JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 26 No. 6, pp. 19-28, June 2021
https://doi.org/10.9708/jksci.2021.26.06.019

A personalized exercise recommendation system using dimension reduction algorithms

Ha-Young Lee*, Ok-Ran Jeong*

*Student, Dept. of AI·SW, Gachon University, Seongnam, Korea
*Professor, Dept. of AI·SW, Gachon University, Seongnam, Korea

[Abstract]

Nowadays, interest in health care is increasing due to Coronavirus (COVID-19), and a lot of people are doing home training as there are more difficulties in using fitness centers and public facilities that are used together. In this paper, we propose a personalized exercise recommendation algorithm using personalized propensity information to provide more accurate and meaningful exercise recommendation to home training users. Thus, we classify the data according to the criteria for obesity with a k-nearest neighbor algorithm using personal information that can represent individuals, such as eating habits information and physical conditions. Furthermore, we differentiate the exercise dataset by the level of exercise activities. Based on the neighborhood information of each dataset, we provide personalized exercise recommendations to users through a dimensionality reduction algorithm (SVD) among model-based collaborative filtering methods. Therefore, we can solve the problem of data sparsity and scalability of memory-based collaborative filtering recommendation techniques and we verify the accuracy and performance of the proposed algorithms.

▸ **Key words**: Health-care, Classification, Recommendation System, Personalized Method, Dimensionality reduction model, SVD

[요 약]

코로나로 인해 건강관리에 대한 관심이 증가하고 있는 요즘, 여러 사람이 함께 이용하는 헬스장이나 공용시설을 이용하는데 어려움이 늘어남에 따라 홈 트레이닝을 하는 이들이 늘어나고 있다. 이에 본 연구에서는 홈 트레이닝 사용자들에게 좀 더 정확하고 의미 있는 운동 추천을 제공하기 위해 개인 성향 정보를 활용한 개인화된 운동 추천 알고리즘을 제안한다. 이를 위해 식습관 정보, 육체적 조건 등 개인을 나타낼 수 있는 개인 성향 정보를 사용해 k-최근접 이웃 알고리즘으로 데이터를 비만의 기준에 따라 분류하였다. 또한, 운동 데이터 셋을 운동의 레벨에 따라 등급을 구별하였으며 각 데이터 셋의 이웃 정보를 바탕으로 모델 기반 협업 필터링 방법 중 차원 축소 모델인 특이값 분해 알고리즘(SVD)을 통해 사용자들에게 개인화된 운동 추천을 제공한다. 따라서 메모리 기반 협업 필터링 추천 기법의 데이터 희소성과 확장성의 문제를 해결할 수 있고, 실험을 통해 본 연구에서 제안하는 알고리즘의 정확도와 성능을 검증한다.

▸ **주제어**: 헬스 케어, 분류, 추천 시스템, 개인화 기법, 차원 축소 모델, 특이값 분해

• First Author: Ha-Young Lee, Corresponding Author: Ok-Ran Jeong
*Ha-Young Lee (hhzet11@gachon.ac.kr), Dept of AI·SW, Gachon University
*Ok-Ran Jeong (orjeong@gachon.ac.kr), Dept of AI·SW, Gachon University
• Received: 2021. 05. 11, Revised: 2021. 05. 31, Accepted: 2021. 05. 31.

http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945

I. Introduction

코로나로 인해, 여러 사람이 함께 이용하는 헬스장 등의 공용 운동 시설을 이용하는 데 어려움이 증가하면서 많은 사람들이 홈 트레이닝을 하는 경우가 늘어나고 있다. 또한, 건강관리에 더욱 신경을 쓰게 되면서 자연스럽게 헬스케어 시스템에 대한 요구도 증가하고 있다. 이에 사람들은 자신의 신체조건 등의 개인 정보에 따라 운동 효율을 높일 수 있는 운동을 찾고자 한다.

그러나 다양한 정보 속에서 사용자들은 자신이 원하는 정보를 손쉽게 찾기는 매우 어렵다. 또한 판매자 입장에서는 개개인의 선호도를 고려해 적절한 아이템을 추천하고 이를 구매로 연결시키는 것은 이윤 창출과 직결되므로 이에 따라 적절한 아이템의 추천의 중요성이 대두되고 있다. 따라서 빅 데이터를 사용해 정보를 추천하는 시스템이 등장하였고[1], 더불어 사용자에게 보다 알맞은 정보를 제공하기 위해서 최근에는 개인 성향과 같은 개인 정보를 활용해 사용자에게 정보를 제공하기도 한다[2].

추천시스템의 개인화 기법으로는 규칙 기반 필터링, 내용 기반 필터링, 협업 필터링 등 다양한 방법이 존재한다. 그 중 가장 대표적인 개인화 기법인 협업 필터링 방법은 데이터 희소성 및 확장성에 문제가 있다. 이 문제를 해결하기 위해 데이터 분석 기법 중 대표 기법인 클러스터링 기법을 적용하여 유사한 선호도를 가진 사용자를 식별해 추천시스템에 활용한다[3].

이에 본 연구에서는 홈 트레이닝을 하는 사용자들에게 개인 맞춤화된 운동 추천을 제공하도록 개인 성향 정보를 활용한 추천시스템을 개발한다. 비만 데이터 셋에서 사용자의 식습관 정보, 육체적 조건 등 개인 성향 정보를 바탕으로 그룹으로 나누는 데이터 분류 알고리즘을 진행하고, 운동 데이터 셋 역시 운동의 레벨에 따라 데이터를 분류한다. 따라서 해당 그룹 정보와 모델 기반 협업 필터링 방법 중 차원 축소 기법인 특이값 분해 알고리즘(SVD)을 사용해 추천 시스템을 구현한다. 그 후, 추천이 제대로 제공되고 있는지 정확도, 재현율, 정밀도 등 값의 계산을 통해 알아본다. 또한 다른 추천 실험 방법과의 비교 실험을 통해 본 연구의 성능을 검증한다.

2장에서는 이 연구에서 제안하는 방법과 관련된 연구들에 대해 소개한다. 3장에서는 이 연구에 활용된 데이터 셋에 대해 설명하고 4장에서는 본 연구에서 제안하는 추천 시스템에 대하여 설명한다. 5장에서는 실험 및 평가를 하였으며, 6장에서는 결론을 서술한다.

II. Related Works

2.1 Data Classification Techniques

데이터를 비슷한 집단으로 묶는 방법으로는 비지도 학습과 지도 학습 각각의 대표적인 방법인 군집 분석(Clustering)과, 분류 방법 (Classification)이 존재한다. 각 방법에 대해 아래에서 소개한다.

2.1.1 Clustering

군집 분석은 비지도 학습 방법으로 각 개체가 소속 집단인 군집에 대한 정보를 몰라 데이터 자체의 특성에 대해 알고자 할 때 사용하는 기법이다. 따라서 데이터 간의 유사도를 정의하고 그 유사도에 가까운 것부터 순서대로 합쳐가는 방식으로 진행되는 알고리즘이다. 여러 협업 필터링 기반 알고리즘은 희소성과 확장성 문제를 완화하기 위해 클러스터링 방법을 통합해 사용하기도 한다. 군집분석 알고리즘의 예시로는 k-means 알고리즘, DBSCAN 알고리즘 등이 존재한다[3]. 대표적인 군집 분석을 사용하는 예시로는 사용자 경험을 개인화하기 위해 비슷한 제품끼리 묶어주는 추천시스템이 있다.

2.1.2 Classification

분류는 지도 학습 방법으로 각 개체별 그룹의 label이 사전에 알려져 있을 때 사용하는 분석 방법이다. 즉, 기존에 존재하는 데이터의 관계를 파악하고, 새롭게 관측된 데이터의 category를 스스로 판별하는 과정으로, 분류 알고리즘의 예시로는 의사 결정 트리(decision tree), k-최근접 이웃(k – nearest neighborhood) 알고리즘 등이 존재한다.

이 중 k-최근접 이웃 방법은 가장 가까운 이웃을 찾아 새로운 사용자에 대한 예측 및 분류 작업을 하는데 사용되는 방법이다. 즉 새로운 사용자에 대해 전체 사용자 자료로부터 가장 가까운 k개의 근접 이웃을 선택하여 다수결 원칙 또는 근접 정도에 따른 가중치 평균으로 분류 또는 예측 값을 계산하는 방법이다.

2.2 Personalized Recommendation Techniques

추천 시스템은 다양한 정보와 제품들 속에서 사용자가 관심 가질 만한 제품을 추천해주는 시스템이다. 추천 시스템의 가장 기본적인 가정은 사용자들의 의견이 선택되고 통합되어 실질적인 사용자의 선호를 기반으로 합리적인 예측을 할 수 있다는 것이다. 이러한 가정은 사용자들이 몇몇 물품에 대해 품질 또는 선호에 대한 내용 간의 공통점이 있다는 것을 전제로 한다[4].

사용자에게 더욱 개인 맞춤화 된 추천을 제공하고자, 다양한 개인화 기법을 추천 시스템에 적용하여 구현한다. 다음에서 관련 연구들에 대해 기술한다.

2.2.1 Rule-based Filtering

규칙 기반 필터링(Rule-based Filtering) 방법은 사용자의 프로파일이나 행동에 근거해 설정한 규칙에 따라 개인에게 맞춤형 추천을 제공하는 방법으로 사용과 이해가 쉽고, 추천에 걸리는 시간이 상대적으로 짧다는 장점이 있다[5]. 그러나 직접 규칙을 생성해야 하며 이를 위해 전문적인 지식이 필요하다. 또한, 데이터의 양이 늘어날수록 효율이 저하된다는 단점도 존재한다. 대표적인 알고리즘으로는 어떤 항목에 대한 관측 값과 목표 값을 연결해주는 예측 모델인 의사 결정 트리(decision tree) 방법이 존재한다[2].

2.2.2 Content-based Filtering

내용 기반 필터링(Content-based Filtering)은 정보 검색 기술에 바탕을 둔 시스템으로, 아이템과 아이템 혹은 아이템과 사용자 선호도 간 유사성을 분석해 이를 기반으로 사용자에게 아이템을 추천하는 방식이다. 과정을 살펴보면 사용자가 직접 입력한 프로파일 정보나 사용자가 아이템에 대해 평가한 점수 혹은 과거 이용 내역 등을 바탕으로 생성된 정보를 통해 선호하는 아이템을 파악한다. 그 다음, 미리 선정된 기준을 통해 분류된 아이템 category와 사용자 선호 아이템 간의 유사도를 측정한다. 그 결과, 유사도가 가장 높게 나타난 category에 해당하는 아이템을 사용자에게 추천하는 방식으로 추천을 제공한다[6]. 위 과정을 그림으로 나타내면 다음 그림 1과 같다.

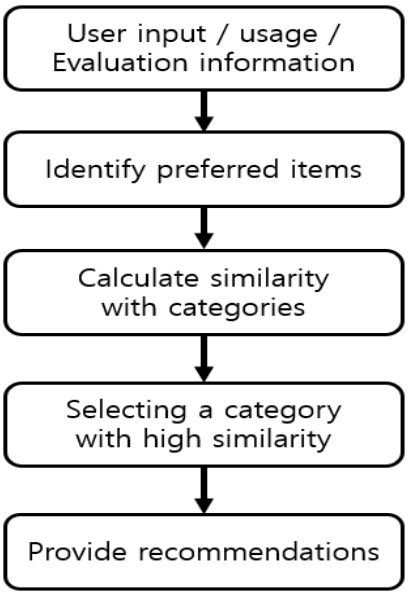

Fig. 1. Content-based Filtering Process

내용 기반 필터링은 분석이 용이하기 때문에 영화, 음악뿐만 아니라 텍스트 기반의 뉴스나 인터넷 기사 등을 추천하는 데 많이 사용된다[7].

이는 기본적인 방법으로 구현이 간단하고 사용자의 명시적인 신호 정보를 직접적으로 반영할 수 있다는 장점이 있지만, 아이템의 내용 기반 정보를 구하기 힘든 경우가 많고, 사용자의 명시적인 프로필을 얻기 어렵다. 특히 사용자의 선호 취향을 특정 단어로 표현하기 매우 어렵다는 단점도 존재한다[8].

2.2.3 Memory-based Collaborative Filtering

협업 필터링(Collaborative-Filtering) 방법은 메모리 기반 알고리즘과 모델 기반 알고리즘으로 분류된다. 메모리 기반 협업 필터링 방법은 추천할 대상 항목에 대한 사용자의 평가를 입력받아 분석한 후, 사용자 간의 유사도를 계산해 이를 바탕으로 이웃을 형성한다. 유사한 성향을 가진 사용자라면 비슷한 성품을 선호할 것이라는 가정을 기반으로 사용자에게 추천을 제공하는 방법이다. 위 과정을 그림으로 나타내면 다음 그림 2와 같다.

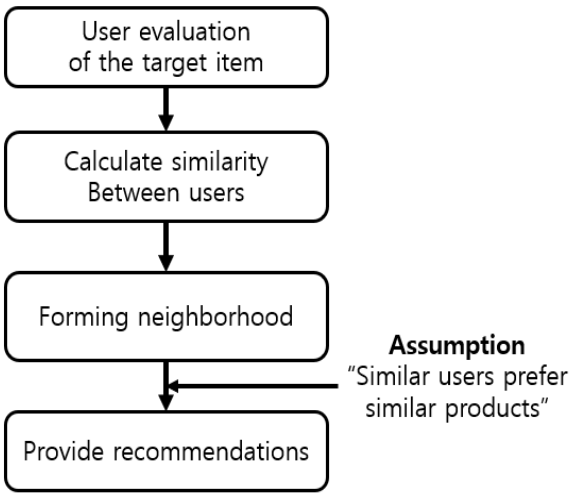

Fig. 2. Memory-based Collaborative Filtering

메모리 기반 협업 필터링 추천 시스템은 평가 정보를 활용하므로 사용자나 항목에 대한 정보가 없어도 추천을 수행할 수 있다는 장점을 갖지만, 데이터 증가에 따른 입력 데이터의 희박성 문제와 시스템 확장성 문제가 존재한다. 따라서 이러한 문제를 해결하기 위해 협업 필터링 시스템은 주로 k-최근접 이웃 알고리즘을 사용하여 사용자들을 k개의 이웃으로 나누어 추천을 진행한다.

2.2.4 Model-based Collaborative Filtering

메모리 기반 협업 필터링에서는 추천 대상 사용자와 성향이 비슷한 사용자를 선정해 관련 추천을 진행하는데 이때, 사용자가 구매한 아이템이 너무 적으면 성향 분석을 통한 추천이 거의 불가능하게 되어 추천 시스템의 성능을 저하시키게 된다. 또한, 사용자와 아이템의 수가 많을 경우 유사도와 선호도의 계산에 있어 많은 계산 비용이 발생하게 된다. 이러한 문제를 해결하기 위해 고차원의 행렬을 저차원의 행렬로 축소시키는 기법으로 추천시스템의 Item-user matrix에서도 적용할 수 있는 기법이 존재한다[6].

이 기법은 모델 기반 Matrix Factorization 중 차원 축소 모델인 특이값 분해 (SVD; Singular Value Decomposition)로 대표적으로 사용되는 방식이다. 이는 하나의 행렬을 여러 행렬의 곱으로 분해하는 방법으로써, 모든 사용자와 아이템에 대한 m × n 크기의 행렬 M을 특이값 분해하면, $M = U\sum V^T$와 같이 2개의 직각 행렬과 1개의 대각 행렬로 구성된 총 세 개의 행렬의 곱으로 나타낼 수 있다. 이때, $U_{m\times m}$은 사용자 행렬을 나타내고, $\sum_{m\times n}$은 특이값을 대각 항으로 가지는 대각 행렬, $V^T{}_{n\times n}$는 운동 행렬을 나타낸다[9].

여러 행렬 분해 모델 중에서 SVD는 명확한 평가 또는 관계 정보가 있을 경우, 이를 바탕으로 행렬을 분해하였을 때, 암묵적 요인(latent factor)을 잘 정의한다고 알려져 있다. 따라서 SVD는 일종의 차원 축소 개념이며 고차원의 행렬을 저차원의 행렬로 축소시켜 분석의 정확성을 높일 수 있다[8].

III. Dataset

실험을 위한 데이터 셋으로는 kaggle에서 제공하는 'Obesity based on eating habits and physical condition'[10]와 'Calories burned during exercise and activities'[11]을 사용하였다.

3.1 Obesity Dataset

이 데이터 셋은 멕시코, 페루, 콜롬비아 인들의 개인 식습관과 신체 상태를 바탕으로 비만의 정도를 추정하기 위한 데이터를 제시한다. 데이터는 17개의 속성, 2111개의 기록을 포함하고 있으며 개인의 비만의 정도를 식별하고 이를 모니터링 하는 추천 시스템을 구축하기 위한 도구를 생성하는 데 사용될 수 있음을 밝히고 있다[12].

3.1.1 Introduction of obesity dataset

해당 데이터 셋에서는 사용자의 비만의 정도를 결정하기 위해, 여러 가지의 개인 성향들을 사용한다. 먼저, 개인 정보로는 성별, 나이, 키, 몸무게, 과체중 가족력 여부, 흡연 여부 등이 있고 식습관 요소로는 고칼로리 음식 섭취여부(FAVC), 야채 섭취 여부(FCVC), 주식 섭취 횟수(NCP), 간식 섭취 여부(CAFC), 일일 물 소비량(CH20), 알코올 소비(CALC) 등이 있다. 또한, 육체적 조건으로는 칼로리 소비 모니터링 여부(SCC), 물리적 활동 빈도(FAF), 전자기기 사용하는 시간(TUE), 사용하는 교통수단(MTRANS) 등이 존재하며, 이 column들을 바탕으로 target column인 비만 정도를 나타내는 column인 NObeyesdad의 값이 결정된다[12].

3.1.2 Reorganize dataset with experiments

NObeyesdad는 Insufficient Weight, Normal Weight, Over Weight Level 1-2, Obesity Type1-3 이렇게 7개의 값을 가진다. 이에 값의 표현을 간단하게 하고, 결과 성능을 향상시키기 위해 제품 범주 속성에 따라[12] 저체중(Insufficient Weight), 정상(Normal Weight), 과체중(Over Weight Level1, Over Weight Level2), 비만(Obesity Type1, Obesity Type2, Obesity Type3) 이렇게 4가지의 값을 가지도록 분류하였다. 또한 앞으로의 실험을 위해 object type의 값을 가지는 NObeyesdad는 저체중을 1로, 정상을 2로, 과체중을 3으로, 비만을 4로 각각 변환해 데이터 셋에 적용한 후 실험을 진행한다.

비만과 관련한 데이터 셋을 살펴보면, 비만의 정도를 결정하는 target column을 제외하고 총 16개의 column이 존재한다. 이에 모델의 복잡도를 줄이고자 여러 변수 중에서 유의미한 변수를 선택해 추출하는 과정이 필요하다. 따라서 여러 개인 성향 요소들이 비만의 정도를 결정하는데 각각 얼마나 영향을 미치는지 알아보기 위해 로지스틱 회귀 실험을 진행한다.

이때, 실험을 위한 준비로서, object column들은 Label Encoder를 이용해 숫자로 바꾸어 사용한다. 또한 모든 value는 정규화 과정을 진행해 평균 0, 표준편차 1로 반환하는 Standard Scaler를 사용해 값을 스케일링한다. 로지스틱 회귀 실험을 위해 데이터 셋을 70%의 train-set과 30%의 test-set으로 나누어 수행한다.

실험 결과, 각 개인 성향 요소 column들과 비만 정도를 나타내는 column인 NObeyesdad와의 상관관계를 계수(coefficient)의 값을 통해 나타내었다. 이때 값이 양수인 경우, 해당 변수와 비만 결정 정도 column은 양의 상

관관계를 가지며, 값이 음수인 경우는 해당 변수와 비만 결정 정도 column은 음의 상관관계를 가진다[14]. 실험 결과는 다음 Table 1과 같다.

Table 1. Result of logistic regression

feature	coefficient
Height	2.651
FCVC	0.337
NCP	0.13
TUE	0.122
CH20	0.083
SCC	0.005
FAF	-0.025
CALC	-0.054
Gender	-0.076
MTRANS	-0.111
FAVC	-0.154
Age	-0.209
family_history_with_overweight	-0.303
CAEC	-0.326
SMOKE	-0.433
Weight	-9.682

로지스틱 회귀 실험 결과에 따라, 상관 계수가 양수 값을 갖는 Height, FCVC, NCP, TUE, CH20, SCC 이렇게 총 6개의 column을 사용해 데이터 셋을 재구성한다. Table 2는 재구성된 비만 데이터 셋의 예이다.

Table 2. Obesity user dataset

	Height	FCVC	NCP	CH20	SCC	TUE	level
0	1.62	2.0	3.0	2.0	0	1.0	3
1	1.52	3.0	3.0	3.0	1	0.0	1
2	1.80	2.0	3.0	2.0	0	1.0	3
3	1.80	3.0	3.0	2.0	0	0.0	1
4	1.78	2.0	1.0	2.0	0	0.0	3

재구성된 데이터 셋으로도 여전히 비만의 정도를 잘 결정하는지 알아보기 위해 k값이 3일 때 나누어진 train-set과 test-set을 사용하여 k-최근접 이웃 실험을 진행한다. 그 결과 약 84.07%의 accuracy 값을 얻을 수 있었고, 이를 통해 재구성된 데이터 셋이 여전히 비만의 정도를 잘 결정하고 있음을 확인할 수 있다.

3.2 Exercise Dataset

운동 데이터 셋은 운동의 이름과 kg당 소모 칼로리 column으로 구성되어 있으며, 248개의 데이터가 존재한다. 비만 데이터 셋과 같이 데이터 셋을 여러 그룹으로 묶어 추천 시스템에 활용하고자 한다. 이에 운동의 레벨을 결정하기 위해, kg 당 소모 칼로리 column의 값을 바탕으로 5개의 등급을 매겨 해당 데이터 셋에 level column을 추가하였다. 따라서 운동 데이터 셋의 level column은 운동의 레벨을 나타내는 column으로 1~5의 값을 가지며, 이를 추천 시스템에 활용한다. Table 3은 운동 데이터 셋의 예이다.

Table 3. Exercise item dataset

	Activity	Calories per kg	level
243	General cleaning	0.721	1
244	Cleaning, dusting	0.515	1
245	Taking out Trash	0.617	1
246	Walking, pushing a wheel chair	0.823	2
247	Teach physical education, exercise class	0.823	2

IV. The Proposed System

본 장에서는 각 데이터 셋의 이웃 정보와 SVD 알고리즘을 사용하여 추천 시스템의 구현 방법에 대해 설명한다. 제안하는 추천 기법은 먼저 Item user matrix를 생성하고 여기에 SVD 알고리즘 적용하여 본 연구에서 소개하는 기준에 따라 추천 시스템을 구현한다.

4.1 Recommendation System Design

메모리 기반 협업 필터링 추천 기법에는 데이터 희소성과 확장성의 문제가 존재한다. 이를 해결하기 위해 주로 데이터를 비슷한 집단으로 묶는 비지도 학습 방법의 군집 분석과, 지도학습 방법의 분류 방법을 사용하곤 한다. 따라서 본 연구에서는 개인 성향 정보 데이터 셋을 데이터 분류 방법 중 하나인 k-최근접 이웃 알고리즘을 사용해 분류하여 데이터 셋을 재구성하고, 이를 통해 추천 시스템을 구현함으로써 데이터 희소성과 확장성의 문제를 해결한다. 더불어, 본 연구에서는 모델 기반 협업 필터링 추천 기법을 사용해 추천을 구현하는데 그 중에서 고차원 행렬을 저차원의 행렬로 축소시켜 암묵적 요인을 잘 정의하고, 정확성을 높이는 차원 축소 모델인 특이값 분해 알고리즘을 사용해 추천 시스템을 구현한다[8]. 제안하는 시스템의 전반적인 구조는 다음 그림 3과 같다.

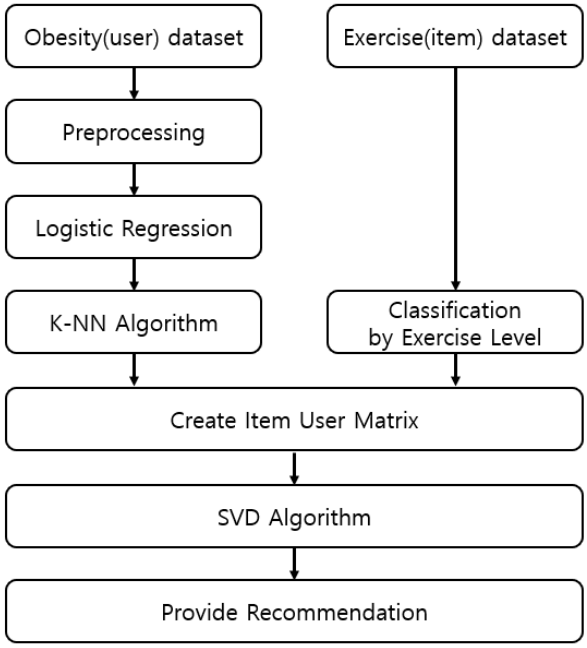

Fig. 3. Recommendation System Architecture

제안하는 시스템의 전반적인 구조인 그림 3에 대해 자세히 살펴보면, 개인 성향 정보를 포함한 사용자 데이터 셋인 비만 데이터 셋을 사용해 먼저 정규화, 전처리 과정을 진행한다. 이후, 로지스틱 회귀를 통해 비만의 정도를 나타내는 column과 양의 상관관계를 갖는 column들을 선택해 데이터 셋을 재구성하고, k-최근접 이웃 알고리즘을 통해 재구성된 데이터 셋이 여전히 비만의 정도를 잘 결정하는지 확인한다. 또한, 아이템 데이터 셋인 운동 데이터 셋을 운동의 레벨에 따라 5개의 그룹으로 분류해 해당 그룹 정보 column을 데이터 셋에 추가해 재구성한다.

재구성된 두 개의 데이터 셋을 바탕으로 Item user matrix를 생성하고 모델 기반 협업 필터링 방법 중 차원 축소 기법인 특이값 분해 알고리즘을 사용해 추천 시스템을 구현해 사용자에게 맞춤화 된 운동 추천을 제공한다.

4.2 Create Item user matrix

사용자에게 운동을 추천하는 시스템을 구현하기 위하여, 재구성된 비만 데이터 셋과 운동 데이터 셋을 사용해 Item-user matrix를 구해야 한다. 이에 비만 데이터 셋에서 비만의 정도를 나타내는 'level' column의 값과 운동 데이터 셋에서 운동의 레벨을 나타내는 'level' column의 값인 이웃의 정보를 활용한다.

한 명의 사용자에 대하여 운동 데이터 셋에 존재하는 모든 운동의 추천 점수를 결정하기 위해 비만의 정도를 나타내는 level 값과 운동 레벨 값의 차의 절댓값을 'rating'의 최댓값인 5에서 빼 구한 값으로 'rating' 값을 구하여 Item-user matrix를 생성한다.

그림 4는 item user matrix를 생성하는 알고리즘이고 Table 4는 비만 데이터 셋 2111개와 운동 데이터 셋 248개를 곱해 총 523,528개의 데이터로 구성된 matrix로 이는 그 예이다.

```
Algorithm Make_Matrix
Input : Item_set : {I_x}; User_set : {U_y};
        Ratings of users for items : {R_x,y},
        (x = 1, 2, ··· , n; y = 1, 2, ··· , m)
Output : Matrix;
1 for i = 1 to n {
2     for j = 1 to m {d
3         u_l ← U_level,x
4         i_l ← I_level,y
5         val ← absolute value of u_l − i_l
6         compute 5 − val
7         insert value into Original_x,y
8     }
9 }
```

Fig. 4. Item User Make Matrix

Table 4. Item user matrix

	user	exercise	rating
0	1	Cycling, mountain bike	5
1	1	Cycling, <10mph, leisure bicycling	4
...	...	...	...
523526	2111	Walking, pushing a wheelchair	5
523527	2111	Teach physical education, exercise class	5

4.3 Apply SVD algorithm

본 연구에서는 모델 기반 협업 필터링 방법을 통해 추천 시스템을 구현한다. 이는 메모리 기반 협업 필터링 방법의 문제점인 사용자가 구매한 아이템이 너무 적을 경우 추천이 불가능해 성능을 저하시키며 반대로 사용자와 아이템의 수가 많을 경우 유사도와 선호도 계산에 많은 계산 비용이 발생하게 되는 문제점을 해결한다. 즉, 고차원의 행렬을 저차원의 행렬로 축소시키는 기법인 차원 축소 기법인 특이값 분해 알고리즘을 사용해 위 문제점을 해결하고[6], 앞서 생성한 Item user matrix에 특이값 분해 알고리즘을 적용하여 사용자에게 운동 추천을 제공할 수 있다[16].

즉, 특이값 분해 알고리즘은 데이터에서의 노이즈(noise)를 제거하고, 희소성 문제를 해결하는데 도움을 준다. 이와

함께 아이템의 개수를 줄여 차원을 축소함으로써 암묵적 요인을 잘 정의해 분석의 정확성을 높이고 빠른 속도로 결과를 도출할 수 있도록 한다. 따라서 다음 장에서 실험을 통해 해당 알고리즘의 정확도에 대해 알아보고자 한다.

V. Experiments

본 연구에서 제안하는 추천 시스템이 잘 구현되었는지 알아보기 위해 분류와 회귀의 평가 척도들 중 정확도, 재현율 등 값의 계산을 통해 이를 확인한다. 또한, 본 연구에서 주장하는 추천 기법과 다른 추천 기법과의 비교 실험을 통해 본 연구의 정확도 및 성능에 대해 검증하고자 한다.

5.1 Experimental Environment

본 논문에서 제안하는 시스템의 실험을 위해 구글 Colab에서 파이썬 Surprise 라이브러리를 사용하여 특이값 분해 알고리즘을 구현하여 성능 계산과 비교 실험을 수행한다.

아이템에 대한 선호 여부를 분류하기 위해서는 먼저 선호 아이템에 대한 기준을 정해야 한다. 평점 범위가 1~5인 데이터에 대한 추천 시스템에서는 일반적으로 평점이 4점 이상인 경우를 선호 아이템의 기준으로 삼으며, 본 논문에서도 이와 같은 기준을 사용해 추천을 진행한다[7]. 그 이유는 아이템에 대한 평가 값을 4 이상으로 한 사용자는 같은 특성을 가진 아이템을 다시 선택할 확률이 높기 때문이다.

이 실험을 위해 Item user matrix를 70%의 train-set과 30%의 test-set으로 나눈다. 먼저 train-set을 특이값 분해 알고리즘에 적용해 훈련시키고 test-set을 통해 알고리즘의 예측 결과를 살펴본다.

또한, 사용자에게 더욱 개인 맞춤화 된 운동 추천을 제공하기 위해 특이값 분해 알고리즘에 사용되는 다양한 파라미터 값을 바꿔가며 각각의 예측 결과를 살펴보고자 한다. 값에 영향을 주는 여러 파라미터들 중에서 본 연구에서는 epoch 수와 learning rate의 값을 바꿔가며 실험을 진행하고자 한다. epoch는 [5, 10, 15, 20]의 값으로 진행하고, learning rate는 [0.0001, 0.0003, 0.0005, 0.0007, 0.0009]의 값으로 실험을 진행한다.

예측의 정확성을 평가하기 위해 먼저, 다양한 분류 문제에서 보편적으로 사용하는 평가 척도인 정밀도(precision), 재현율(recall), F1 measure, 정확도(accuracy) 값을 사용한다.

정밀도는 추천된 항목들 중에서 사용자가 선호하는 항목들의 비율을 나타내는 값으로 다음 식 1과 같다.

$$\mathrm{P}recision = \frac{True\,Positive}{True\,Positive + False\,Positive} \quad (1)$$

재현율은 테스트 데이터에서 실제로 선호된 아이템 중에서 예측 결과, 선호 아이템으로 분류된 아이템의 비율이며[15] 이를 계산하는 식은 다음 식 2와 같다.

$$Recall = \frac{True\,Positive}{True\,Positive + False\,Negative} \quad (2)$$

F1 measure는 precision과 recall을 통합한 가중치로, 추천 상품 수가 커질수록 recall 값은 증가하지만, precision값은 감소하게 된다. 즉, 두 값은 서로 상호 상충 관계에 있기 때문에 F1을 사용해 분류의 효율성을 평가하는 척도로 사용한다. 따라서 F1 measure 값이 1에 근접할수록 recall값과 precision 값 모두 높다는 것을 의미하며 0에 근접할수록 둘 중 하나의 값이 상대적으로 낮은 값임을 의미한다[6]. F1 measure를 계산하는 식은 다음 식 3과 같다.

$$F1\,measure = \frac{2 \times precision \times recall}{precision + recall} \quad (3)$$

정확도는 전체 테스트 데이터 중에서 선호여부를 정확하게 분류한 데이터의 비율이며[15] 이를 계산하는 식은 다음 식 4와 같다.

$$Accuracy = \frac{True\,Positive + True\,Negative}{Number\,of\,Test\,Dataset} \quad (4)$$

다음으로 다양한 회귀 문제에서 주로 사용하는 평가 척도 중 RMSE(Root Mean Squared Error) 값을 사용해 예측의 정확성을 평가하며, 이를 구하는 식은 다음 식 5와 같다.

$$RMSE = \sqrt{\frac{1}{N}\sum (p_{i,j} - r_{i,j})^2} \quad (5)$$

여기서 $p_{i,j}$는 예측 점수이고, $r_{i,j}$는 실제 점수며, N은 총 데이터의 개수이다. 회귀 평가 척도 중 하나인 MSE는 예측 점수와 실제 점수의 차이를 제곱한 후 평균한 값으로, 오차에 제곱을 취함으로써 오차가 큰 값에 대해 가중치를 높게 부여하는 평가 방법입니다. RMSE는 이 MSE의 제곱근으로써, MSE 값에 비해 예측 오차가 큰 관측 치에 대해 상대적으로 적은 가중치를 부여한다[6]. 따라서

RMSE의 값이 작을수록 추천 시스템의 예측 정확성이 우수함을 나타낸다.

5.2 The result of the experiment

앞에서 제시한 분류 평가 척도인 precision, recall, F1, 그리고 accuracy를 사용해 본 연구에서 제안하는 추천 시스템의 예측 성능을 평가해 보았다.

먼저, learning rate 값을 0.0005로 고정해두고, epoch 값을 다르게 해 실험을 진행해 본 결과, 이는 다음 Table 5와 같다.

Table 5. Result of different n_epochs

epoch	precision	recall	F1	accuracy
5	0.965	0.526	0.68	0.668
10	0.957	0.682	0.797	0.766
15	0.98	0.781	0.869	0.842
20	0.999	0.824	0.904	0.882

위 Table 5를 바탕으로 그래프로 표현해보면, 다음 그림 5과 같이 나타낼 수 있다.

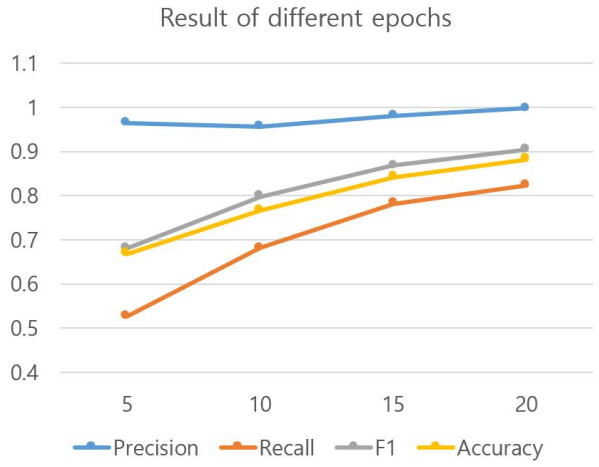

Fig. 5. Result of different epochs

이번엔 반대로, epoch 수를 10으로 고정하고 learning rate를 다르게 해 실험을 진행해본 결과 는 다음 Table 6과 같다.

Table 6. Result of different learning rate

lr_all	precision	recall	F1	accuracy
0.0001	0.921	0.425	0.582	0.59
0.0003	0.967	0.562	0.711	0.693
0.0005	0.957	0.682	0.797	0.766
0.0007	0.978	0.77	0.862	0.834
0.0009	0.998	0.802	0.889	0.866

마찬가지로, 위 Table 6을 바탕으로 그래프로 표현해보면, 다음 그림 6와 같이 나타낼 수 있다.

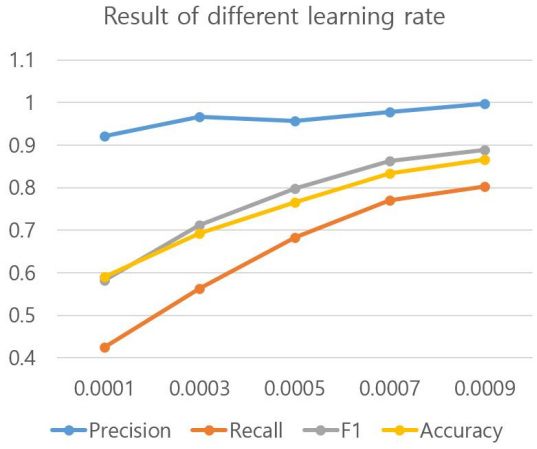

Fig. 6. Result of different learning rate

5.3 Comparison experiments

5.3.1 Compare Precision, Recall, F1 values

본 연구에서의 실험이 메모리 기반 협업 필터링 추천 시스템의 문제점을 해결하고 개인 맞춤화 된 추천을 제공하고 있음을 증명하기 위해 다른 추천 기법을 참조해 본 연구에서 제안하는 추천 기법과 비교해 보기로 한다. 먼저, 개인화 요소를 사용한 헬스 케어 관련 추천과 관련한 실험 방법 중 rule - based 추천 방법을 사용한 방법 (Personalization + Rule-based) [5]과 DBSCAN 클러스터링 알고리즘을 사용해 데이터를 분류한 후, 메모리 기반 협업 필터링 방법을 통해 알맞은 영화를 추천하는 방법 (Cluster + CF) [3], 그리고 본 연구에서 제안하는 추천 기법을 비교하는 실험을 진행한다. 이는 분류 평가 척도인 정밀도, 재현율, F1 measure 값을 사용해 추천 기법의 성능을 비교하고, 그 결과는 다음 Table 7과 같다.

Table 7. Comparison with other techniques by precision, recall, f1 score values

System	Precision	Recall	F1
Proposed System	0.99	0.8	0.89
Personalization + Rule-Based [5]	0.73	0.98	0.84
Cluster+CF [3]	0.86	0.76	0.81

개인 성향 정보를 추천에 사용하였지만, 데이터 셋을 분류하지 않은 [5]의 추천 방법에 비해 본 연구는 precision 부분에서 약 26%, F1 부분에서 약 5% 더 나은 성능을 보인다. 또한 클러스터링 알고리즘을 사용해 데이터를 분류

하였지만, 개인 성향 정보를 추천에 사용하지 않은 [3]의 추천 방법과 비교했을 때, 이 논문에서 제안하는 방법이 precision 부분에서 약 13%, recall 부분 약 4%, 그리고 f1 부분에서 약 8% 더 나은 성능을 보임을 확인할 수 있다.

5.3.2 Compare RMSE values

이번엔 본 연구에서와 같이 모델 기반 협업 필터링 방법 중 차원 축소 모델인 특이값 분해 알고리즘을 사용한 다른 모델과 비교한다. 비교할 모델은 Item-based 협업 필터링 모델을 기반으로 특이값 분해 알고리즘을 적용한 추천 방식으로[4] 본 연구에서 제안한 시스템과 비교 실험을 진행한다. 이때 회귀 문제 평가 척도인 RMSE값을 사용해 성능을 비교하고, 결과는 다음 Table 8과 같다.

Table 8. Comparison with other model by RMSE values

System	RMSE
Proposed System	0.6044
IBCF-SVD [4]	0.8199

따라서 두 방법 모두 같은 특이값 분해 차원 축소 알고리즘을 이용해 추천 시스템을 구현하였지만, 개인화 요소를 사용하고 데이터 분류 작업을 거친 본 연구에서 진행한 모델이 [4]의 추천 방법보다 약 0.2155만큼 더 나은 성능을 보여주고 있음을 확인 할 수 있다.

VI. Conclusions

본 연구에서는 비만 데이터 셋인 사용자 관련 데이터 셋과 운동 데이터 셋인 아이템 관련 데이터 셋을 사용해 사용자들에게 의미 있는 운동 추천을 제공하고자 한다. 이때, 더욱 개인 맞춤화된 운동을 사용자에게 추천하기 위해 신체적 조건, 식습관 정보 등의 개인 성향 정보를 사용하여 비만의 정도를 구별하였다. 즉, 협업 필터링의 단점인 데이터 희소성과 확장성의 문제를 해결하기 위해, 데이터를 분류하는 Classification 기법 중 k-최근접 이웃 알고리즘을 사용하여 사용자들을 k개의 이웃으로 나누어 추천을 진행한다.

또한, 사용자가 구매한 아이템이 너무 적을 때 성향 분석을 통한 추천이 거의 불가능하고, 반대로 사용자와 아이템의 수가 너무 많을 경우 유사도와 선호도의 계산에 있어 많은 계산 비용이 발생하는 문제인 메모리 기반 협업 필터링 방법의 문제점을 해결하고자, 본 논문에서는 모델 기반 협업 필터링 방법으로 추천을 진행한다.

위의 문제를 해결하기 위해 고차원의 행렬을 저차원의 행렬로 축소시키는 기법인 특이값 분해 SVD 알고리즘을 사용하여 추천을 구현한다. 차원을 축소시켜 분석의 정확성을 높이며 계산 속도도 줄일 수 있는 SVD 알고리즘을 통해 본 연구에서 정확도, 정밀도, 재현율, F1 measure 값, 그리고 RMSE 값까지 모두 우수한 성능을 나타냄을 확인할 수 있다. 또한, 다른 추천 기법과 비교 실험을 진행해 본 논문에서 제안하는 시스템이 충분한 성능을 발휘하고 있음을 알 수 있다.

향후 연구에서는 운동 추천뿐만 아니라 사용자들의 건강에 영향을 미칠 수 있는 다양한 요소들을 추천할 수 있도록 확장하고자 한다.

ACKNOWLEDGEMENT

This research was supported by the Basic Science Research Program through the NRF (National Research Foundation of Korea), funded by the MSIT (Ministry of Science and ICT), Korea (Grant No. NRF2019R1A2C1008412)

REFERENCES

[1] Jong-Chan Lee, and Moon-Ho Lee, “Big data-based information recommendation system,” Journal of the Korea Institute of Information and Communication Engineering, Vol. 22, No. 3, pp. 443-450, Mar. 2018.

[2] S. B. Ahire and H. K. Khanuja, "A Personalized Framework for Health Care Recommendation," 2015 International Conference on Computing Communication Control and Automation, pp. 442-445, Feb. 2015. doi: 10.1109/ICCUBEA.2015.92.

[3] J. Das, P. Mukherjee, S. Majumder and P. Gupta, "Clustering-based recommender system using principles of voting theory," 2014 International Conference on Contemporary Computing and Informatics (IC3I), pp. 230-235, Nov. 2014. doi: 10.1109/IC3I.2014.7019655.

[4] Dong-Wook Kim, Sung-Geun Kim, and Ju-Young Kang, “An Empirical Study on Hybrid Recommendation System Using Movie Lens Data,” Korea Bigdata Society, Vol. 2, No. 1, pp.41-48, Feb. 2017.

[5] J. Kim, K. Lee, D. Park and E. Jung, "Context-Aware U-Health Service: Identification of Exercise Recommendation Factors and Creation of Decision-Making Model Using Association Rule,"

2013 International Conference on Information Science and Applications (ICISA), pp. 1-4, Aug. 2013. doi: 10.1109/ICISA.2013.6579439.

[6] Jieun Son, Seoung Bum Kim, Hyunjoong Kim, and Sungzoon Cho, “Review and Anaylsis of Recommender Systems,” Journal of the Korean Institute of Industrial Engineers, Vol. 41, No. 2, pp. 185-208, April. 2015.

[7] Hyemin Ko, Serim Kim, and Namhi Kang, “Design and Implementation of Smart-Mirror Supporting Recommendation Service based on Personal Usage Data,” KIISE Transactions on Computing Practices, Vol. 23, No. 1, pp. 65-73, Jan. 2017.

[8] Seung-Yoon Jeong, and Hyoung Joong Kim, “A Recommender System Using Factorization Machine,” Journal of Digital Contents Society, Vol. 18, No. 4, pp.707-712, 7. 2017.

[9] Jiyeon Hyun, Sangyi Ryu, and Sang-Yong Lee, “How to improve the accuracy of recommendation systems : Combining ratings and review texts sentiment scores,” Journal of Intelligence and Information Systems, Vol. 25, No. 1, pp. 219-239, Mar. 2019.

[10] Kaggle, “Obesity based on eating habits & physical cond.”, https://www.kaggle.com/ankurbajaj9/obesity-levels

[11] Kaggle, “Calories Burned During Exercise and Activities”, https://www.kaggle.com/aadhavvignesh/calories-burned-during-exercise-and-activities

[12] Fabio Mendoza Palechor, and Alexis de la Hoz Manotas, “Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from Colombia, Peru and Mexico,” Data in Brief, Vol. 25, Aug. 2019.

[13] Geetha G, Safa M, Fancy C, and Saranya D, “A Hybrid Approach using Collaborative filtering and Content based Filtering for Recommender System,” National Conference on Mathematical Techniques and its Applications (NCMTA 18), Vol. 1000, Jan. 2018. doi :10.1088/1742-6596/1000/1/012101

[14] Soyeon Jung, and Keumjin Lee, “Prediction Model with a Logistic Regression of Sequencing Two Arrival Flows,” Journal of Korean Society for Aviatioin and Aeronautics, Vol. 23, No. 4, pp. 42-48, Dec. 2015.

[15] Chan-soo Park, Taegyu Hwang, Junghwa Hong, and Sung Kwon Kim, “Recommendation Algorithm by Item Classification Using Preference Difference Metric,” KIISE Transactions on Computing Practices, Vol. 21, No. 2, pp. 121-125, Feb. 2015.

[16] In-Jeong Jeong, Bo-Mi Kim, Su-Kyung Kim, Kyeonah Yu, “News Recommendation System Based on Text and Image Tag Data,” Journal of Digital Contents Society, Vol. 21, No. 3, pp. 479-486, Mar. 2020.

Authors

Ha-Young Lee is an undergraduate student of AI·SW at Gachon University, Korea. She is interested in health-care, recommendation system, machine learning, deep learning, and artificial intelligence.

Ok-Ran Jeong received Ph.D. degrees in Computer Science and Engineering from Ewha Womans University, Korea, in 2005. She was a postdoctoral researcher at the University of Illinois at Urbana-Champaign, USA and Seoul National University, Korea. Dr. Jeong joined the faculty of the Department of Software Design & Management at Gachon University, Seongnam, Korea, in 2009. She is currently an associate Professor in the Department of AI·SW, Gachon University. She is interested in big data mining, machine learning, deep learning and applications of artificial intelligence.