JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 26 No. 5, pp. 47-53, May 2021
https://doi.org/10.9708/jksci.2021.26.05.047

Applying Different Similarity Measures based on Jaccard Index in Collaborative Filtering

Soojung Lee*

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

[Abstract]

Sparse ratings data hinder reliable similarity computation between users, which degrades the performance of memory-based collaborative filtering techniques for recommender systems. Many works in the literature have been developed for solving this data sparsity problem, where the most simple and representative ones are the methods of utilizing Jaccard index. This index reflects the number of commonly rated items between two users and is mostly integrated into traditional similarity measures to compute similarity more accurately between the users. However, such integration is very straightforward with no consideration of the degree of data sparsity. This study suggests a novel idea of applying different similarity measures depending on the numeric value of Jaccard index between two users. Performance experiments are conducted to obtain optimal values of the parameters used by the proposed method and evaluate it in comparison with other relevant methods. As a result, the proposed demonstrates the best and comparable performance in prediction and recommendation accuracies.

▸ **Key words**: Collaborative Filtering, Recommender System, Similarity Measure, Data Sparsity, Jaccard index

[요 약]

희소한 평가 데이터는 사용자들 간의 신뢰할만한 유사도 산출을 저해하기 때문에 추천 시스템을 위한 메모리 기반의 협력 필터링 기법의 성능을 저하시킨다. 기존 연구의 많은 결과물은 이 데이터 희소성 문제를 해결하기 위해 개발되었으며, 가장 단순하고 대표적인 업적은 자카드 계수를 활용하는 방법들이다. 이 계수는 두 사용자의 공통 평가 항목수를 반영하며, 그들 간의 유사도를 보다 정확하게 계산하기 위해 전통적인 유사도 척도와 통합된다. 그러나, 그러한 통합은 데이터 희소성의 정도를 고려하지 않은 매우 단순한 방법이다. 본 연구는 두 사용자의 자카드 계수값에 의거하여 다른 유사도 척도를 적용하는 새로운 아이디어를 제시한다. 제안 방법에서 사용하는 파라미터의 최적값을 구하기 위하여 성능 실험을 진행하였고, 다른 관련된 방법들과 비교 평가하였다. 결과로서, 제안 방법은 예측 정확도와 추천 정확도에 있어서 가장 우수하거나 대등한 성능을 보였다.

▸ **주제어**: 협력 필터링, 추천 시스템, 유사도 척도, 데이터 희소성, 자카드 계수

• First Author: Soojung Lee, Corresponding Author: Soojung Lee
*Soojung Lee (sjlee@gin.ac.kr), Dept. of Computer Education, Gyeongin National University of Education
• Received: 2021. 04. 21, Revised: 2021. 05. 19, Accepted: 2021. 05. 19.

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

I. Introduction

인터넷 정보의 과부하로 인해 상업용 시스템에서 사용자들이 선호하는 상품의 정보 검색은 더욱 어려워지고 있다. 이러한 추세에서 추천 알고리즘은 정보 검색 시간을 절약해 주고 선호 상품을 보다 용이하게 찾아내기 위한 방법이다. 현재 다양한 분야의 상업용 사이트에서 성공적으로 활용되고 있으며 학계의 활발한 연구가 진행되고 있다. 최근에는 인공지능의 발달로 인해 딥러닝 알고리즘을 추천 시스템 분야에 적용하여 복잡한 사용자의 상품 선호 특성을 학습하여 반영하기 위한 연구 경향도 주목할 만하다[1][2].

이와 같은 노력에도 불구하고 추천 시스템의 성능을 저하시키는 가장 큰 문제점은 데이터 희소성(Data Sparsity)이라고 할 수 있다[3]. 즉, 사용자가 시스템을 사용한 이력 정보가 불충분하여, 이를 이용한 상품 선호 특성을 파악하기 힘들기 때문에, 추천의 성능이 저하되는 문제이다. 물론, 내용 기반 필터링(content-based filtering, CBF) 방식[3]을 취함으로써 사용자의 프로필 정보 또는 시스템에서 관리하는 상품들 정보를 유지 관리하고, 이를 기초로 하여 과거에 사용자가 선택한 상품들과 유사한 특성을 가진 상품들을 추천할 수 있다. 그러나 이러한 방식은 필요 정보의 관리 비용 및 정보 획득의 어려움이 매우 크고, 과거 선호 상품과는 다른 특성의 새로운 상품의 발견이 어려운 치명적인 문제(serendipity problem)를 갖는다[4][5]. 이러한 이유로 인해 현재 상업계의 추천 시스템은 대개 협력 필터링(collaborative filtering, CF) 방식을 기본 원리로 채택한다.

CF 기반의 추천 시스템은 현재 영화, 서적 등 다양한 품목을 다루는 상업 시스템에서 사용되고 있다. 2006년에 개시된 넷플릭스 챌런지는 대용량의 영화 데이터셋에서 가장 높은 추천 성능을 가져오는 기법을 경쟁하는 대회로서, 많은 연구자들의 관심을 이끌었다[3][4]. CF 방법은 과거에 유사한 취향을 보였던 사용자들 간에는 향후에도 유사한 항목 선호도를 나타낼 것이라는 가정을 기반으로 한다. 현 사용자가 미평가한 항목들의 집합을 I라고 할 때, 그를 위한 추천 항목 리스트를 결정하기 위한 절차는 다음과 같다. 현 사용자와 유사한 선호 이력을 보였던 다른 사용자(nearest neighbor, 인접 이웃)들을 산출한 후, I에 속한 항목들 중에서 이웃들이 선호하였던 항목에 대해 현 사용자가 부여할 평가치를 예측한다. 현 사용자를 위한 추천 리스트는 높은 평가 예측치의 항목 순으로 제공된다. 이러한 과정에서 유사도 산출은 시스템의 성능에 매우 중요한 역할을 한다.

기존 연구에서 두 사용자 간의 유사도는 대개 그들이 공통으로 평가한 항목들에 대해 부여했던 평가치들로부터 산출한다. 대표적인 방법으로서 상관도 기반과 벡터 코사인 기반이 있다. 전자의 예로는 피어슨 상관(Pearson correlation), 스피어만 순위 상관(Spearman rank correlation), 켄델 타우 상관(Kendall's tau correlation) 등이 포함되며, 후자는 코사인 유사도(cosine similarity), 조정된 코사인 유사도(adjusted cosine similarity) 등이 포함된다[3]. 앞서 언급하였듯이, 이러한 방식들은 평가 데이터가 희소한 경우에 유사도 값의 산출이 아예 불가능하거나, 또는 산출된 유사도 값의 신뢰도가 저하되는 문제점을 내포한다. 이 문제의 극복을 위해 특이성 분해(singular value decomposition)와 같은 차원 감소 기법, 주성분 분석(principle component analysis) 등의 요인 분석 기법이 개발되었으나, 이들은 주요한 추천 정보를 간과하여 저하된 성능을 초래한다는 보고가 있다[4].

본 연구에서는 데이터 희소성 문제의 최근 해결책으로서 매우 간단하면서도 효과적인 자카드 계수(Jaccard index)를 활용한 유사도 산출 방법의 단점을 개선하기 위한 연구를 진행하였다. 자카드 계수는 두 사용자 간의 공통평가항목수를 반영한 척도로서, 주로 다른 척도와 결합하여 유사도 값을 산출하기 위해 사용되었다[6][7]. 그러나, 공통평가항목개수의 희소 정도를 고려하지 않은 결합 방식으로 인해 성능 저하를 초래할 가능성이 존재한다. 본 논문에서는 이러한 문제점에 대한 상세한 분석을 제시하고, 이를 개선한 방법을 소개한다. 또한 실험 평가를 통하여 개선 방법이 주요 성능 척도에 있어서 우수함을 확인하였다.

논문의 구성은 다음과 같다. 2절에서는 자카드 계수와 이를 활용한 기존 문헌의 유사도 척도를 소개한다. 3절에서는 제안 방법을 설명하고 4절에서 성능 측정 실험 결과를 제시하며, 5절에서 논문의 결론을 맺는다.

II. Related Works

자카드 계수는 두 사용자가 평가한 모든 항목들 중에서 공통으로 평가한 항목수의 비율을 말한다[8]. 구체적으로, 사용자 u가 평가한 항목들의 집합을 I_u라고 할 때, 사용자 u와 v 간의 자카드 계수는 다음과 같이 정의된다.

$$Jaccard(u,v) = \frac{|I_u \cap I_v|}{|I_u \cup I_v|} \quad \cdots\cdots\cdots\cdots\cdots\cdots(1)$$

자카드 계수는 평가치의 크기를 고려하지 않은 채 공통 평가항목 개수의 상대적 비율만을 반영하므로, 두 사용자의 평가치 이력에 기반한 유사도를 산출하기엔 부적합하다. 그러나, 희소한 평가 데이터 환경에서 기존 유사도 척도의 단점을 보완하여 성능 개선을 가져온다는 다수의 연구가 발표되었다[6][9]. [10]에서는 자카드 계수가 어떠한 기존 유사도 척도와 결합했을 때 어떠한 데이터 환경에서 성능 개선 효과를 나타내는지를 분석하였다. 결과적으로, 자카드 계수만을 유사도 척도로 사용하는 방식은 희소 데이터 환경에서 전통 척도들보다 우수한 성능을 보이며, 결합한 유사도 척도는 대개의 경우 성능 향상의 결과를 보이는 것으로 밝혔다.

Bobadilla 외 2인은 평균자승차이(mean squared differences, MSD)와 자카드 계수를 결합한 유사도 척도를 제안하고, 개선된 성능 실험 결과를 보고하였다[6]. 또한 MSD 이외에 피어슨 상관도 등의 척도와도 자카드 계수를 통합하여 그 성능을 비교 분석하였으나, 평균자승차이와 통합한 척도의 성능이 가장 우수하였음을 밝혔다. Saranya 외 2인은 피어슨 상관도와 결합했을 때의 성능 개선 효과를 분석하였다[11]. Iftikhar 외 4인은 트라이앵글 유사도(triangle similarity)에 기반한 상품 추천 방법을 제시하였는데, 공통 평가 항목만을 고려하는 트라이앵글 유사도의 단점을 개선하여 두 사용자의 비공통 항목들의 평가치를 모두 반영한 방법을 개발하였고, 추가적으로 사용자 평가 선호 행태를 고려하여 개발한 유사도 척도를 보완하였다[12]. 이밖에도 트라이앵글 유사도를 자카드 계수와 통합하여 예측 오류를 개선하려는 연구 결과가 발표되었다[7].

데이터 희소성 문제를 해결하기 위해 자카드 계수와 더불어 헬링거 거리(Hellinger Distance)[13]를 활용한 연구가 Mu 외 4인에 의해 발표되었다[14]. 이 연구에서 이 척도들은 공통 피어슨 상관계수(Common Pearson Correlation Coefficient, COPC)와 결합되었는데, 그 동기로서 공통항목 평가개수의 영향을 낮추고 피어슨 상관도의 단점을 극복하고자 하였다.

[6]의 연구에서와 마찬가지로 [8]에서도 자카드 계수와 평균자승차이를 결합한 방식의 유사도 척도를 제안하였다. 즉, 보다 높은 관련성의 이웃들을 선정하기 위하여 사용자의 모든 평가 벡터를 활용한 유사도 모델을 개발하였다. 결과적으로 제안한 척도는 MovieLens 데이터셋에서 기존의 전통적인 유사도 척도들보다 정확도와 효율성 면에 있어서 높은 추천 결과를 나타냄을 보였다.

이밖에도 [15]의 연구에서는 자카드 계수 자체의 단점을 보완하기 위하여, 특정한 평가범위 영역별로 공통평가개수를 별개로 산출하고, 각 영역의 최적 크기를 유전자 알고리즘을 활용하여 결정함으로써 성능 개선 효과를 가져올 수 있음을 보고하였다.

이와 같이 자카드 계수는 평가 데이터 희소성 문제를 극복하기 위한 방안으로써 주로 전통적인 유사도 척도와 결합하는 방식으로 흔히 사용되어 왔고, 특히 [6]의 제안 척도는 이후 많은 연구자들에 의해 비교 분석 대상이었다[16]. MSD는 공통평가항목 각각에 대해 두 평가치 차이의 제곱의 평균값으로서 사용자 u와 v에 대해 아래와 같이 산출한다.

$$MSD(u,v) = \frac{1}{I_{u,v}} \sum_{i \in I_{u,v}} (r_{u,i} - r_{v,i})^2$$

.....(2)

위 식에서 $I_{u,v}$는 두 사용자의 공통평가항목집합이고 $r_{u,i}$는 사용자 u가 항목 i에 대해 부여한 평가치이다. 각 항목에 대한 평가치 차이가 작을수록 유사도는 커지므로, 두 사용자간 유사도는 1-MSD로 정의한다.

III. Proposed Methodology

공통평가항목수를 기반으로 유사도를 산출하는 전통적 유사도 척도(SIM)와 자카드 계수(JAC)가 통합된 유사도 척도는 두 방법을 각기 독립적으로 사용하는 경우보다 우수한 추천 성능을 가져온다고 알려져 있다[6][11][16]. 본 연구에서는 과거 몇몇 연구에서 토대로 하였던 JAC*SIM과 같은 단순 통합의 유사도 척도를 각각의 값의 크기에 따라 다음과 같이 분류하여 문제점을 분석한다.

(경우 1) JAC과 SIM의 값이 모두 작은 경우: JAC*SIM의 값은 각각보다 더욱 작아지고, 공통평가항목수의 비율이 작으므로, 그보다 작은 값의 JAC*SIM는 유사도로 채택하기에 타당한 것으로 판단한다.

(경우 2) JAC 값은 작고 SIM의 값이 큰 경우: 공통평가항목수의 비율이 작은데도 불구하고 SIM 값은 크므로, 신뢰하기 어려운 SIM 값이 산출된 경우이다. 따라서, JAC*SIM 값의 유사도는 타당하지 않은 것으로 판단한다.

(경우 3) JAC 값은 크고 SIM의 값이 작은 경우: 공통평가항목수의 비율이 크므로, 기존 척도에 따라 두 사용자간 유사도 값을 계산하기에 충분하며, 따라서 산출된 SIM의 값이 타당한 것으로 판단한다.

(경우 4) JAC과 SIM의 값이 모두 큰 경우: (경우 3)과 같은 근거로 인해 산출된 SIM의 값이 타당한 것으로 판단한다.

두 사용자 간의 유사도 값을 결정하기 위하여, 위 (경우 1)과 (경우 2)는 JAC 값에 비례하는 작은 값을, (경우 3)과 (경우 4)는 각각에서 산출된 SIM 값을 택하기로 한다. 즉, 본 연구의 주요 아이디어는, 경우에 따라 다른 유사도 척도를 적용하여 두 사용자 간의 유사도 값을 산출하는 것이다. 결과적으로, 제안 척도 PROP(u,v)는 아래와 같이 간단한 공식에 의거한다. 이 때, J_θ와 α는 실험적으로 결정해야 할 파라미터들이다.

$$PROP(u,v) = \begin{cases} JAC^{\alpha}(u,v), & \text{if } JAC(u,v) < J_\theta \\ SIM(u,v), & otherwise \end{cases}$$

.....(3)

위에서 JAC^{α}는 (경우 1)과 (경우 2)에 적용되며, SIM은 (경우 3)과 (경우 4)에 적용된다. α값은 산출된 유사도 값이 작아지도록 해야 하므로, 1 보다 큰 값을 갖도록 범위를 설정한다. SIM은 임의의 유사도 척도를 모두 대신할 수 있으므로, 제안 척도의 유연성의 폭이 매우 넓어지는 잇점이 있다.

제안 척도를 기존의 통합 척도인 JAC*SIM과 비교하자면, (경우 2)와 (경우 3)에 있어서 차이를 보일 수 있다. 즉, 두 경우 모두에서 JAC*SIM의 산출 값은 정반대의 크기를 가진 각각의 값이 상호 보완적으로 작용하지만, PROP에서는 두 경우 모두 JAC과 SIM 중에서 작은 값에 의해 유사도가 결정되므로, 차이가 존재한다.

IV. Performance Experiments

1. Design of Experiments

제안 척도의 성능을 비교 분석하기 위하여 사용자들의 평가항목개수의 분산이 충분히 커서 식 (3)의 각 조건에 해당하는 경우의 수가 많은 실험 데이터셋을 선정하였다. 관련 연구에서 흔히 사용되어 온 MovieLens는 이에 합당한 데이터셋 중의 하나로서[11][16], 영화에 대한 사용자들의 평가치 정보를 포함한다. 각 사용자 당 20개 이상의 평가개수, 총 3952개의 영화 항목, 1부터 5까지의 정수 평가치를 갖는다. 평가데이터 량이 희소할수록 임의의 두 사용자 간의 자카드 계수는 낮아지는데, 본 연구의 실험에 사용한 MovieLens의 희소성 수준(sparsity level)은 0.980582로서, 이는 전체사용자×전체항목의 행렬(user-item matrix) 내 평가가 매겨지지 않은 요소의 비율을 의미한다. 따라서, 100개의 셀 중에서 약 98개가 미평가된 항목에 해당하므로, 실험에 사용한 MovieLens는 매우 희박한 데이터셋임을 알 수 있다.

성능 평가 척도로써 가장 널리 활용되는 대표적인 예측 정확도 지표인 평균절대오차(Mean Absolute Error, MAE)와 추천 정확도 지표인 F1을 선정하였다. MAE는 현 사용자가 미평가한 항목 x에 대해 시스템이 얼마나 정확한 예측 평가치를 산출하는지를 나타내며, 현 사용자의 인접 이웃들 중에서 x를 평가한 이웃사용자들의 평가치에 대해 현 사용자와의 유사도에 기반한 가중 합(weighted sum)을 적용하여 계산하였다[3]. 본 실험을 위하여 유사도 산출을 위한 훈련 데이터 집합을 80%, 성능 평가를 위한 시험 데이터 집합을 20%로 전체 데이터셋을 분할하였다. 사용자 u를 위한 항목 x에 대한 예측 평가치를 $\tilde{r}_{u,x}$, 시험 데이터 집합 내 실제 평가치를 $r_{u,x}$, 전체 사용자 집합을 U, 시험 데이터집합을 IT라고 표기할 때 MAE는 다음과 같이 산출한다.

$$MAE = \sum_{u \in U} \sum_{x \in I_T} |r_{u,x} - \tilde{r}_{u,x}| \quad .....(4)$$

F1은 시스템이 산출한 추천 리스트가 얼마나 사용자의 선호에 적합한지를 가늠하는 지표이다. 정밀도(precision)와 재현율(recall)을 동일한 비중으로 결합한 조화평균으로 계산하는데[3], 본 실험에서는 평가치가 4점 이상이면 선호하는 것으로 간주하였다.

성능 평가의 비교를 위하여, 본 연구의 목적에 부합되는 기존 유사도 척도를 선정하였는데, 자카드 계수(JAC), 평균자승차이(MSD), 그리고 JMSD[6]로서, 이들 척도들 대비 성능 개선 효과를 살펴보았다. 본 연구의 제안 척도는 PROP으로 표기하였고, 식 (3)에서 SIM 유사도 산출을 위하여 MSD를 사용함으로써 비교 대상 척도들과의 형평성을 유지하였다.

2. Effect of Parameters

본 절에서는 제안 척도의 J_θ와 α의 값이 변화함에 따라 성능 변화의 결과를 제시한다. 이는 다음 절에서 실험에 사용한 유사도 척도들과의 성능 비교 분석을 위해 우선 제안 척도 파라미터의 최적값을 구하기 위함이다.

그림 1은 α 값을 1.1부터 3.9까지 변화한데 따른 MAE 성능 차이를 나타낸다. 평가 예측치를 산출할 인접이웃 수를 증가시킴에 따라 성능은 점차적으로 고정 값에 수렴하는 양상을 보이는데, 이러한 현상은 α 값에 무관하게 관찰되었다. 다만, 인접이웃수가 40일 때 가장 좋은 성능을 보

였고, 더 이상 이웃수를 증가하는 것은 성능 향상에 도움이 되지 않음을 알 수 있다. 이는 데이터셋이 희박하므로 더 이상의 유효한 인접이웃이 존재하지 않음을 의미한다.

그림 1에서 가장 작은 두 개의 α 값, 즉, α=1.1과 1.5에 대하여 가장 저조한 성능 결과를 보이며, 1.9 이상일 때는 MAE 성능 차이가 나타나지 않았다. 따라서, 식 (3)에서 두 사용자간 공통평가항목수의 비율이 작을 때는 유사도 값을 기하급수적으로 작게 해야 함을 실험적으로 증명한 것이다. 또한, 특정 하한선 이하의 낮은 유사도 값을 부여하는 것은 더 이상의 성능 향상을 가져오지 않음을 확인하였다. 본 연구에서는 α=2.1로 고정하여 이후 실험을 진행하였다.

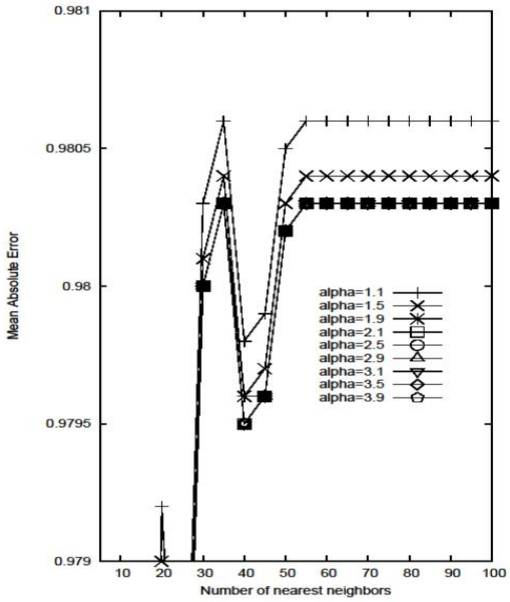

Fig. 1. Mean absolute error for varying alpha

그림 2는 최적의 J_θ 값을 알아내기 위하여 실험한 결과이다. α 값을 변화할 때보다 J_θ 값의 변화에 따라 MAE 성능 차이의 폭이 더욱 큰 것을 알 수 있다. 대체로 J_θ 값이 클 경우에 성능이 저조하고, 값이 매우 작을 때보다는 중간 값 정도에서 가장 좋은 성능을 보였다. J_θ 값이 크면 식 (3)에서 유사도 값이 JAC^α 식에 의해 산출될 가능성이 더 높아지는데, 실제로는 공통항목수가 신뢰할만한 유사도 값을 산출해내기에 충분히 많으므로, 기존 척도인 SIM을 사용하는 것이 더 타당하다. 이러한 이유로 인해 큰 J_θ 값에 대해 저조한 성능을 보이는 것으로 해석할 수 있다. 그림에서 0.05일 때 성능이 가장 우수함을 확인하였으므로, 이후 실험에서 제안 척도를 위한 파라미터 값은 J_θ=0.05, α=2.1로 정하였다.

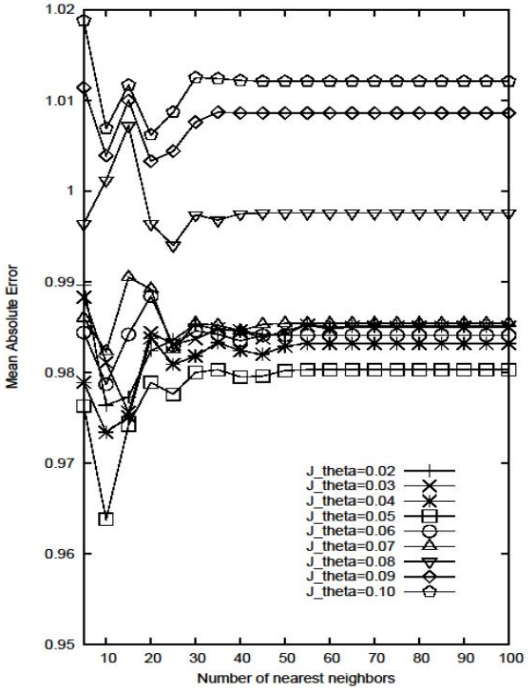

Fig. 2. Mean absolute error for varying J_θ

3. Performance Comparison

표 1은 인접 이웃 수(NN)의 변화에 따라 네 가지 유사도 척도의 평균절대오차 성능 변화를 제시한다. 각 인접 이웃 수에 해당하는 컬럼 당 가장 좋은 성능을 나타낸 MAE 결과는 굵은 표시, 그 반대의 경우는 밑줄 표시를 하여 식별하기 용이하도록 나타냈다.

인접이웃수와 무관하게 대체로 MSD는 가장 저조한 성능을 보였는데, 특히 두 사용자의 평가치를 전혀 반영하지 않는 JAC 보다도 뒤떨어진 성능을 보이는 것은 주목할 만하다. 이는 실험에 사용한 데이터셋이 매우 희소하여 공통항목의 평가치를 기준으로 하는 것은 효율성이 거의 없음을 의미한다.

Table 1. Results of mean absolute error with varying nearest neighbors

number of NNs / Similarity Measure	5	10	15	20	25	30	35	40	45
MSD	0.9933	0.9795	0.9806	0.9856	0.9866	0.9878	0.9873	0.9872	0.9868
JAC	0.9828	0.9843	0.9817	0.9840	0.9849	0.9844	0.9821	0.9816	0.9809
JMSD	0.9807	0.9768	0.9808	0.9851	0.9847	0.9825	0.9821	0.9811	0.9805
PROP	**0.9763**	**0.9638**	**0.9742**	**0.9789**	**0.9776**	**0.9800**	**0.9803**	**0.9795**	**0.9796**

Table 2. F1 results with varying nearest neighbors

number of NNs / Similarity Measure	5	10	15	20	25	30	35	40	45
MSD	0.2567	0.2978	0.2518	0.2518	0.2518	0.2518	0.2518	0.2518	0.2518
JAC	0.2092	0.2894	0.2585	0.2585	0.2585	**0.2585**	0.2585	0.2585	0.2585
JMSD	0.2097	**0.3088**	**0.3006**	0.2585	0.2585	**0.2585**	0.2585	0.2585	0.2585
PROP	**0.2629**	0.2678	0.2623	**0.2623**	**0.2623**	0.2561	**0.2623**	**0.2623**	**0.2623**

MSD의 저하된 성능은 JMSD에 영향을 주는데, 표에서 볼 수 있듯이 JMSD의 성능은 JAC의 결과를 매우 약간 개선시킨 것에 불과하다. 그러나 [6]에서 언급한대로 두 척도를 통합시킴으로써 성능 개선 효과는 가져옴을 확인할 수 있다. PROP의 성능은 모든 경우에 있어서 가장 우수한 결과를 보였다. 이로써 3절에서 기술한 각 경우의 문제점 분석 및 해결책으로써 제시한, 희소성에 따른 멀티 유사도 척도의 적용 방법이 어느 정도 타당한 것으로 판단할 수 있다.

표 2에서 각 유사도 척도의 F1 결과를 제시하여 추천 성능을 살펴보았다. 표 1의 MAE 결과와 거의 유사한 양상을 보이는데, MSD는 대체로 가장 낮은 성능을 나타냈고, PROP은 가장 우수한 성능 결과를 보였다. 다만, MAE에서처럼 PROP이 모든 인접 이웃 수에 대해 최고의 F1 성능을 보이지 않는 이유는 앞 절에서 파라미터 J_{θ}와 α의 최적값을 정하기 위하여 MAE 결과를 기준으로 하였기 때문인 것으로 해석된다. 만약 두 파라미터의 최적값을 F1 결과에 따라 결정한다면 PROP의 F1 결과는 다른 척도들에 비해 더욱 우수할 것으로 예상한다.

V. Conclusions

데이터 희소성 문제는 협력 필터링 기반의 추천 시스템에서 매우 치명적인 문제로서 성능 저하의 주요 요인이다. 본 연구에서는 자카드 계수를 이용하여 희소성 문제를 다루었던 기존 연구 결과의 단점을 분석하고 새로운 유사도 척도를 제안하였다. 제안 방법의 주요 아이디어는 평가데이터 희소 기준에 따른 멀티 유사도 척도의 적용으로서, 사용자 개인별로 유사한 평가이력의 인접 이웃을 산출하기 위한 척도가 달라질 수 있다. 따라서 사용자의 평가이력 밀도와 상관없이 항상 동일한 하나의 유사도 척도를 적용한 기존 관련 연구 결과와 크게 차별화된다. 이러한 제안 척도의 유연성은 실험 결과 그 성능의 우수성이 입증되었는데, 예측 성능과 추천 성능 측면에서 기존 유사도 척도들을 능가하였다.

제안 척도는 희소한 평가 데이터 환경에서 유용하다. 실제로 현재 상업적으로 성행 중인 협력 필터링을 활용한 추천 시스템들은 가입 사용자들이 평가 값을 부여하는 경우가 많지 않은 반면에 시스템이 제공하는 항목 수는 매우 크므로, 매우 희소한 데이터 환경을 가질 수밖에 없으며, 따라서 본 연구의 유사도 척도가 유용하게 활용될 수 있다. 그러나, 만약 사용자의 평가 데이터가 충분한 경우에는 기존 척도를 활용하는 것과 동일한 성능을 가져온다. 또한 본문에서는 단일 데이터셋만으로 실험 환경을 구축하였는데, 이와는 다른 특성의 데이터셋들을 활용하여 제안 척도의 성능을 검증할 필요가 있다. 본문에서 정의한 유사도 척도 공식에서 도입한 두 가지 파라미터의 최적값을 구하기 위해 임의의 여러 실험치를 대입하여 성능 검증을 실시하였으나, 다양한 종류의 진화 알고리즘 등을 이용하여 최적값 산출이 가능하다. 결과적으로, 본 연구 결과를 통해 자카드 계수를 활용한 유사도 산출 방법의 새로운 측면을 제시하였으며, 향후 또 다른 특성의 데이터셋을 이용하여 실험을 확장 수행하고 여러 가지 성능 지표를 통한 검증 작업이 필요하다.

REFERENCES

[1] Z. Batmaz, A. Yurekli, A. Bilge, and C. Kaleli, “A Review on Deep Learning for Recommender Systems: Challenges and Remedies,” Artificial Intelligence Review, Vol. 52, No. 1, pp. 1-37, 2019. DOI: 10.1007/s10462-018-9654-y

[2] R. A. Mancisidor, M. Kampffmeyer, K. Aas, and R. Jenssen, “Learning Latent Representations of Bank Customers with the Variational Autoencoder,” Expert Systems with Applications, Vol. 164, 2021. DOI: 10.1016/j.eswa.2020.114020

[3] M. Aamir and M. Bhusry, “Recommendation System: State of the Art Approach,” International Journal Computer Applications, Vol. 120, No. 12, pp. 25-32, 2015. DOI: 10.5120/21281-4200

[4] 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

[5] 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

[6] J. Bobadilla, F. Serradilla, and J. Bernal, “A New Ccollaborative Filtering Metric that Improves the Behavior of Recommender Systems,” Knowledge-Based Systems, Vol. 23, No. 6, pp. 520-528, 2010. DOI: 10.1016/j.knosys.2010.03.009

[7] S.-B. Sun, Z.-H. Zhang, X.-L. Dong, H.-R. Zhang, T.-J. Li, L. Zhang, and F. Min, “Integrating Triangle and Jaccard Similarities for Recommendation,” PLoS ONE, Vol. 12, No. 8, 2017. DOI: 10.1371/journal.pone.0183570

[8] 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

[9] S. Kosub, “A Note on the Triangle Inequality for the Jaccard Distance,” Pattern Recognition Letters, Vol. 120, pp. 36-38, 2019. DOI: 10.1016/j.patrec.2018.12.007

[10] S. Lee, “Performance Analysis of Similarity Reflecting Jaccard Index for Solving Data Sparsity in Collaborative Filtering,” The Journal of Korean Association of Computer Education, Vol. 19, No. 4, pp. 59-66, July 2016.

[11] K. G. Saranya, G. S. Sadasivam, and M. Chandralekha, "Performance Comparison of Different Similarity Measures for Collaborative Filtering Technique,“ Indian Journal of Science and Technology, Vol. 9, No. 29, Aug. 2016. DOI: 10.17485/ijst/2016/v9i29/91060

[12] 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. DOI: 10.1109/ACCESS.2020.3005953

[13] J. Guo, J. Deng, X. Ran, Y. Wang, and H. Jin, “An Efficient and Accurate Recommendation Strategy using Degree Classification Criteria for Item-based Collaborative Filtering,” Expert Systems with Applications, Vol. 164, 2021. DOI: 10.1016/j.eswa.2020.113756

[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. DOI: 10.1016/j.procs.2019.01.258

[15] S. Lee, “Improving Jaccard Index using Genetic Algorithms for Collaborative Filtering,” Lecture Notes on Computer Science, 10385, pp. 378–385, 2017. DOI: 10.1007/978-3-319-61824-1_41

[16] B. Zhu, R. Hurtado, J. Bobadilla, and F. Ortega, “An Efficient Recommender System Method Based on the Numerical Relevances and the Non-Numerical Structures of the Ratings,” IEEE Access, Vol. 6, pp. 49935-49954, 2018. DOI: 10.1109/ACCESS.2018.2868464

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.