JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 24 No. 12, pp. 93-100, December 2019
https://doi.org/10.9708/jksci.2019.24.12.093

Using Genre Rating Information for Similarity Estimation in Collaborative Filtering

Soojung Lee*

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

[Abstract]

Similarity computation is very crucial to performance of memory-based collaborative filtering systems. These systems make use of user ratings to recommend products to customers in online commercial sites. For better recommendation, most similar users to the active user need to be selected for their references. There have been numerous similarity measures developed in literature, most of which suffer from data sparsity or cold start problems. This paper intends to extract preference information as much as possible from user ratings to compute more reliable similarity even in a sparse data condition, as compared to previous similarity measures. We propose a new similarity measure which relies not only on user ratings but also on movie genre information provided by the dataset. Performance experiments of the proposed measure and previous relevant measures are conducted to investigate their performance. As a result, it is found that the proposed measure yields better or comparable achievements in terms of major performance metrics.

▸ **Key words**: Collaborative Filtering, Recommender System, Similarity Measure, Data Sparsity Problem, Cold-start problem

[요 약]

유사도 계산은 메모리 기반 협력필터링 시스템의 성능에 매우 중요하다. 이 시스템들은 사용자 평가치들을 이용하여 온라인 상업 사이트에서 고객들에게 상품을 추천한다. 더욱 적합한 추천을 위해 현 사용자와 가장 유사한 사용자들을 선정하여 참조한다. 기존 문헌에는 많은 유사도 척도들이 개발되었는데, 이들은 대개 데이터 희소성이나 완전 시작 문제를 내포하고 있다. 본 논문에서는 기존 척도들과는 달리 사용자 평가치들로부터 선호 정보를 최대한 추출함으로써 희소한 데이터 조건에서도 더욱 신뢰할 수 있는 유사도값을 산출하고자 한다. 사용자 평가치 뿐만 아니라 데이터셋이 제공하는 영화장르 정보를 이용하는 새로운 유사도 척도를 제시한다. 본 척도와 기존의 관련된 척도들의 성능 실험을 하였고, 그 결과, 제안 척도는 주요 성능 평가기준 상으로 더욱 우수하거나 유사한 성능 결과를 보임을 확인하였다.

▸ **주제어**: 협력 필터링, 추천 시스템, 유사도 척도, 데이터 희소성 문제, 완전 시작 문제

• First Author: Soojung Lee, Corresponding Author: Soojung Lee
*Soojung Lee (sjlee@gin.ac.kr), Dept. of Computer Education, Gyeongin National University of Education
• Received: 2019. 09. 17, Revised: 2019. 11. 21, Accepted: 2019. 11. 23.

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

I. Introduction

오늘날 온라인 상거래 시스템의 사용자들은 시스템에서 제공하는 다양한 상품 추천 기능의 도움으로 원하는 정보를 보다 용이하게 획득할 수 있다. 이러한 기능의 특징은 사용자가 선호한다고 판단되는 상품들 또는 구매하였던 상품들의 기록을 보관하고 이와 연관된 제품들을 추천함으로써 고객들의 소비행위를 유도한다. 도서, 음악, 영화, 가구, 여행 등 다양한 제품군이 추천의 대상이 될 수 있으며, 아마존, 이베이 등과 같은 시스템이 매우 활발하게 서비스되고 있다[1][2].

추천 시스템의 분류는 어떠한 정보를 활용하여 추천하는가를 기준으로 볼 때 크게 내용 기반(content-based), 협력 필터링(collaborative filtering, CF), 하이브리드(hybrid), 인구통계 기반(demographic-based filtering)[3], 신뢰 기반(trust-based) 등으로 나뉜다[4][5]. 각 방식의 특징과 장단점이 존재하나, 많은 인터넷 상업 시스템으로 실제 구현되어 가장 잘 알려진 방식은 협력 필터링이다[4]. 협력 필터링의 기본적인 원리는 다른 사용자들의 상품 평가를 종합하여 추천 여부를 결정하는 것이다. 따라서, 사용자의 프로필 정보나 상품 특성 및 내용 등의 어떠한 문맥 정보도 필요로 하지 않는 대신 직접적 또는 간접적 평가 기록만을 추천에 활용한다.

본 연구는 현 사용자와 유사한 다른 사용자들의 평가 기록을 참조하는 사용자 기반의 협력 필터링(user-based collaborative filtering) 기법에 초점을 둔다. 이 기법은 항목 기반(item-based) CF와 더불어 이제까지 많은 연구자들이 관심을 보여 왔다. 유사성의 측정은 CF 시스템의 성능에 매우 중요한 문제인데, 이는 유사한 사용자들, 즉, kNN(k-Nearest Neighbors)[5], 의 평가 기록으로부터 특정 항목에 대한 추천 여부가 결정되기 때문이다.

CF 시스템에서 유사도 측정에 사용되는 상관 기반과 벡터 기반의 전통적인 유사도 척도 외에 척도 개발을 위한 많은 연구가 진행되어 왔다. 이 척도들은 대개 사용자들의 평가등급을 활용하되 이에 더하여 평가등급을 이용한 다양한 휴리스틱 척도를 추가하여 유사도를 측정하였다. 그러나, CF 시스템의 가장 큰 단점인 평가데이타 희소성 문제(rating data sparsity problem)[1]를 극복하기엔 한계가 있다. 이 문제는 대부분의 사용자들이 시스템에서 제공하는 전체 항목들 중 극히 일부분의 항목들에 대해서만 평가 등급을 부여함으로써 발생한다. 또한 새로운 사용자나 항목에 대하여는 평가 기록이 거의 전무하므로 문제가 발생한다(new user/item problem). 이러한 경우에는 산출된 유사도값의 신뢰도는 저하될 수 밖에 없다.

본 연구에서는 사용자 평가등급 외에 영화 관련 추천 시스템에서 추가적으로 제공하는 영화 특성 정보를 활용하여 유사도를 산출하는 방법을 제안한다. 제안 방법은 평가 기록으로부터 사용자가 어떠한 장르의 영화들을 선호 또는 비선호하는지를 유추하므로, 데이터 희소성 문제를 해결하는데 도움이 될 수 있다. 다양한 실험을 통해 기존 유사도 척도들과 성능을 비교하였고, 그 결과 제안 척도는 여러 가지 성능 기준에 의하면 우수하거나 대등한 성능을 나타냈다.

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

사용자 기반의 협력 필터링 시스템은 사용자-항목 평가 매트릭스(user-item rating matrix)를 구축하여 각 사용자의 각 항목에 대한 평가등급을 저장한다. 두 사용자 간 유사도는 이 매트릭스에 포함된 정보를 기반으로 측정된다. 만약 매트릭스의 크기가 매우 크면 유사도를 측정하기 위한 소요 시간과 공간이 막대하므로, 협력필터링 시스템의 또다른 단점인 확장성 문제(scalability problem)가 야기된다[5]. 본 연구에서는 데이터 희소성 문제를 해결하여 유사도값의 신뢰성을 높일 수 있는 방안에 초점을 둔다.

유사한 사용자들을 구하기 위한 전통적인 유사도 척도로는 피어슨 상관계수(Pearson correlation coefficient), 코사인(cosine), 유클리디안(Euclidean), 평균자승차이(Mean squared differences) 등이 있다[6]. 또한 이들의 확장된 척도로서 조정 코사인(adjusted cosine), constrained Pearson correlation 등이 개발되었다[1]. 피어슨 상관 계수는 가장 많이 연구되고 사용되는 효과적인 척도로 알려져 있는데[1][5], 두 사용자 u와 v 간의 유사도값은 다음과 같은 식으로 구한다. 아래 식에서 Iu,v는 두 사용자의 공통평가항목 집합이며, $r_{u,I}$는 사용자 u의 항목 i에 대한 평가치, $\overline{r_u}$는 Iu,v에 속한 항목들에 대한 사용자 u의 평균 평가치를 나타낸다.

$$COR(u,v)=\frac{\sum_{i\in I_{u,v}}(r_{u,i}-\overline{r_u})(r_{v,i}-\overline{r_v})}{\sqrt{\sum_{i\in I_{u,v}}(r_{u,i}-\overline{r_u})^2}\sqrt{\sum_{i\in I_{u,v}}(r_{v,i}-\overline{r_v})^2}} \quad .....(1)$$

상관도가 아닌 사용자 평가등급들을 벡터로 취급하여 유사도값을 산출하는 코사인 유사도 척도는 다음과 같이 산출된다.

$$COS(u,v)=\frac{\sum_{i\in I_{u,v}} r_{u,i}r_{v,i}}{\sqrt{\sum_{i\in I_{u,v}} r_{u,i}^2}\sqrt{\sum_{i\in I_{u,v}} r_{v,i}^2}} \quad \text{.....(2)}$$

평균자승차이(MSD)는 원 평가치의 정규화값을 기초로 하는데, $r_{u,i}$의 정규화값 $r'_{u,i}$은 시스템에서 허용한 원 평가치 범위 [r_{min}, r_{max}]를 [0, 1] 이내의 값으로 변환시킨 것이다. 즉, $r'_{u,i}=(r_{u,i}-r_{\min})/(r_{\max}-r_{\min})$이며 MSD는 다음과 같이 계산한다.

$$MSD(u,v)=1-\frac{1}{|I_{u,v}|}\sum_{i\in I_{u,v}}(r'_{u,i}-r'_{v,i})^2 \quad \text{.....(3)}$$

위 척도들은 두 사용자 간에 공통된 평가항목이 존재할 때에만 의미가 있으므로, 그 수가 매우 적을 때에는 낮은 신뢰도의 유사도값이 산출된다. 언급한 척도들의 장단점 및 성능 비교 결과는 Saranya 외 2인의 연구 결과를 참조할 수 있다[6].

공통평가항목수가 적은 경우에 유사도 산출값이 부정확할 가능성이 크므로, 공통평가 항목들에 대한 평가치 이외에, 평가치로부터 다른 정보를 유추하여 유사도 산출에 반영하려는 시도가 행해졌다. Hafshejani 외 2인은 사용자의 개인적 특성을 클러스터링하는 방법으로 데이터 희소성을 극복하려는 연구를 행하였고[7], Koohi와 Kiani는 평가치를 선호 범주부터 비선호 범주까지의 네 가지로 분류하여 클러스터링함으로써 희소성과 고차원의 문제를 해결하려 하였다[8]. Zhu 외 3인은 평가치들의 수치적 관련성 외에도 비수치적 구조 정보를 알아내어 유사도 산출에 활용하였다[9].

한편 사용자 평가치들의 엔트로피를 참조함으로써 유사도값의 산출 정확도를 개선하려는 노력이 시도되었는데, 대개 전통적 유사도 척도와 결합하는 방식으로 사용되었다. 예를 들어, 각 사용자 기준으로 엔트로피를 산출한 Li와 Zheng의 연구에서는 피어슨 상관도와 결합하였고[10], Wang 외 2인의 연구에서도 사용자 엔트로피의 상대적 차이를 피어슨 상관도와 결합하였다[11]. 반면에 각 항목에 부여된 평가치들의 엔트로피를 구하여 유사도 측정에 활용한 Lee의 연구에서는 각 항목에 대한 전체 평가치들의 패턴을 반영하여 전통적 유사도 척도의 성능 향상을 도모하였다[12].

그러나, 데이터 희소성을 극복하기 위한 가장 단순하고 효율적인 방법은 자카드(Jaccard) 계수[13]를 이용한 유사도 척도의 개발이라고 할 수 있다. 이 계수는 유사도 측정 대상인 두 사용자의 전체 평가항목개수 대비 공통평가항목개수의 비율을 말한다. 대부분의 기존 연구에서는 이 계수를 전통적인 유사도 척도와 결합하여 시스템의 성능 향상을 시도하였다. Bobadilla 외 2인이 개발한 유사도 척도는 자카드 계수와 평균자승차이를 결합한 형태이고 기존 척도들에 비해 성능이 향상되었음을 보고하였다[14]. Zhu 외 3인은 자카드 계수의 개념을 확장하여 평가치들로부터 비수치적 구조 정보를 알아내고, 이들 정보를 결합한 새로운 유사도 척도를 제안하였다[9]. 또한 Lee의 연구에서는 자카드 계수 자체의 성능을 향상시킨 개선된 계수를 개발하여 실험을 통하여 성능을 입증하였는데, 이 계수의 근본 아이디어는 자카드 계수를 각 평가치 구간별로 별도로 계산하여 결합한 것이다[15].

III. The Proposed Scheme

1. Assumptions

본 연구는 사용자가 추천 시스템에서 제공하는 항목들에 대해 직접적으로 평가등급을 부여하는 경우를 가정한다. 평가등급에 관해서는 정수 또는 실수 등 모든 타입과 범위를 허용한다. 만약 사용자의 직접적인 평가등급의 입력 기능을 제공하지 않는 시스템이라면, 사용자의 항목에 대한 선호 여부를 파악하기 위한 다양한 방법을 활용하여 평가등급을 유추할 수 있다. 예를 들어, 사용자의 구매 이력, 항목 클릭 여부 등 사용자의 온라인 행태로부터 유추한다[1].

일반적으로 추천 시스템의 항목 종류는 음악, 도서, 영화, 전자제품 등 매우 다양한데, 본 연구에서는 항목들을 장르별로 구분할 수 있으며, 그와 같은 정보를 제공하는 시스템을 가정한다. 예로서, 영화 추천 시스템이라면, 특정 영화가 코미디, 로멘틱, 판타지, 공포, 액션 등의 장르들 중에서 어느 장르에 속하는지의 정보를 유지관리할 수 있어야 한다. 결론적으로, 본 연구에서 활용하는 정보는 사용자 ID, 항목 ID, 장르 ID, 평가등급이다.

2. Genre Similarity

사용자 u가 평가한 항목은 한 개 또는 여러 개의 장르에 속할 수 있다. 만약 사용자가 '액션' 장르에 속한 영화를 '판타지' 영화보다 더 많이 평가했다면, 이 사용자는 당연히 '액션' 장르에 대한 선호도가 더 높다고 할 수 있다. 따라서, 본 연구에서는 각 장르별 선호도를 해당 장르에 속한 항목들의 평가개수로 간주하고 이를 사용자 간의 유사도 산출에 활용한다.

사용자 u의 장르 g에 대한 상대평가도수를 $f_{u,g}$라고 표기하자. 전체항목집합을 I, 사용자 u의 항목 i에 대한 평가

치를 $r_{u,i}$, 그리고 전체 장르 집합을 G라고 할 때, $f_{u,g}$는 다음과 같이 정의한다.

$$f_{u,g} = \frac{|\{i \in I | r_{u,i} \neq NULL \text{이며} i \text{는 장르} g \text{에 속함.}\}|}{\sum_{g' \in G} |\{i \in I | r_{u,i} \neq NULL \text{이며} i \text{는 장르} g' \text{에 속함.}\}|} \quad \text{.....(4)}$$

표 1은 위 식에 대한 간단한 예시 자료를 제시하고 있다. 사용자 u는 세 항목, i1, i2, i3를 평가하였으며, g1~g5의 다섯 개 장르 중에서 각 항목이 속한 장르는 ∨ 표시로 나타냈다. 이 예시에서 사용자는 g1을 가장 선호하며 g2를 가장 비선호함을 알 수 있다.

Table 1. Illustration of relative genre frequency

item \ genre	g1	g2	g3	g4	g5
i1	∨		∨		
i2	∨			∨	
i3	∨			∨	∨
fu,g	3/7	0/7	1/7	2/7	1/7

위와 같이 정의한 평가장르 상대도수를 기반으로 두 사용자 간의 평가장르 유사도를 측정한다. 이 유사도를 GPSIM'이라고 표기하고, 각 장르별 상대도수 제곱차이의 평균으로 정의한다. 구체적으로, Gu를 사용자 u가 평가한 장르 집합이라고 할 때, 다음 식을 이용한다.

$$GPSIM'(u,v) = 1 - \frac{1}{|G_u \cup G_v|} \sum_{g \in G_u \cup G_v} (f_{u,g} - f_{v,g})^2 \text{....(5)}$$

위 식에서는 각 장르를 동일하게 취급하였으나, 만일 한 사용자가 특정 장르를 다른 장르보다 더욱 선호한다면, 그 선호 장르에 대한 도수 차이가 작을수록 평가장르 유사성 측면에서 더욱 유사한 사용자라고 말할 수 있다. 즉, 선호 장르에 대한 가중치를 상대적으로 높게 책정함이 타당할 것이다. 따라서, 가중치가 부과된 평가장르 유사성은 다음과 같이 정의하며, 이 때 가중치는 해당 장르에 대한 상대 평가도수로 정한다.

$$GPSIM(u,v) = 1 - \frac{\sum_{g \in G_u \cup G_v} \max(f_{u,g}, f_{v,g}) * (f_{u,g} - f_{v,g})^2}{\sum_{g \in G_u \cup G_v} \max(f_{u,g}, f_{v,g})} \quad \text{.....(6)}$$

3. User Similarity

본 절에서는 앞 절에서 정의한 장르 유사도를 전통적 유사도 척도와 결합한, 새로운 사용자 간 유사도 척도를 제안한다. 이 때 두 구성요소의 상대적인 중요도를 결정하는 파라미터를 활용하고, 그 최적값을 실험을 통하여 알아보기로 한다. 장르 유사도 GPSIM과 피어슨 상관도(COR)를 결합한 척도를 GPSIM*COR로 표기하고 아래와 같이 정의하였다.

$$GPSIM^* COR(u,v) = \alpha^* GPSIM(u,v) + (1-\alpha) * COR(u,v),\ 0 < \alpha < 1 \quad \text{.....(7)}$$

이외에도 [14]에서 제안한 JMSD와 결합한 척도를 GPSIM*JMSD로 표기하고 다음과 같이 정의한다.

$$GPSIM^* JMSD(u,v) = \beta^* GPSIM(u,v) + (1-\beta) * JMSD(u,v),\ 0 < \beta < 1 \quad \text{.....(8)}$$

IV. Performance Experiments

1. Experimental Background

성능 실험을 위한 데이터셋은 관련 연구에서 널리 활용되는 MovieLens 1M(http://www.grouplens.org)를 택하였다. 이 데이터셋은 3952개의 영화에 대한 사용자들의 1부터 5까지의 정수 평가등급을 제공한다. 또한 사용자 평가등급 외에 영화를 18 가지의 장르로 분류하여 정보를 제공하므로, 본 연구에서 제안한 유사도 척도의 성능 실험에 적합하다. 원래는 6040명의 사용자들의 평가 데이터를 갖고 있으나, 본 실험에 사용된 PC의 용량 및 처리 속도 등이 수용할 수 있는 최대 크기라고 판단되는 3000명의 사용자들에 대해서만 실험을 진행하였다. 이와 같이 축소 데이터를 활용하는 실험 방법은 기존 연구에서도 행해져 왔던 것으로서, 이를테면 Salehi 외 2인에서는 과거 연구들이 적절한 시간 내에 결과를 추출하기 위하여 원데이터셋의 0.03~8% 만을 활용하여 실험을 진행했다고 언급하였다[16]. 본 연구에서 선정한 3000명의 실험대상자들은 원데이터에서 제공된 사용자 아이디 1~3000번까지이며, 이들 아이디의 부여 방법 상에는 어떠한 특정 조건도 적용되지 않는다. 따라서 3000명의 선정된 사용자들은 임의 선정이라고 할 수 있으므로, 실험 결과의 정당성을 보장할 만하다. 이에 더하여 선정 인원들을 제외한 나머지 데이터로도 실험을 진행하였고, IV.3 절에서 기술할 결과들과 유사한 양상을 보임을 확인하였다.

유사도 척도의 성능을 확인하기 위하여 수행하는 실험 절차는 다음과 같다. 첫째, 전체 평가데이타 집합을 훈련데이타 집합과 시험데이타 집합으로 나눈다. 본 실험에서는 통상적인 8:2의 비율을 적용하였다. 둘째, 모든 사용자 쌍 간의 유사도값을 미리 선정한 유사도 척도에 의거하여 산출한다. 셋째, 시험데이타 집합 내의 각 사용자 u에 대해 두 번째 단계에서 구한 유사도값의 내림차순 정렬 결과로서, 가장 유사한 이웃

(최인접 이웃, Nearest Neighbors)들을 구한다. 넷째, 사용자 u가 미평가한 항목 x에 대해 평가 예측치를 구한다. 본 실험에서는 예측치 산출을 위한 대표적인 weighted sum 방법으로서, Resnick's formula[1]를 활용하였다. 이 방법은 최인접이웃들의 x에 대한 평가치를 합하여 평가 예측치를 구하는데, 이 때 사용자 u와 이웃의 유사도가 클수록 해당 이웃의 평가치에 더 높은 가중치를 부여한다. 결과적으로 평가예측치가 정확할수록, 즉, 실제 평가치에 근접할수록, 선정한 유사도 척도의 성능이 우수한 것으로 판단한다.

2. Performance Metrics

구체적인 성능 평가 기준으로서, MAE(Mean Absolute Error, 평균절대오차)와 F1을 선택하였다. 이 두 가지는 관련 연구에서 많이 활용되는 대표적인 기준이다[6][17]. MAE는 미평가 항목에 대해 산출한 평가 예측치의 정확성을 판단하는 기준으로서, 예측치와 실제치 차이의 절대값 평균, 즉, $\frac{1}{n}\sum_{u}\sum_{i}|r_{u,i}-r'_{u,i}|$로 구한다 ($r'_{u,i}$는 예측치).

협력 필터링 시스템에서는 현 사용자가 미평가한 항목에 대해 평가예측치를 구한 후, 예측치가 추천 기준값(recommendation threshold) 보다 크면 이 항목을 추천한다. MovieLens 데이터셋에서 허용한 평가값이 1부터 5 사이의 정수임을 감안하여, 본 실험에서 사용한 추천 기준값은 4로 정하였다.

두 번째 성능평가 기준인 F1은 정밀도(precision)와 재현율(recall)의 조화평균이다. 정밀도란 시스템이 추천한 모든 항목들 중에서 사용자의 실제 평가치가 추천 기준값 이상인 항목들의 비율이다. 재현율은 사용자의 실제 평가치가 추천 기준값 이상인 모든 항목들 중에서 시스템이 실제로 추천한 항목들의 비율을 말한다.

성능을 비교할 유사도 척도들은 본 연구의 제안 방법의 기반이 되는 척도들을 선정하여, 성능 향상 정도를 파악하였다. 따라서, 피어슨 상관도(COR), 평균자승차이(Mean Squared Differences), JMSD[14], GPSIM*COR, GPSIM*JMSD를 각각 적용한 협력 시스템의 성능을 측정하였다.

3. Results

3.1 Effect of Parameters

본 연구의 제안 척도에서는 파라미터를 이용하여 장르 선호 유사도와 전통 유사도 척도의 비중을 결정하였다. 따라서, 파라미터, 즉, α와 β 값의 변화에 따라 성능이 어떻게 달라지는지 관찰하였다.

그림 1에서는 α값을 0.1부터 0.9까지 변화시킨 후 측정한 GPSIM*COR 방법의 성능을 제시하였다. 최인접 이웃 수가 커짐에 따라 MAE는 α값과 무관하게 점차적으로 감소하여 안정된 상태에 도달함을 알 수 있다. 또한 F1 값은 점차 증가하는 양상을 나타내므로, 더욱 많은 인접 이웃들을 참조할수록 시스템의 추천 성능은 향상됨을 보인다. α값의 변화에 따른 성능을 살펴보면, 대체로 α값이 클수록 두 가지 측면의 성능은 매우 근소한 차이로 더욱 우수하나, α=0.9일 때 상대적으로 좀 더 큰 차이를 보였다. 특히 MAE에서보다 F1 성능에서 α값의 변화가 더욱 큰 영향을 주었다. 따라서 식 (7)로부터, GPSIM이 COR 보다 성능 향상에 더 크게 기여함을 알 수 있다. 이는 GPSIM이 식 (6)에서와 같이 사용자들의 평가 절대치가 아닌 평가 장르 빈도수만을 입력값으로 함에도 불구하고 얻어진 것이므로 예상 밖의 주목할만한 결과이다.

그림 2는 GPSIM을 JMSD와 각각의 비중을 달리하여 결합했을 때의 성능 결과이다. 그림 1의 결과에서와 거의 마찬가지로, β값의 차이에 따른 성능 차이는 무시할 만하다. 다만, β=0.9일 때 다른 결과들에 비해 약간의 차이를 보이는데, 특히 F1의 성능에서 더욱 그러하며 더 우수한 성능을 보임을 알 수 있다.

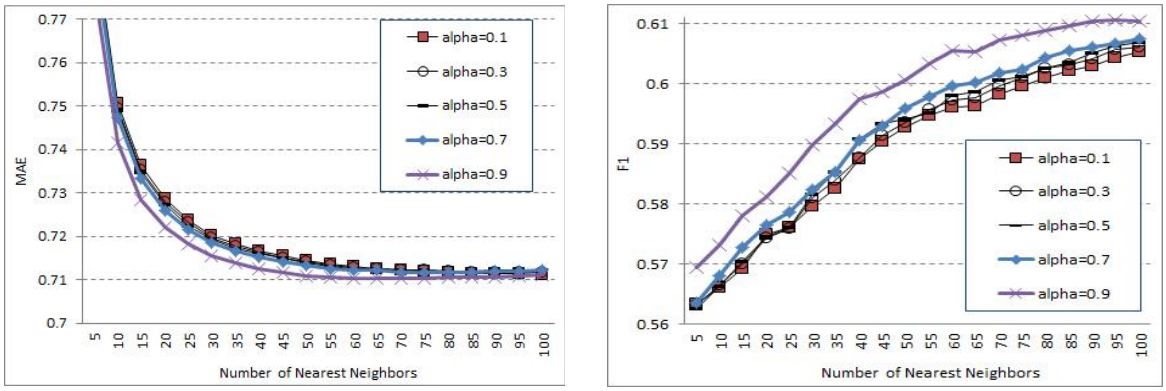

Fig. 1. Performance of GPSIM*COR with varying alphas.

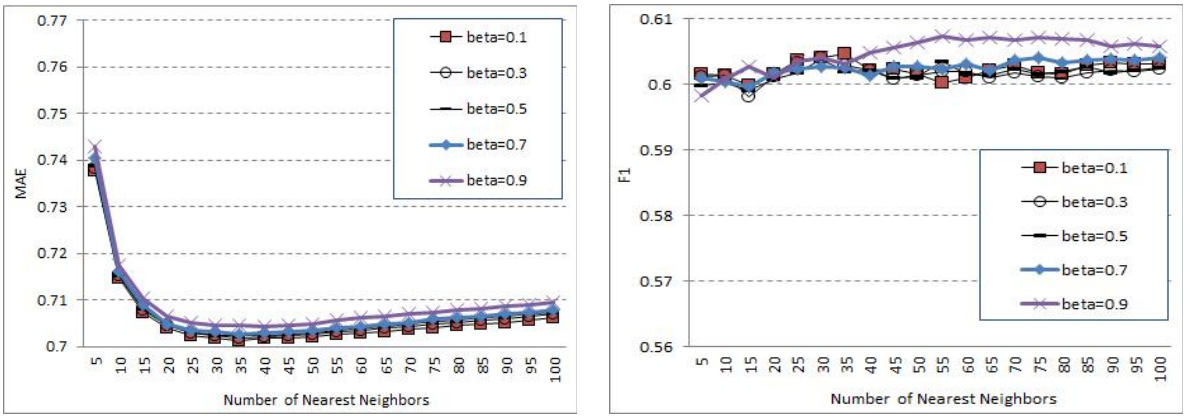

Fig. 2. Performance of GPSIM*JMSD with varying betas.

그러나, MAE 성능은 가장 큰 β값을 이용하였을 때 가장 저조하였고, 그 차이는 매우 근소하다. 결론적으로, GPSIM*JMSD 방법에서도 GPSIM의 성능 기여도가 JMSD 보다 더욱 큰 것으로 확인되었다.

위와 같은 결과를 토대로, 다음 절에서는 가장 높은 성능치를 보여준 파라미터값들을 활용하여 실험을 진행하였다.

3.2 Performance of Similarity Measures

본 실험에서 선택한 비교 대상의 유사도 척도들을 각기 적용한 협력 필터링 시스템의 성능을 그림 3에 제시하였다. 우선 MAE 결과는 크게 세 그룹으로 나뉠 수 있는데, MSD가 가장 낮은 성능을 보이며, 그 다음으로 COR와 GPSIM*COR, 그리고 마지막 그룹인 GPSIM*JMSD와 JMSD이다. 관련 연구 분야에서 널리 활용되는 대표적인 척도인 COR는 MSD 보다 월등히 좋은 결과를 보였다. 또한 장르 선호 유사도를 결합시킴으로써 성능 개선 효과를 가져올 수 있었으나, 이러한 효과는 평가치를 참조한 인접 이웃수가 증가할수록 점차적으로 감소하였다. 이로써 사용자의 평가 절대치를 활용하지 않는 장르 선호도 측정방법이 기존의 COR의 MAE 성능을 넘어서는 예측 정확도를 달성하지는 못하는 것으로 판단된다.

그림 3에서 JMSD는 [14]에서 보고하였듯이 MSD의 성능을 매우 혁신적으로 개선한 것을 알 수 있다. 이는 실험에 사용된 MovieLens 데이터셋의 특성상 각 사용자 당 평가개수가 19~1517개로서 변동이 매우 크고, 평균 154개의 사용자 당 항목 평가개수를 가지므로 전체 항목수에 비해 매우 적지만, 단순 평가치 차이만을 활용하는 MSD 척도는 근본적으로 이같은 특성을 반영하지 못하므로 적절한 인접 이웃을 구하기에 부적합하기 때문이다. 따라서 두 사용자 간 공통평가항목수를 반영한 자카드 계수를 결합한 JMSD 방식은 이와 같은 단점을 보완하기에 매우 유효한 것으로 나타났다.

JMSD의 성능이 GPSIM과 결합함으로써 더욱 개선될 수 있는지를 알아보기 위하여, GPSIM*JMSD의 결과를 산출하였다. 그림 3에서 보듯이, 성능 개선은 이루어지지 않았으며, 오히려 더많은 인접이웃들의 평가치를 참조할수록 극도로 미세하지만 MAE 성능의 저하가 나타났다. 그러므로 실험 데이터셋과 같은 매우 희소한 데이터 환경에서는

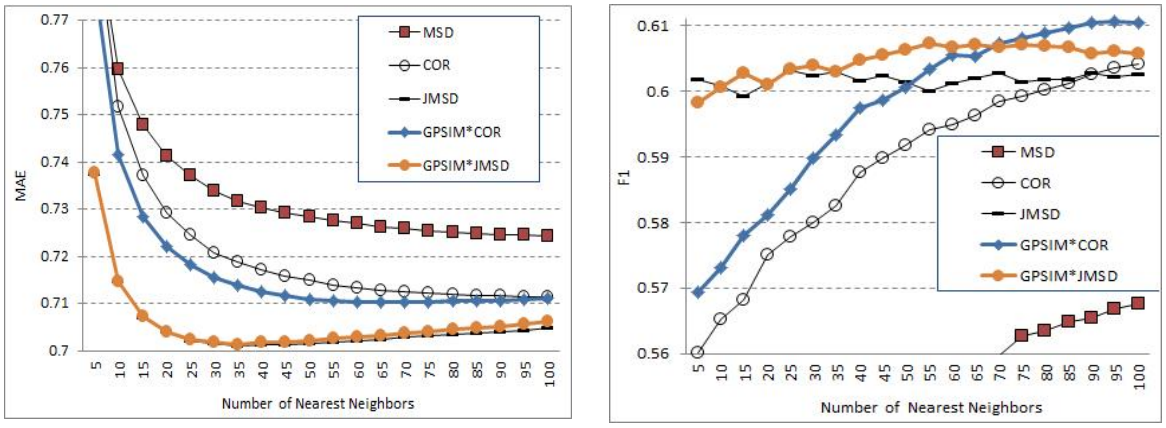

Fig. 3. Performance of various similarity measures.

예측치 정확도의 측면에서는 자카드 계수의 영향이 크다는 것을 확인하였다.

그림 3에 제시한 F1 결과를 살펴보면, MAE 결과와 거의 유사한 양상을 보인다. 즉, MSD가 상당히 저조한 성능을 나타내며 COR와 GPSIM*COR가 그 뒤를 잇는다. 또한, 이들 세 가지 방법은 인접이웃수가 증가함에 따라 꾸준히 상승하는 추천 성능을 보였다. 그러나 나머지 두 방법들, 즉, JMSD와 GPSIM*JMSD은 인접이웃수와 거의 무관하게 일정한 추천 성능을 나타냈다. 이는 공통평가항목수의 반영으로 말미암아 보다 정확한 예측치를 구하는데 도움이 되는 인접이웃들을 구할 수 있기 때문인 것으로 판단된다. 그러나, 인접이웃수가 55명 이상일 때 JMSD는 GPSIM*COR보다 뒤처지며, 90명을 넘을 때는 COR 보다도 낮은 성능을 보인다. 그러므로, JMSD의 예측 성능, 즉, MAE는 가장 우수하지만, 정확한 예측치를 필요로 하지 않고, 추천 하한선값만을 기준으로 하는 F1의 경우에는 COR 관련 척도가 더 우세함을 알 수 있다. 또한, MAE 결과에서와는 다르게, GPSIM은 JMSD의 성능을 개선하는데 큰 도움이 됨을 확인할 수 있다. 결론적으로, 충분한 수의 인접이웃을 참조한다면 장르 선호도를 반영한 피어슨 상관도가 가장 우수한 추천 성능을 나타냈다.

V. Conclusions

본 연구에서는 메모리 기반의 협력 필터링 시스템의 성능을 향상시키기 위한 새로운 유사도 척도를 제안하였다. 사용자들의 평가등급 기록으로부터 평가등급 간의 유사성 및 영화 장르 선호도를 파악하여 유사도값을 산출하였다. 이 방법은 데이터 희소성 문제를 해결하는데 도움을 줄 뿐만 아니라 사용자 간에 공통평가항목수가 적은 경우에도 보다 신뢰할만한 유사도값을 산출할 수 있다. 실험을 통한 성능 분석에 따르면, 제안 척도는 평가치의 예측 정확도에서 기존과 대등한 우수한 성능 결과를 나타냈고, 추천 성능 측면에서는 가장 우수한 성능을 나타냈다.

제안 방법은 장르 정보를 제공하는 데이터 환경을 가정하였기 때문에, 본문에서 실험한대로 영화 추천 시스템에서 활용 가능하지만, 서비스 항목의 종류가 다를지라도 유사한 항목 특성 정보를 제공하는 모든 시스템에 적용 가능하다. 또한, 장르 특성만을 이용하여 선호도를 측정하였으나, 다른 특성 정보들도 마찬가지 방식으로 선호도 측정 대상이 될 수 있으므로, 제안 방법은 확장성 측면에서 매우 유용하다고 할 수 있다.

REFERENCES

[1] X. Su and T.M. Khoshgoftaar, “A Survey of Collaborative Filtering Techniques,” Advances in Artificial Intelligence, 2009. DOI:10.1155/2009/421425

[2] 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. DOI: 10.1007/s12652-018-1081-z

[3] J. Gupta and J. Gadge, “Performance Analysis of Recommendation System based on Collaborative Filtering and Demographics,” International Conference on Communication Information & Computing Technology, pp. 1-6, 2015. DOI: 10.1109/ICCICT.2015.7045675

[4] 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

[5] 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

[6] 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. DOI: 10.17485/ijst/2016/v9i29/91060

[7] Z. Y. Hafshejani, M. Kaedi, and A. Fatemi, “Improving Sparsity and New User Problems in Collaborative Filtering by Clustering the Personality Factors,” Electronic Commerce Research, Vol. 18, No. 4, pp. 813-836, 2018. DOI: 10.1007/s10660-018-9287-x

[8] Koohi and K. Kiani, “A New Method to Find Neighbor Users that Improves the Performance of Collaborative Filtering,” Expert Systems With Applications, Vol. 83, pp. 30-39, 2017. DOI: 10.1016/j.eswa.2017.04.027

[9] 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

[10] 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. DOI: 10.2991/lemcs-15.2015.255

[11] 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

[12] S. Lee, “Using Entropy for Similarity Measures in Collaborative Filtering,” Journal of Ambient Intelligence and Humanized Computing, Feb. 2019. DOI: 10.1007/s12652-019-01226-0

[13] 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.

[14] 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. DOI: 10.1016/j.knosys.2010.03.009

[15] S. Lee, “Improving Jaccard Index for Measuring Similarity in Collaborative Filtering,” Lecture Notes in Electrical Engineering, Vol. 424, pp. 799-806, 2017. DOI: 10.1007/978-981-10-4154-9_93

[16] 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. DOI: 10.1504/IJBIS.2013.054465

[17] 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. DOI: 10.1145/1921591.1921593

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.