JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 29 No. 11, pp. 243-250, November 2024
https://doi.org/10.9708/jksci.2024.29.11.243

Optimizing Similarity for User-based Collaborative Filtering

Soojung Lee*

*Professor, Dept. of Computer Education, Gyeongin National University of Education, Anyang, Korea

[Abstract]

Collaborative filtering is one of the most widely known implementation methods of recommender systems, which recommends items that similar users have preferred in the past. Therefore, similarity measurement is a very important factor that determines the performance of the system. In this study, in order to solve the shortcomings of the existing single or integrated heuristic similarity measures, the genetic algorithm was used to calculate the optimal similarity between users per item genre. In addition, in order to solve the data scalability problem, the number of users for calculating similarity for each genre was limited according to a preset threshold, and the average of the ratings of the items was used to solve the data sparsity problem. Through performance experiments, the optimal probabilities of the genetic operators were obtained and the prediction accuracy performance was analyzed. As a result, it was confirmed that the performance of the proposed method was superior to the existing methods, especially in a sparse data environment.

▸ **Key words**: Similarity Measure, Genetic Algorithm, Data Sparsity Problem, Data Scalability Problem, Collaborative Filtering, Recommender System

[요 약]

협력 필터링은 가장 널리 알려진 추천 시스템의 구현 방식들 중의 하나로서, 유사한 사용자들이 과거에 선호하였던 항목들을 추천한다. 따라서 유사도 측정은 시스템의 성능을 좌우하는 매우 중요한 요소이다. 본 연구에서는 기존의 단일의 또는 통합된 휴리스틱 유사도 척도의 단점을 해결하기 위한 목적으로, 유전 알고리즘을 활용하여 항목 장르별로 사용자 간 최적의 유사도를 산출하였다. 또한 데이터 확장성 문제를 해결하기 위하여 미리 설정한 임계치에 따라 각 장르별 유사도 산출 대상 사용자 수를 제한하였고, 데이터 희소성 문제 해결을 위하여 항목의 평가치 평균을 활용하였다. 성능 실험을 통하여 유전 연산자 확률의 적정값을 구하였고, 예측 성능을 분석한 결과, 제안 방법의 성능이 기존보다 우수하고, 특히 희소 데이터 환경에서 더욱 우수함을 확인하였다.

▸ **주제어**: 유사도 척도, 유전 알고리즘, 데이터 희소성 문제, 데이터 확장성 문제, 협력 필터링, 추천 시스템

• First Author: Soojung Lee, Corresponding Author: Soojung Lee
*Soojung Lee (sjlee@gin.ac.kr), Dept. of Computer Education, Gyeongin National University of Education
• Received: 2024. 10. 04, Revised: 2024. 10. 31, Accepted: 2024. 11. 01.

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

I. Introduction

추천 시스템은 사용자의 개인화된 경험을 제공하기 위해 다양한 알고리즘과 기술을 활용하는 필수 도구로 자리 잡고 있다. 이러한 시스템은 정보 과잉 시대에서 특히 중요한 역할을 하며, 사용자에게 맞춤형 콘텐츠를 추천하여 만족도를 높이고, 지속적인 사용을 유도하는 데 기여한다.

추천 시스템의 설계 및 구현에서 기본이 되는 두 가지 접근 방식은 협력 필터링(collaborative filtering, CF)과 내용 기반 필터링이다(content-based filtering, CBF)[1][2]. CF는 사용자의 과거 행동과 선호도를 분석하여 유사한 취향을 가진 다른 사용자들이 선호하는 항목을 추천하는 방식이다. 이 방법의 주요 장점은 사용자의 경험에 대한 직관적인 이해를 가능하게 하며, 즉각적인 추천을 제공하는 데 있다. 그러나 데이터 희소성 문제와 대규모 사용자 집합에 대한 스케일링의 한계에 직면해 있다. 특히 신규 사용자나 항목이 추가될 경우 초기 데이터 부족으로 인해 유사도 계산이 어려워지는 문제가 있다[3].

반면, CBF는 추천하려는 항목의 속성과 사용자의 과거 선호도를 기반으로 추천을 생성한다[1]. 이 방식은 사용자에게 개인화된 추천을 제공할 수 있지만, 이전과 전혀 다른 새로운 내용의 추천이 어렵다는 단점이 있다. 즉, 사용자가 이전에 경험해 보지 않은 장르의 항목은 추천 리스트에서 제외될 가능성이 크다. 이러한 특성 때문에 내용 기반 접근은 종종 추천의 다양성이 부족할 수 있으며, 이는 사용자 경험을 저해하는 요인이 된다.

하이브리드 필터링은 이러한 두 접근 방식을 결합하여 서로의 단점을 보완하는 방법이다. 이 방법은 CF와 CBF의 장점을 활용하여 추천의 정확성과 다양성을 향상시킬 수 있다[4]. 그러나 하이브리드 시스템 또한 구현의 복잡성과 계산 비용 증가라는 새로운 문제에 직면할 수 있다.

최근 CF 연구는 딥러닝 기술과 같은 인공지능의 도입으로 급속히 발전하고 있다. 특히 신경망 기반 방법은 사용자와 항목 간의 복잡한 상호작용을 효과적으로 모델링할 수 있는 능력을 갖고 있다[5]. 이러한 연구는 추천 시스템이 사용자 행동을 정교하게 이해할 수 있도록 지원하지만, 모델이 산출한 추천 결과의 해석이 어렵다는 단점이 있다. 또한 소셜 네트워크와 같은 외부 데이터를 활용하는 연구도 활발하며[6], 이는 사용자 간의 관계와 상호작용을 고려하여 추천의 질을 향상시키는 데 기여한다.

본 연구에서는 메모리 기반 CF 시스템을 위한 사용자 간 최적의 유사도를 산출한다. 특히, 시스템에서 제공하는 항목 장르 정보를 활용하여 각 장르별로 별도의 유전 알고리즘(genetic algorithm, GA)을 수행함으로써 사용자 간 유사도를 산출한다. 이러한 방법은 유사도 측정을 위해 대개 휴리스틱 또는 이들의 통합 방식을 제안한 기존의 방법들과 대비된다. 또한 본 연구에서는 데이터 확장성 문제를 해결하기 위해 미리 설정한 임계치에 따라 각 장르별 유사도 산출 대상 사용자 수를 제한하는 방안을 제시하였고, 데이터 희소성 문제를 해결하기 위해 항목의 평가치 평균을 사용하였다. 두 종류의 공개 데이터셋을 이용한 성능 평가 결과, 제안 방법은 예측 성능 면에서 우수하였는데, 특히 희소 데이터셋에서 기존 방법들을 크게 능가하였다.

논문의 나머지 구성은 다음과 같다. 2절에서는 본 주제와 연관된 기존 연구 성과를 소개한다. 3절에서는 제안 방법을 설명하고, 4절에서는 성능 실험 결과를 제시하며, 5절에서는 논문의 결론을 맺는다.

II. Related Works

메모리 기반 CF에서 유사도 측정은 현 사용자의 인접 이웃들을 구하고 이들로부터 추천 리스트를 얻기 위한 절차로서 시스템 성능에 큰 영향을 미치므로 이 분야의 핵심 연구 주제 중 하나이다. 사용자-항목 행렬(user-item matrix)을 기반으로 평가치에 대한 사용자 간 또는 항목 간의 유사성을 계산한다. 메모리 기반 방법은 모델 기반 방법의 단점인 막대한 계산 비용과 과적합 문제 등을 해소하며, 매우 간단하고 직관적이라는 특징이 있다[2][7].

전통적인 유사도 산출 방법으로서 피어슨 상관(Pearson Correlation), 코사인 유사도(Cosine Similarity), 평균자승차이(Mean Squared Difference), 자카드 유사도(Jaccard Similarity), 유클리드 거리(Euclidean Distance) 등이 사용되었으나, 이들은 주로 공통으로 평가한 항목 수에 의존한 계수로서, 희소한 데이터 환경에서 추천 성능이 저하될 수 있다. 이에 따라 이들을 융합하고 개선한 방법들도 개발되었다[1][3][7][8]. 최근 Abdalla 외 4인은 기존 방법을 단일로 사용하는 경우의 문제를 해결하기 위해 자카드 계수와 다른 유사도 방법들을 통합한 방법을 제안하였다[9]. 자카드 유사도의 여러 변형 방식도 개발되었는데 Bag 외 2인은 관련 자카드 유사도를 제안하였고, 이는 희소 데이터 환경에서 사용자의 모든 평가 벡터를 고려한 방법이다[10].

진화 알고리즘은 메타휴리스틱 기법으로, 초기 해의 진화 과정을 통해 점진적으로 최적해를 얻는 방법으로 CF 연구에 도입되었다[11][12]. 이들 중 유전 알고리즘

(Genetic Algorithm, GA)은 CF 시스템에 가장 많이 활용된 기법 중 하나이다. 피어슨 상관과 벡터 코사인 유사도와 같은 잘 알려진 유사도 척도를 사용하지 않고 사용자 간의 최적의 유사도 값을 계산하기 위해 유전 알고리즘을 활용한 연구 방안이 제안되었다[13][14]. 특히 [13]에서는 제안 방법의 예측 품질과 성능이 기존에 비해 30~40% 향상되었다고 보고하였다. [15]에서는 사용자 간의 유사도 값을 미평가 항목에 대한 예측치 산출에 활용하기 전에 이를 정제하기 위한 유전 알고리즘 방식을 제시하였고, 다양한 수의 이웃에 대해 수행한 통계 분석 결과, 전통적인 유사도 척도들에 비해 예측 오류를 상당히 감소시킨 것으로 확인되었다. 한편, 여러 다른 적합도 함수들을 사용하여 항목 자체가 아닌 추천 목록을 계층적으로 평가하여 최적의 추천 목록을 발견하기 위해 유전 알고리즘을 활용한 방안도 제시되었는데[16], 구체적으로 적합도 함수는 항목 간의 의미적 유사성의 강도, 사용자 간의 만족도 유사성, 그리고 예측 평가치와 관계된 필터링 레벨로 제안되었다.

[17]에서는 최상의 항목 리스트를 추천하기 위해 퍼지-유전적 협업 필터링 접근법을 제시하고 연속 유전 알고리즘에서 퍼지 유사도를 최적화하며 퍼지 예측에 사용함으로써 항목을 추천하였다. 이밖에 유전 알고리즘을 유사도 최적값을 구하는 것이 아닌 다른 목적으로 활용한 연구도 시도되었는데, [18]의 연구에서는 시스템 확장성 문제를 언급하고 최적의 사용자 클러스터링을 형성하기 위해 유전 알고리즘을 활용하였다. 또한 [19]의 연구에서는 유사한 사용자 집합을 기반으로 상관관계가 있는 항목의 하위 그룹을 형성하여 관련 항목에 대한 예측만을 얻는 기술을 제안하였다. 이는 각 하위 그룹의 사용자가 하위 그룹에 포함된 항목 하위 집합에 대한 선호도가 비슷하다는 주장에 근거하여, 미평가 항목에 대한 평가를 예측하기 위해 상관관계가 높은 사용자-항목 하위 그룹을 선택하는 다양한 방법을 탐색하였다. 유사도 전이(similarity transitivity) 개념의 실현을 위하여 유전 알고리즘을 활용한 연구도 실행되었는데, [20]에서는 클러스터링 방법을 통해 부정확한 유사성을 필터링한 후 적절한 교차 임계값을 활용하여 이를 전이 유사성으로 대체함으로써, 최적화 및 검색 문제에 대한 유용한 솔루션을 생성하기 위해 유전 알고리즘 기반 유사성 전이를 제안하였다. [21]에서는 기존의 연관 규칙과 입자군 최적화(particle swarm optimization)와 같은 진화 알고리즘들의 성능 및 정확도를 향상시키기 위해 유전 알고리즘을 기반으로 더 높은 성능의 신뢰 연관 규칙을 생성하는 방법을 제안하였다. 이와 같이 유전 알고리즘은 다양한 방식으로 메모리 기반 CF 시스템에서 활용되었으나, 그 연구 결과는 대체로 많지 않으며, 유사도의 최적화를 위해 사용된 예는 매우 드문 것으로 파악되었다.

III. Proposed Methodology

본 연구의 주요 아이디어는 전통적인 휴리스틱 방식에서 탈피하여 시스템 성능을 극대화할 수 있는 사용자 간 유사도 값의 최적화를 추구하는 것이다. 이를 위해 유전 알고리즘을 채택하여 활용하며, [13]의 방법처럼 사용자 간 포괄적인 최적의 유사도 산출 시도를 향상시키는 새로운 방식을 제안한다. 즉, 사용자 간의 유사도 최적값은 여러 가지 요인 및 상황에 따라 달라질 수 있다는 가정을 두며, 그 요인들 중 하나로 항목의 장르를 채택한다. 다시 말해, 항목의 장르별로 사용자 간 유사도는 다를 수 있으며, 예를 들어 로맨틱 영화에 대해서는 두 사용자가 매우 유사한 평가 시각을 가질 수 있지만, 공포 영화에 대해서는 이들 간의 유사도가 다를 수 있음을 가정한다. 물론 이러한 아이디어를 구현하기 위해 시스템에서는 사용자 ID, 항목 ID, 평가치, 그리고 항목의 장르 정보를 제공해야 한다.

CF 시스템의 원리에 따라 현 사용자가 미평가한 항목들에 대한 예측치를 구한 후, 시스템은 예측치가 높은 항목들 중심으로 추천 항목 리스트를 제공해야 한다. 본 연구에서 예측치를 산정하는 구체적인 절차는 다음과 같다.

1. 각 장르별로 해당 장르에 속한 항목을 평가한 이력이 있는 사용자들을 구한다.
2. 1단계에서 구한 사용자의 장르 평가 빈도(Genre Rating Frequency, GRF)가 임계치 이상인 사용자들을 추출한다.
3. 2단계에서 추출한 사용자들 간의 유사도를 GA를 활용하여 산출하고, 각 장르별로 최적화된 유사도값을 얻는다.
4. 현 사용자 u가 미평가한 항목 x가 장르 g에 속한다고 할 때, 이 장르에서의 x에 대한 예측치는 다음 식으로 산출한다.

$$\bar{r}_u + \frac{\sum_v sim_g(u,v)(r_{v,x} - \bar{r}_v)}{\sum_v |sim_g(u,v)|}$$

위 식에서 $sim_g(u,v)$는 3단계에서 산출한 장르 g에서의 두 사용자 u와 v 간의 최적 유사도이며, $\bar{r}_u$와 $r_{v,x}$는

각각 u의 평균 평가치와 v의 x에 대한 평가치이다. 만약 항목 x가 둘 이상의 장르에 속한다면 이들 장르에서의 유사도를 모두 고려하여 다음 식과 같이 예측치를 산출한다.

$$\bar{r}_u + \frac{\sum_g \sum_v sim_g(u,v)(r_{v,x} - \bar{r}_v)}{\sum_g \sum_v |sim_g(u,v)|}$$

또한 만약 항목 x를 평가한 이웃이 부재할 경우, 위 식으로는 예측치를 산출할 수 없으므로, x를 평가한 모든 사용자들의 평가치 평균값을 예측치로 결정한다.

사용자의 GRF는 사용자의 해당 장르에 대한 선호도를 나타내는 지수로서, 만약 장르 g에 대한 GRF가 1이라면, 이 사용자가 평가한 모든 항목이 장르 g에 속했음을 의미한다. 반대로, GRF가 0에 가까울수록 해당 장르의 선호도는 점점 더 낮아진다. GRF 임계치는 모든 장르에 대해 동일한 값으로 설정하며, 이 값이 크면 2단계에서 각 장르별로 추출되는 사용자 수가 적어지므로, 3단계에서 GA 계산 복잡도가 낮아져 데이터 확장성 문제 해결에 유리하지만, 인접 이웃 수가 감소하므로 정확도 성능이 저하될 수 있다. 적절한 GRF 임계치의 설정은 시스템 성능에 중요한 요소이며, 본 연구에서는 실험을 통해 적정값을 정하였다.

GA의 각 해는 모든 두 사용자 간의 유사도로 구성되며, 이는 사용자수×사용자수 개의 0과 1 사이의 실수값으로 초기화한다. 이러한 해들의 집합에 대해 세 가지 유전 연산(genetic operators), 즉, 선택(selection), 교차(crossover), 변이(mutation)를 적용하였다. 알고리즘은 종료 조건이 만족될 때까지 반복 수행되는데, 목표로 하는 최적의 해를 발견하거나 특정 반복 실행 회수에 도달하면 종료된다. 적합도 함수(fitness function)로는 CF 연구의 예측 정확도 척도로 널리 사용되는 평균 절대 오차(Mean Absolute Error, MAE)를 채택하였다.

각 반복 회차마다 알고리즘은 유전 연산을 통해 새로운 세대를 만들어낸다. 구체적으로, 먼저 적합도 값에 따른 확률을 적용하여 현재 세대에서 두 개의 해를 선택하고, 교차 확률값에 따른 교차 연산을 수행함으로써 두 개의 새로운 자손 해를 생성한다. 이들 각 자손 해에 대해서는 변이 확률값에 따른 변이 연산을 수행한다. 이러한 과정을 반복하여 생성된 자손 해의 개수가 기존 부모 세대의 개수와 동일하게 되면 한 세대에 대한 알고리즘 실행이 완료된다. 이와 같이 생성한 새로운 자손 세대가 특정 차수에 이르게 되면 전체 알고리즘을 종료한다.

IV. Performance Experiments

1. Experimental Background

제안 방법의 성능 평가를 위해 사용자와 항목 ID, 항목 평가치, 장르 정보를 포함하는 공개 연구용 데이터셋을 선정하였다. 관련 연구에서 널리 활용되어 온 MovieLens(https://movielens.org) 1M과 CiaoDVD 데이터셋(https://dvd.ciao.co.uk)은 이들 정보를 모두 포함하므로 제안 방법의 실험 조건에 적합하여 본 실험에 활용하였다.

MovieLens 1M은 원래 6040명의 사용자와 이들의 항목 평가치 정보를 포함하고 있으며, CiaoDVD는 17615명의 사용자들의 평가 관련 데이터를 포함한다. 본 연구의 실험은 Intel Core i5의 윈도우 운영체제에서 진행되었으며, 유전 알고리즘을 사용한 제안 방법의 성능 결과를 제한된 시간 내에 얻기에는 컴퓨터 용량 및 처리 속도에 한계가 있어, 원래의 규모를 축소하여 각 데이터셋 당 임의의 1000명의 사용자와 이들의 평가 데이터를 추출하여 실험하였다. 이와 유사하게, 기존의 여러 관련 연구에서도 1000명 미만의 사용자들을 포함한 데이터로 GA 활용 CF 성능 실험을 수행하였다[13][15][17][19][22]. 표 1은 실험에 사용한 각 데이터셋의 특징을 나타낸다. 이들 간의 희소성 수준 차이가 작지 않으므로, 실험 방법들의 성능 비교를 다양한 데이터 환경에서 점검할 수 있다.

Table 1. Description of datasets used for experiments

	MovieLens	CiaoDVD
No. of users	1000	1000
No. of items	3952	16121
Rating range	1~5, integer	1~5, integer
Sparsity level	0.96099	0.99877
No. of genres	18	17

실험 방법은 다음과 같다. 전체 데이터 중 80%를 훈련용 데이터로, 나머지 20%를 테스트용 데이터로 설정하였다. 훈련 데이터를 이용하여 사용자 간 유사도를 측정하고, 나머지 데이터를 현 사용자가 미평가한 항목의 데이터로 간주하여 시스템에서 예측치를 구한다. 이를 위해 현 사용자와의 유사도값 순으로 인접 이웃들을 구하고, 이들로부터 미평가 항목에 대한 평가치의 가중합을 계산한다. 따라서 예측치와 테스트 데이터에 포함된 실제 평가치 간의 차이가 작을수록 시스템의 예측 성능이 우수하다고 판단할 수 있다. 즉, 구해진 인접 이웃이 적절하고 유사도 측정 방법이 더욱 정확하다는 것을 의미한다.

본 실험에서는 CF 연구에서 주로 사용되는 예측 성능 평가 척도로서, MAE(Mean Absolute Error, 평균절대오차), RMSE(Root Mean Square Error, 평균 제곱근 오차), Coverage(커버리지)를 도입하였다[1][3][7]. 이는 제안 방법의 GA에서 적합도 함수로 MAE를 사용하여 예측 성능의 최적화를 목표로 하였기 때문이다.

본 연구의 주제는 최적의 유사도 값을 산출하는 것이므로, 성능 비교 대상으로 전통적인 대표적 유사도 척도들과 이를 개선한 방법들을 선정하였다. 구체적으로, 피어슨 상관도(COR), 코사인 유사도(COS), 평균자승차이(MSD), 그리고 GA를 활용하여 최적 유사도를 구하는 [13]의 방법(SGA)을 포함하였다. 추가로 데이터 희소성 문제를 해결하기 위해 개발된 Jaccard 계수(JAC)[23], Jaccard와 MSD를 결합한 방법(JMSD)[10]의 성능 실험도 진행하였다. 본 논문의 제안 방법은 SGA_GR로 표기하였다.

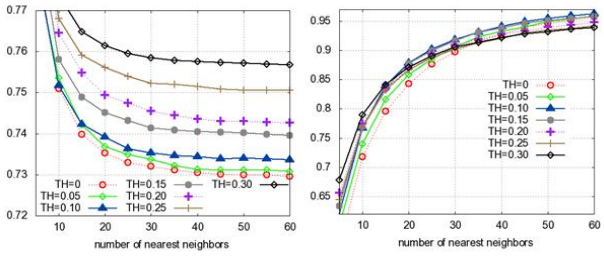
2. Results of Experiments

2.1 Effect of Genre Rating Frequency Threshold

본 제안 방법에서는 사용자의 특정 장르에 대한 평가 빈도(GRF)가 낮으면 해당 사용자를 그 장르에 대한 임의의 인접 이웃 리스트에서 배제하므로, GRF 임계치의 적정값 설정은 시스템 성능에 중요한 역할을 한다. 따라서, 임계치의 변화에 따른 성능을 점검하는 실험을 진행하였으며, 그 결과를 그림 1과 2에 나타냈다.

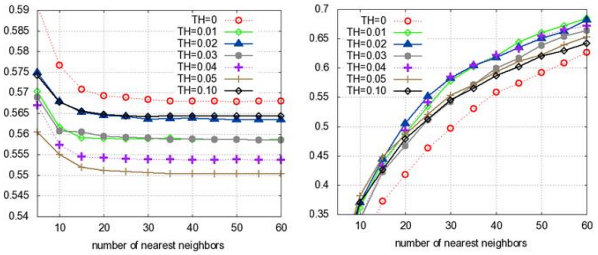
그림 1은 MovieLens를 활용한 다양한 GRF 임계치에 따른 성능 결과를 나타낸다. 임계치가 0인 경우(TH=0)에는 가장 좋은 MAE를 보였다. 이는 사용자 간 유사도 계산 시 제외되는 사용자가 없기 때문이며, 동일한 이유로 임계치가 커질수록 예측 성능은 낮아졌다. 커버리지 결과는 MAE와 다소 차이를 보이는데, 이는 임계치가 작을수록 그에 따라 결정되는 인접 이웃들이 반드시 현 사용자의 미평가 항목을 평가한다는 보장이 없기 때문이다. 이러한 실험 결과를 바탕으로 본 연구의 실험에서는 MovieLens에 대한 GRF 임계치를 0.05로 설정하였다.

그림 2는 CDVD를 활용하여 GRF 임계치의 성능 영향을 조사한 결과를 나타낸다. TH=0일 때 가장 저조한 MAE와 커버리지를 보였으며, 제안 방법에서는 미평가 항목을 평가한 인접 이웃이 없는 경우 해당 항목의 평균 평가치를 예측치로 설정하기 때문에, 희소 데이터 환경에서 GRF 임계치가 0보다 클 때 더 우수한 성능을 나타내는 것으로 판단된다. 이러한 결과를 바탕으로 GRF 임계치를 0.05로 설정하였다.



(a) (b)

Fig. 1. (a) MAE and (b) coverage using MovieLens



(a) (b)

Fig. 2. (a) MAE and (b) coverage using CiaoDVD

표 2는 데이터셋의 각 장르별로 GRF 임계치에 따라 포함되는 사용자 수의 비율을 나타낸다. 예를 들어, MovieLens에서 표 2의 장르 1(gr1)은 임계치가 0.05일 때 0.97로, 이는 장르 1에 속한 항목 평가 비율이 5% 이상인 사용자가 전체의 97%임을 의미한다. 임계치가 클수록 사용자 수의 비율은 감소하므로, 유사도 계산에 필요한 비용도 줄어든다. 본 실험에서 설정한 임계치인 0.05의 경우, MovieLens는 평균적으로 55%, CiaoDVD는 21%의 사용자들만을 대상으로 유사도를 산출하면 되므로 시스템 확장성에 큰 이익이 될 수 있다.

2.2 Effect of Genetic Operator Probability

이 절에서는 본 연구의 GA에서 도입한 연산의 확률값에 따른 CF 성능의 변화를 살펴보았다. 그림 3은

Table 2. Ratio of users for varying GRF threhold by genre(gr)

Dataset	GRF Threshold	gr1	gr2	gr3	gr4	gr5	gr6	gr7	gr8	gr9	gr10	gr11	gr12	gr13	gr14	gr15	gr16	gr17	gr18	avg.
MovieLens	0.05	0.97	0.81	0.26	0.49	0.99	0.71	0.02	1.00	0.26	0.08	0.44	0.25	0.20	0.93	0.82	0.92	0.70	0.08	0.55
	0.10	0.86	0.55	0.08	0.20	0.98	0.31	0.01	0.98	0.07	0.02	0.17	0.10	0.05	0.70	0.60	0.78	0.29	0.02	0.38
CiaoDVD	0.01	0.82	0.05	0.08	0.16	0.07	0.01	0.02	0.39	0.14	0.00	0.32	0.17	0.67	0.55	0.26	0.50	0.30		0.25
	0.05	0.78	0.03	0.02	0.08	0.03	0.00	0.01	0.34	0.07	0.00	0.26	0.09	0.66	0.54	0.18	0.48	0.24		0.21

MovieLens를 활용한 경우, 인접 이웃 수에 따른 MAE 성능 결과를 나타낸다. 각 교차 연산과 변이 연산의 확률값이 큰 MAE 차이를 보이지는 않았으나, 가장 우수한 성능을 보인 교차 연산 확률 0.6과 변이 연산 확률 0.2를 채택하여 이후 실험을 진행하였다. CiaoDVD를 활용한 실험에서는 유전자 연산 확률값의 차이가 가져오는 성능 차이가 미미하게 나타났으므로, MovieLens에서와 동일한 확률값을 사용하였다.

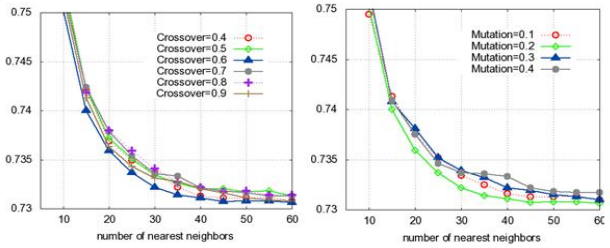


Fig. 3. MAE for varying genetic operator probability using MovieLens

2.3 Comparison of Performance Results

그림 4는 MovieLens를 활용하여 각 방법의 예측 성능을 조사한 결과를 나타낸다. MAE와 RMSE 결과는 매우 유사하며, GA를 활용한 SGA와 SGA_GR은 다른 방법들에 비해 확연히 우수한 성능을 보인다. 이는 유사도 최적화 기법이 매우 효과적임을 보여준다. 특히, 미세한 차이가 있지만 SGA_GR이 약간 더 우수한 결과를 나타낸 것으로 보이며, 이는 인접 이웃이 부재할 경우 제안 방법의 처리 절차 덕분인 것으로 판단된다.

희소 데이터 문제를 해결하기 위한 방안인 JAC과 JMSD는 기존 방법들의 성능을 매우 약간 개선하였다. 이를 통해 본 실험에서 MovieLens의 데이터 희소성 문제는 심각하지 않은 것으로 판단된다. 주목할 점은 JAC이 공통 항목 개수만을 기준으로 유사도를 결정함에도 불구하고 기존 척도들과 견줄 만한 성능을 보인다는 것이다. 기존 방법들 중에서 COR의 RMSE 성능은 가장 저조하였으며, 이는 COR이 예측치 오차의 차이가 크다는 사실을 보여준다.

그림 4에서 커버리지 결과는 MAE 또는 RMSE와는 다소 다른 양상을 보였다. SGA와 SGA_GR은 거의 가장 우수한 축에 속하였으며, JAC과 JMSD 또한 가장 뛰어난 성과를 보였다. 반면에 전통적 기법들은 상대적으로 크게 낮은 성능을 보였는데, 특히 COS는 가장 저조한 결과를 나타내어 현 사용자의 미평가 항목을 평가한 인접 이웃들이 부재한 경우가 많음을 시사한다.

그림 5는 CiaoDVD를 활용한 성능 결과를 나타낸다. MovieLens를 활용했을 때보다 방법들 간의 성능 차이가 매우 컸다. 이는 CiaoDVD가 더욱 희소한 데이터셋이며, 이러한 특성에 따른 영향이 다르게 나타났기 때문으로 보인다. MAE 결과를 살펴보면, COR는 희소 데이터의 영향을 가장 크게 받았고, JAC과 JMSD 또한 예상과는 달리 영향을 크게 극복하지 못하였다. 반면에 SGA는 이들보다 월등히 우수한 성능을 보였으며, 이는 희소 데이터 환경에서 유사도 최적화 작업이 큰 효과를 발휘함을 시사한다. 그러나, SGA_GR의 성능은 다른 방법들과 비교하여 매우 차별화된 우수성을 보였다. 이는 희소 데이터 환경에서 현 사용자의 미평가 항목을 평가한 인접 이웃이 부재할 경우가 많고, 이에 대한 제안 방법의 처리 방안이 매우 효과적임을 나타낸다. RMSE 결과에서는 MAE와는 달리 JAC과 JMSD가 매우 우수한 성과를 보였으며, 특히 SGA 보다도 더 나은 성능을 나타내었다. 이는 예측치의 오차 크기가 전통적 방법들보다 크지 않음을 의미한다.

그림 5에서 커버리지는 MAE 결과와 유사한 패턴을 보이며, 가장 낮은 성능의 COR, 가장 우수한 SGA_GR, 그리고 나머지 방법들로 구분할 수 있다. 특히 주목할 점은 SGA의 결과가 전통적 방법들과 거의 대등한 성과를 보인 것이다. 이는 최적화된 유사도 값을 통해 최적의 인접 이웃들을 도출하였음에도 불구하고, 현 사용자의 미평가 항목을 평가한 인접 이웃이 부재한 경우가 많음을 시사한다.

결론적으로, 제안 방법은 항목 장르별로 최적화된 인접 이웃들을 구함으로써 다양한 측면에서 예측 성능을 크게

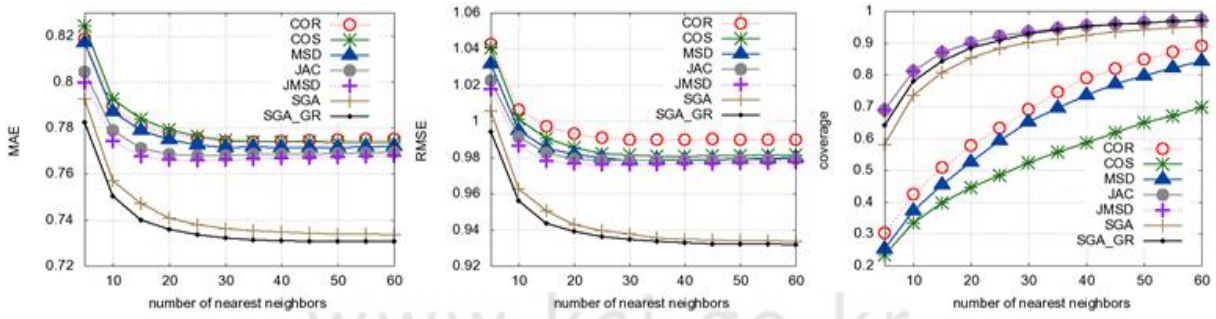


Fig. 4. Performance results using MovieLens

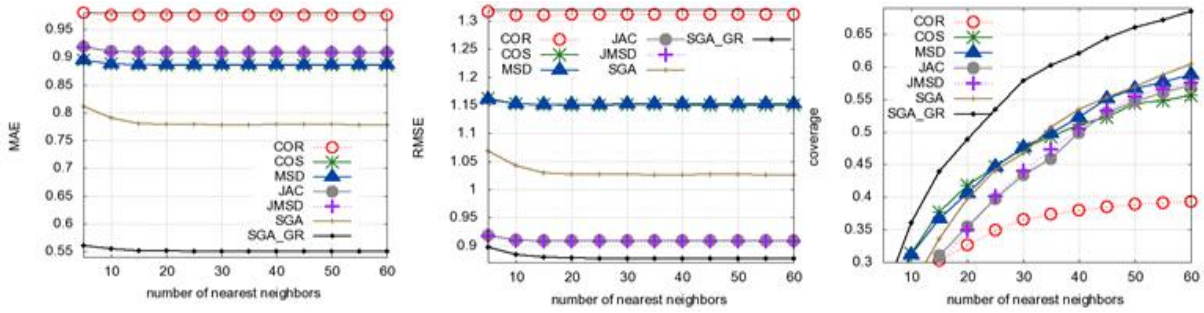


Fig. 5. Performance results using CDVD

향상시켰다. 특히 희소한 데이터 환경에서의 개선 효과가 두드러지며, 커버리지 성능이 뛰어나기 때문에 추천 리스트의 다양성도 더욱 높아질 것으로 기대된다.

V. Conclusions

추천 시스템은 다양한 알고리즘과 접근 방식을 통해 사용자에게 맞춤형 선호 항목들을 제공하는 중요한 도구이다. 메모리 기반의 협력 필터링 기법에서 사용자 간의 정확한 유사도 측정을 위해 많은 연구가 진행되었으나, 다양하고 복잡한 데이터 환경에서 시스템 성능에 제약을 가져왔다. 본 논문에서는 기존의 휴리스틱 방식에서 벗어나 최적의 유사도 산출을 목표로 유전 알고리즘을 활용하였으며, 사용자 간의 유사도는 항목 장르별로 다를 수 있다는 가정하에 각 장르별로 최적 유사도를 산출하였다. 또한, 각 장르별 계산 복잡도를 줄이기 위한 임계치 기반의 사용자 수 제한 방안을 제시하고, 데이터 희소성 문제를 해결하기 위해 항목의 평균 평가치를 예측치로 활용하였다. 두 종류의 공개 데이터셋을 활용한 성능 실험 결과, 제안 방법의 예측 성능은 기존 방법들을 크게 능가하였으므로 추천의 질을 향상시키는 데 기여할 것으로 기대된다.

본 연구는 기존의 협력 필터링 기법에 대한 새로운 시각을 제공하며, 장르별 유사도 측정의 중요성을 강조함으로써 향후 연구에 있어 더 다양한 데이터 환경을 고려한 모델링의 필요성을 제기한다. 또한 제안된 임계치 기반 사용자 수 제한 방안은 데이터 처리의 효율성을 높여 추천 시스템의 실제 적용 가능성을 크게 향상시킬 수 있다. 이로 인해 기업은 사용자 경험을 개선하고 더욱 개인화된 서비스 제공을 통해 고객 만족도를 증가시킬 수 있을 것이다.

향후 연구에서는 보다 다양한 데이터 환경에서 제안 방법의 성능을 검토하고, 항목 장르 정보를 제공하지 않는 시스템에서의 구현 기술에 대해서도 연구할 계획이다. 또한, 예측 성능 뿐만 아니라 다양한 척도를 기준으로 성능 비교를 실시하여 각 실험 방법의 장단점 및 특성을 파악할 필요가 있다. 이번 실험에서는 메모리 기반 기법들만을 대상으로 하였으나, 기타 모델 기반 알고리즘과의 비교 분석도 향후 연구 과제 중 하나로 고려하고 있다.

REFERENCES

[1] 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. DOI: 10.1109/ACCESS.2018.2883742

[2] 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. DOI:10.1016/j.eswa.2020.113764

[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. DOI: 10.1016/j.jksuci.2021.09.014

[4] B. Walek and P. Fajmon, “A Hybrid Recommender System for an Online Store using a Fuzzy Expert System,” Expert Systems with Applications, Vol. 212, 2023. DOI:10.1016/j.eswa.2022.118565

[5] H. Zhou, F. Xiong, and H. Chen, “A Comprehensive Survey of Recommender Systems Based on Deep Learning,” Applied Sciences, Vol. 13, No. 20, 2023. DOI:10.3390/app132011378

[6] R. Chen, Q. Hua, Y. -S. Chang, B. Wang, L. Zhang and X. Kong, "A Survey of Collaborative Filtering-based Recommender Systems: From Traditional Methods to Hybrid Methods based on Social Networks," IEEE Access, Vol. 6, pp. 64301-64320, 2018. DOI: 10.1109/ACCESS.2018.2877208

[7] H. Khojamli and J. Razmara, “Survey of Similarity Functions on Neighborhood-based Collaborative Filtering,” Expert Systems with Applications, Vol. 185, 2021, Article Number 115482, DOI: 10.1016/j.eswa.2021.115482

[8] A. A. Amer, and L. Nguyen, “Combinations of Jaccard with Numerical Measures for Collaborative Filtering Enhancement: Current Work and Future Proposal,” ArXiv. /abs/2111.12202, 2021. DOI: 10.48550/arXiv.2111.12202

[9] H.I. Abdalla, Y.A. Amer, L. Nguyen, A.A. Amer, and 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. DOI: 10.1007/978-3-031-43247-7_20

[10] 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. DOI: 10.1016/j.ins.2019.01.023

[11] B. Alhijawi and A. Awajan, “Genetic Algorithms: Theory, Genetic Operators, Solutions, and Applications,” Evolutionary Intelligence, Vol. 17, pp. 1245-1256, 2024. DOI: 10.1007/s12065-023-00822-6

[12] A. Livne, E. S. Tov, A. Solomon, A. Elyasaf, B. Shapira, and L. Rokach, “Evolving Context-aware Recommender Systems with Users in Mind,” Expert Systems with Applications, Vol. 189, 2022, Article Number 116042, DOI: 10.1016/j.eswa.2021.116042

[13] 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. DOI: 10.1109/ICIS.2016.7550751

[14] F.H. Nanehkaran, S.M. Lajevardi SM, and M.M. Bidgholi, “Nearest Neighbors Algorithm and Genetic-based Collaborative Filtering,” Concurrency and Computation: Practice and Experience, Vol. 34, No. 1, 2022. DOI: 10.1002/cpe.6538

[15] Y. Ar and E. Bostanci, “A Genetic Algorithm Solution to the Collaborative Filtering Problem,” Expert Systems with Applications, Vol. 61, pp. 122-128, 2016. DOI: 10.1016/j.eswa.2016.05.021

[16] B. Alhijawi and Y. Kilani, “A Collaborative Filtering Recommender System using Genetic Algorithm,” Information Processing & Management, Vol. 57, No. 6, 2020. DOI: 10.1016/j.ipm.2020.102310

[17] F.H. Nanehkaran, S.M. Lajevardi SM, 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. DOI: 10.1111/exsy.12893

[18] S. Lee, “Collaborative Filtering System Using Clustering and Genetic Algorithms,” Communications in Computer and Information Science, Vol. 1071, 2019. DOI: 10.1007/978-981-32-9563-6_16

[19] 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. DOI: 10.1007/s10489-019-01495-4

[20] P.A. Khodke and P.B. Rathod, “Genetic Algorithm Based Similarity Transitivity in Collaborative Filtering,” International Journal of Engineering Research and Technology, Vol. 2, No. 12, pp. 2933-2936, 2013.

[21] B.S. Neysiani, N. Soltani, R. Mofidi, and M.H. Nadimi-Shahraki, “Improve 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. DOI: 10.5815/ijitcs.2019.02.06

[22] 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. DOI: 10.1016/j.jmsy.2020.12.019

[23] G. Koutrica, B. Bercovitz, and H. Garcia, "FlexRecs: Expressing and Combining Flexible Recommendations,” Proc. the ACM SIGMOD International Conference on Management of Data, pp. 745-758, 2009. DOI: 10.1145/1559845.1559923

Authors

Soojung Lee received the B.S. degree in Mathematics Education from Ewha Woman’s University, Korea in 1985. She received M.S. and Ph.D. degrees in Computer Science from Texas A&M University in 1990 and 1994, respectively. Dr. Lee joined the faculty of the Department of Computer Education at Gyeongin National University of Education, Gyunggi-do, Korea, in 1998, as a professor. She is interested in recommender systems, information filtering, data mining techniques, and computer education.