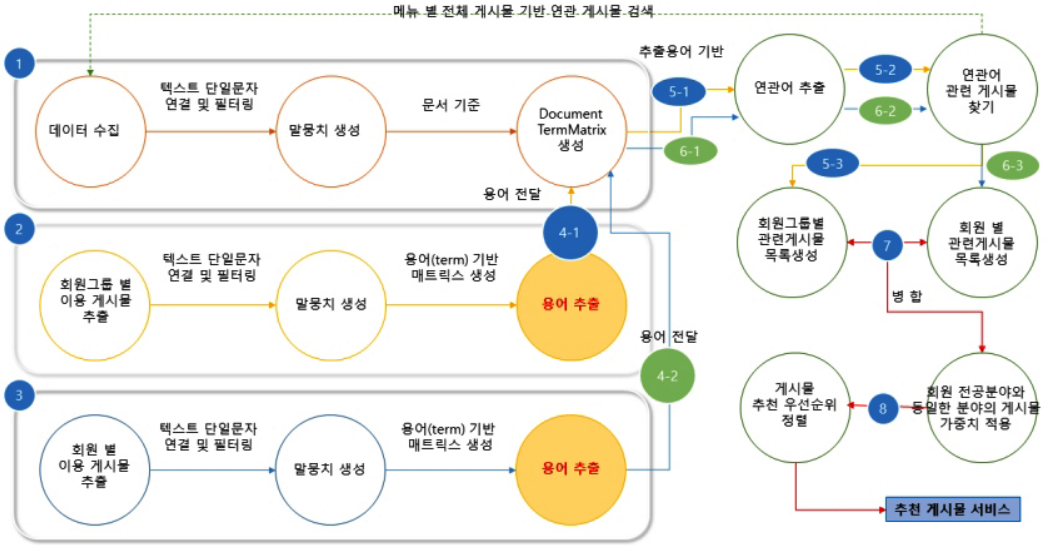
로그인 시점에 카드 서비스의 하단에 관심 서비스 설정 기능이 활성화 되고, 활성화 이후 개인 사용자의 설정에 의해 11개 메뉴 중에서 본인이 선호하는 서비스를 추가로 설정 할 수 있다. 단, 관리자가 추천을 위해 최우선 순위로 설정한 메뉴에 대해서는 변경이 불가능하며 주기적 분석에 의해 추천되는 서비스의 순서는 변경 적용된다. 개인화가 적용된 카드식 서비스 예시는 〈그림 12〉와 같다.

4.6 콘텐트 추천 알고리즘 설계

우선 리스트식과 카드식의 메뉴별 전체 게시물을 대상으로 게시물명과 본문내용, 게시물 분야 등을 텍스트 마이닝 기법 적용을 위한 데이터 항목을 선정하고 수집하였다. 이후 메뉴별 전체 게시물 정보를 식별할 수 있는 정보를 기반으로 언어의 표본을 추출한 집단(말뭉치)를 생성하고 텍스트 문서를 기반으로 단어/용어의 표현빈도를 매트릭스 형태로 생성하고 용어추출시 TF-IDF(TF: 문서별 단어 빈도, IDF: 역문서 빈도)방식을 사용하여 알고리즘을 구성하였다. 이를 통한 추천 리스트는 분석 시스템을 통해 추천되는 각 서비스 별 서브 페이지 상에서 사용자가 이용하지 않았던 게시물 중에서 유저 그룹이 사용했던 정보를 기반으로 추천하는 기능으로 구성되어 서비스된다. 설계된 콘텐츠 추천 알고리즘 모식도는 〈그림 13〉과 같다.

5. 결론 및 제언

과학기술정보는 전문성 및 신속성을 요구하기 때문에, 정보의 홍수 시대에 우수한 휴먼네트워크의 구축 및 활용이 매우 중요하다 할 수 있겠다. 연구자들에게 지식을 습득하여 연구 활동에 도입하는데 걸리는 소요시간을 단축하는 것은 연구생산성 향상에 필수적인 요소라고 할 수 있다.



〈그림 13〉 콘텐츠 추천 알고리즘 모식도

본 연구에서는 사용자 기반 개인화 서비스 추천 시스템 설계에 중점을 두었다. 회원 서비스 이용 활성화를 위해 회원정보와 활동 로그 정보를 분석하여 그룹별 추천 서비스를 제공하였으며, 회원들의 선호도에 따라 노출방식을 변경하여 원하는 정보를 직관적으로 확인할 수 있도록 하였다. 또한, 신규 서비스는 우선 노출하되 회원이 서비스의 설정을 변경할 수 있는 기능도 제공되고 있다.

본 연구의 목적은 한민족과학기술자네트워크(KOSEN) 사용자들의 정보 이용 패턴을 군집화하고 그룹화 된 사용자들에게 맞는 개인화 추천서비스 알고리즘의 최적화 방안을 제안하는 것이다. 사용자들의 연구활동과 이용정보에 기반하여 적합한 서비스와 콘텐츠를 식별한 후 빅데이터 분석 기술을 적용하여 개인화 추천 알고리즘을 도출하였다. 이 추천 알고리즘은 사용자의 정보검색에 소요되는 시간을 절약하고 적합한 정보를 찾아내는데 도움을 줄 수 있다. 향후 지속적인 서비스 변수별 테스트와 고도화 작업을 통해서 사용자별로 선호모델을 재해석하고 가중치의 조절 등을 통한 고객 맞춤형 콘텐츠 제공과 함께 사용자 모델 알고리즘 집합을 통한 복잡한 데이터 사이의 연관관계를 추측하여 다양한 패턴에서의 의미의 정보를 추출해 낼 것이다.

콘텐츠의 양과 질은 회원 참여도와 밀접한 관계가 있으므로 사용자 기반의 알고리즘 적용과 더불어 한인과학기술자들을 통한 다양한 콘텐츠 구축을 위해서는 온·오프라인에서의 다양한 홍보와 함께 국제 공동연구를 통한 해외 각지의 한인과학기술자들과의 유기적인 관계 형성이 함께 해야 할 것이다. 과학기술자들을 위한 맞춤형 개인화 서비스는 향후 과학기술자들에게 도움이 되는 과학기술 정보서비스 플랫폼의 방향을 설정하는 데에 큰 도움이 되리라 기대된다.

참 고 문 헌

김용, 문성빈. 2007. 대용량 음악콘텐츠 환경에서의 데이터마이닝 기법을 활용한 추천시스템에 관한 연구.『정보관리학회지』, 24(2): 89-104.

박성은, 황윤영, 윤정선. 2017. 과학 학술정보 서비스 플랫폼에서 개인화를 적용한 콘텐츠 추천 알고리즘 최적화를 통한 추천 결과의 성능 평가.『한국콘텐츠학회』, 17(11): 183-191.

오수영, 오연희, 한성희, 김희정. 2011. 사용자 소비이력기반 방송 콘텐츠 추천 시스템.『방송공학회』, 17(1): 129-139.

유영석, 김지연, 손방용, 정종진. 2017. 온라인 음악 콘텐츠 추천 시스템 구현을 위한 협업 필터링 기법들의 비교 평가.『대한전기학회』, 66(7): 1083-1091.

윤정선, 정혜주, 한선화. 2010. 사용자 참여형 데이터베이스 구축 연구: 코센 오픈랩 운영사례를 중심으로.『정보관리연구』, 41(2): 95-110.

정선양, 김기동. 2008. 산학연 협력의 새로운 방향: 산학연 협력연구실 구축을 중심으로.『기술혁신연구』, 16(2): 17-40.

정인용, 양새동, 정회경. 2015. 혼합 필터링 기반의 영화 추천 시스템에 관한 연구.『한국정보통신학회논문지』, 19(1): 113-118.

정재화. 2016. Spark SQL 기반 고도 분석 지원 프레임워크 설계.『한국정보처리학회』, 5(10): 477-482.

최가현, 황윤영, 윤정선. 2017. 회원정보 활용 그룹별 추천 서비스 적용에 관한 연구 - 국내외 한인과학자들을 중심으로.『2017년 한국컴퓨터종합학술대회 논문집』, 2017년 06월 18-20일. 제주: 전자정보연구센터. 1466-1468.

최성우, 한성희, 정병희. 2014. 협업 필터링 기반의 콘텐츠 추천 시스템과 빅데이터 처리 솔루션을 이용한 상용화 개발 방향.『방송공학회지』, 19(4): 50-59.

한선화. 2005. 휴먼네트워크를 활용한 정보 수집 및 분석.『지식정보인프라』, 18(4): 51-55.

KOSEN homepage. [online]. [cited 2017.12.15]. 〈http://kosen21.org〉.

• 국문 참고자료의 영어 표기
(English translation / romanization of references originally written in Korean)

Choi, GaHyun, Yun-Young Hwang, and JungSun Yoon. 2017. "A Study on the Application of Recommendation Services to Member Information Usage Groups." Jeju: 2017 Electronic & Information Research Information Center. 2017. June 18-20. 1466-1468.

Choi, S. W., S. H. Han, and B. Jung. 2014. "Content recommendation system based on the collaborative filtering and big-data solutions for its commercialization." *Journal of broadcast*

engineering, 19(4): 50-59.

Chung, Jaehwa. 2016. "Design of Spark SQL Based Framework for Advanced Analytics." *KIPS transactions on software and data engineering*, 5(10): 477-482.

Hahn, Sun-Hwa. 2005. "Collection and analysis of information using human network." *Journal of scientific & technological knowledge infrastructure*, 18(4): 51-55.

Jeong I., X. Yang, and H. Jung. 2015. "A Study on Movies Recommendation System of Hybrid Filtering-Based." *Journal of the Korea Institute of Information and Communication Engineering*, 19(1): 113-118.

Jung, S. Y. and K. D. Kim. 2008. "The New Approach to the Collaboration Among Academia, Industry, and Public Research Sector: Focussing on Building a Collaboration Research Center." *Technical Innovation Research*, 16(2): 17-40.

Kim, Yong and Sung-Been Moon. 2007. "A Study on Recommendation System Using Data Mining Techniques for Large-sized Music Contents." *Journal of the Korean society for information management*, 24(2): 89-104.

KOSEN homepage. [online]. [cited 2017.12.15]. 〈http://kosen21.org〉.

Oh, Soo-Young et al. 2011. "Broadcast Content Recommender System based on User's Viewing History." *Journal of broadcast engineering*, 17(1): 129-139.

Park, Seong-Eun, Yun-Young Hwang, and Jungsun Yoon. 2017. "Performance Evaluation of Recommendation Results through Optimization on Content Recommendation Algorithm Applying Personalization in Scientific Information Service Platform." *Journal of Contents Association*, 17(11): 183-191.

Yoo, Youngseok et al. 2017. "Evaluation of Collaborative Filtering Methods for Developing Online Music Contents Recommendation System." *The Transactions of the Korean Institute of Electrical Engineers*, 66(7): 1083-1091.

Yoon, Jungsun, Hye-Ju Jung, and Sun-Hwa Hahn. 2010. "Study on the Building up the Laboratory Database: Case Study from the KOSEN OpenLab Service." *Journal of information management*, 41(2): 95-110.