Journal of The Korea Society of Computer and Information
Vol. 24 No. 9, pp. 77-82, September 2019
www.ksci.re.kr
https://doi.org/10.9708/jksci.2019.24.09.077

Using User Rating Patterns for Selecting Neighbors in Collaborative Filtering

Soojung Lee*

Abstract

Collaborative filtering is a popular technique for recommender systems and used in many practical commercial systems. Its basic principle is select similar neighbors of a current user and from their past preference information on items the system makes recommendations for the current user. One of the major problems inherent in this type of system is data sparsity of ratings. This is mainly caused from the underlying similarity measures which produce neighbors based on the ratings records. This paper handles this problem and suggests a new similarity measure. The proposed method takes users rating patterns into account for computing similarity, without just relying on the commonly rated items as in previous measures. Performance experiments of various existing measures are conducted and their performance is compared in terms of major performance metrics. As a result, the proposed measure reveals better or comparable achievements in all the metrics considered.

▸Keyword: Collaborative Filtering, Recommender System, Similarity Measure, Entropy

I. Introduction

협력 필터링(collaborative filtering, CF)은 추천 시스템(recommender system)의 주요 구현 기술로서 현재 상업계에서 널리 활용되어 서비스되고 있다. 관련 시스템의 주목적은 서적, 음악, 영화, 등의 다양한 분야에서 사용자들이 선호할만한 항목들을 추천함으로써 검색 편의를 제공하는 것이다[1].

CF의 원리는 다음과 같다. 시스템 사용자가 살펴 보았던 항목들에 대한 모든 평가 기록은 보관되며, 만약 직접적인 평가치를 입력할 수 없는 시스템이라면 사용자의 클릭수, 화면체류시간 등의 간접적 정보로부터 평가치를 유추할 수 있다[2]. 이와 같이 누적된 사용자 평가 기록으로부터, 현 시스템 사용자와 유사한 기록을 가진 사용자들을 구한 후, 유사한 사용자들이 높은 평가치를 부여했던 항목들을 추천한다. 이러한 방식의 기술을 사용자 기반 협력필터링(user-based CF)이라고 하며, 저장된 평가 기록을 이용한다는 점에서 볼 때 메모리 기반(memory-based) 시스템이라고 부른다[2][3].

메모리 기반 CF를 구현하는 또다른 방식인 항목 기반(item-based) 기술은 임의의 항목 x의 추천 여부를 결정하기 위해서, 이 항목에 대한 평가기록과 유사한 평가기록을 가진 다른 항목들을 찾은 후, 이들 유사항목들에게 현 사용자가 높은 평가치를 부여했다면, x를 추천하는 것으로 한다[4][5]. 일반적으로 한 사용자의 평가 개수보다 한 항목에 대한 평가개수가 더욱 많기 때문에, 항목 기반 기술은 데이터 희소성에 따른 문제에 대해 상대적으로 낮은 민감도를 갖는다[6]. 본 연구에서는 그동안 학계의 많은 연구가 이루어져 왔던 사용자 기반 CF 시스템에 초점을 둔다.

유사한 사용자들의 집합 구성원에 따라 시스템의 성능은 좌우될 수 밖에 없다. 따라서 기존 연구에서 다양한 유사도 척도들의 개발을 위해 많은 노력이 시행되었는데, 이들은 크게 상관도 기반과 벡터 코사인 기반으로 분류된다[2][7][8]. 전자의 대표적 척도로는 피어슨 상관도(Pearson correlation), 후자는 코사인 유사도(cosine similarity) 등이 있다. 이들 척도들은 사용자들의 평가기록을 근거로 하는데, 실제 시스템들은 대개 매우 많은 항목들을 유지관리하기 때문에, 각 사용자 당 평가 개수는 상대적으로 매우 적어서 신뢰할만한 유사 사용자들을 구하는 것은 어려운 일이다.

본 연구에서는 위에서 언급한 문제점들을 해결하기 위하여,

• First Author: Soojung Lee, Corresponding Author: Soojung Lee
*Soojung Lee (sjlee@ginue.ac.kr), Dept. of Computer Education, Gyeongin National University of Education
• Received: 2019. 07. 29, Revised: 2019. 09. 02, Accepted: 2019. 09. 03.

기존의 척도들이 근거로 삼았던 사용자 평가 기록 뿐만 아니라, 평가치들로부터 사용자 평가 패턴을 유추하여 이들을 활용한 새로운 유사도 척도를 제안한다. 또한 제안 방식을 여러 성능 평가 척도를 기준으로 하여 기존 방식들과 비교 실험하여 그 우수성을 확인하였다.

논문의 구성은 다음과 같다. 2절에서는 관련 연구에 대해 기술한다. 3절에서는 제안 방법을 설명한 후 4절에서 성능 측정 실험 결과를 제시하고, 5절에서 논문의 결론을 맺는다.

II. Related Works

기존의 유사도 척도들은 주로 두 사용자가 공통으로 평가한 항목들에 대한 평가치를 기반으로 유사도를 산출하기 때문에, 공통평가항목수가 적을 때, 신뢰할만한 유사도값을 산출하지 못할 수 있다는 큰 단점을 갖고 있다[9][10]. 이러한 접근 방법으로는 cold-start 문제[11]나 data sparsity 문제[2][6]를 해결하기 어렵다. 이와 같은 단점을 극복하기 위하여 공통항목에 대한 평가치 이외에 다른 정보 요소들을 추가하여 두 사용자 간의 유사도값을 산출한 연구들이 진행되었다[8][10]. 가장 간단하면서 효율적인 방법으로서 자카드 계수(Jaccard Coefficient)[12]가 개발되었는데, 이는 단순히 두 사용자의 전체 평가개수 대비 공통평가개수의 비율을 말한다.

[13]의 연구에서는 자카드 계수와 전통적 유사도 척도인 평균자승차이(mean squared differences)를 결합한 새로운 유사도 척도를 제안하였고, 기존에 비해 성능이 향상되었음을 보고하였다. [14]의 연구에서는 자카드 계수의 확장 개념을 개발하였는데, 평가치들의 수치적 관련성과 비수치적 구조 정보를 결합하기 위해 사용하였다. 또한 Lee는 평가치 구간별로 자카드 계수를 각각 별도로 계산하여 결합함으로써 기존 계수의 성능을 향상시킨 결과를 보고하였다[15].

한편 평가치 정보의 엔트로피를 활용하여 유사도 측정의 정확성을 향상하려는 연구가 시도되었다. Li와 Zheng은 피어슨 상관도와 엔트로피 등을 결합하여 유사도를 측정하였다[16]. Wang 외 2인의 연구에서도 사용자 엔트로피의 상대적 차이를 피어슨 상관도와 결합하였다[17]. Kwon 외 3인은 두 사용자의 엔트로피 차이값의 역수를 기존의 전통적인 유사도 척도에 곱한 결과를 산출하고, 이 값을 그들 간의 유사도로 결정함으로써, 엔트로피 차이가 작을수록 유사도가 커지는 것으로 정의하였다[18]. [19]에서는 사용자 엔트로피를 활용하지 않고, 각 항목에 부여된 평가치들로써 엔트로피를 구하여, 엔트로피값이 고려된 항목 평가치들을 유사도 측정에 활용함으로써, 각 항목에 대한 전체 평가치들의 패턴을 반영하여 전통적 유사도 척도의 성능 향상을 도모하였다.

위와 같은 연구들은 모두 추가적인 정보를 휴리스틱하게 전통 기법과 결합하여 성능 개선을 달성하려 하였으나, 각 사용자의 전체적 평가 행태를 반영하려는 노력은 미흡하였다. 본 연구에서는 두 사용자의 공통 평가항목들의 평가치 이외에 그들의 평가 패턴을 기준으로 하여 유사도 값을 산출하는 방법을 제안한다.

III. The Proposed Scheme

1. User Rating Pattern

두 사용자 간의 유사도 측정을 위하여 제안 방법은 각 사용자의 평가치 부여 패턴을 추가적으로 활용한다. 이러한 방식은 전통적인 유사도 측정 방법과는 다르게 공통평가항목들에 대한 평가치에 전적으로 의존하지 않으므로, 특히 공통평가항목이 희소할 경우에 측정된 유사도값의 신뢰도를 향상시키는 효과를 가져올 수 있다.

사용자의 평가치 패턴을 정의하기 위하여 다음의 사항들을 고려하였다. 첫째, 패턴의 정의를 범용적으로 이용할 수 있도록 특정한 데이터셋에서만 제공하는 특성 정보는 제외한다. 또한 사용자나 항목의 정보도 일부 데이터셋에서만 제공될 수 있기 때문에 제외하였다. 따라서, 각 사용자 평가치 패턴은 두 가지 요소로 구성하는데, 첫째는 시스템에서 정한 평가치 범위 내에서 사용자가 각 평가치를 부여한 확률이며, 두 번째는 사용자가 부여한 평가치들의 엔트로피이다.

그림 1은 관련 연구에서 널리 활용되는 Jester 데이터셋(http://eigentaste.berkeley.edu/)에서 수집한 각 사용자의 평가치 엔트로피이다. 이 데이터셋은 -10.0부터 +10.0까지의 실수값의 평가치 범위를 사용하지만, 엔트로피 계산을 위하여 해당 범위 내의 이산값으로 반올림하여 산출하였다. 그림에서 1000명의 사용자들의 엔트로피는 매우 광범위하게 분포되어 있음을 알 수 있다. 따라서, 이러한 정보를 유사도 산출에 반영한다면 서로 다른 사용자의 평가 행태가 유사도 측정의 정확성에 기여할 수 있으리라 판단된다.

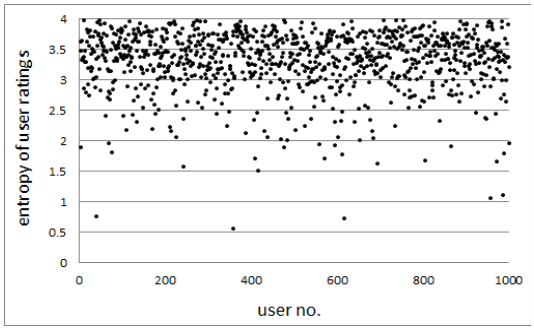

Fig. 1. Entropy of ratings of each user

2. Formulation of the Proposed Similarity Measure

본 절에서는 앞에서 기술한 사용자의 평가치 패턴을 구체적으로 정의한다. 이에 필요한 각종 기호들의 의미는 표 1에 제시하였다.

www.kci.go.kr

사용자 평가치 패턴을 구성하는 첫 요소인 각 평가치의 확률은 아래와 같다. 사용자 u가 부여한 전체 평가개수들 중에서 평가치 r을 부여하는 확률은

$$PR_u(r) = \frac{|\{i \in I | r_{u,i} = r\}|}{|\{i \in I | r_{u,i} \neq NULL\}|}$$

Table 1. Notations and descriptions of symbols

notation	description
$r_{u,i}$	Rating of user u given to item i. If i is not rated, NULL is assigned.
I	The set of all items
I_u	The set of items for which user u has rated.
$r_{\min}$	The minimum rating allowed in the system
$r_{\max}$	The maximum rating allowed in the system

두 사용자 u와 v의 평가치 확률에 대한 패턴 차이는 각 평가치 r에 대한 확률값의 자승차이의 평균으로서 아래와 같이 정의하였다.

$$PSD(u,v) = 1 - \frac{1}{|R(u,v)|} \sum_{r \in R(u,v)} (PR_u(r) - PR_v(r))^2$$

$$R(u,v) = |\{r | PR_u(r) > 0 \text{ 또는 } PR_v(r) > 0\}|$$

사용자 평가치 패턴을 구성하는 두 번째 요소인 엔트로피는 각 평가치에 대한 확률을 기반으로 하여 사용자 u에 대해서 다음과 같이 정의한다.

$$E(u) = -\sum_{r=r_{\min}}^{r_{\max}} PR_u(r) \log_2(PR_u(r))$$

단, 유사도값 산출을 위하여 사용자 평가치 패턴의 구성 요소의 모든 값을 0과 1 사이의 값으로 변환하기 위하여, 사용자 엔트로피값 x에 대하여 아래의 대표적인 정규화 함수를 사용하기로 한다[18].

$$sig(x) = \frac{1}{1+e^{-x}}$$

$$\tanh(x) = \frac{e^{2x}-1}{e^{2x}+1}$$

정규화 변환된 엔트로피 값을 E(u)'라고 할 때, 두 사용자 u와 v 간의 엔트로피 차이값을 아래와 같이 정의하여, 후에 유사도값 산출을 위해 엔트로피 차이가 작을수록 큰 값이 산출되도록 한다.

$$ESD(u,v) = 1 - (E(u)' - E(v)')^2$$

위와 같은 사용자 평가치 패턴을 위한 정의들을 전통적인 유사도 척도인 평균자승차이(Mean Squared Differences, MSD)[2]와 결합하여 최종적으로 두 사용자 u와 v 간의 유사도값 산출 방법을 아래와 같이 제안하고, 이를 URP_A(u, v)로 표기한다.

$$URP_A(u,v) = \begin{cases} \frac{1}{2}(ESD(u,v) + PSD(u,v)), & \text{if } I_u \cap I_v = \varnothing \\ \frac{1}{3}(ESD(u,v) + PSD(u,v) + MSD(u,v)), & otherwise \end{cases}$$

MSD는 원 평가치의 정규화값을 기초로 다음과 같이 계산한다[1].

$$MSD(u,v) = 1 - \frac{1}{|I_u \cap I_v|} \sum_{i \in I_u \cap I_v} (r'_{u,i} - r'_{v,i})^2$$

위 식에서 $r'_{u,i} = (r_{u,i} - r_{\min})/(r_{\max} - r_{\min})$으로서 원 평가치 범위를 [0, 1] 이내의 값으로 정규화한 값이다.

URPA는 세 요소들의 산술평균으로서 유사도값을 측정하지만, 사용자 평가치 패턴을 기존 유사도값에 가중치로 반영하여 유사도를 산출하는 또다른 방식을 제안한다. 이를 URP_B로 명명하고 아래와 같이 정의한다.

$$URP_B(u,v) = \begin{cases} \frac{1}{2}(ESD(u,v) + PSD(u,v)), & \text{if } I_u \cap I_v = \varnothing \\ \frac{1}{2}(ESD(u,v) + PSD(u,v)) * MSD(u,v), & otherwise \end{cases}$$

마지막으로, Jaccard를 이용하는 기존 방법과[13]의 성능 비교를 위해 다음의 척도를 제안하고 다음 절에 실험 결과를 제시하였다.

$$URP_C(u,v) = \begin{cases} \frac{1}{2}(ESD(u,v) + PSD(u,v)), & \text{if } I_u \cap I_v = \varnothing \\ \frac{1}{3}(ESD(u,v) + PSD(u,v) + J(u,v) * MSD(u,v)), & otherwise \end{cases}$$

Jaccard 계수, J는 다음과 같이 산출한다.

$$J(u,v) = \frac{|I_u \cap I_v|}{|I_u \cup I_v|}$$

IV. Performance Experiments

1. Preparations for Experiments

제안 척도의 성능을 조사하기 위하여 관련 연구 분야에서 빈번하게 활용되는 개방형 데이터셋인 Jester (http://eigentaste.berkeley.edu/)를 이용하였다. 이 데이터셋은 사용자 ID, 항목 ID, 평가등급으로 구성되어 있으며, 표 2에 제시한 바와 같이 평가치 범위가 크고, 사용자의 평가 개수가 많아, 사용자 평가 패턴이 보다 명확히 드러날 수 있다. 희소성 수준(sparsity level)은 1-총평가개수/행렬크기로 산출한다.

Table 2. Characteristics of Jester dataset

Matrix size (users×items)	998×100
Rating scale	-10.0 ~ +10.0
Total number of ratings	70,502
Number of ratings per user (average)	36~100 (70.6)
Sparsity level	0.2936
Recommendation threshold	3.0

실험을 위하여, 전체 평가데이터를 8:2의 비율로 훈련 데이터와 시험 데이터로 나누었다. 협력 필터링 시스템의 성능 조사

www.kci.go.kr

절차는 다음과 같다. 우선, 시스템에서 채택한 유사도 척도에 의해 각 사용자와 유사도가 높은 순서대로 사용자들을 정렬한다. 이들을 해당 사용자의 인접이웃집합이라 부른다. 채택한 유사도 척도의 성능을 검증하는 것이 목적이므로 인접이웃들은 훈련 데이터로부터 구하고, 이들이 제대로 선정되었는지를 시험 데이터를 통하여 성능 기준을 근거로 검증하였다. 구체적으로, 현 사용자가 미평가한 항목에 대한 예측 평가치를 인접이웃들로부터 구하고, 이 예측치가 얼마나 정확한지를 검증한다. 실험 비교할 유사도 척도로는 MSD(Mean Squared Differences), JMSD[13], KWON[18], 그리고 3장에서 제안한 방법들이다. KWON 방식의 실험을 위해 전통 유사도 척도로는 MSD를 사용하였는데, 이는 실험에 사용된 다른 척도들과의 공정한 평가를 위함이다.

기존 연구에서는 성능 평가를 위하여 주로 예측 성능(prediction quality)과 추천 성능(recommendation quality)의 두 가지 관점을 사용해 왔다[7][8]. 예측 성능이란 현 사용자가 미평가한 항목에 대한 시스템의 예측 평가치가 실제치에 얼마나 근접하는지를 평가하는 것으로서, MAE(Mean Absolute Error, 평균절대오차)가 대표적이다. 이 척도는 예측치와 실제치 차이의 절대값 평균으로 구한다. 따라서, 값이 작을수록 시스템 성능이 우수하다.

추천 성능의 대표적인 척도로서 대개 정밀도(precision)와 재현율(recall)을 이용한다. 정밀도는 시스템에서 추천한 항목들, 즉, 예측치가 높은 항목들 중에서 사용자가 실제로 높은 평가치를 부여한 항목들의 비율이다. 따라서 이 값의 산출을 위하여, 기준값(recommendation threshold)이 필요한데, 표 2에서 제시한대로 3.0을 기준으로 하였다. 재현율은 사용자가 기준값보다 높은 평가치를 부여한 항목들 전체 중에서 시스템이 추천한 항목들의 비율이다. 그러므로, 정의에 따르면 추천 항목수가 많아질수록 정밀도와 재현율은 서로 상반된 결과값을 보이게 된다.

본 실험에서는 성능 평가를 위한 또다른 척도로서 순위 기반의 nDCG(normalized discounted cumulative gain)[20]를 사용한다. 이 척도는 추천된 항목들 중에서 사용자의 실제 선호도가 높은 항목들이 얼마나 높은 순위에 올라 있는지를 알아보는 것으로서, 값이 클수록 좋은 성능을 나타낸다.

2. Performance Results

표 3은 인접 이웃 수를 변화하여 실험한 유사도 척도들의 MAE 성능 결과를 나타낸다. 즉, 현 사용자의 미평가항목을 평가한 인접이웃들의 평가치들을 유사도값을 가중치로 하는 weighted sum[2] 방식으로 결합하여 현 사용자의 평가예측치를 산출한 것이다. 평가치 참조를 위해서 현 사용자와의 유사도 값이 높은 순으로 인접이웃들을 선정하는데, 표 3에서 이 숫자를 늘릴수록 모든 척도들의 MAE 결과값은 작아지므로, 보다 많은 이웃의 평가치를 참조하는 것이 바람직함을 알 수 있다. 전반적으로 URP-A가 가장 좋은 결과를 보이며, 근소한 차이지만 URP-B와 MSD가 그 뒤를 이었다. 반면에 공통평가항목 개수를 반영한 Jaccard 지수를 접목했을 때의 성능은 오히려 원래의 MSD 성능보다 저하되었는데, 이는 JMSD와 URP-C의 결과에서 확인할 수 있다. 그 이유는 Jester는 표 2에 제시하였듯이 각 사용자의 평가개수가 매우 많은 밀집된 데이터셋이기 때문인 것으로 판단된다. KWON 방식은 본 연구의 방법과 마찬가지로 사용자 엔트로피를 이용하였지만, 표 3에서 보듯이 예상 밖으로 매우 큰 MAE 결과를 보였다.

그림 2는 각 방법의 추천 성능 결과를 비교하여 나타낸다. 추천 항목수가 변화함에 따라 정밀도 수치는 정의에 따라 감소함을 보이는 반면에 재현율은 증가함을 확인할 수 있다. 두 가지 평가 척도에서 모두 KWON은 MAE 결과에서와 마찬가지로 가장 낮은 성능 결과를 보이고, 이를 다소 개선한 결과를 JMSD와 URP-C에서 볼 수 있다. 따라서 사용자 엔트로피 차이를 결합한 방식 보다는 밀집 데이터셋일지라도 Jaccard 지수를 접목

Table 3. Mean Absolute Error (MAE) of the methods for varying number of nearest neighbors

	5	10	15	20	25	30	35	40	45	50	55	60	65	70
MSD	3.403	3.349	3.334	3.329	3.325	3.323	3.322	3.322	3.321	3.321	3.322	3.324	3.323	3.321
JMSD	3.582	3.456	3.416	3.393	3.383	3.374	3.367	3.361	3.354	3.352	3.349	3.347	3.346	3.344
KWON	3.957	3.829	3.775	3.74	3.719	3.701	3.685	3.674	3.662	3.654	3.648	3.641	3.635	3.629
URP-A	**3.387**	**3.326**	**3.314**	**3.305**	**3.302**	**3.3**	**3.3**	**3.305**	**3.302**	**3.304**	**3.305**	**3.305**	**3.305**	**3.304**
URP-B	3.388	3.338	3.316	3.318	3.309	3.308	3.308	3.31	3.311	3.312	3.314	3.313	3.311	3.31
URP-C	3.587	3.446	3.41	3.388	3.377	3.367	3.362	3.355	3.35	3.348	3.343	3.342	3.341	3.339

Table 4. Normalized Discounted Cumulative Gain (nDCG) of the methods for varying ranks

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
MSD	.800	.807	.817	.829	.838	.850	.854	.859	.863	.869	.873	.877	.883	**.885**	**.890**	**.892**
JMSD	.792	.800	.809	.820	.830	.843	.845	.850	.853	.861	.865	.869	.875	.876	.882	.886
KWON	.761	.770	.779	.791	.804	.817	.820	.826	.830	.835	.837	.845	.847	.853	.856	.864
URP-A	.800	**.809**	.818	.829	.839	**.851**	**.856**	**.861**	**.864**	**.871**	**.874**	**.878**	**.883**	**.885**	**.890**	**.892**
URP-B	**.803**	.807	**.819**	**.830**	**.840**	.850	.85	.860	**.864**	.870	.873	**.878**	.882	**.885**	.889	.892
URP-C	.791	.798	.807	.821	.831	.843	.846	.851	.854	.861	.866	.870	.875	.877	.883	.886

www.kci.go.kr

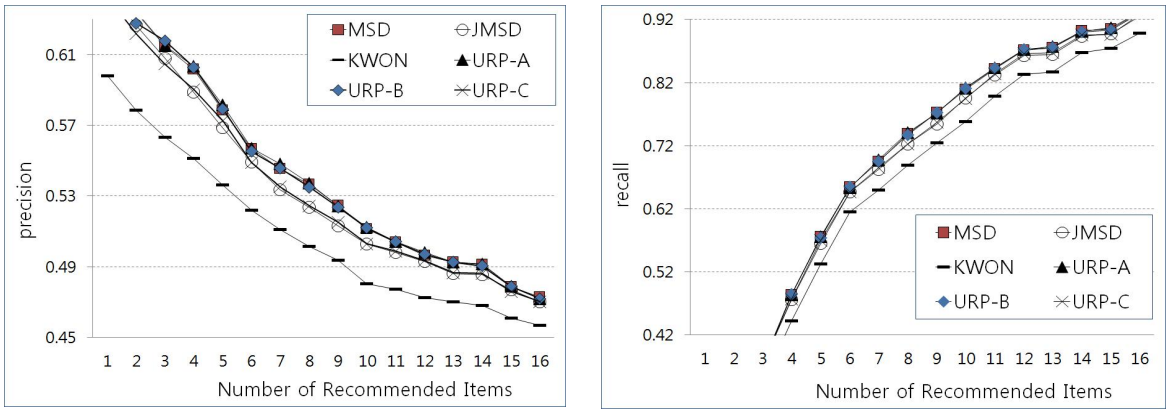

Fig. 2. Recommendation quality of various measures

했을 때의 성능이 더욱 우수함을 알 수 있다. MSD, URP-A, URP-B의 추천 성능은 거의 동일한 결과를 보이며 다른 방법들보다 가장 우수하다.

비록 추천 성능이 거의 동일할지라도, 추천 순위 측면에서 어떠한 방법이 우세한지를 살펴볼 필요가 있다. 이에 표 4는 추천 항목 개수를 변화시켰을 때 각 방법의 nDCG 결과를 나타낸다. 예를 들어, 5개의 추천 항목에서는 매우 근소한 차이지만 URP-B가 .840으로서 가장 우수한 성능을 보였다. 대체로 URP-A의 순위 성능이 다른 척도들을 미세하나마 능가하는데, 이러한 결과는 MAE와 마찬가지이다. 그러므로, URP-A는 MSD나 URP-B보다 사용자 선호에 더욱 부합하는 순위의 항목들을 추천함을 알 수 있다. 또한 KWON은 이 평가 척도에서도 가장 낮은 성능을 보임을 알 수 있다.

결론적으로 본 연구의 제안 방법은 두 사용자의 공통평가항목개수가 적다고 할지라도 신뢰할만한 유사도값을 산출하는 것으로 실험을 통해 입증되었다. 즉, 공통평가항목수를 접목하여 기존의 문제점을 해결하고자 한 JMSD 보다 여러 가지 성능 평가 기준에서 우수하였고, 엔트로피를 이용한 KWON 보다도 월등히 좋은 결과를 보였다. 특히, 본 제안 방법 또한 KWON과 마찬가지로 엔트로피를 이용함에도 불구하고 이러한 결과를 초래한 것은 주목할 만하다. 제안 방법은 사용자의 평가등급을 입력받는 모든 협력 필터링 기반의 추천 시스템에 적용할 수 있으며, 직접적인 입력 기능을 제공하지 않는 경우, 사용자의 화면 지체 시간, 클릭 수 등의 간접적 의도로부터 평가등급을 파악하여 적용할 수 있다. 다만, 본 실험에서 사용한 Jester 데이터셋은 매우 밀집된 평가데이타를 갖고 있어서, 사용자 평가 패턴을 위한 확률이나 엔트로피 산출이 유효할 수 있으므로, 다른 특성의 데이터셋을 활용하여 제안 방법의 성능 실험이 추가적으로 필요하다.

V. Conclusions

본 논문에서는 협력 필터링을 활용한 추천 시스템의 성능을 개선하기 위해 인접 이웃들을 선정하는 새로운 유사도 척도를 제안하였다. 사용자들의 공통 평가 항목들에 대한 평가치를 주요 기준으로 하는 기존의 유사도 척도와는 달리, 제안 방법은 사용자의 평가치 패턴을 기준으로 하기 때문에, 본질적으로 데이터 희소성 문제를 갖는 메모리 기반의 협력 필터링 기법의 단점을 해소하기에 보다 유리하다. 제안 척도의 성능을 파악하기 위하여, 관련 연구에서 널리 사용되는 데이터셋을 이용하여 실험하였고, 다양한 성능 평가 척도들을 기준으로 하여 기존의 유사도 척도들의 성능과 비교하였다. 결과적으로, 예측 성능과 순위 성능 측면에서 제안 방법은 가장 우수하였고, 다른 측면에서는 기존과 대등한 성능 결과를 보였다. 향후, 데이터 밀집 정도가 다른 데이터셋을 이용하여 다른 종류의 유사도 척도들과도 비교 실험하는 연구가 필요할 것으로 본다.

REFERENCES

[1] S. Du, H. Zhang, H. Xu, J. Yang, and O. Tu, “To Make the Travel Healthier: A New Tourism Personalized Route Recommendation Algorithm,” Journal of Ambient Intelligence and Humanized Computing, Vol. 10, No. 9, pp. 3551-3562, 2019.

[2] X. Su and T.M. Khoshgoftaar, “A Survey of Collaborative Filtering Techniques,” Advances in Artificial Intelligence, 2009.

[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.

[4] U. A. Anusha and S. Biradar, “Recommender Systems: A Survey,” International Journal of Latest Technology in Engineering, Management & Applied Science, Vol. V, No. I, pp. 42-45, 2016.

[5] R. Prasad and V. V. Kumari, “A Categorical Review of Recommender Systems,” Inernational Journal of Distributed and Parallel Systems, Vol. 3, No. 5, pp. 73-84. 2012.

www.kci.go.kr

[6] K. Madadipouya and S. Chelliah, “A Literature Review on Recommender Systems Algorithms, Techniques and Evaluations,” BRAIN: Broad Research in Artificial Intelligence and Neuroscience, Vol. 8, No. 2, pp. 109-124, 2017.
[7] F. Cacheda, V. Carneiro, D. Fernandez, and V. Formoso, “Comparison of Collaborative Filtering Algorithms: Limitations of Current Techniques and Proposals for Scalable, High-performance Recommender Systems,” ACM Transactions on the Web, Vol. 5, No. 1, pp. 1-33, 2011.
[8] 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, 2016.
[9] J. Bobadilla, F. Ortega, and A. Hernando, “A Collaborative Filtering Similarity Measure based on Singularities,” Information Processing and Management, Vol. 48, No. 2, pp. 204-217, 2012.
[10] 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.
[11] C.C. Chen, Y.H. Wan, M.C. Chung, and Y.C. Sun, “An Effective Recommendation Method for Cold Start New Users using Trust and Distrust Networks,” Information Sciences, Vol. 224, pp. 19-36, 2013.
[12] G. Koutrica, B. Bercovitz, and H. Garcia, “FlexRecs: Expressing and Combining Flexible Recommendations,” Proc. of the ACM SIGMOD Int'l Conf. on Management of Data, pp. 745-758, 2009.
[13] J. Bobadilla, F. Serradilla, and J. Bernal, “A New Collaborative Filtering Metric that Improves the Behavior of Recommender Systems,” Knowledge Based Systems, Vol. 23, No. 6, pp. 520-528, 2010.
[14] 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.
[15] S. Lee, “Improving Jaccard Index for Measuring Similarity in Collaborative Filtering,” Lecture Notes in Electrical Engineering, Vol. 424, pp. 799-806, 2017.
[16] M. Li and K. Zheng, “A Collaborative Filtering Algorithm Combined with User Habits for Rating,” International Conference on Logistics Engineering, Management and Computer Science, pp 1279-1282, 2015.
[17] W. Wang, G. Zhang, and J. Lu, “Collaborative Filtering with Entropy-driven User Similarity in Recommender Systems,” Interational Journal of Intelligent Systems, Vol. 30, No. 8, pp. 854-870, 2015.
[18] H.-J. Kwon, T.-H. Lee, J.-H. Kim, and K.-S. Hong, “Improving Prediction Accuracy using Entropy Weighting in Collaborative Filtering,” Symposia and Workshops on Ubiquitous, Autonomic and Trusted Computing, pp. 40-45, 2009.
[19] S. Lee, “Using Entropy for Similarity Measures in Collaborative Filtering,” Journal of Ambient Intelligence and Humanized Computing, Feb. 2019.
[20] L. Baltrunas, T. Makcinskas, and F. Ricci, “Group Recommendation with Rank Aggregation and Collaborative Filtering,” Proceedings of the ACM Conference on Recommender Systems, pp 119-126, 2010.

Authors

Soojung Lee received the B.S. degree in Mathematics Education from Ewha University, Korea in 1985. She received M.S. and Ph.D. degrees in Computer Science from Texas A&M University, U.S.A, 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