J. Internet Comput. Serv.
ISSN 1598-0170 (Print) / ISSN 2287-1136 (Online)
http://www.jics.or.kr

협력 필터링에서 유전자 알고리즘을 활용한 최적 유사도 산출

Optimal Similarity Calculation using Genetic Algorithms in Collaborative Filtering

이 수 정1*
Soojung Lee

요 약

협력 필터링 기반의 추천 시스템은 현 사용자를 위한 추천 항목들을 제공함에 있어서 유사한 인접 이웃들이 선호한 항목들을 우선적으로 고려하는 방식이다. 시스템의 성능을 위하여 유사도 척도는 매우 중요한데, 본 연구에서는 유전자 알고리즘을 활용하여 최적 성능을 가져오는 사용자 간 유사도 값을 산출하였으며, 특히 평가 항목 특성별로 유전자 알고리즘을 별도 실행하여 예측 정확도 성능을 높이고자 하였다. 성능 실험을 통하여 유전자 알고리즘의 연산 확률 적정값을 구하였고, 두 종류의 공개 데이터셋을 활용한 실험 결과로서 제안 방법의 예측 성능이 기존 방법들보다 우수하고, 특히 희소 데이터 환경에서 더욱 우수함을 확인하였다. 본 연구 결과는 개인화된 추천의 정확성을 개선하고, 대규모 사용자 및 항목 데이터가 존재하는 실세계 애플리케이션에서 유용하게 활용될 수 있다.

☞ 주제어 : 추천 시스템, 협력 필터링, 유사도 척도, 유전자 알고리즘

ABSTRACT

A collaborative filtering-based recommender system is a method that gives priority to items preferred by similar neighbors when providing recommended items for the current user. The similarity measure is very important for the performance of the system. In this study, a genetic algorithm was used to calculate the similarity value between users that results in optimal performance. In particular, the genetic algorithm was run separately for each rated item feature to improve prediction accuracy performance. Through performance experiments, the optimal probabilities of the genetic algorithm operators were obtained, and as a result of experiments using two types of public datasets, it was confirmed that the prediction performance of the proposed method was superior to that of existing methods, especially in a sparse data environment. The results of this study can improve the accuracy of personalized recommendations and be effectively applied in real-world applications with large-scale user and item data.

☞ keyword : recommender system, collaborative filtering, similarity measure, genetic algorithm

1. 서 론

전자 상거래 사용자들은 압도적으로 많은 상품과 서비스로 인해 선택의 어려움을 겪고 있다. 추천 시스템은 의사 결정을 돕기 위한 목적으로 사용자의 선호를 예측함으로써 개인화된 콘텐츠를 제공한다. 현재 전자상거래 영역에서는 추천 시스템을 널리 사용하고 있는데, 일반적으로 가장 유명한 분야는 음악, 뉴스, 쇼핑, 서적 등이다[1].

추천 시스템의 구현을 위하여 여러 가지 방식들이 연구되어 왔다. 대표적으로, 내용 기반 필터링(content-based filtering, CBF)과 협력 필터링(collaborative filtering, CF)을 들 수 있다[2]. CBF는 사용자 선호 항목들에 대한 이력을 바탕으로 선호를 분석하여 항목을 추천한다. 예로서, 문서 내용, 뉴스 정보, 웹 로그, 항목 특징 등을 분석한다. 따라서, 선호 정보가 없는 초기 단계에는 추천 정확도를 보장할 수 없다. 또한, 사용자가 과거 선호했던 내용만의 추천 리스트를 지속적으로 제공한다는 문제점도 있다.

CF는 이러한 내용 기반 필터링의 문제점들을 해결 가능한데, 그 이유는 사용자들의 선호도 간의 유사성을 기반으로 하기 때문이다[2]. 예를 들어, 두 사용자 A와 B의 선호 항목들이 매우 유사할 때, 만약 사용자 B가 항목 k를 선호하였다면, 이를 사용자 A에게 추천하는 방식이다. 따라서, 유사도 측정은 CF 시스템의 성능에 매우 중요한 역할을 하며 다양한 척도들이 개발되어 왔다[3][4].

CF는 위와 같은 이점으로 인하여 많은 학계 및 상업계의 관심을 받아 왔으며, 크게 메모리 기반 기술과 모델 기반 기술로 분류된다. 특히, 모델 기반 기술을 위하여

[1] Dept. of Computer Education, Gyeongin National University of Education, Anyang, 13910, Korea.
* Corresponding author (sjlee@gin.ac.kr)
[Received 8 July 2024, Reviewed 25 July 2024(R1 20 September 2024), Accepted 4 October 2024]

다양한 종류의 세부 알고리즘들이 개발되었다. 일례로, 베이지안 네트웍 모델, 클러스터링 모델, 마르코프 결정 프로세스 모델 등이 있다[2]. 최근 딥러닝 기술을 활용한 시도가 활발한데 신경망을 이용한 잠재 요인 모델링은 사용자와 항목 간의 복잡한 상호작용을 효과적으로 학습하여 비선형적인 관계를 파악하는 장점을 가진다[5].

메모리 기반 방법에서는 사용자-항목 행렬을 기반으로 사용자 간 또는 항목 간의 유사성을 계산하여 추천을 생성한다. 이 방법은 간단하고 직관적이며 모델 기반 방법의 여러 단점, 즉, 막대한 계산 비용, 과적합 문제, 추천 결과 해석의 어려움 등을 해결한다. 유사도 산출 방법은 시스템 성능을 좌우하는 매우 중요한 요소이다. 가장 기본적으로 피어슨 상관(Pearson Correlation), 코사인 유사도(Cosine Similarity), 자카드 유사도(Jaccard Similarity)나 유클리드 거리(Euclidean Distance) 등이 사용되었으며, 이들을 융합 및 개선한 방법들도 개발되었다[3][6][7][8][9]. 그러나, 각각의 유사도 측정 방법은 데이터의 특성과 목표에 따라 다른 성능을 보이며, 적절한 방법을 선택하기 위하여 휴리스틱에 의존하는 경우가 대부분이다.

본 연구에서는 메모리 기반 CF 시스템의 최적 성능을 가져오는 사용자 간 유사도 값을 산출한다. 이를 위하여 기존과 같은 휴리스틱이 아닌 유전자 알고리즘(Genetic Algorithm, GA)을 활용하며, 사용자의 미평가 항목에 대한 예측 성능을 적합도 함수로 설정한다. GA를 추천 시스템에 활용하는 기존 연구는 많지 않으며, 더군다나 최적의 유사도 산출을 위한 목적으로 활용된 예는 극히 드물다[10]. 본 연구의 또다른 특징은 사용자 간 유사도는 평가한 항목의 특성에 따라 다를 수 있다는 가정 하에, 각 항목 특성별로 GA를 활용하여 유사도를 산출한다. 성능 실험을 위하여, 학계에서 널리 활용되는 공개 데이터셋을 이용하였으며, 실험 결과 제안 방법은 예측 성능 면에 있어서 기존 방법들을 능가하였으며, 특히 희소 데이터 환경에서 더욱 우수한 성능 향상을 보였다.

논문의 구성은 다음과 같다. 2장에서는 CF의 유사도 산출 및 GA와 관련된 기존 연구 성과를 소개한다. 3장에서는 제안 방법을 기술하고 4장에서 성능 실험 결과를 비교 제시하며, 5장에서 논문의 결론을 맺는다.

2. 관련 연구

메모리 기반의 CF 시스템은 크게 항목 기반 필터링(item-based filtering)과 사용자 기반 필터링(user-based filtering)의 두 유형으로 구성된다. 전자의 경우 항목 간의 유사도를 측정하여 사용자가 선호하는 특정 항목과 유사한 항목을 추천하며 확장성 문제(scalability problem)와 새로운 사용자 문제(new user problem)를 해결하는 장점이 있지만, 사용자가 기선호한 항목들과 유사한 항목들만을 추천하므로 새로운 항목 추천에 한계가 있을 수 있다[1][4]. 한편 후자는 개인화된 추천을 제공하는 강점이 있지만, 확장성과 데이터 희소성 문제를 겪을 수 있다[11].

유사도 측정은 시스템 성능에 매우 큰 영향을 주므로, 많은 연구자들의 주요 관심 분야가 되어 왔다. Abdalla 외 4인은 기존의 유사도 산출 방법을 단독 사용할 때의 문제점을 언급하고, 효과적인 유사성 측정을 위해 자카드와 여러 다양한 방법들의 결합을 제안하였다[12]. Bag 외 2인은 희소 데이터 환경에서 적절한 유사 사용자들을 식별하기 위해 사용자의 모든 평가 벡터를 고려하여 새로운 유사도 측정 모델인 관련 자카드 유사도를 제안하였다[6]. [3]에서는 항목 기반과 사용자 기반 CF의 결합 모델을 제안하였는데, 최적 효율을 위한 사용자의 유사 이웃 수를 추정했으며, 콜드 스타트(cold-start) 및 데이터 희소성 문제의 처리 방법도 제안하였다. [13]의 연구에서는 유사도 척도가 충족해야 할 직관적이고 질적인 조건을 수학 방정식으로 변환하고, 유사도 척도의 커널 함수 달성을 위해 방정식을 해결하는 수학적 표현을 제안하였다.

한편 Iftikhar 외 4인의 연구에서는 삼각형 유사성을 기반으로 CF를 위한 제품 추천 방법을 소개하였으며, 이를 더욱 개선하여 사용자들의 공통 평가 항목들 뿐만 아니라 비공통 평가 항목들도 고려하여 사용자의 평가 선호 행태로서 제안 유사도를 보완하였다[7]. 이 밖에 [8]의 연구에서는 대개의 기존 시스템에서 추구하는 단일 종류의 성능 목표를 넘어서서 인기도와 다양성의 두 가지 성능 목표에 대한 솔루션을 얻기 위해 다중 목표 최적화를 추구하였는데, 새로운 유사도 모델로서 기존의 비선형 유사도 측정 모델과 Bhattacharyya 계수의 통합을 제시하여 기존 등급 평가 방법의 예측 정확도를 높이고자 하였다.

기존의 단일 유사도 측정 방법은 서로 다른 특성의 데이터셋에 모두 적합한 성능을 보이지 않으므로, [14]의 연구에서는 피어슨 상관도를 개선하기 위하여, 희소성 문제의 완화와 공통 평가 항목수의 영향 감소를 위해 새로운 거리 함수를 전역 유사도 척도로 제안하였다. 한편 [9]에서는 기존의 메모리 기반 CF는 최첨단 모델 기반 CF보다 우수한 예측 정확도를 보이지 않았음을 주목하고, 구조적 유사성과 등급 기반 유사성 측정을 결합한 새로운 측정 방식을 제안하였으며, 이 방식이 모델 기반 CF보다 예측 성능이 우수하고 메모리와 시간을 절약함을 보였다.

최근 인구 기반 메타휴리스틱 최적화 기법인 진화 알고리즘 중 하나인 유전자 알고리즘(Genetic Algorithm, GA)을 CF 시스템에 활용한 연구가 시도되어 왔다. GA 방법은 초기 해집단에서 출발하여 선택, 교차, 돌연변이 과정을 반복함으로써 점진적으로 최적 솔루션에 도달한다. [10]에서는 세 가지 적합도 함수가 있는 유전 알고리즘을 채택하였는데, 이들은 의미적 상관관계, 만족도 기반, 그리고 평가 예측과 관련된 필터링 레벨이며, 제안 알고리즘은 높은 시간 복잡도를 가지는 것으로 판명되었다. [15]에서는 전통적인 유사도 측정 도구를 사용하지 않고 GA를 활용한 최적의 사용자 간 유사도 값을 계산하였으며, 100명의 사용자들로 구성된 합성 데이터를 활용하여 성능을 분석하였다. Jain 외 2인은 제안된 다중 목표 추천 필터링에서 다중 부모 교차 메커니즘을 제시하였는데, 이는 인기 있고 다양한 항목을 추천할 때 우수한 객관성을 위해 부모 유전자의 순서와 빈도를 유지하는 방식이다[8]. [16]에서는 사용자 간의 유사도 값을 직접 사용하는 기존 방법과 달리, 이들 값을 예측 과정에 사용하기 전에 GA를 활용하여 정제하였으나, 전통적 유사도 척도 외에 기존의 다른 GA 기반의 CF 방법과는 성능 비교 실험을 수행하지 않았다. 최근 불필요한 항목들에 대한 추천을 배제하기 위한 상관된 항목들의 부분 그룹을 기반으로 하는 방법이 대두되었는데, [17]의 연구에서는 GA의 활용 목적이 최적의 유사도값을 구하기 위한 것이 아니라, 유사한 사용자 집합을 기반으로 상관 항목의 부분 그룹을 형성하여 관련 항목에 대해서만 예측을 얻는 네 가지 새로운 기술을 제안하였다.

한편 Liu 외 3인은 제조 서비스 분야에서 제조 서비스 구성의 개인화된 추천을 위한 하이브리드 방식을 제안하였으며, QoS 목표 속성과 고객 선호 속성을 종합적으로 고려하여 제조 서비스 구성 최적화의 부족한 개별화 결함을 해결하고 고객 선호 속성 순위를 통해 최적의 솔루션을 추천하고자 하였다[18]. [19]의 연구에서는 퍼지 유사도 산출 및 퍼지 예측을 통해 추천 항목을 결정하였는데, 퍼지 유전 CF 방식을 활용하여 퍼지 유사도 값의 최적화를 시도하였으며, 우수한 성능의 추천 결과를 산출한 반면에 콜드 스타트 문제에 대한 해결책은 제시하지 않았다. 이밖에 [20]에서는 연관 규칙과 입자 군집 최적화(particle swarm optimization)와 같은 진화 알고리즘을 사용한 기존 방법의 정확성 및 런타임 성능의 부족함을 언급하고, GA를 기반으로 더 높은 성능을 갖는 연관 규칙을 생성하는 효율적인 방법을 제안하였으므로 본 연구 목표와 같이 최적의 유사도값을 산출하기 위한 것과는 차이가 있다.

이상과 같이 메모리 기반 CF 시스템에서 유사도의 측정은 중요한 주제이나, 진화 알고리즘을 활용한 연구 결과는 많지 않으며 더욱이 유사도 값 자체의 최적화 작업은 거의 시도되지 않은 것으로 확인되었다. 진화 알고리즘은 해집합 크기, 후보해 선택과 변이 파라미터의 조정 등을 통하여 지역 최적값에 갇히는 문제를 해결하며 비선형적이고 다차원적인 최적화 문제인 최적의 유사도 값을 산출하기에 적합하다. 또한 본 연구에서는 항목 특성별로 최적값을 구하므로 대규모 데이터셋의 경우에 병렬처리를 통한 알고리즘의 효율화가 가능하다.

3. 제안 방법

3.1 연구 동기

본 아이디어는 [15]의 연구를 기반으로 수립되었다. [15]에서는 임의의 두 사용자 간의 최적의 유사도를 산출하려는 연구 목적을 위하여 GA를 활용하였다. GA의 매 반복 회차마다 구해진 유사도 값들을 기반으로 시스템에서 산출한 사용자의 미평가항목들의 예측 평가치를 구한 후, 이들의 실제 평가치와의 차이를 최소화하도록 최적화 함수를 설정하였다.

이와는 대조적으로 본 연구에서는 사용자 간의 유사도가 항목의 특성에 따라 좌우될 수 있다고 가정한다. 예를 들어, 두 사용자가 A 특성의 항목들에 대하여 매우 유사한 평가치를 부여함으로써 유사도가 0.9이고, B 특성의 항목그룹에 대하여는 별로 유사하지 않은 평가를 함으로써 유사도가 0.3이라고 하자. [15]의 연구에서 GA는 항목 특성과 무관하게 전체 항목에 대해 최적의 유사도를 산출하므로, 이 예의 경우에 0.9와 0.3의 평균값인 0.6의 유사도가 산출된다면 이 값을 기준으로 미평가항목들의 예측 평가치를 구하기 때문에, A와 B 특성의 항목들에 대한 예측치 모두 정확도가 낮아진다. 따라서 전체 항목이 아닌 각 특성의 항목 그룹별로 이들을 평가한 사용자 간의 최적화된 유사도를 산출하는 것이 바람직하다.

3.2 항목 특성별 사용자 간 유사도 산출

시스템에서 제공하는 항목들은 특성별로 분류될 수 있다고 가정한다. 예를 들어, 영화 데이터셋에서 특성을 장르로 정의하며 각 영화 항목은 하나 또는 여러 개의 장르에 속할 수 있다. 이러한 배경하에, 본 연구에서 제안하는 CF 시스템은 다음과 같은 절차를 갖는다.

1. 항목 평가 기록이 있는 사용자들을 각 항목 특성 f_1, f_2, ..., f_k별로 나눈다.
2. 각 항목 특성별로 별개의 GA를 수행하여 사용자 간 유사도를 산출한다. 결과적으로 임의의 항목 특성 f에 대하여 사용자 u와 v 간 최적의 유사도 $sim_f(u, v)$를 얻는다.
3. 현 사용자 u가 미평가한 항목 x에 대한 평가 예측치를 구하기 위하여, x가 속한 특성(들) f에 대하여 2에서 산출한 유사도 $sim_f(u, v)$를 활용한다. x의 평가 예측치는 아래와 같이 산출하는데, $\bar{r}_u$와 $r_{v,x}$는 각각 u의 평균 평가치, v의 x에 대한 평가치이다.
$$\overline{r_u} + \frac{1}{\sum_f \sum_v |\, sim_f(u,v)\,|} \sum_f \sum_v sim_f(u,v)(r_{v,x} - \overline{r_v})$$
4. 평가 예측치가 높은 순위대로 해당 항목들을 현 사용자에게 추천한다.

본 연구의 GA에서 하나의 해(solution)는 모든 두 사용자 간의 유사도, 즉, 사용자수×사용자수 개의 실수값으로 구성한다. 적합도 함수(fitness function)로서 CF 연구의 예측 정확 척도로 널리 사용되는 평균 절대 오차(Mean Absolute Error, MAE)를 채택하였다. 구체적인 GA 절차는 다음과 같다.

1. 각각의 해(solution)를 구성하는 모든 유사도 값을 0~1 사이의 실수값으로 초기화한다.
2. 각각의 해에 대하여, 적합도 함수를 적용한다.
3. 다음의 절차에 따라 새로운 세대를 형성한다.
   3.1 적합도 값에 따른 확률을 적용하여 현 세대에서 두 개의 해를 선택하고, 교차 확률값에 따른 교차(crossover) 연산을 수행하여 두 개의 새로운 자손 해를 생성한다. 이들 각 자손 해에 대하여 변이(mutation) 확률값에 따른 변이 연산을 수행한다.
   3.2 생성된 자손 해의 개수가 기존 부모 세대의 개수와 동일하게 되도록 3.1 절차를 반복한다.
4. 목표로 한 최적의 해를 발견하거나 또는 특정 반복 실행 회수에 도달할 때까지 2~3의 절차를 반복한다.

위 알고리즘에서 알 수 있듯이, GA를 통하여 최적의 결과를 얻기 위하여는 다양한 종류의 매개변수값의 적절한 설정이 중요하다. 즉, 유전 연산 확률값, 적합도 함수의 결정, 해의 개수, 알고리즘 실행 회수 등의 결정은 어려운 문제이다. 관련 연구에서는 대개 이들 값의 다양한 조합에 대하여 각각 성능 실험 결과를 얻은 후 최선의 결과를 가져온 값의 조합으로 결정한다. GA의 또다른 단점은 최적의 결과를 얻기 위한 진화 절차에 상당한 시간이 소요된다는 사실이다. 그러나, 현 멀티코어 또는 멀티프로세서 시스템은 병렬 처리를 제공하므로, 이를 이용하여 시간 비용을 감소시킬 수 있다.

4. 성능 실험

4.1 실험 데이터

본 연구의 성능 실험을 위한 데이터셋은 사용자, 항목, 항목 평가치, 항목 특성 등의 정보를 포함하여야 한다. 이를 만족하는 공개 데이터셋으로서 추천 시스템 관련 연구에서 널리 활용되어 온 MovieLens 1M과 CiaoDVD를 선정하였다. 이들은 각각 영화 항목이 속하는 장르 정보를 제공하므로 이를 항목 특성으로 간주하였다.

MovieLens와 CiaoDVD는 원래 6040명과 17615명 사용자들의 평가데이터를 각각 포함하지만, 본 연구의 GA를 수행하는 컴퓨팅 자원의 용량 및 처리 속도를 고려하고, 과거 연구의 대부분이 원래 데이터 세트의 0.03~8%만 사용했다고 보고된 점을 고려하며[21], 성능 결과를 현실적 시간 내에 얻기 위해 각각 임의로 추출한 1000명의 사용자들과 그들의 평가데이터로 국한하여 실험하였다. 이와 유사한 규모의 데이터 환경은 GA를 활용한 기존의 여러 CF 연구에서도 도입하였는데, 구체적으로 943명의 사용자와 1682개의 항목수를 가진 MovieLens 100K를 활용하였다[15][16][17][19]. 특히 [18]에서는 200명의 사용자와 500개의 제조서비스와 관련된 평가레코드를 활용하여 그 규모가 매우 작았다. 본 연구의 실험 데이터셋의 특성은 표 1과 같다.

(표 1) 데이터셋의 특성
(Table 1) Characteristics of the dataset

	MovieLens	CiaoDVD
사용자수	1000	1000
항목수	3952	16121
평가치 범위	1~5의 정수값	1~5의 정수값
희소성 수준	0.96099	0.99877
장르 개수	18	17

4.2 성능 평가 척도

성능 비교를 위한 시스템의 평가 기준으로서, 본 연구에서는 예측 정확도인 MAE(Mean Absolute Error, 평균절대오차), RMSE(Root Mean Square Error, 평균 제곱근 오차), 그리고 커버리지(coverage)를 선택하였다. 이들은 관련 연구에서 많이 활용되는 대표적인 척도이다. MAE는 시스템이 미평가 항목에 대하여 산출한 평가 예측치의 정확도를 측정하는데, 예측치와 실제치 차이의 절대값 평균으로 구한다. 즉, $r'_{u,x}$를 사용자 u의 항목 x에 대한 예측치라고 할 때, $\frac{1}{n}\sum_{u}\sum_{x} | r_{u,x} - r'_{u,x}$로써 산출한다.

예측 정확도를 판단하는 또다른 척도인 RMSE는 오차의 차이가 클수록 정확도 성능을 더욱 저하시키는 방법이다. 산출식은 $\sqrt{\frac{1}{n}\sum_{u}\sum_{x}(r_{u,x} - r'_{u,x})^2}$ 와 같다. 또한, 커버리지를 활용하여 예측 정확도를 판단할 수 있는데, 전체 미평가 항목들 중에서 시스템이 예측치를 산출할 수 있는 항목들의 비율을 말한다. CF 정의에 따르면, 유사한 이웃 사용자들이 미평가 항목을 평가한 이력이 있을 경우에만 예측치를 계산할 수 있기 때문에, 시스템이 선정한 이웃 사용자들의 구성과 밀접한 관련이 있다.

성능 비교 대상의 유사도 척도로서 본 제안 방법의 기반이 되는 척도들을 선정하여 성능 향상 정도를 파악하였다. 즉, 피어슨 상관도(COR), 코사인 유사도(COS), 평균자승차이(MSD), 그리고 [15]의 방법(SGA)이다. 제안 방법은 SGRGA로 표기하였다.

4.3 성능 결과

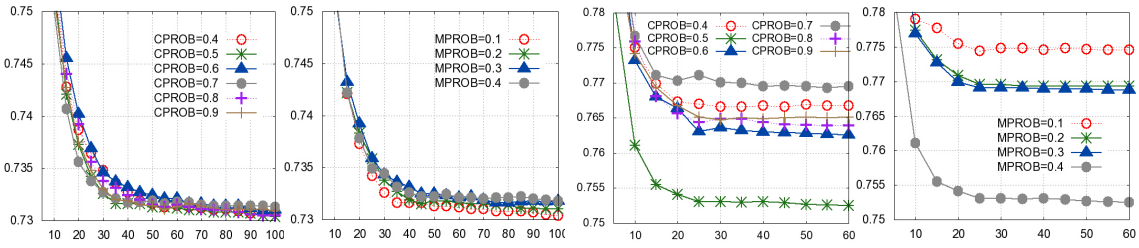
4.3.1 유전자 알고리즘 확률 변화에 따른 성능

GA 연산 확률은 새로운 세대의 해, 즉, 유사도 값을 좌우하므로, 시스템 성능에 영향을 미친다. 이에 제안 방법에서 활용한 교차 연산 확률(CPROB)과 변이 연산 확률(MPROB) 변화에 따른 성능을 관찰하였다. 그림 1(a)와 (b)는 MovieLens를 활용하여 연산 확률의 영향을 살펴보기 위하여 인접이웃수의 변화에 따라 MAE를 산출한 것이다. 두 실험 결과 모두에서 성능 차이는 미미하므로 확률의 영향은 크지 않음을 알 수 있다. 그러나 가장 우수한 결과를 가져온 확률값을 추후 실험 조건으로 하였다 (CPROB=0.5, MPROB=0.1).

CiaoDVD의 활용 결과는 MovieLens 보다 확률의 영향을 크게 받음을 알 수 있는데(그림 1(c)와 (d)), 이는 전자의 데이터셋 희소성이 훨씬 크므로 임의의 이웃 사용자와의 유사도값 변화가 시스템 성능에 크게 영향을 미치기 때문인 것으로 파악된다. 이러한 결과를 토대로 추후 실험은 CPROB=0.5, MPROB=0.4로 진행하였다.

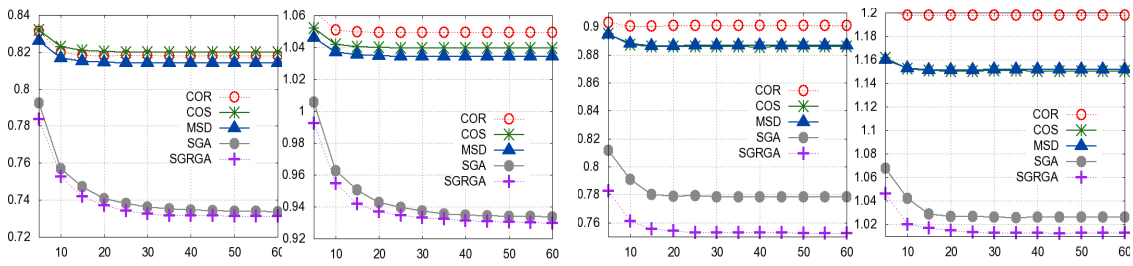
4.3.2 예측 정확도

그림 2는 인접 이웃 수의 변화에 따른 MAE와 RMSE 성능을 통한 실험 방법들의 예측 정확도를 나타낸다. 방법들 간에 일부 결과가 주목할 만한 차이를 보였는데, 두 데이터셋 모두에서 대체로 COR, COS, MSD보다 GA를 활용하여 사용자 간의 유사도를 산출하는 SGA와 SGRGA가 월등히 우수한 결과를 가져왔다. 이러한 현상은 특히 CiaoDVD에서 더욱 확연히 나타나는데, 이는 희소성이 매우 크므로 기존의 유사도 척도의 효율성은 더욱 저하되며, 최적의 유사도값 산출을 통하여 인접 이웃의 중요도를 결정 및 반영하는 SGA와 SGRGA 방법이 유



(그림 1) 유전자 알고리즘 연산 확률값에 따른 MAE 성능: MovieLens((a), (b))와 CiaoDVD((c), (d))

(Figure 1) Performance of MAE by the probability of the GA operators: MovieLens((a), (b)) and CiaoDVD((c), (d))



(a) (b) (c) (d)

(그림 2) 예측 정확도: MovieLens ((a) MAE, (b) RMSE)와 CiaoDVD ((c) MAE, (d) RMSE)
(Figure 2) Prediction accuracy: MovieLens ((a) MAE, (b) RMSE) and CiaoDVD ((c) MAE, (d) RMSE)

효함을 알 수 있다.

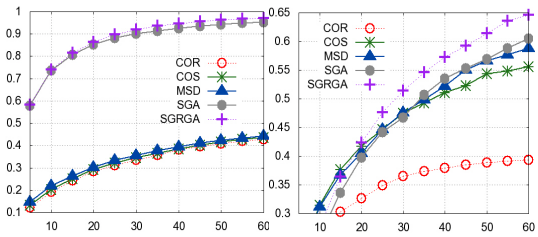
RMSE 성능 결과는 기존 척도들 간의 우열을 보다 명확하게 구분한다. 즉, MovieLens에서 COR의 성능이 가장 저조하였고 이는 CiaoDVD를 활용한 결과에서 더욱 확연하였다. 그 이유는 매우 희소한 데이터 환경에서 COR의 예측 정확도는 저하된다고 보고되었으므로 미평가 데이터를 채우기 위한 imputation-boosted 기술 등을 활용한 사전 작업이 해결 방안 중의 하나이다[1][2].

SGRGA는 모든 경우에 가장 월등한 성능을 보였는데 특히 희소 데이터 환경에서 더욱 우수하였다. 이는 다른 사용자와의 유사도가 항목 장르에 따라 다른 값을 갖는다는 것을 의미하며 이러한 사실은 희소 환경에서 더욱 확연함을 의미한다. 따라서 항목 장르별로 다른 유사도값을 반영하는 제안 방법이 성능 면에서 유리하였다. 또한, GA 적합도 함수는 MAE로 하였으나, MAE 뿐만 아니라 RMSE에 있어서도 GA를 활용한 두 방법, 즉, SGA와 SGRGA의 성능이 더 우수하였다. 그 이유는 MAE와 RMSE 모두 예측치와 실제치의 차이를 측정함으로써 서로 유사한 공식을 사용하기 때문이다.

4.3.3 커버리지

그림 3은 인접 이웃 수의 변화에 따른 커버리지 결과로서, 앞 절의 예측 정확도 결과처럼 MovieLens에서는 두 그룹으로 명확하게 구분되었다. 그룹 간에 매우 큰 성능 차이를 보이는데, 참조 이웃 수가 60일 때 약 0.544의 커버리지 차이가 발생하였다. 따라서, GA를 활용하여 유사도를 산출하는 본 제안 방법은 현 사용자의 미평가항목들에 대한 예측치 산출 가능 비율이 월등히 높으므로, 추천 리스트의 다양성이 더욱 클 것으로 예상된다.

CiaoDVD 활용한 실험에서는 COR 성능이 다른 방법들에 비해 주목할 만하게 낮았으며 MSD 성능은 이웃 수



(a) (b)

(그림 3) 커버리지: (a) MovieLens와 (b) CiaoDVD
(Figure 3) Coverage:(a) MovieLens and (b) CiaoDVD

가 증가함에 따라 COR와 COS 보다 우수하였다. SGRGA는 가장 뛰어난 커버리지를 보였는데, 특히 참조 이웃 수가 60일 때 COR와의 차이가 약 0.253로서 매우 컸다.

5. 결론 및 향후 연구과제

메모리 기반 협력 필터링의 핵심 원리는 유사한 취향을 가진 사용자 그룹을 찾아내어 그들의 선호도를 기반으로 현 사용자를 위한 추천 리스트를 생성하는 것이다. 따라서 사용자 간 유사도를 어떻게 정의하고 산출하느냐가 시스템 성능을 결정짓는 중요한 요소이다. 본 연구에서는 시스템의 최적 성능을 가져오는 사용자 간 유사도 값을 산출하기 위하여 유전자 알고리즘을 활용하였으며, 특히 평가 항목 특성별로 유전자 알고리즘을 별도로 실행하여 예측 정확도 성능을 높이고자 하였다. 성능 실험을 통하여 유전자 알고리즘 연산 확률의 적정값을 구하였고, 실험 결과로서 제안 방법의 예측 성능이 기존 방법들을 능가하고, 특히 희소 데이터 환경에서 더욱 우수함을 확인하였으므로, 항목 특성별로 유사한 이웃 사용자들을 구하는 것이 성능 면에서 유리함을 증명하였다.

유전자 알고리즘은 본 연구의 활용 목적 외에 사용자 간 유사도 산출의 가중치나 파라미터를 최적화하는 데 사용될 수 있다. 본 연구에서는 데이터셋에서 제공한 항목에 대한 장르 정보를 활용하였으나, 또다른 항목 특성을 선택하여 성능을 검토할 필요성이 있다. 이에 더하여 다양한 데이터 및 실험 환경, 데이터 밀집도, 기존의 여러 유사도 척도 등 조건을 다양화하여 연구해볼 가치가 있다.

본 연구에서는 예측 정확도를 중점으로 성능 비교를 수행하였으나, 기타 척도를 기준으로 성능 비교를 하기 위하여는 유전자 알고리즘의 적합도 함수를 그러한 기준의 척도로 설정하여 재실험하여야 하며, 이는 여러 관련 이슈를 포함하고 있는 방대한 내용으로서 추가 연구 주제 중 하나이다. 제안 방법은 관련된 기존 연구들 중에서 최적 유사도 산출을 목표로 하는 방법을 개선하는데 목표를 둔 메모리 기반 알고리즘이며, 모델 기반 또는 하이브리드 시스템 등의 다른 범주의 알고리즘을 성능 비교 대상으로 포함하지 않았다. 각 범주에 속하는 최신 알고리즘들의 성능 비교 및 장단점 분석은 별도의 연구가 필요하다. 만약 데이터셋이 항목 특성 정보를 제공하지 않거나 뚜렷한 특성의 항목들을 유지 관리하지 않는 경우에는 제안 방법의 적용이 불가하므로, 이를 해결하기 위한 별도의 연구가 필요하다. 또한 본 연구에서는 최적 유사도를 얻기 위하여 유전자 알고리즘을 활용하였는데 항목 특성별로 수행함으로써 시간 비용 소모가 클 수 있으므로, 이와 더불어 장르 개수 또는 사용자수의 증가에 따른 시스템 확장성 문제를 해결하는 것은 향후 연구 과제 중의 하나이다.

참고문헌(Reference)

[1] B. Shao, X. Li, and G. Bian, “A survey of research hotspots and frontier trends of recommendation systems from the perspective of knowledge graph,” Expert Systems with Applications, vol. 165, 2021. https://doi.org/10.1016/j.eswa.2020.113764

[2] M. Jalili, S. Ahmadian, M. Izadi, P. Moradi, and M. Salehi, “Evaluating collaborative filtering recommender algorithms: a survey,” IEEE Access, vol. 6, pp. 74003-74024, 2018. https://doi.org/10.1109/ACCESS.2018.2883742

[3] F. Fkih, “Similarity measures for collaborative filtering-based recommender systems: review and experimental comparison,” Journal of King Saud University - Computer and Information Sciences, vol. 34, no. 9, pp. 7645-7669, 2022. https://doi.org/10.1016/j.jksuci.2021.09.014

[4] H. Khojamli and J. Razmara, “Survey of similarity functions on neighborhood-based collaborative filtering,” Expert Systems with Applications, vol. 185, 2021. https://doi.org/10.1016/j.eswa.2021.115482

[5] H. Zhou, F. Xiong, and H. Chen, “A comprehensive survey of recommender systems based on deep learning,” Applied Sciences, vol. 13, no. 20: 11378, 2023. https://doi.org/10.3390/app132011378

[6] S. Bag, S.K. Kumar, and M.K. Tiwari, “An efficient recommendation generation using relevant Jaccard similarity,” Information Sciences, vol. 483, pp. 53-64, 2019. https://doi.org/10.1016/j.ins.2019.01.023

[7] A. Iftikhar, M. A. Ghazanfar, M. Ayub, Z. Mehmood and M. Maqsood, “An improved product recommendation method for collaborative filtering,” IEEE Access, vol. 8, pp. 123841-123857, 2020. https://doi.org/10.1109/ACCESS.2020.3005953

[8] A. Jain, P.K. Singh, and J. Dhar, “Multi-objective item evaluation for diverse as well as novel item recommendations,” Expert Systems with Applications, vol. 139, 2020. https://doi.org/10.1016/j.eswa.2019.112857

[9] D. Wang, Y. Yih, and M. Ventresca, “Improving neighbor-based collaborative filtering by using a hybrid similarity measurement,” Expert Systems with Applications, vol. 160, 2020. https://doi.org/10.1016/j.eswa.2020.113651

[10] B. Alhijawi and Y. Kilani, “A collaborative filtering recommender system using genetic algorithm,” Information Processing & Management, vol. 57, no. 6, 2020. https://doi.org/10.1016/j.ipm.2020.102310

[11] F. Fkih, “Enhancing item-based collaborative filtering by users’ similarities injection and low-quality data handling,” Data & Knowledge Engineering, vol.144, 2023.

https://doi.org/10.1016/j.datak.2022.102126

[12] H.I. Abdalla, Y.A. Amer, L. Nguyen, A.A. Amer, B.M. Al-Maqaleh, “Numerical similarity measures versus Jaccard for collaborative filtering,” Proceedings of the 9th Int’l Conf. Advanced Intelligent Systems and Informatics, 2023.
https://doi.org/10.1007/978-3-031-43247-7_20

[13] A. Gazdar and L. Hidri, “A new similarity measure for collaborative filtering based recommender systems,” Knowledge-Based Systems, vol. 188, 2020.
https://doi.org/10.1016/j.knosys.2019.105058

[14] Y. Mu, N. Xiao, R. Tang, L. Luo, and X. Yin, “An efficient similarity measure for collaborative filtering,” Procedia Computer Science, vol. 147, pp. 416-421, 2019.
https://doi.org/10.1016/j.procs.2019.01.258

[15] B. Alhijawi and Y. Kilani, “Using genetic algorithms for measuring the similarity values between users in collaborative filtering recommender systems,” IEEE/ACIS 15th Int’l Conf. on Computer and Information Science, 2016.
https://doi.org/10.1109/ICIS.2016.7550751

[16] Y. Ar and E. Bostanci, “A genetic algorithm solution to the collaborative filtering problem,” Expert Systems with Applications, vol. 61, pp. 122-128, 2016.
https://doi.org/10.1016/j.eswa.2016.05.021

[17] A. Laishram and V. Padmanabhan, “Discovery of user-item subgroups via genetic algorithm for effective prediction of ratings in collaborative filtering,” Applied Intelligence, vol. 49, pp. 3990-4006, 2019.
https://doi.org/10.1007/s10489-019-01495-4

[18] Z. Liu, L. Wang, X. Li, and S. Pang, “A multi-attribute personalized recommendation method for manufacturing service composition with combining collaborative filtering and genetic algorithm,” J. of Manufacturing Systems, vol. 58, pp. 348-364, 2021.
https://doi.org/10.1016/j.jmsy.2020.12.019

[19] F.H. Nanehkaran, S.M. Lajevardi, and M.M. Bidgholi, “Optimization of fuzzy similarity by genetic algorithm in user-based collaborative filtering recommender systems,” Expert Systems, vol. 39, no. 4, 2022.
https://doi.org/10.1111/exsy.12893

[20] B.S. Neysiani, N. Soltani, R. Mofidi, and M.H. Nadimi-Shahraki, “Improving performance of association rule-based collaborative filtering recommendation systems using genetic algorithm,” Int’l J. of Information Technology and Computer Science, vol. 11, no. 2, 2019.
https://doi.org/10.5815/ijitcs.2019.02.06

[21] M. Salehi, I.N. Kamalabadi, and M.B. Ghaznavi-Ghoushchi, “Attribute-based collaborative filtering using genetic algorithm and weighted C-means algorithm,” International Journal of Business Information Systems, vol. 13, no. 3, pp. 265-283, 2013.
https://doi.org/10.1504/IJBIS.2013.054465

◐ 저 자 소 개 ◑

이 수 정(Soojung Lee)
1985년 이화여자대학교 수학교육과 (이학사)
1990년 미국 Texas A&M 대학교 컴퓨터과학과 (공학석사)
1994년 미국 Texas A&M 대학교 컴퓨터과학과 (공학박사)
1994년～1998년 삼성전자 통신개발실 선임연구원
1998년～현재 경인교육대학교 컴퓨터교육학과 교수
관심분야 : 정보필터링, 추천시스템, 컴퓨터교육
E-mail : sjlee@gin.ac.kr