JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 31 No. 11, pp. 369-378, November 2026
https://doi.org/10.9708/jksci.2026.31.11.369

Tail Probability-Based Singularity for Collaborative Filtering Similarity Measures

Soojung Lee*

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

[Abstract]

In memory-based collaborative filtering, recommendation accuracy is determined by the similarity measure between users. Singularity-based measures, reflecting the rarity of a rating, split ratings into positive and negative categories and define singularity as the proportion of users within each category. Thus the ratings of the same category share the same value, and rater's tendency is not reflected. This paper defines singularity as the self-information of the tail probability that a rating equal to or more extreme is observed under the rating distributions of the user and the item, so that each rating carries its own value, and integrates and normalizes it from the two perspectives by a weighted geometric mean. We propose HSING which combines this singularity with the mean squared difference by multiplication, and JHSING which adds the Jaccard index to HSING by weighted summation. Five-fold cross validation on MovieLens 100K and 1M shows that JHSING is the best on all three performance metrics for 1M, has generally the lowest mean absolute error for 100K, and that the proposed methods outperform existing singularity-based measures with the same combination structure on both datasets.

▸ **Key words**: Similarity Measure, Singularity, Self-Information, Collaborative Filtering

[요 약]

메모리 기반 협업 필터링에서 추천 정확도는 사용자 간 유사도 척도에 의해 좌우된다. 평가 희귀성을 반영하는 특이성 기반 척도들은 평가치를 긍정과 부정으로 이분한 뒤 각 범주에 속한 사용자 수의 비율로 특이성을 정의한다. 따라서 같은 범주의 평가들은 동일한 값을 공유하며, 평가자의 성향은 반영되지 않는다. 본 논문은 사용자와 항목의 평가치 분포에서 그 평가와 같거나 더 극단적인 평가가 관측될 꼬리 확률의 자기 정보량으로 특이성을 정의하여 각 평가가 고유한 값을 갖도록 하고, 두 관점의 희귀성을 가중 기하평균으로 통합 및 정규화한다. 이 특이성을 평균 제곱차이와 곱셈으로 결합한 HSING과, 여기에 자카드 지수를 가중 덧셈으로 더한 JHSING을 제안한다. MovieLens 100K와 1M을 대상으로 5-겹 교차 검증을 수행한 결과, JHSING은 1M에서 세 가지 성능 지표 모두에서 가장 우수하였고 100K에서도 평균 절대 오차가 대체로 가장 낮았으며 제안 방법들은 두 데이터셋 모두에서 결합 구조가 동일한 기존 특이성 기반 척도보다도 우수하였다.

▸ **주제어**: 유사도 척도, 특이성, 자기 정보량, 협업 필터링

• First Author: Soojung Lee, Corresponding Author: Soojung Lee
*Soojung Lee (sjlee@gin.ac.kr), Dept. of Computer Education, Gyeongin National University of Education
• Received: 2026. 09. 18, Revised: 2026. 11. 06, Accepted: 2026. 11. 10.

Copyright © 2026 The Korea Society of Computer and Information
http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945
www.kci.go.kr

I. Introduction

추천 시스템은 사용자가 접근할 수 있는 항목의 수가 개인이 검토할 수 있는 범위를 넘어선 환경에서 선택의 부담을 덜어주는 수단이다. 그 가운데 협업 필터링은 항목의 내용 정보를 요구하지 않고 사용자-항목 평가 행렬만으로 동작한다는 점에서 가장 널리 사용된다. 협업 필터링은 메모리 기반 방식과 모델 기반 방식으로 구분된다. 메모리 기반 방식은 구현이 단순하고 추천의 근거를 해석하기 쉬우며, 희소한 데이터 환경에서 모델 기반 방식보다 적합하다는 보고도 있다[1].

메모리 기반 협업 필터링의 예측 정확도는 이웃 선정에 좌우되고, 이웃 선정은 사용자 간 유사도 척도에 의해 결정된다. 유사도의 정의가 이 방식의 성능을 규정하는 이유가 여기에 있다. 전통적으로 피어슨 상관계수[2]와 코사인 유사도가 사용되어 왔고, 평가치의 차이를 직접 이용하는 평균 제곱 차이와 평가 성향의 편차를 보정하는 제약 피어슨 상관계수[3], 그리고 평균 제곱 차이에 공통 평가 항목의 비율을 반영하는 자카드 지수를 결합한 척도[4]가 제안되었다. 이들의 상대적 우열은 데이터의 희소도와 실험 설정에 따라 달라진다[5].

전통적 척도들은 개별 평가치를 모두 동등한 정보로 취급한다는 한계를 공유한다. 다수가 평가한 항목에서의 일치와 소수만이 평가한 항목에서의 일치가 같은 무게로 계산된다. 전자는 항목의 대중성에서 비롯될 수 있으나, 후자는 두 사용자가 취향을 공유한다는 근거로서 더 강하다. 평가 행렬이 희소할수록 이 차이는 이웃 선정의 품질에 직접 반영된다[6].

이러한 문제의식에서 평가치가 지니는 희소함의 정도, 즉 특이성을 유사도에 반영하려는 시도가 제시되었다. Bobadilla 등은 항목에 대한 평가를 긍정과 부정으로 구분한 뒤 각 범주에 속한 사용자 수의 비율로 특이성을 정의하였고[7], Liang 등은 여기에 미평가 사용자 집합을 더하여 세 범주의 특이성으로 확장하였으며[8], 이후 특이성을 타니모토 계수와 결합한 연구도 제시되었다[9]. 이 계열의 접근은 평가치의 이분화, 항목 측면으로의 한정, 경계값의 고정이라는 세 가지 제약을 갖는다.

한편 평가치의 확률분포를 유사도 산출에 활용하는 연구가 별도의 흐름을 이루어 왔다. Kullback-Leibler 발산[10]과 α-발산[11]을 이용한 척도는 공통 평가 항목에 대한 의존을 줄이기 위해 평가치의 확률밀도분포를 비교한다. 평가가 사용자 또는 항목의 평균에서 얼마나 벗어나 있는가를 유사도의 신호로 삼은 척도들도 제시되었다[12][13]. 그러나 이들 연구에서 분포는 두 대상 사이의 거리를 재거나 벗어남의 정도를 환산하는 도구이며, 개별 평가가 지니는 희귀성을 확률로 정의하지는 않았다.

본 논문은 각 평가가 고유한 특이성 값을 갖도록, 그 값을 꼬리 확률의 자기 정보량으로 정의한다. 해당 사용자의 평균보다 높은 평가는 그 값 이상을, 낮은 평가는 그 값 이하를 부여할 확률로 측정하며, 항목의 평가치 분포에도 같은 방식을 적용한다. 확률이 낮을수록 큰 값을 갖는 자기정보량의 성질은 특이성의 개념과 그대로 일치한다. 평가치를 범주로 축약하지 않고, 방향을 가르는 기준이 고정된 경계값이 아니라 사용자와 항목 각각의 평균이므로, 위에서 언급한 세 가지 제약을 해소할 수 있다.

이 특이성을 기존 유사도와 결합하는 방식으로 두 가지를 제안한다. 첫 번째 방법은 평균 제곱 차이와 특이성을 곱셈으로 결합하며, 이는 [7]이 채택한 결합 구조와 동일하다. 두 번째 방법은 첫 번째 방법의 유사도 척도와 자카드 지수를 가중 덧셈으로 결합하며, 이는 [8]의 결합 구조와 동일하다. 결합 구조를 비교 대상 문헌과 일치시킨 이유는 성능의 차이가 결합 방식이 아니라 특이성의 정의 차이에서 비롯된 것임을 확인하기 위함이다.

제안 방법은 MovieLens 100K와 1M을 대상으로 평균 절대 오차(MAE), 평균 제곱근 오차(RMSE), Mean Average Precision(MAP)으로 평가하였다. 실험 결과, 두 번째 제안 척도의 개선 폭이 가장 컸으며, MovieLens 1M에서는 세 지표 모두 가장 우수하였고 MovieLens 100K에서도 MAE가 대체로 가장 낮았다.

본 논문의 구성은 다음과 같다. 2장에서 유사도 척도 관련 연구를 정리한다. 3장에서 제안 방법으로서 꼬리 확률 기반 특이성의 정의와 이를 접목한 두 유사도 척도를 제시한다. 4장에서 실험 내용을, 5장에서 결론을 기술한다.

II. Related Study

1. Traditional Similarity Measures

메모리 기반 협업 필터링에서 사용자 간 유사도는 두 사용자가 함께 평가한 항목 위에서 계산된다. 피어슨 상관계수는 각 사용자의 평가치를 자신의 평균으로 중심화한 뒤 선형 상관을 구하며[2], 코사인 유사도는 평가치 벡터 사이의 각도를 이용한다. 두 척도는 공통 평가 항목의 개수를 고려하지 않는다. 공통 항목이 하나뿐일 때 피어슨 상관계수는 분모가 0이 되어 정의되지 않고, 코사인 유사도는 평가치의 차이와 무관하게 1이 된다.

www.kci.go.kr

평균 제곱 차이는 공통 평가 항목에 대한 평가치 차이의 제곱 평균을 척도 범위로 정규화하여 유사도로 환산한다. 제약 피어슨 상관계수는 사용자 평균 대신 평가 척도의 중앙값을 기준으로 중심화하여, 두 사용자의 평가가 척도상에서 어느 쪽에 놓이는지를 반영한다[3]. Bobadilla 등은 평균 제곱 차이에 자카드 지수를 곱한 JMSD를 제시하여 평가치의 근접성과 공통 평가 비율을 함께 반영하였다[4]. 이들 척도의 상대적 우열은 데이터의 밀도에 따라 달라지며[5], 희소 데이터에서는 가장 널리 쓰이는 피어슨 상관계수와 코사인 유사도가 최상위 성능군에 들지 않는다는 보고도 있다[14]. 공통 평가 항목에 대한 의존을 낮추려는 시도는 최근까지 이어지고 있다[6].

2. Singularity-based Similarity Measure

특이성은 하나의 항목이 얼마나 적은 사용자에 의해 평가되었는가를 수치화한 개념이다. Bobadilla 등은 항목에 대한 평가를 긍정과 부정으로 구분한 뒤, 각 범주에 속한 사용자 수가 전체 사용자 수에서 차지하는 비율을 1에서 뺀 값으로 특이성을 정의하였다[7]. 소수만이 평가한 항목일수록 특이성이 높으며, 이를 평균 제곱 차이와 곱셈 결합하여 유사도를 정의하였다. 또한 평가치를 두 범주가 아닌 여러 등급으로 나누는 일반화 모형도 함께 제시하였다.

Liang 등은 이 개념을 두 방향으로 확장하였다[8]. 하나는 개선된 피어슨 상관계수로, 각 항목을 평가한 사용자 수의 비율을 평가치에 곱한 뒤 상관을 계산하였다. 다른 하나는 개선된 자카드 지수로, 긍정 및 부정 특이성에 더하여 항목을 평가하지 않은 사용자 집합에 기반한 공백 특이성을 도입하고, 두 사용자의 평가가 일치하는 경우와 상반되는 경우, 한쪽만 평가한 경우에 대응하는 특이성의 합으로 유사도를 정의하였다. 두 척도는 곱셈과 가중 덧셈의 두 방식으로 결합하였으며, 실험에서는 가중 덧셈 방식이 더 나은 결과를 보였다.

Selvi와 Sivasankar는 타니모토 계수가 공통 평가 항목의 개수만 사용하고 실제 평가치를 버린다는 점에 주목하여, 타니모토의 분자와 분모에 특이성을 가중한 개선된 타니모토를 제안하였다[9]. 특이성의 정의 자체는 [7]을 따르되, 두 사용자의 평가 조합을 긍정-긍정, 부정-부정, 긍정-부정, 긍정-공백, 부정-공백의 다섯 경우로 나누고 각 경우에 대응하는 두 특이성의 곱을 실제 평가치의 곱에 가중하였다. 공백이 개입하는 경우를 분모에서 차감하므로 한쪽 사용자만 평가한 항목까지 유사도에 반영된다. 이 논문은 평가치를 n 개의 등급으로 나누는 일반화 모형도 제시하였으나, 실험적 검증은 제시하지 않았다.

이 계열의 연구는 특이성을 산출하는 방식에서 공통된 구조를 지닌다. 평가치를 긍정과 부정으로 구분한 뒤, 각 범주에 속한 사용자 수가 전체 사용자 수에서 차지하는 비율만으로 특이성 값을 정한다는 것이다. 이에 따라 세 가지 제약이 따른다. 첫째, 특이성이 항목과 범주의 조합에 따라 결정되므로, 한 항목에 긍정적 평가를 부여한 사용자들은 값의 크기와 무관하게 동일한 특이성을 공유하고 척도 위에서 평가치가 차지하는 위치는 반영되지 않는다. 둘째, 특이성이 항목 측면에서만 산출되므로 같은 평가치를 부여한 두 사용자의 평가 성향 차이는 구분되지 않는다. 셋째, 긍정과 부정을 가르는 경계값이 데이터셋 전체에 대해 하나로 고정되어, 후하게 평가하는 사용자와 박하게 평가하는 사용자가 같은 기준으로 다루어진다. [7][9]가 제시한 일반화 모형은 등급의 수를 늘려 첫째 제약을 완화하지만, 특이성이 항목 단위의 비율로 정의되고 등급의 경계 또한 고정되므로 나머지 두 제약은 그대로 남는다.

3. Information-Theoretic Approaches

평가치의 분포를 확률로 다루는 연구도 협업 필터링에 접목되었다. Deng 등은 두 대상의 평가치 분포 사이의 Kullback-Leibler 발산을 이용하여 유사도를 정의하고, 희소한 데이터에서 공통 평가 항목에 대한 의존을 낮추었다[10]. Wang 등은 이를 α-발산으로 일반화하여 항목 간 유사도를 산출하였다[11]. 두 연구에서 확률분포는 사용자 또는 항목이라는 대상 전체를 대표하며, 유사도는 두 분포 사이의 거리로 정의된다.

정보이론의 엔트로피를 유사도 산출에 직접 도입한 연구도 이어져 왔다. Kwon 등은 사용자 선호의 정보 엔트로피를 가중치로 삼아 전통적 유사도 값을 보정하였다[15]. Wang 등은 평가치 사이의 상대적 차이를 엔트로피로 측정하는 사용자 유사도 모형을 제시하고, 이를 피어슨 상관계수와 가중 평균으로 결합하였다[16]. Lee는 항목에 대한 전역적 평가 행태가 반영되도록 사용자 평가치의 정보 엔트로피를 기존 척도에 접목하고, 규모가 작고 희소한 데이터에서 개선 효과가 크다고 보고하였다[17]. 이들 연구에서 엔트로피는 한 사용자 또는 한 항목의 평가 분포 전체를 하나의 수치로 요약하여 유사도에 가중치로 작용한다. 엔트로피가 분포의 요약 통계로 쓰일 뿐, 분포 위에 놓인 개별 평가치에 부여되는 양으로는 쓰이지 않는다는 점에서 앞의 발산 기반 척도와 성격을 같이한다.

4. Positioning of This Work

기존 척도들도 하나의 평가가 사용자 또는 항목의 평균

www.kci.go.kr

에서 얼마나 벗어나 있는가를 유사도의 신호로 삼아 왔다. Manochandar와 Punniyamoorthy는 근접성, 영향도, 인기도의 세 요소를 각각 정규화하고 가중하여 결합하는 척도를 제시하였는데, 이 가운데 인기도는 항목의 평균 평가치에서 멀리 떨어진 평가에 더 큰 비중을 둔다[12]. Liu 등의 척도는 두 평가가 다른 평가들로부터 벗어난 정도를 반영하는 항과, 각 사용자 평가치의 평균과 분산으로 평가 성향을 보정하는 항을 결합하였다[13]. 두 연구 모두 벗어남을 평균으로부터의 거리로 환산한다. 거리는 평가 척도의 범위에 의존하는 양이므로, 같은 크기의 이탈이라도 그것이 해당 사용자나 항목의 분포에서 얼마나 드문 일인지는 반영하지 못한다.

앞 절의 두 연구 분야, 즉 특이성 계열과 확률분포 계열이 서로 독립적으로 발전해 온 것에 반해, 본 연구는 이 둘을 하나의 정의 안에서 결합한다. 2절의 특이성 계열은 평가의 희귀성을 유사도에 반영하였으나, 그 희귀성을 항목의 평가 범주별 사용자 수의 비율로 측정하였다. 3절의 확률분포 및 정보이론 계열은 평가치의 분포를 다루었으나, 그 분포를 대상 전체의 요약이나 두 대상 사이의 거리로만 사용하였다. 본 연구는 희귀성을 각 평가의 속성으로 유지하면서, 그 값을 해당 사용자와 항목의 평가치 분포 위에서 확률로 측정한다. 이로써 2절에서 지적한 세 가지 제약이 해소될 수 있다.

구체적으로, 제안 방법은 기존 특이성 기반 척도와 다음 세 가지 점에서 다르다. 첫째, 특이성을 범주별 사용자 비율이 아니라 개별 평가의 꼬리 확률로 정의하므로 같은 범주에 속한 평가라도 평가치의 크기에 따라 서로 다른 값을 갖는다. 둘째, 항목의 평가치 분포 뿐 아니라 사용자의 평가치 분포에서도 희귀성을 측정하므로, 같은 평가치를 부여한 사용자들의 평가 성향 차이가 반영된다. 셋째, 평가의 방향을 가르는 기준을 데이터셋 전체에 고정된 경계값이 아니라 해당 사용자와 항목 각각의 평균에 둔다. 아울러 특이성을 자기 정보량으로 정의하므로 두 관점의 희귀성을 정보량의 가중 평균으로 결합할 수 있다.

III. Proposed Methodology

1. Overview

본 연구에서 제안하는 특이성 산출 방법의 착안점은 다음과 같다. 하나의 평가가 지니는 정보의 크기는 그 평가가 얼마나 드문가에 달려 있으며, 드문 정도는 그 평가에 관계된 사용자와 항목에 대응하는 평가 이력을 기준으로 판단한다. 평소에 4점 이하만 주던 사용자가 부여한 5점과, 대부분의 평가가 5점인 사용자가 부여한 5점은 동일한 값이지만 같은 의미를 갖지 않는다. 항목의 입장에서도 마찬가지이다. 따라서 제안 방법은 각 평가에 대하여 해당 사용자의 평가치 분포와 해당 항목의 평가치 분포에서 그 평가가 얼마나 극단적인 위치에 있는지를 확률로 측정하고, 그 확률을 자기 정보량으로 변환하여 특이성을 산출한다. 이렇게 얻은 특이성은 두 사용자의 유사도를 계산할 때 각 공통 평가 항목의 기여도를 조정하는 가중치로 작용한다.

2. Rating Singularity Based on Tail Probability

사용자 집합을 U, 항목 집합을 I라 하고, 사용자 u가 항목 i에 부여한 평가치를 $r_{u,i}$로 표기한다. 사용자 u가 평가한 항목들의 집합을 I(u), 항목 i를 평가한 사용자들의 집합을 U(i)라 하며, 두 집합에서 관측된 평가치의 평균을 각각 μ(u)와 μ(i)로 표기한다.

하나의 평가치가 지니는 희귀성은 그 값이 기준보다 얼마나 극단적인가에 따라 판단한다. 평균보다 높은 평가에 대해서는 그 값 이상의 평가가, 평균보다 낮은 평가에 대해서는 그 값 이하의 평가가 얼마나 드문지를 파악한다. 이러한 관점을 반영하기 위하여 사용자 u의 평가치 분포에 대한 평가치 r의 꼬리 확률을 다음과 같이 정의한다.

$$P_u(r) = \begin{cases} prob(R \geq r|u), & r > \mu(u) \\ 1, & r = \mu(u) \\ prob(R \leq r|u), & r < \mu(u) \end{cases} \quad \text{....식(1)}$$

항목 i의 평가치 분포에 대해서도 평가치 r의 꼬리 확률을 같은 방식으로 다음과 같이 정의한다.

$$P_i(r) = \begin{cases} prob(R \geq r|i), & r > \mu(i) \\ 1, & r = \mu(i) \\ prob(R \leq r|i), & r < \mu(i) \end{cases} \quad \text{.....식(2)}$$

두 확률은 각각 I(u)와 U(i)에서 관측된 평가치의 상대빈도로 추정한다. 이 확률은 단순한 발생 확률이 아니라 관측된 평가와 같거나 더 극단적인 평가가 나타날 확률, 즉 한쪽 꼬리 누적 확률이다. 평가치가 평균과 같을 때 확률이 1인 것은, 자신의 평균에 해당하는 평가가 그 사용자 또는 그 항목에게 가장 전형적이므로 희귀성을 부여하지 않는다는 의미이다. 평가치가 정수일 때 평균은 대개 정수가 아니므로, 평가치가 평균과 정확히 같은 경우는 드물다. 이때는 평균에 가장 가까운 두 평가치가 전형적인 평가의 역할을 한다. 예를 들어 평균이 3.6이면 모든 평가는 3 이하이거나 4 이상이므로 3의 꼬리 확률과 4의 꼬리 확률을 더하면 1이 되고, 둘 중 적어도 하나는 0.5 이상의 꼬리 확률을 받아 낮은 특이성을 갖는다.

확률이 낮은 사건일수록 큰 값을 갖는 자기 정보량은 특

www.kci.go.kr

이성의 개념과 일치하므로, 평가치 r에 대한 사용자 관점과 항목 관점의 특이성을 각각 식(3)과 식(4)로 정의한다.

$$S_u(r) = -\log P_u(r) \quad \text{.....식(3)}$$

$$S_i(r) = -\log P_i(r) \quad \text{.....식(4)}$$

기존 연구의 특이성이 사용자와 무관하게 항목마다 평가치 구간에 따라 동일한 값을 갖는 것과 달리, 제안 방법은 같은 항목에 대한 평가라도 그 값과 그 값을 부여한 사용자에 따라 서로 다른 특이성을 갖는다. 관측된 평가치는 꼬리에 포함되므로 n개의 평가로 추정한 꼬리 확률은 1/n 보다 크거나 같고 자기 정보량은 log n을 넘지 않는다. 평가가 하나뿐인 항목에서는 그 평가가 곧 평균이므로 꼬리 확률이 1이 되어 특이성이 0이 된다. 즉 평가 수가 적은 분포에서 희귀성이 과대평가되지 않는 대신, 평가치 간 특이성의 변별력이 줄어든다.

3. Integrated Singularity

앞 절에서 정의한 사용자와 항목 분포 관점의 희귀성을 하나의 값으로 통합하기 위하여, 항목 관점의 영향력을 1로 고정하고 사용자 관점의 영향력을 α로 조절하여, 평가치 r에 대한 통합된 자기 정보량을 식(5)로 정의한다.

$$S'_{u,i}(r) = S_i(r) + \alpha S_u(r) \quad \text{.....식(5)}$$

여기에서 α는 0 이상 1 이하의 실수이며, α가 0.5이면 사용자 관점의 기여도가 항목 관점의 절반으로 반영된다. 항목 관점에 더 큰 비중을 두는 이유는 하나의 평가가 지니는 의미가 일차적으로 그 항목에 대한 다른 사용자들의 평가 분포에 의해 규정된다고 보았기 때문이다.

식(5)는 확률이 0에 근접하면 발산하여 상한이 없다. 유사도 척도의 구성 요소로 쓰려면 유계인 값이 필요하므로 지수 변환 $1-\exp(-S'_{u,i}(r))$를 적용한다. 식(5)를 대입하고 두 지수의 합이 1이 되도록 정규화하여, 사용자 u의 항목 i에 대한 평가치 r의 통합 특이성을 아래와 같이 정의한다.

$$S_{u,i}(r) = 1 - P_i(r)^{1/(1+\alpha)} P_u(r)^{\alpha/(1+\alpha)} \quad \text{....식(6)}$$

식(6)에서 두 지수의 비가 1대 α이므로 항목 관점과 사용자 관점의 영향력 비율이 그대로 유지되고, 두 지수의 합이 1이며 $0<P_i(r)\le 1$과 $0<P_u(r)\le 1$이므로 $0\le S_{u,i}(r)\le 1$이 항상 성립한다. 두 확률이 모두 1인 가장 전형적인 평가에서 $S_{u,i}(r)$는 0이 된다. 통합 특이성이 0과 1 사이로 정규화되므로 기존 유사도와 결합할 때 별도의 척도 조정이 필요하지 않다.

식(6)의 두 번째 항은 두 꼬리 확률의 가중 기하평균이다. 만약 정규화하지 않으면 α를 크게 하는 것만으로 특이성 값이 커진다. 예를 들어 두 관점에서 모두 평범한 $P_i(r)=P_u(r)=0.5$인 평가는 α=0일 때 0.5이지만 α=1일 때는 0.75가 된다. 그러나 두 지수의 합을 1로 맞추면 이 평가는 α와 상관 없이 0.5가 되며, α는 특이성의 크기를 바꾸지 않고 두 관점 중 어느 쪽에 더 비중을 둘지만 조절하는 역할을 한다.

통합 특이성의 예는 다음과 같다. $P_i(r)=0.1$, $P_u(r)=0.01$인 평가에 대하여 $S_{u,i}(r)$는 α=0.2일 때 약 0.932, α=0.5일 때 약 0.954, α=1일 때 약 0.968이 된다. α가 커질수록 사용자 관점의 희귀성이 통합 특이성에 더 크게 반영된다. α=0.5에서 항목 관점에서만 희귀한 $P_i(r)=0.01$, $P_u(r)=0.5$인 평가의 $S_{u,i}(r)$는 약 0.963이며, 사용자 관점에서만 희귀한 $P_i(r)=0.5$, $P_u(r)=0.01$인 평가의 $S_{u,i}(r)$는 약 0.864이다. 즉 항목 관점의 영향이 더 크다는 설계 의도가 드러난다. α의 값은 실험을 통하여 결정한다.

4. Combining Singularity with Similarity Measures

산출된 특이성을 기존 유사도와 결합하는 방식으로 두 가지를 제안한다. 두 사용자 u와 v의 공통 평가 항목 집합을 I(u,v), 그 크기를 n이라 하고, 평가치를 척도 범위로 정규화한 값을 r'로 표기한다. 즉, 척도의 최소값과 최대값을 각각 r_{min}과 r_{max}라고 할 때 $r'=(r-r_{min})/(r_{max}-r_{min})$으로 산출한다. 첫 번째 유사도 척도인 HSING은 평균 제곱 차이와 특이성을 곱셈으로 결합하며, 다음 식과 같이 정의한다.

$$HSING(u,v) = \frac{1}{n}\sum_{i\in I(u,v)} (1-(r'_{u,i}-r'_{v,i})^2) S_{u,i}(r_{u,i}) S_{v,i}(r_{v,i})$$

.....식(7)

여기에서 $S_{u,i}(r)$와 $S_{v,i}(r)$는 각각 사용자 u와 v가 항목 i에 부여한 평가의 통합 특이성이다. 각 공통 평가 항목의 기여도는 두 정규화 평가치의 근접도에 두 평가의 특이성을 곱한 값이며, 근접도와 두 특이성이 모두 0과 1 사이에 놓이므로 HSING의 값 또한 그 범위를 벗어나지 않는다. 두 사용자가 드문 평가를 서로 가깝게 부여한 항목일수록 기여도가 커지고, 평가치가 어긋난 항목은 근접도 값이 작아져 기여도가 줄어든다. 즉 특이성은 평가치의 일치 정도에 가중치를 부여하는 인자로 작동하므로, 흔한 평가로만 이루어진 항목은 두 평가치가 일치하더라도 높은 유사도를 갖지 않는다. 이 곱셈 결합 구조는 [7]이 채택한 방식과 동일하다.

두 번째 척도인 JHSING은 HSING과 자카드 지수를 가중 덧셈으로 결합한다. 자카드 지수는 식(8)과 같이 두 사

www.kci.go.kr

용자의 공통 평가 항목 수를 전체 평가 항목 수로 나눈 값이며, JHSING은 식(9)로서 정의한다.

$$Jaccard(u,v) = \frac{|I(u) \cap I(v)|}{|I(u) \cup I(v)|} \quad \text{.....식(8)}$$

$$JHSING(u,v) = \beta \bullet HSING(u,v) + (1-\beta) \bullet Jaccard(u,v) \quad \text{.....식(9)}$$

β는 0과 1 사이의 실수이며 실험을 통하여 결정한다. HSING은 공통 평가 항목 위에서 평가치의 근접성과 특이성만을 반영하는 데 비해, JHSING은 두 사용자가 공유하는 평가 항목의 비율을 함께 고려한다. 이 가중 덧셈 구조는 [8]이 채택한 방식과 동일하다.

식(1)부터 식(6)까지의 계산은 평가 행렬이 주어지면 사용자별, 항목별로 한 차례 수행하여 저장할 수 있으므로, 추가되는 연산은 전처리 단계에 한정되고 이웃 선정 단계의 계산량은 기존 척도와 같은 수준으로 유지된다. 즉, 전처리 단계의 계산은 사용자와 항목별 평가치 빈도표로부터 수행되므로 평가 수 $|R|$에 대해 $O(|R|)$의 전처리로 끝난다. 이후 두 사용자의 유사도는 공통 평가 항목 수 n에 대해 O(n)으로 계산되므로 MSD 및 Jaccard와 점근적 복잡도가 같다.

IV. Performance Experiments

1. Experimental Setup

실험에는 MovieLens 100K와 MovieLens 1M을 이용하였으며, 두 데이터셋의 특성은 표 1과 같다. 평가치는 1~5의 정수이며, 두 데이터셋 모두 각 사용자가 최소 20개 항목을 평가하였다. 두 데이터셋은 규모와 밀도가 서로 다른데, 사용자당과 항목당 평가 개수에서 알 수 있듯이 평가는 고르게 분포하지 않으며 한 건뿐인 항목이 있는 반면 수천 건이 집중된 항목도 있으므로, 규모와 평가 분포가 다른 두 환경에서 제안 척도의 성능을 비교할 수 있다.

Table 1. Characteristics of the MovieLens datasets

	ML100K	ML1M
No. of users	943	6,040
No. of items	1,682	3,706
total # of ratings	100,000	1,000,209
Rating range	1~5	1~5
Density	6.30%	4.47%
Ratings per user(min/max/avg)	20/737/106.1	20/2,314/165.6
Ratings per item(min/max/avg)	1/583/59.5	1/3,428/269.9

모든 유사도 척도에 대하여 동일한 조건에서 5-겹 교차검증을 수행하였다. 각 데이터셋의 평가를 무작위로 섞어 평가 단위로 크기가 같은 다섯 개의 부분집합으로 분할한 뒤, 네 개를 학습에 사용하고 나머지 하나를 시험에 사용하는 과정을 다섯 차례 반복하였다. 꼬리 확률을 추정하는 평가치 분포와 사용자 및 항목의 평균, 자카드 지수 등 모든 통계량은 학습에 사용한 평가만으로 계산하였다. 분할에 따른 편차를 줄이기 위하여 서로 다른 분할 세트를 두 벌 마련하여 이 과정을 반복하였으므로, 척도마다 열 차례의 측정이 이루어졌다. 보고하는 값은 그 평균이다.

사용자 u의 항목 i에 대한 예측 평가치는, 항목 i를 평가한 사용자 가운데 u와의 유사도가 높은 상위 k명을 이웃 $N_k(u,i)$로 선정하여 식(10)으로 산출하였다. 이웃이 없으면 μ(u)로 예측하였다.

$$\hat{r}_{u,i} = \mu(u) + \frac{\sum_{v \in N_k(u,i)} sim(u,v)(r_{v,i} - \mu(v))}{\sum_{v \in N_k(u,i)} |sim(u,v)|} \quad \text{.....식(10)}$$

MAE와 RMSE는 이웃의 수를 5에서 50까지 5 간격으로 변화시키며 측정하였고, MAP은 이웃의 수를 50으로 고정한 뒤 추천 목록의 길이를 변화시키며 측정하였다.

비교 대상은 여덟 가지 척도이다. 전통적 척도로는 피어슨 상관계수(COR)[2], 코사인 유사도(COS), 평균 제곱 차이(MSD)와 제약 피어슨 상관계수(CPC)[3], 그리고 평균 제곱 차이에 자카드 지수를 결합한 JMSD[4]를 포함하였다. 특이성 기반 척도로는 Bobadilla 등의 척도(BS)[7], Liang 등의 척도(LIANG)[8], Selvi와 Sivasankar의 척도(SELVI)[9]를 포함하였다. 비교 척도는 원 논문의 식과 매개 변수를 따라 같은 실험 틀 안에서 구현하였으며, 제안 척도인 HSING과 JHSING을 이들과 같은 실험 조건에서 비교하였다.

성능은 세 가지 척도로 측정하였다. 평균 절대 오차(MAE)는 예측 평가치와 실제 평가치 차이의 절댓값에 대한 평균이고, 평균 제곱근 오차(RMSE)는 그 차이의 제곱에 대한 평균의 제곱근이다. 두 값은 작을수록 예측이 정확함을 뜻하며, RMSE는 큰 오차에 더 민감하다. Mean Average Precision(MAP)은 각 사용자에게 제시된 추천 목록에서 관련 있는 항목이 상위에 배치된 정도를 평가한다. 추천 목록은 각 사용자의 시험 항목을 예측 평가치의 내림차순으로 정렬한 상위 N개로 구성하였다. 관련 항목의 기준은 4.0으로 두어 4점 이상의 평가치를 받은 항목을 관련 있는 항목으로 간주하였다. MAP은 관련 항목이 있는 사용자들의 평균 정밀도(AP)를 평균한 값이다.

www.kci.go.kr

2. Performance Comparison

먼저 제안 척도의 매개변수 값을 학습 데이터를 활용하여 결정하였다. 표 2는 ML100K에서 α를 0.0부터 1.0까지 변화시키며 JHSING의 MAE를 측정한 결과이다. 사용자 관점을 반영하지 않는 α=0.0에서 오차가 가장 크고 α가 커질수록 대체로 미세하게 줄어들어, 항목 관점만으로 특이성을 정의하는 것보다 사용자 관점을 함께 반영하는 편이 유리함을 보인다. 이 결과에 따라 ML100K의 α를 0.8로 고정하였다. ML1M에 대해서도 같은 방식으로 실험을 수행한 결과(표 3), 대체로 α=0.4에서 MAE가 가장 낮았고 그보다 커지면 소폭 증가하여 α를 0.4로 고정하였다. 두 데이터셋의 α가 다른 것은 항목 분포 추정을 위한 표본 크기의 차이에 기인한 것일 수 있다. 항목당 평균 평가수는 ML100K에서 59.5로 사용자당 106.1보다 적지만, ML1M에서는 269.9로 사용자당 165.6보다 많다. 따라서 사용자 분포가 더 안정적인 ML100K에서는 큰 α가, 항목 분포가 더 안정적인 ML1M에서는 작은 α가 선정되었다.

Table 2. MAE of JHSING with varying α on ML100K

α \ NN	10	20	30	40	50
0.0	0.7659	0.7482	0.7452	0.7446	0.7444
0.2	0.7622	0.7472	0.7443	0.7434	0.7437
0.4	0.7616	0.7470	0.7439	0.7430	0.7437
0.6	0.7621	0.7469	0.7435	0.7432	0.7437
0.8	0.7615	0.7464	0.7432	0.7430	0.7435
1.0	0.7614	0.7468	0.7432	0.7429	0.7435

Table 3. MAE of JHSING with varying α on ML1M

α \ NN	10	20	30	40	50
0.0	0.7383	0.7184	0.7126	0.7105	0.7094
0.2	0.7362	0.7171	0.7116	0.7094	0.7088
0.4	0.7352	0.7163	0.7111	0.7093	0.7087
0.6	0.7350	0.7163	0.7111	0.7095	0.7089
0.8	0.7350	0.7166	0.7114	0.7098	0.7092
1.0	0.7351	0.7167	0.7117	0.7101	0.7095

JHSING의 결합 가중치 β는 [8]이 두 데이터셋에 대하여 제시한 최적값을 그대로 사용하여 ML100K에서 0.7, ML1M에서 0.6으로 두었다. 결합 구조가 [8]과 동일하므로 같은 가중치를 적용함으로써, 가중치 탐색의 이점이 제안 척도에만 주어지지 않도록 하였다.

그림 1은 ML100K에 대한 결과이다. 이웃 수 50에서 MAE는 JHSING과 COR이 약 0.744로 가장 낮고, JMSD가 약 0.748, MSD와 CPC가 약 0.750으로 뒤를 잇는다. LIANG, HSING, COS, SELVI는 약 0.753~0.754이며, BS가 약 0.758로 가장 높다. JHSING은 이웃 수 20 이상에서 최저값을 보였고 이웃 수 30~40에서 약 0.744로 수렴한다. 이웃 수가 작을수록 방법 간 격차가 커서, 이웃 수 20에서 COR는 약 0.752인데 반해 JHSING은 약 0.746로서 적은 수의 이웃만으로도 안정된 예측을 얻는다.

RMSE에서는 순위가 일부 바뀐다. JHSING이 이웃 수 20~40에서 가장 낮으나 45 이상에서는 COR이 근소하게 앞서, 이웃 수 50에서 COR이 약 0.948, JHSING과 MSD와 CPC가 약 0.950이다. 특이성 기반 척도는 HSING이 약 0.959, BS가 약 0.960, SELVI가 약 0.962이며 LIANG은 약 0.966으로 비교 대상 가운데 가장 높다. MAE에서 중위권에 있던 LIANG이 RMSE에서 최하위인 것은 평균적 오차는 억제하되 소수의 큰 오차는 억제하지 못하였음을 뜻한다.

MAP에서는 JHSING이 중간 수준에 놓인다. 추천 목록의 길이 50에서 COR이 약 0.770으로 가장 높고 MSD와 COS가 약 0.766, JHSING과 JMSD가 약 0.761, CPC가 약 0.760으로 뒤를 잇는다. 또한 BS가 약 0.754, LIANG과 HSING이 약 0.746, SELVI가 약 0.735로 이어진다. 전통적 척도들이 상위를 차지하였으나 특이성 기반 척도 가운데에서는 JHSING이 가장 높다. ML100K에서는 예측 오차에서의 우위가 순위 품질로 그대로 이어지지는 않았다.

그림 2는 ML1M에 대한 결과이다. 세 지표 모두에서 JHSING이 가장 우수하다. 이웃 수 50에서 MAE는 JHSING이 약 0.710, JMSD가 약 0.718, COR이 약 0.722이고, LIANG과 HSING이 약 0.728, MSD가 약 0.736으로 이어지며 SELVI가 약 0.755로 가장 높다. RMSE의 순서도 이와 거의 같아 JHSING이 약 0.908, COR이 약 0.916, JMSD가 약 0.918, LIANG이 약 0.921, MSD가 약 0.926, HSING이 약 0.928, SELVI가 약 0.971이다. MAP에서도 JHSING이 모든 목록 길이에서 가장 높아 길이 5에서 약 0.893, 50에서 약 0.816을 기록하였다. ML1M에서는 예측 정확도에서의 우위가 순위 품질로 그대로 이어졌다.

표 4와 5는 JHSING의 예측 정확도 결과가 통계적으로 유의한지를 검정한 결과이다. 성능 결과가 안정적으로 수렴하는 결과를 나타내는 이웃 수(ML100K 30, ML1M 40)를 기준으로 각 척도의 10회 측정에 대한 평균과 표준편차를 제시하였고 각 비교 척도의 차이는 같은 분할에서 얻은 측정값을 짝지은 대응 표본 t-검정으로 확인하였다. 그 결과 JHSING은 두 데이터셋에서 다른 모든 척도들보다 $p<0.01$ 수준에서 유의하게 우수하였다.

www.kci.go.kr

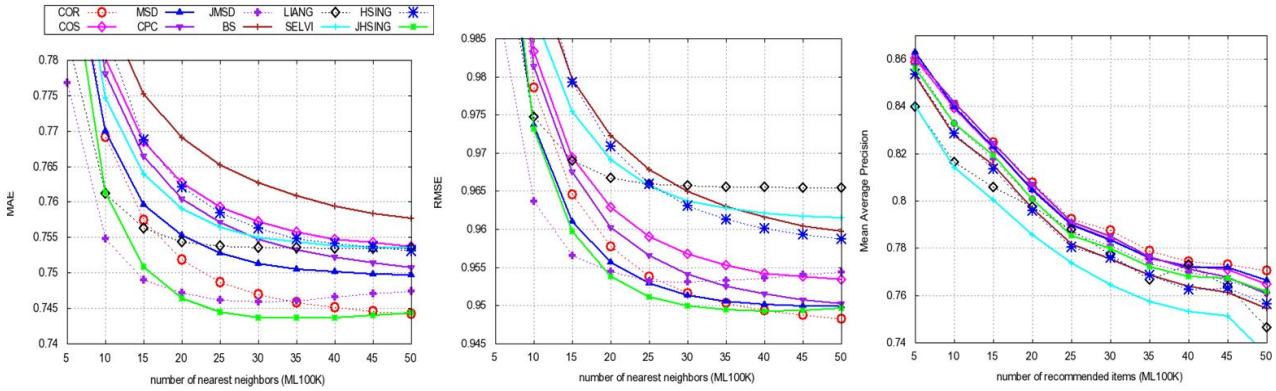

Fig. 1. Performance results with ML100K

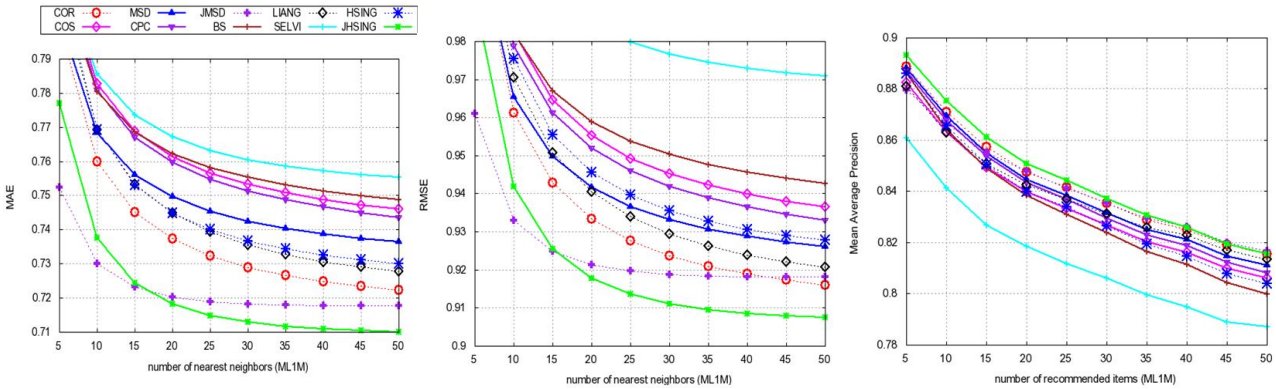

Fig. 2. Performance results with ML1M

Table 4. Mean±standard deviation over 10 runs on ML100K (number of neighbors=30)

Measure	MAE	RMSE
COR	0.7470±0.004396**	0.9517±0.005442**
COS	0.7572±0.004675**	0.9567±0.005386**
MSD	0.7513±0.004713**	0.9514±0.005581**
CPC	0.7547±0.004718**	0.9541±0.005294**
JMSD	0.7459±0.005298**	0.9531±0.006440**
BS	0.7627±0.003926**	0.9650±0.004581**
LIANG	0.7536±0.004754**	0.9657±0.005464**
SELVI	0.7549±0.004630**	0.9637±0.005331**
HSING	0.7563±0.005309**	0.9631±0.006158**
JHSING	0.7437±0.005244	0.9499±0.006108

**: p<0.01

Table 5. Mean±standard deviation over 10 runs on ML1M (number of neighbors=40)

Measure	MAE	RMSE
COR	0.7248±0.000881**	0.9189±0.000921**
COS	0.7487±0.001035**	0.9399±0.001250**
MSD	0.7386±0.000952**	0.9288±0.001221**
CPC	0.7467±0.001013**	0.9365±0.001223**
JMSD	0.7178±0.000888**	0.9183±0.001199**
BS	0.7513±0.000902**	0.9456±0.001092**
LIANG	0.7306±0.000919**	0.9239±0.001110**
SELVI	0.7572±0.001440**	0.9730±0.001684**
HSING	0.7325±0.001071**	0.9307±0.001244**
JHSING	0.7108±0.001030	0.9085±0.001110

**: p<0.01

특이성을 사용한 척도들 사이의 비교는 결합 구조가 같은 쌍을 놓고 볼 때 분명해진다. HSING과 BS는 특이성을 평균 제곱 차이와 곱셈으로 결합하는 동일한 구조를 가지므로, 두 척도의 차이는 특이성의 정의에서만 비롯된다. ML100K의 MAE에서 HSING이 약 0.753, BS가 약 0.758이고, ML1M에서는 각각 약 0.729와 약 0.749로 격차가 더 벌어진다. ML1M의 RMSE에서도 HSING이 약 0.927, BS가 약 0.943이다. 같은 관계가 JHSING과 LIANG 사이에도 성립한다. 두 척도는 가중 덧셈 결합을 공유하는데, ML100K의 MAE에서 JHSING이 약 0.744, LIANG이 약 0.753이고 ML1M에서는 각각 약 0.710과 약 0.728이다. 결합 구조를 고정한 상태에서 특이성의 정의만 바꾸었을 때 두 쌍 모두에서 개선이 나타났으므로, 성능의 차이를 특이성의 정의에 귀속시켜 해석할 수 있다.

한편 특이성 기반 척도들이 전통적 척도를 일관되게 앞서지는 못하였다. BS는 ML100K의 MAE에서 비교 대상 가운데 가장 높은 값을 보였고, SELVI는 ML1M의 세 지표 모두에서 가장 낮은 성능을 기록하였다. SELVI가 공통 평가가 없는 항목까지 유사도 산출에 반영함에도 이러한 결과가 나타난 것은, 반영하는 정보의 양보다 그 정보에 부여하는 가중치의 정의가 더 결정적임을 시사한다.

www.kci.go.kr

LIANG은 데이터셋에 따라 위치가 크게 달라져 ML100K의 RMSE에서는 최하위였으나 ML1M에서는 상위권으로 올라섰다. 이에 비해 JHSING은 두 데이터셋의 세 지표 모두에서 특이성 계열의 최상위를 유지하였다.

두 데이터셋에서 JHSING의 우위가 나타난 정도는 서로 달랐다. 이러한 차이의 한 가지 원인으로 분포 추정에 쓰이는 표본의 크기를 생각해 볼 수 있다. ML1M은 밀도가 4.47%로 ML100K의 6.30%보다 낮지만, 사용자당과 항목당 평균 평가 수는 각각 165.6과 269.9로서 ML100K의 106.1과 59.5보다 크다. 제안 방법은 각 사용자와 각 항목의 평가치 분포로부터 꼬리 확률을 추정하므로, 분포를 추정할 표본이 충분해야 특이성이 안정적으로 산출된다. ML1M에서 세 지표가 모두 개선된 결과는 이러한 조건과 부합한다. 다만 본 연구에서는 평가 수에 따른 성능을 별도로 분석하지 않았으므로, 이는 결과와 부합하는 하나의 해석일 뿐 검증된 원인은 아니며, 두 데이터셋의 규모나 사용자 구성 등 다른 요인이 작용했을 가능성도 있다.

HSING과 JHSING의 차이도 주목할 만하다. HSING은 두 데이터셋에서 모두 중위권에 머물러, ML1M의 MAE에서 LIANG과 비슷한 수준으로 약 0.728이었다. 같은 특이성을 사용하면서도 자카드 지수를 가중 덧셈으로 더한 JHSING은 약 0.710을 기록하여 뚜렷한 차이를 보였다. 즉, 특이성만으로는 공통 평가 항목의 비율이라는 정보를 대체하지 못하며, 두 정보를 함께 반영하였을 때 비로소 개선이 나타난다. 이는 [8]이 곱셈 결합보다 가중 덧셈 결합에서 더 나은 결과를 얻은 것과 같은 양상이며, [4]의 JMSD가 두 데이터셋에서 전통적 척도 가운데 상위를 유지한 것과도 맥을 같이한다.

V. Conclusions

본 논문은 메모리 기반 협업 필터링의 유사도 척도에 반영되는 특이성을 확률의 관점에서 재정의하였다. 기존 연구는 평가치를 긍정과 부정으로 이분한 뒤 각 범주에 속한 사용자 수의 비율로 특이성을 정의하여, 항목의 같은 범주에 속한 평가들이 모두 동일한 값을 공유하였다. 이에 비해 제안 방법은 각 평가가 고유한 특이성 값을 갖도록 그 값을 꼬리 확률의 자기 정보량으로 정의하였다. 평가의 방향을 가르는 기준을 데이터셋 전체에 대해 고정된 경계값이 아니라 해당 사용자와 해당 항목의 평균에 둠으로써, 평가 성향의 차이가 특이성 산출 단계에서 흡수되도록 하였다. 사용자 관점과 항목 관점에서의 희귀성은 가중 기하평균으로 통합하여 0과 1 사이로 정규화하였다. 이렇게 산출한 특이성을 기존 유사도와 결합하여 HSING과 JHSING 두 척도를 제시하였다. 결합 구조를 각각 [7]과 [8]이 채택한 방식과 일치시킴으로써, 성능 차이를 결합 방식이 아니라 특이성의 정의에 귀속시켜 해석할 수 있도록 하였다.

성능 실험에서 JHSING은 MovieLens 1M의 세 지표 모두에서 가장 우수하였고 MovieLens 100K에서도 대체로 우수한 MAE 성능을 보였다. 같은 결합 구조의 쌍을 비교하면 HSING이 BS보다, JHSING이 LIANG보다 두 데이터셋 모두에서 성능이 더 우수하였다. 특이성 정의를 바꾼 것이 성능 개선으로 이어졌음을 뒷받침하는 결과이다. 다만 이러한 결과는 1~5의 정수 평가 척도를 갖는 MovieLens 계열의 두 데이터셋에서 확인된 것으로 평가 척도의 범위와 희소도가 다른 환경에서도 같은 결과가 나타나는지는 추가 검증이 필요하다.

다만 MovieLens 100K에서 이웃 수가 커질 때 RMSE가 COR에 근소하게 뒤졌고 MAP은 중간 수준에 머물렀다. 제안 방법이 각 사용자와 각 항목의 평가 이력으로부터 꼬리 확률을 추정하는 만큼, 표본이 적은 환경에서는 추정의 분산이 커질 수 있으나, 본 연구에서는 이를 실험적으로 확인하지 않았다. 따라서 평가 수에 따른 성능의 분석을 통한 이 해석의 검증, 평가 수가 적은 사용자와 항목에 대하여 전체 분포를 함께 고려하는 평활 기법의 적용, 평가 척도의 범위와 희소도가 다른 데이터셋으로의 확장과 최근 제안된 유사도 척도와의 비교, 그리고 매개변수 값을 데이터로부터 자동 결정하는 방법이 향후 과제이다.

REFERENCES

[1] S.-A. Lapadaki, J. Nanos, D. Margaris, C. Vassilakis, and D. Spiliotopoulos, "Optimizing collaborative filtering for accurate rating predictions in very sparse datasets," Future Internet, Vol. 18, No. 2, 114, 2026. DOI: 10.3390/fi18020114

[2] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl, "GroupLens: An open architecture for collaborative filtering of netnews," in Proc. of the ACM Conference on Computer Supported Cooperative Work, pp. 175-186, 1994. DOI: 10.1145/192844.192905

[3] U. Shardanand and P. Maes, "Social information filtering: Algorithms for automating word of mouth," in Proc. of the SIGCHI Conference on Human Factors in Computing Systems, pp. 210-217, 1995. DOI: 10.1145/223904.223931

[4] J. Bobadilla, F. Serradilla, and J. Bernal, "A new collaborative filtering metric that improves the behavior of recommender

www.kci.go.kr

systems," Knowledge-Based Systems, Vol. 23, No. 6, pp. 520-528, 2010. DOI: 10.1016/j.knosys.2010.03.009

[5] 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

[6] J. Guan, B. Chen, and S. Yu, "A hybrid similarity model for mitigating the cold-start problem of collaborative filtering in sparse data," Expert Systems with Applications, Vol. 249, 123700, 2024. DOI: 10.1016/j.eswa.2024.123700

[7] J. Bobadilla, F. Ortega, and A. Hernando, "A collaborative filtering similarity measure based on singularities," Information Processing & Management, Vol. 48, No. 2, pp. 204-217, 2012. DOI: 10.1016/j.ipm.2011.03.007

[8] S. Liang, L. Ma, and F. Yuan, "A singularity-based user similarity measure for recommender systems," International Journal of Innovative Computing, Information and Control, Vol. 11, No. 5, pp. 1629-1638, 2015.

[9] C. Selvi and E. Sivasankar, "A novel singularity based improved Tanimoto similarity measure for effective recommendation using collaborative filtering," in Proc. of the 8th International Conference on Cloud Computing, Data Science & Engineering (Confluence), pp. 256-262, 2018. DOI: 10.1109/CONFLUENCE.2018.8442697

[10] J. Deng, Y. Wang, J. Guo, Y. Deng, J. Gao, and Y. Park, "A similarity measure based on Kullback-Leibler divergence for collaborative filtering in sparse data," Journal of Information Science, Vol. 45, No. 5, pp. 656-675, 2019. DOI: 10.1177/0165551518808188

[11] Y. Wang, P. Wang, Z. Liu, and L. Y. Zhang, "A new item similarity based on α-divergence for collaborative filtering in sparse data," Expert Systems with Applications, Vol. 166, 114074, 2021. DOI: 10.1016/j.eswa.2020.114074

[12] S. Manochandar and M. Punniyamoorthy, "A new user similarity measure in a new prediction model for collaborative filtering," Applied Intelligence, Vol. 51, No. 1, pp. 586-615, 2021. DOI: 10.1007/s10489-020-01811-3

[13] H. Liu, Z. Hu, A. Mian, H. Tian, and X. Zhu, "A new user similarity model to improve the accuracy of collaborative filtering," Knowledge-Based Systems, Vol. 56, pp. 156-166, 2014. DOI: 10.1016/j.knosys.2013.11.006

[14] K. Sgardelis, D. Margaris, D. Spiliotopoulos, and C. Vassilakis, "An evaluation review of user similarity metrics in sparse collaborative filtering datasets," International Journal of Data Science and Analytics, Vol. 20, No. 7, pp. 6665-6693, 2025. DOI: 10.1007/s41060-025-00846-4

[15] H.-J. Kwon, T.-H. Lee, and K.-S. Hong, "Improving prediction accuracy using entropy weighting in collaborative filtering," in Proc. of Symposia and Workshops on Ubiquitous, Autonomic and Trusted Computing, pp. 40-45, 2009. DOI: 10.1109/UIC-ATC.2009.50

[16] W. Wang, G. Zhang, and J. Lu, "Collaborative filtering with entropy-driven user similarity in recommender systems," International Journal of Intelligent Systems, Vol. 30, No. 8, pp. 854-870, 2015. DOI: 10.1002/int.21735

[17] S. Lee, "Using entropy for similarity measures in collaborative filtering," Journal of Ambient Intelligence and Humanized Computing, Vol. 11, No. 1, pp. 363-374, 2020. DOI: 10.1007/s12652-019-01226-0

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.

www.kci.go.kr