Journal of KIIT. Vol. 16, No. 6, pp. 91-99, Jun. 30, 2018. pISSN 1598-8619, eISSN 2093-7571 91
http://dx.doi.org/10.14801/jkiit.2018.16.6.91

# 키워드 기반 추천시스템 데이터 셋 구축 및 분석

배은영*, 유석종**

# Keyword-based Recommender System Dataset Construction and Analysis

Eun-Young Bae*, Seok-Jong Yu**

## 요 약

추천시스템(Recommender System)이란 대상자가 좋아할 만한 무언가를 추천하는 시스템을 일컫는다. 아마존의 상품 추천, 페이스북이나 링크드인의 친구 추천, 넷플릭스와 왓차의 영화 추천, 행태 기반 광고, 뉴스 추천 등 여러 분야에서 추천시스템은 이미 널리 활용 중이며 진화 중에 있다. 학계에서도 추천시스템에 대한 관심과 연구는 꾸준하게 증가를 하고 있으며, 이 분야의 논문 수 또한 해마다 증가하고 있고, 연구 분야 또한 점점 세분화되고 다양해지고 있다. 본 논문에서는 추천시스템에 관한 주요 연구 주제나 적용 대상 도메인, 선호 기법, 연구 주제에 대한 트렌드 등을 파악하기 위하여 IEEE Xplore 전자 도서관 및 ACM 전자 도서관으로부터 얻은 추천시스템 논문 관련 자료를 토대로 데이터 셋을 구축하고 분석을 진행하였다.

## Abstract

Recommender system is a system that recommends something that a person would like. Recommender systems are widely used and evolving in various fields such as Amazon product recommendation, Facebook or Linkedin friend recommendation, Netflix and Watcha movie recommendation, behavior based advertisement, news recommendation. In the academia, interest and research on recommender systems are steadily increasing, and the number of papers in this field is increasing year by year, and the field of research is becoming more and more fragmented and diversified.

In this paper, we constructed a dataset based on dissertation data of recommender system obtained from IEEE Xplore Digital Library and ACM Digital Library, and proceeded with the analysis in order to grasp major research topics, application domains, preference techniques, and trends in research topics.

Keywords
recommender system, collaborative filtering, recommender trend

* 숙명여자대학교 소프트웨어학부
- ORCID: https://orcid.org/0000-0002-7253-0778
** 숙명여자대학교 소프트웨어학부(교신저자)
- ORCID: https://orcid.org/0000-0002-1631-4034

· Received: Mar. 09, 2018, Revised: Jun. 19, 2018, Accepted: Jun. 22, 2018
· Corresponding Author: Seok-Jong Yu
Department of Software, Sookmyung Womens's University, Korea
Tel.: +82-2-710-9831, Email: sjyu@sm.ac.kr

# Ⅰ. 서 론

추천시스템은 사용자가 아이템에 대해 어떻게 평가할 지를 예측하는 정보 필터링 시스템의 일종이다. “과거가 미래다”라는 말이 있듯이 과거의 데이터를 기반으로 하여 사용자가 원하는 최적화된 아이템이 무엇인가를 예측하고 추천하는 일은 많은 산업군에서 필요한 부분이며, 이미 아마존의 상품 추천, 페이스북이나 링크드인의 친구 추천, 넷플릭스와 왓차의 영화 추천, 행태 기반 광고, 뉴스 추천, 쇼핑몰의 상품 추천 등 다양한 영역에서 활용되고 있다. 추천시스템에 대한 관심 및 활용도가 높아지면서 학계의 연구 방향도 다양해지고 있다. 1998년부터 등장하기 시작한 추천 시스템 관련 논문은 그 수가 해마다 증가하고 있고, 연구 분야 또한 세분화되고 다양한 형태로 변화되고 있다. 본 논문에서는 IEEE 및 ACM의 전자 도서관에서 제공되는 논문 관련 자료를 가지고 데이터 셋을 구축하고, 논문의 키워드를 중심으로 과거부터 현재까지 추천시스템의 주요 연구 주제, 주요 이슈의 변화에 대한 흐름 등을 확인해 보고자 하였다.

# Ⅱ. 추천시스템의 이해

## 2.1 추천시스템 개요

1990년대 중후반 협업필터링에 관한 연구 논문[1]이 등장한 이후, 추천시스템은 학계에서 주요한 연구 분야가 되었다. 추천시스템은 선호도 또는 사용자에 의해 아이템에 주어지는 평점을 예측하는 정보 필터링 시스템이라고 할 수 있으며, 추천시스템의 목표는 사용자가 관심을 가질만한 아이템이나 제품들에 대해 유용하고도 분별 있는 추천을 해주는 것이다.

## 2.2 추천시스템 분류

추천시스템은 사용하는 정보 등에 따라 보통 아래와 같이 분류가 된다[2]-[7].

- 내용 기반 필터링(Content based Filtering)

내용 기반 필터링은 아이템 자체를 분석하여 추천을 구현하는 기법으로, 내용 기반 필터링을 위해서는 아이템을 분석한 프로파일(Item Profile)과 사용자의 선호도 프로파일(User Profile)을 추출하여 이의 유사성을 계산한다. 이 기법은 아이템의 내용을 분석해야 하므로 아이템 분석 알고리즘이 핵심적이며, 이를 위해 군집분석(Clustering Analysis), 인공신경망(Artificial Neural Network), tf-idf(term frequency inverse document frequency) 등의 기술이 주로 사용된다.

- 협업 필터링(CF, Collaborative Filtering)

협업 필터링이란 사용자의 행위를 모델링하고 유사도가 높은 사용자들의 행위를 예측하는 방식으로, 비슷한 패턴을 가진 사용자나 항목을 추출하는 기술이 핵심이며 주로 행렬분해(Matrix Factorization), k-최근접 이웃 알고리즘(kNN, k-nearest Neighbor Algorithm) 등의 방법이 사용된다. 협업 필터링을 위해서는 반드시 기존 자료를 활용해야 하고, 사용자의 행위로 분석하기 때문에 높은 정확도를 위해서는 많은 데이터가 있어야 한다. 협업 필터링 방식은 메모리 기반 협업 필터링(Memory based CF)과 모델 기반 협업 필터링(Model based CF)이 있다. 메모리 기반 협업 필터링은 사용자와 아이템에 대한 액션을 모두 메모리 위에 올려두고 사용자-아이템 간의 관계를 계산하는 방법이며, 모델 기반 협업 필터링은 베이즈 네트워크(Bayesian Network)나 인공 신경망 등 다양한 머신러닝 기법을 통해서 추천을 해주는 방식이다.

- 하이브리드 추천 시스템(Hybrid Recommender system)

내용 기반 필터링과 협업 필터링을 결합하는 등 다양한 추천 방식을 조합하여 추천하는 시스템이다.

## 2.3 추천시스템 주요 연구 주제

추천시스템 연구에서 다루어지는 주요 연구 주제에는 초기 사용자 문제(Cold Start), 확장성(Scalability),

희소성(Sparsity), 다양성(Diversity), 예측 가능성(Predictability), 동의어(Synonymy), 악의적 평가 점수 입력(Shilling Attack), 선호 특성 항목의 반복 추천(Overspecialization), 비 일관적 평가 점수 입력(Grey Sheep), 관심 적은 다수 항목 추천(Long Tail Item Recommendation), 참신성(Novelty), 개인 정보 보호(Privacy), 오류 및 악의적인 데이터(Erroneous and Malicious Data), 빅 데이터 등이 있으며 이 중 몇 가지를 소개하면 아래와 같다[4]-[8].

- 초기 사용자 문제 : 새로운 항목이 추가되는 경우 이를 추천할 수 있는 정보가 쌓일 때까지는 정확한 추천이 어려워진다는 것을 의미한다.

- 희소성 : 막대한 양의 아이템에 대비하여 사용자의 평점의 양은 아주 희소할 수 있으며, 이 문제로 인해 예측의 정확도가 감소할 수 있다. 희소성 문제를 해결하기 위하여 특이값 분해(SVD, Singular Value Decomposition), 인구통계학적 필터링, 내용기반 필터링 기법 등이 제안되고 있다.

- 동의어 : “comedy movie”와 “comedy film”과 같이 유사한 아이템에 대하여 다른 이름으로도 불리어지는 경우를 의미하며, 이 문제를 해결하기 위하여 온톨로지(Ontologies), 특이값 분해, LSA(Latent Semantic Analysis) 등이 사용된다.

- 악의적 평가 : 악의적인 사용자가 아이템에 대하여 의도적으로 잘못된 평점을 넣는 경우를 의미한다.

- 비 일관된 평가 : 독특한 취향의 사용자에 의한 추천을 의미하며, k-mean 클러스터링 등의 방법을 통해 다른 사용자들로부터 독특한 사용자를 분리해내는 기법을 사용한다.

### 2.4 연구 배경

추천시스템은 서비스에 큰 영향을 미치는 요소이기 때문에 오래 전부터 많은 학자들과 기업들에서 연구를 하고 있는 주요한 분야 중의 하나이다.

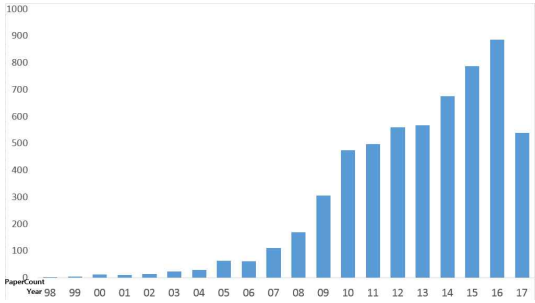


그림 1. 연도별 논문편수
Fig. 1. Number of articles per year

그림 1은 IEEE 데이터 셋과 ACM 데이터 셋을 합산한 추천시스템 논문 편수이다. 1998년부터 등장하기 시작한 추천시스템 관련 논문은 해마다 증가를 하고 있으며, 특히 2009년경부터 급격한 증가추세를 보이고 있음을 알 수 있다.

본 본문에서는 매년 증가하고 있는 추천시스템에 대한 주요 연구 주제가 무엇인지, 연구 도메인은 무엇인지, 연구 주제는 어떠한 방향으로 다양해지고 있는지 등에 대해 알아보고자 하였으며, 이를 위하여 IEEE 및 ACM 전자 도서관에서 제공되는 논문 정보를 활용하여 키워드를 중심으로 분석을 진행하기 위하여 데이터 셋을 구축하였고, 분석을 시도해 보았다.

## III. 추천시스템 키워드 기반 데이터 셋 구축

### 3.1 추천시스템 주제 분석을 위한 데이터 셋 구축 절차

추천시스템 연구 주제 분석을 위한 데이터 셋 구축 절차는 그림 2와 같으며, 이를 위하여 본 논문에서는 IEEE 및 ACM의 전자도서관에서 제공되는 자료를 활용하였다.

전처리 작업을 통하여 결측값을 제거하고, 논문들에서 제시되는 키워드 묶음을 각각의 키워드로 분리한 후 논문번호와 년도에 일대일 매핑작업을 진행하였고, 유사 의미를 갖는 키워드에 대해서는 그룹핑 작업을 수행하였다. IEEE와 ACM의 자료는 합치지 않고 각각의 자료를 별도로 하여 분석을 진행하였다.

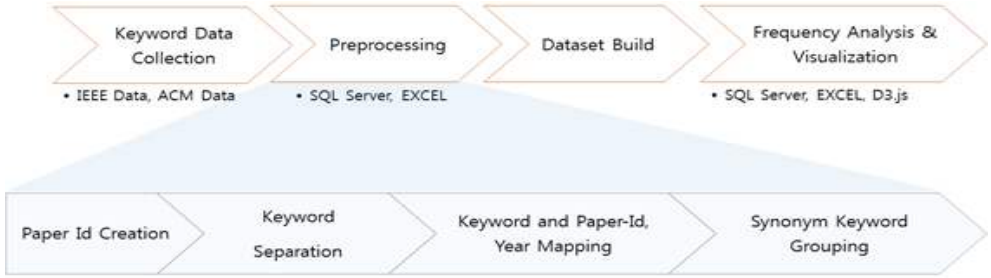


그림 2. 추천시스템 주제 동향 분석을 위한 데이터 셋 구축 절차 및 전처리 단계
Fig. 2. Dataset construction procedure and pre-processing step

### 3.2 데이터 셋 구축을 위한 전처리 및 데이터 매핑

IEEE 및 ACM의 전자도서관으로부터 얻은 자료는 xls파일과 csv파일이며, 그림 2와 같은 절차로 전처리 작업을 진행하였다.

IEEE 자료의 경우는 IEEE Xplore 전자도서관[9]로부터 2017년 12월 기준, 1998년부터 2017년까지의 추천시스템 논문 자료 5839건을 xls파일 형태로 받은 후 전처리 작업을 진행하였다. 제공된 총 31개의 필드 중에서 데이터 셋 구축을 위한 용도로 논문제목, 저자가 제시한 키워드, 논문발행년도 필드를 사용하였으며, 저자가 제시한 키워드가 제시되어 있지 않은 논문에 대해서는 논문의 내용이 비교적 구체적으로 표현되어 있는 INSPEC non-controlled Terms을 사용하였다. 사용된 필드 및 데이터 값의 예는 그림 3과 같다. IEEE 자료에서는 논문을 식별하는 고유번호가 제공이 되지 않았기 때문에 각 논문마다 논문번호(Id)를 생성하였으며, 논문들에서 제시된 키워드의 묶음을 각각의 키워드로 분리를 하고, 분리된 각 키워드들은 그림 5와 같이 논문번호 및 논문발행년도에 일대일 매핑시켰다.

ACM 자료의 경우에는 ACM 전자도서관[10]로부터 2017년 12월 기준, 1998년부터 2017년까지 추천시스템에 대한 논문 자료 2007건을 csv파일 형태로 받은 후 전처리 작업을 진행하였다. 제공된 총 27개의 필드 중에서 논문번호, 논문제목, 논문발행년도, 키워드 필드를 사용하였으며, IEEE 자료처리처럼 키워드 묶음을 각각의 키워드로 분리하고 분리된 각 키워드마다 논문번호와 논문발행년도를 매핑하는 작업을 진행하였다. ACM 데이터 셋 구축을 위하여 사용된 필드 및 데이터 예는 그림 4와 같다.

| IEEE DataSet Field | Data |
|---|---|
| Document Title | Cross-Domain Recommender Systems |
| Authors | P. Cremonesi; A. Tripodi; R. Turrin |
| Author Keywords | cross-domain;neighborhood-based;recommender |
| INSPEC Non-Controlled Terms | Netflix dataset;cross domain collaborative recommender system;direct graph;live TV programs;mobile applications;music;neighborhood collaborative filtering;on- |
| Year | 2011 |

그림 3. IEEE 데이터 셋의 필드와 데이터 예시
Fig. 3. IEEE dataset field & sample data

| ACM Dataset Field | Data |
|---|---|
| type | article |
| id | 3110321 |
| title | Towards a Deep Learning Model for Hybrid |
| author | Gabriele Sottocornola and Fabio Stella and Markus |
| year | 2017 |
| keywords | deep learning, hybrid recommendation system, recommendation systems |

그림 4. ACM 데이터 셋의 필드와 데이터 예시
Fig. 4. ACM dataset field & sample data

| Paper ID | Keyword | Year |
|---|---|---|
| 189 | visualization | 2009 |
| 543 | visualization | 2012 |
| 1167 | visualization | 2014 |
| 1270 | visualization | 2014 |
| 312 | visualization of recommendations | 2012 |
| 1458 | visualization recommendation | 2015 |
| 1780 | visualization recommender | 2016 |
| 1986 | visualizations | 2017 |
| 267 | voting | 2010 |
| 1458 | wandering | 2015 |
| 1654 | watson analyitcis | 2016 |

그림 5. 논문 키워드와 논문id, 논문발행년도 매핑 예시
Fig. 5. Example of keyword, id, and year mappings

### 3.3 유사 키워드 그룹핑

일대일 매핑 작업을 한 데이터는 MS-SQL 서버를 사용하여 데이터베이스로 구축한 후 그룹핑 작업을 하였다[그림 6]. 먼저 데이터가 제대로 들어가 있지 않은 레코드를 삭제한 후에 동의어 처리를 위해서 모든 문자는 소문자로 변환하고, 복수 형태의 단어는 단수 형태로 통일화하였다.

| GroupingKeyword | Keyword |
|---|---|
| content-based filtering | content-based approach |
| | content-based features |
| | content-based filtering |
| | content-based recommendation |
| cold-start | cold-start problem |
| | cold-start problems |
| | cold-start recommendation |
| similarity | similarity measure |
| | similarity measurer |
| | similarity measures |
| SVM | support vector machine |
| | SVM |

그림 6. 유사 키워드 그룹핑 예
Fig. 6. Synonym grouping example

| Rank | Keyword | Count |
|---|---|---|
| 1 | collaborative-filtering | 1134 |
| 2 | personalized recommenda | 356 |
| 3 | social-network | 314 |
| 4 | e-commerce | 280 |
| 5 | user preference | 207 |
| 6 | movielens | 178 |
| 7 | cold-start | 171 |
| 8 | sparsity | 147 |
| 9 | data-mining | 145 |
| 11 | hybrid recommendation | 112 |
| 12 | content-based filtering | 112 |
| 13 | association rule | 111 |
| 15 | user profile | 105 |
| 16 | context-aware recommend | 102 |
| 17 | information filtering | 98 |
| 18 | decision-making | 98 |
| 20 | matrix-factorization | 94 |
| 21 | machine-learning | 93 |
| 22 | item-based collaborative-f | 91 |
| 23 | content-based recommend | 81 |
| 24 | user-based collaborative-f | 80 |

그림 7. 키워드 누적빈도(IEEE)
Fig. 7. Cumulative frequency per keyword

또한 복합어로 구성된 키워드나 유사 의미이면서 다르게 표현된 키워드에 대해서는 동일한 형식으로 표준화하였다. SVM와 support vector machine의 예처럼 같은 용어인데 다르게 표현된 키워드는 모두 약어로 통일화하여 진행을 하였다. ACM 자료에 대한 전처리 작업도 IEEE 자료와 유사한 방법으로 진행하였다.

## IV. 키워드 중심의 빈도 분석 및 시각화

### 4.1 IEEE 데이터 셋을 활용한 분석

IEEE논문의 경우, 전처리 작업을 통해 그룹핑 작업을 진행했음에도 불구하고 키워드의 종류 수가 현저히 많았다. 키워드 당 빈도수를 확인해 본 결과 총 3132의 유일한 키워드 중에서 그림 7에서 보이는 것처럼 협업 필터링이 압도적으로 많이 나왔으며, personalized recommendation, e-commerce, social network, user preference 등의 용어가 많이 등장하였다.

각 키워드에 대해 연도 별로 빈도를 확인해 본 바, 추천시스템 연구의 주요 이슈인 초기 사용자 문제나 희소성은 여전히 주요 연구 이슈로 다루고 있으며, e-commerce나 hybrid recommendation에 대한 연구도 증가 추세임을 확인할 수 있었다. 또한 data mining, association rule이나 machine learning 등의 키워드의 빈도로 보아 추천시스템의 연구 방향이 좀 더 다양해지고 있음도 그림 8을 통해 확인할 수 있었다. 한편, 영화 데이터 셋인 movielens 키워드가 상위에 랭크되어 있는데 이는 데이터 확보가 관건인 추천시스템 연구에서 공개용 데이터 셋을 활용한 영역이 논문에서 많이 활용이 되고 있음을 보인다.

그림 9는 연구 대상 도메인별로 키워드를 비교해 본 것으로서 movie와 music 영역으로 부터 추천시스템에 대한 연구가 시작되었고, 현재까지도 주요 연구 대상 도메인임을 알 수 있다. 추천시스템 연구에서는 데이터 확보가 중요한 관건이기 때문에 Netflex, Movielens, Last.fm 등 데이터 셋이 공개되어 있는 도메인에 대한 추천시스템 연구가 좀더 활발히 진행되고 있음을 유추할 수 있으며, 근래 들어서는 travel, social network, job 등 추천시스템의 연구 적용분야가 점점 다양해지고 있음도 확인할 수 있다.

| Keyword \ Year | 1999 | 2000 | 2001 | 2002 | 2003 | 2004 | 2005 | 2006 | 2007 | 2008 | 2009 | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 | 2016 | 2017 | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| collaborative-filtering | 0 | 1 | 3 | 1 | 5 | 6 | 17 | 8 | 22 | 46 | 59 | 104 | 85 | 114 | 90 | 156 | 158 | 161 | 98 | 1134 |
| personalized recommend | 0 | 0 | 0 | 0 | 1 | 0 | 2 | 0 | 5 | 19 | 19 | 46 | 36 | 42 | 35 | 44 | 36 | 43 | 28 | 356 |
| social-network | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 2 | 1 | 11 | 23 | 40 | 27 | 30 | 53 | 52 | 51 | 22 | 314 |
| e-commerce | 0 | 1 | 3 | 1 | 4 | 2 | 12 | 4 | 13 | 19 | 23 | 31 | 26 | 30 | 21 | 14 | 21 | 38 | 17 | 280 |
| user preference | 0 | 0 | 0 | 2 | 2 | 0 | 3 | 3 | 3 | 7 | 9 | 13 | 28 | 20 | 18 | 27 | 28 | 30 | 14 | 207 |
| movielens | 0 | 0 | 0 | 0 | 2 | 0 | 2 | 1 | 3 | 5 | 4 | 14 | 10 | 20 | 29 | 26 | 24 | 28 | 10 | 178 |
| cold-start | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 | 1 | 4 | 12 | 14 | 20 | 20 | 19 | 25 | 27 | 24 | 171 |
| sparsity | 0 | 0 | 0 | 0 | 1 | 0 | 2 | 1 | 2 | 8 | 15 | 5 | 8 | 13 | 13 | 23 | 18 | 22 | 16 | 147 |
| data-mining | 0 | 0 | 0 | 1 | 0 | 2 | 7 | 3 | 4 | 4 | 13 | 14 | 11 | 10 | 6 | 8 | 23 | 20 | 19 | 145 |
| hybrid recommendation | 0 | 1 | 0 | 0 | 2 | 1 | 1 | 0 | 2 | 2 | 3 | 6 | 4 | 13 | 14 | 15 | 16 | 24 | 8 | 112 |
| content-based filtering | 0 | 1 | 1 | 1 | 0 | 3 | 4 | 1 | 3 | 3 | 4 | 13 | 3 | 8 | 15 | 12 | 10 | 18 | 12 | 112 |
| association rule | 0 | 0 | 0 | 1 | 0 | 2 | 6 | 1 | 1 | 3 | 6 | 11 | 15 | 10 | 11 | 13 | 14 | 15 | 2 | 111 |
| user profile | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 3 | 5 | 9 | 7 | 12 | 14 | 18 | 14 | 8 | 7 | 105 |
| context-aware recomme | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 2 | 2 | 1 | 3 | 10 | 12 | 18 | 13 | 18 | 15 | 7 | 102 |
| information filtering | 1 | 1 | 0 | 0 | 1 | 2 | 2 | 4 | 4 | 6 | 14 | 6 | 10 | 8 | 6 | 5 | 9 | 13 | 6 | 98 |
| decision-making | 0 | 0 | 0 | 0 | 2 | 0 | 1 | 3 | 1 | 2 | 1 | 8 | 9 | 12 | 4 | 15 | 15 | 17 | 8 | 98 |
| matrix-factorization | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 3 | 0 | 2 | 7 | 3 | 9 | 14 | 21 | 21 | 13 | 94 |
| machine-learning | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 2 | 7 | 4 | 5 | 5 | 6 | 12 | 16 | 18 | 15 | 93 |
| item-based collaborative | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 2 | 2 | 4 | 10 | 7 | 8 | 7 | 9 | 14 | 7 | 13 | 5 | 91 |
| content-based recomme | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 4 | 3 | 6 | 4 | 9 | 13 | 12 | 13 | 9 | 4 | 81 |
| user-based collaborative | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 3 | 4 | 8 | 3 | 6 | 12 | 16 | 7 | 16 | 4 | 80 |

그림 8. 연도별 키워드 빈도(IEEE)
Fig. 8. Keyword frequency by year

| Domain \ Year | 2002 | 2003 | 2004 | 2005 | 2006 | 2007 | 2008 | 2009 | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 | 2016 | 2017 | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| book | | | | | | | 1 | 1 | 1 | 8 | 6 | 9 | 10 | 13 | 15 | 1 | 65 |
| cooking | | | | | | | | | 1 | 1 | | 1 | 2 | 4 | 6 | 3 | 18 |
| e-learning | | | | | | | 2 | 2 | 1 | | | 1 | 1 | 4 | 1 | 2 | 14 |
| health | | | | | | | | 1 | 2 | 3 | 2 | 4 | 3 | 6 | 5 | 1 | 27 |
| hotel | | | | | | | 1 | | 4 | | | 2 | | 4 | 1 | 1 | 13 |
| job | | | | | | 1 | | | | | 1 | 2 | 7 | 5 | 12 | 6 | 34 |
| location | | | | | | | | 1 | | 1 | 2 | 2 | 6 | 3 | 7 | 1 | 23 |
| movie | | 2 | 1 | 4 | | 1 | 2 | 4 | 9 | 11 | 19 | 19 | 24 | 19 | 21 | 15 | 151 |
| music | 1 | | | 3 | 2 | 6 | 9 | 8 | 5 | 13 | 9 | 16 | 9 | 24 | 18 | 12 | 135 |
| news | | | | | 1 | | 2 | 1 | 8 | 5 | 7 | 7 | 17 | 14 | 16 | 3 | 81 |
| travel | | | | | | 2 | 2 | 1 | 6 | 9 | 7 | 6 | 14 | 14 | 16 | 10 | 87 |

그림 9. 추천시스템 연구 도메인별 빈도(IEEE)
Fig. 9. Frequency per domains

## 4.2 ACM 데이터 셋을 활용한 분석

ACM 데이터 셋에서 가장 많이 나온 키워드 역시 협업 필터링이었다. 오랜 시간이 소요되어 실시간 추천이 필요한 현실에 적용하기 어려운 단점이 있음에도 불구하고 좋은 성능을 발휘한다는 장점으로 인해 학계에서는 여전히 협업필터링에 대하여 많은 관심을 보이고 있음을 알 수 있으며, machine learning에 관련된 키워드에 대해서도 deep learning, neural network, recurrent neural network 등이 나타난 것으로 보아 기계학습이 추천시스템에도 활용이 되고 있음을 확인할 수가 있다.

한편 협업 필터링뿐만 아니라 e-commerce, social network, matrix factorization도 꾸준하게 많이 등장하는 주제어이고, 다양성도 2011년 이후 꾸준하게 많이 연구가 이루어지는 주제어였다. 또한 personalization에 관한 주제도 꾸준히 증가하고 있는 연구 주제였으며, 추천의 다양성을 높이고자 하는 연구도 증가하고 있는 추세임을 확인할 수 있다.

그림 10은 연도별로 키워드 당 빈도율을 확인해 본 것이다. 키워드 빈도율은 해당년도의 키워드 빈도를 해당년도의 논문수로 나누어 계산하였다. 협업 필터링의 경우 2004~2006년도에는 논문 중 45%정도가 협업 필터링을 키워드로 하고 있으나, 점차 협업 필터링 키워드 등장 비율이 줄고 있음을 알 수 있다. 협업 필터링 키워드의 등장 비율이 점점 낮아지는 이유는 추천시스템에 대한 연구가 점차 세분화됨에 따라 논문에서 제시되는 키워드의 수와 종류가 다양해지기 때문이다.

그림 11은 2008년부터~2017년까지 주요키워드의 순위 변화를 비교해 본 것이다. 협업 필터링, matrix factorization, social network, personalization, 초기 사용자 문제, 다양성 등은 매해 많은 논문에서 등장하는 키워드이며, machine learning관련 키워드들도 등장횟수가 점점 많아지고 있음을 보인다. content based filtering이나 context aware recommendation, group recommendation 등도 매해 자주 등장하는 키워드임을 확인할 수 있다.

| Rank | Year | 1999 | 2000 | 2001 | 2002 | 2003 | 2004 | 2005 | 2006 | 2007 | 2008 | 2009 | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 | 2016 | 2017 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | PaperCount | 3 | 7 | 9 | 7 | 14 | 13 | 24 | 24 | 58 | 94 | 125 | 157 | 163 | 185 | 201 | 208 | 186 | 225 | 285 |
| | KeywordCount | 14 | 35 | 29 | 24 | 59 | 52 | 87 | 94 | 209 | 352 | 478 | 631 | 616 | 705 | 764 | 797 | 693 | 875 | 1079 |
| | Keyword | | | | | | | | | | | | | | | | | | | |
| 1 | collaborative-filtering | 33.3 | 28.6 | 33.3 | 42.9 | 35.7 | 46.2 | 41.7 | 45.8 | 20.7 | 25.5 | 31.2 | 21.7 | 21.5 | 16.8 | 11.4 | 11.5 | 16.1 | 15.6 | 10.9 |
| 2 | social-network | | | | | 7.1 | | | 4.2 | 3.4 | 3.2 | 9.6 | 8.9 | 8.6 | 5.4 | 7.5 | 5.8 | 5.9 | 5.3 | 2.5 |
| 3 | personalization | | | | 14.3 | 7.1 | 7.7 | 8.3 | 12.5 | 5.2 | 7.4 | 6.4 | 8.9 | 4.3 | 2.7 | 6.0 | 1.9 | 5.4 | 4.0 | 4.6 |
| 4 | matrix-factorization | | | | | | | | | | 4.3 | 3.2 | 1.9 | 7.4 | 2.7 | 6.5 | 7.7 | 6.5 | 4.0 | 4.6 |
| 5 | context-aware recommendatio | | | | | | | 4.2 | | | | 1.6 | 2.5 | 8.0 | 6.5 | 5.5 | 5.3 | 3.2 | 2.7 | 1.8 |
| 6 | content-based filtering | | | | | 7.1 | 7.7 | | | 1.7 | 1.1 | 4.0 | 1.3 | 3.1 | 1.6 | 3.5 | 2.4 | 1.6 | 4.4 | 1.8 |
| 7 | diversity | | | | | | | 4.2 | 4.2 | | 1.1 | 2.4 | | 2.5 | 2.2 | 3.5 | 3.4 | 2.7 | 3.1 | 3.2 |
| 8 | cold-start | | | | | | | | | 3.4 | 2.1 | 2.4 | 1.9 | 3.7 | | 3.0 | 3.8 | 2.2 | 1.8 | 3.2 |
| 9 | machine learning | | 14.3 | 11.1 | | 14.3 | 7.7 | 4.2 | 4.2 | 3.4 | 2.1 | 2.4 | | | 1.6 | 3.0 | 1.4 | 1.6 | 3.1 | 2.5 |
| 10 | group recommendation | | | | | | | | | | | 0.8 | 3.8 | 3.7 | 3.2 | 0.5 | 2.4 | 1.6 | 1.8 | 2.8 |
| 11 | e-commerce | | | | | 7.1 | | 12.5 | 12.5 | | 1.1 | | 1.3 | 1.8 | | 2.0 | 1.9 | 2.7 | 2.7 | 1.4 |
| 12 | hybrid recommender system | | | | | | 7.7 | | | | 2.1 | 0.8 | 1.3 | 1.2 | 2.7 | 3.5 | 1.0 | 1.1 | 1.3 | 1.8 |
| 13 | implicit feedback | 33.3 | | | | | | | | 1.7 | | | 1.3 | 1.2 | 3.2 | 1.5 | 1.4 | 1.1 | 0.4 | 2.8 |
| 14 | similarity | | | | | 14.3 | 7.7 | | 4.2 | | 5.3 | 0.8 | 1.9 | 1.2 | 0.5 | 1.5 | 1.0 | | 0.9 | 2.1 |

그림 10. 키워드 당 빈도율(%)(ACM)
Fig. 10. Frequency rate per keyword

| Rank | 2008 | 2009 | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 | 2016 | 2017 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | collaborative-filtering | collaborative-filtering | collaborative-filtering | collaborative-filtering | collaborative-filtering | collaborative-filtering | collaborative-filtering | collaborative-filtering | collaborative-filtering | collaborative-filtering |
| 2 | personalization | social-network | personalization | social-network | evaluation | social-network | matrix factorization | matrix factorization | social-network | deep-learning |
| 3 | clustering | personalization | social-network | context-aware recomm | context-aware recomm | matrix factorization | evaluation | social-network | content-based filtering | matrix factorization |
| 4 | similarity | trust | evaluation | matrix factorization | social-network | evaluation | social-network | personalization | matrix factorization | personalization |
| 5 | matrix factorization | content-based filtering | context | personalization | group recommendatio | personalization | context-aware recomm | context-aware recomm | personalization | user modeling |
| 6 | evaluation | preference elicitation | group recommendatio | cold-start | implicit feedback | context-aware recomm | cold-start | evaluation | diversity | cold-start |
| 7 | lifelong learning | matrix factorization | web 2.0 | context | clustering | content-based filtering | twitter | diversity | machine learning | diversity |
| 8 | long tail | ranking | folksonomy | evaluation | hybrid recommender s | diversity | diversity | e-commerce | context-aware recomm | group recommendation |
| 9 | multimedia | tagging | social tagging | group recommendatio | matrix factorization | hybrid recommender s | social media | big data | e-commerce | implicit feedback |
| 10 | shilling | web 2.0 | user profile | content-based filtering | personalization | news | content-based filtering | cold-start | topic model | recurrent neural networ |
| 11 | social-network | blogs | user study | context-aware | ranking | cold-start | group recommendatio | ontology | user modeling | machine-learning |
| 12 | web mining | clustering | context-aware recomm | movie recommendatio | user modeling | machine-learning | e-commerce | scalability | user profile | social-network |
| 13 | cold-start | cold-start | music recommendatio | user preference | user profile | social media | mobile application | sparsity | evaluation | clustering |
| 14 | collaborative tagging | diversity | ontology | diversity | diversity | user modeling | personalization | user modeling | top-n recommendatio | similarity |
| 15 | complex network ana | explanation | cold-start | novelty | music recommendatio | information retrieval | ranking | content-based filtering | adaptive system | neural network |
| 16 | flickr | information retrieval | explanation | clustering | social tagging | privacy | active learning | crowdsourcing | cold-start | privacy |
| 17 | hybrid recommender s | machine-learning | matrix factorization | e-commerce | twitter | trust | application | data mining | data mining | content-based filtering |
| 18 | information loss | social software | semantic web | metrics | user experience | big data | benchmarking | deep-learning | deep-learning | context-aware recomme |
| 19 | machine-learning | top-n recommendatio | similarity | multidimensional | user study | e-commerce | big data | emotion | emotion | evaluation |
| 20 | manipulation-resistan | user study | user experience | tags | content-based filtering | linked data | data mining | group recommendatio | explanation | health recommender sy |
| 21 | multi-agent system | context | user modeling | usage log | data mining | ranking | decision making | location-based service | group recommendatio | hybrid recommender sy |
| 22 | novelty | context-aware recomm | active learning | user experience | folksonomy | topic model | implicit feedback | machine-learning | recurrent neural netw | convolutional neural ne |
| 23 | ontology | detection | challenge | adaptation | information retrieval | benchmarking | information retrieval | nutrition | social recommender sy | e-commerce |
| 24 | open source software | digital library | collaborative tagging | context modeling | knowledge manageme | conversational recomm | machine-learning | planning | transactional memory | factorization machine |
| 25 | plsa | folksonomy | content-based filtering | contextual information | learning to rank | decision making | novelty | sentiment analysis | human computer inte | sentiment analysis |
| 26 | popularity | markov models | context-aware | decision making | machine-learning | decision tree | sentiment analysis | smart health | hybrid recommender s | social recommender sys |
| 27 | random walks | minimax regret | e-commerce | emotion | metrics | implicit feedback | | algorithm | music recommendatio | topic model |
| 28 | ranking | novelty | explicit feedback | hybrid recommender s | serendipity | learning to rank | social web | android | natural language proc | user experience |
| 29 | reinforcement learning | ontology | hybrid recommender s | implementation | software development | location based social n | text mining | anomaly detection | neural network | big data |

그림 11. 2008~2017년 주요키워드 순위 변화(ACM)
Fig. 11. Keyword ranking changes

그림 12는 추천시스템 연구의 적용 분야 비교를 자바스크립트 기반의 반응형 시각화 라이브러리인 D3.js를 활용하여 표현한 것이다. 추천시스템에 대한 연구가 활발해지던 시기인 초창기와 비교하여 근래에 들어서는 다양한 분야를 대상으로 추천시스템에 대한 연구가 진행이 되고 있으며, 특히 social network나 health에 대한 연구가 점차 활발해 지고 있음을 알 수 있다.

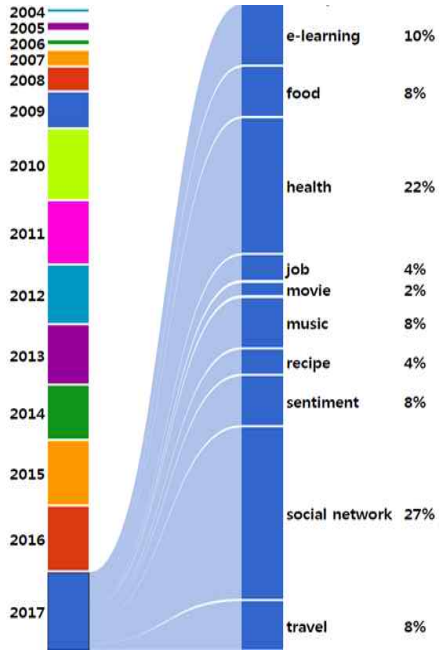


그림 12. 추천시스템 연구 도메인 빈도 비교 시각화 예 (ACM, 2017년)
Fig. 12. Domain frequency comparison visualization example

| Rank | Domain(Count) | | Recommender Technique Keyword | | Research Topic | |
|---|---|---|---|---|---|---|
| 1 | music | 302 | collaborative filtering | 1508 | cold start | 227 |
| 2 | movie | 180 | context aware recommendation | 371 | sparsity | 177 |
| 3 | travel | 173 | content based filtering | 365 | implicit feedback | 72 |
| 4 | book | 104 | machine learning | 247 | diversity | 66 |
| 5 | health | 70 | matrix factorization | 189 | scalability | 36 |
| 6 | food | 51 | hybrid recommender system | 145 | novelty | 28 |
| 7 | e-learning | 43 | association rules | 115 | grey sheep | 7 |

그림 13. 분석 영역별 키워드 순위 (1998~2017, IEEE+ACM)
Fig. 13. Keyword ranking by analysis area (1998~2017, IEEE+ACM)

### 4.3 분석 정리

두 개의 학술사이트로부터 얻은 논문관련 정보 중 키워드 분석을 통해 논문의 주요 연구 이슈들에 대한 빈도를 그림 13과 같이 정리해 본 결과에 의하면, 추천시스템에 관한 논문 중에는 music을 적용 대상 도메인으로 연구를 진행한 논문이 가장 많았다. social network와 e-commerce의 경우, 연구 적용 대상 도메인이기도 하지만 적용대상과는 별개로 키워드 목록에 포함이 되는 부분이 있어서 적용 대상 도메인 순위에서는 생략을 하였다. 추천 기법에 대해서는 협업 필터링에 대한 연구가 많았으며, 주요 연구 이슈로는 초기 사용자 문제나 희소성에 관한 연구가 많이 이루어지고 있음을 알 수 있었다.

## V. 결 론

현재 각 산업 군에서 추천시스템은 다양하게 활용이 되고 있으며, 학계에서도 추천시스템에 대한 연구는 여전히 활발하게 이루어지고 있다.

본 논문에서는 IEEE와 ACM 전자 도서관으로부터 얻은 자료들을 활용하여 데이터 셋을 구축하고, 구축된 데이터 셋을 통해 추천시스템에 관한 주요 연구 주제나 적용 대상 도메인, 선호 기법 등에 대하여 확인해 보았다.

키워드 빈도를 활용한 분석을 통하여, 추천시스템이 등장한 초반에 비해 연구 영역이나 대상 등이 다양해지고 세분화되어지고 있음을 알 수 있었고 개인화 추천, 머신러닝, 데이터 마이닝을 활용한 추천시스템에 대한 연구들이 점차 늘어나고 있음도 확인할 수 있었다.

다른 관점에서 의미 있는 결과를 도출해 보고자 추천시스템 관련 논문들에 제시되어 있는 키워드 묶음을 사용한 분석을 시도해 보았지만, 유사 키워드를 그룹핑하는 데 있어서 인간의 주관적 지식이 개입될 수밖에 없는 한계가 있음을 확인할 수 있었다. 논문 키워드를 통한 분석에서 정확도가 높은 결과를 얻어내기 위해서는 다양한 단어로 표현되어 있는 키워드의 표준화가 중요한 작업이고, 이를 위해서는 관련 전공 용어에 대하여 온톨로지 구축, 메타데이터 레지스트리 구축 등의 방법 등을 사용하여 동일 개념을 식별할 수 있도록 하는 것이 필요하겠다.

또한 다양한 산업 군에서 추천시스템이 활발하게 활용되고 있음에도 불구하고 연구 논문에서는 music이나 movie 등 연구 대상 도메인이 다소 제한적인데, 이는 데이터 확보에 기인하기 때문이라 판단이 되며, 넷플릭스 등과 같이 추천시스템이 유용하게 활용될 수 있는 다양한 산업 군에서 데이터 셋을 공개해준다면 추천시스템에 대한 학계의 연구

가 더욱 활발하게 진행될 수 있을 것이다.

논문에 제시된 키워드와 논문 주제에 대한 관련성을 놓고 보았을 때, 키워드를 중심으로 한 분석 방법은 어느 정도 의미가 있을 것이며, 연구자의 연구 방향 설정 등에도 다소 도움이 될 수 있겠다.

## References


[1] Paul Resnick, Neophytos Iacovou, Mitesh Suchak, and Peter Bergstrom, "An open architecture for collaborative filtering of netnews", Proceedings of the 1994 ACM conference on Computer supported cooperative work, pp. 175-186, Oct. 1994.

[2] Francesco Ricci, Lior Rokach, and Bracha Shapira, "Introduction to Recommender Systems Handbook", Springer, pp. 25-47, 2011.

[3] Isinkaye, Folajimi, and Ojokoh, "Recommendation systems: Principles, methods and evaluation", Egyptian Informatics Journal, Vol. 16, No. 3, pp. 261-273, Jun. 2015.

[4] Shah Khusro, Zafar Ali, and Irfan Ullah, "Recommender Systems; Issues, Challenges, and Research Opportunities", A Comparative Study on Hough Transform Segmentation Approach for Outlier Iris Image, pp. 1179-1189, Feb. 2016.

[5] Sarika Jain, Anjali Grover, Praveen Singh Thakur, and Sourabh Kumar Choudhary, "Trends, Problems And Solutions of Recommender System", International Conference on Computing, Communication and Automation, pp. 955-958, May 2015.

[6] Soanpet. Sree Lakshmi, and Adi Lakshmi, "Recommendation Systems; Issues and challenges", International Journal of Computer Science and Information Technologies, Vol. 5, No. 4, pp. 5771-5772, May 2014.

[7] Omkar S. Revankar and Y. V. Haribhakta, "Survey on Collaborative Filtering Technique in Recommendation System", International Journal of Application or Innovation in Engineering & Management, Vol. 4, No. 3, pp. 85-91, Mar. 2015.

[8] Gediminas Adomavicius and Alexander Tuzhilin, "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, Vol. 17, No. 6, pp. 734-749, Apr. 2005.

[9] IEEE Xplore Digital Library(http://ieeexplore.ieee.org/search/advsearch.jsp) [Accessed: Feb. 01. 2018]

[10] ACM Digital Library(https://dl.acm.org/) [Accessed: Feb. 01. 2018]


## 저자소개


배 은 영 (Eun-Young Bae)

1993년 2월 : 숙명여자대학교 이과대학 통계학과(이학사)
2003년 2월 : 서강대학교 정보통신대학원 정보처리전공(공학사)
2014년 3월 ~ 현재 : 숙명여자대학교 컴퓨터과학과 일반대학원 박사과정
관심분야 : 추천시스템, 정보시각화, 협업필터링, 컴퓨터교육

유 석 종 (Seok-Jong Yu)

1994년 2월 : 연세대학교 컴퓨터과학과(이학사)
1996년 2월 : 연세대학교 대학원 컴퓨터과학과(이학석사)
2001년 2월 : 연세대학교 대학원 컴퓨터과학과(공학박사)
2005년 ~ 현재 : 숙명여자대학교 소프트웨어학부 교수
관심분야 : HCI, 그래픽스, 추천시스템, 협업필터링