JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 25 No. 5, pp. 99-106, May 2020
https://doi.org/10.9708/jksci.2020.25.05.099

Fuzzy Clustering with Genre Preference for Collaborative Filtering

Soojung Lee*

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

[Abstract]

The scalability problem inherent in collaborative filtering-based recommender systems has been an issue in related studies during past decades. Clustering is a well-known technique for handling this problem, but has not been actively studied due to its low performance. This paper adopts a clustering method to overcome the scalability problem, inherent drawback of collaborative filtering systems. Furthermore, in order to handle performance degradation caused by applying clustering into collaborative filtering, we take two strategies into account. First, we use fuzzy clustering and secondly, we propose and apply a similarity estimation method based on user preference for movie genres. The proposed method of this study is evaluated through experiments and compared with several previous relevant methods in terms of major performance metrics. Experimental results show that the proposed demonstrated superior performance in prediction and rank accuracies and comparable performance to the best method in our experiments in recommendation accuracy.

▸ **Key words**: Collaborative Filtering, Recommender System, Similarity Measure, Scalability Problem, Fuzzy Clustering

[요 약]

협력 필터링 기반의 추천 시스템에 내재된 확장성 문제는 지난 수십년간 관련 연구의 이슈가 되어 왔다. 클러스터링은 이 문제를 해결하는 유명한 기술인데 낮은 성능으로 인하여 활발히 연구되어 오진 않았다. 본 논문에서는 협력 필터링 시스템의 고질적인 단점인 확장성 문제를 극복하기 위하여 클러스터링 기법을 채택하였다. 또한 클러스터링을 적용함으로 인해 초래되는 성능 저하 문제를 개선하기 위해, 두 가지 전략을 사용하였는데, 첫째는 퍼지 클러스터링이며, 둘째는 영화 장르에 대한 사용자 선호도에 기반한 유사도 측정 방법을 제안하고 이를 적용하였다. 본 연구에서의 제안 방법을 기존의 여러 관련 방법들과 비교 실험을 통해 다양한 주요 성능 척도에 의거하여 평가하였는데, 실험 결과 제안 방법은 예측과 순위 정확도 측면에서 더 우수한 성능을 보였고, 추천 정확도 측면에서는 실험 대상 중 최상의 방법과 대등한 성능을 나타냈다.

▸ **주제어**: 협력 필터링, 추천 시스템, 유사도 척도, 확장성 문제, 퍼지 클러스터링

• First Author: Soojung Lee, Corresponding Author: Soojung Lee
*Soojung Lee (sjlee@gin.ac.kr), Dept. of Computer Education, Gyeongin National University of Education
• Received: 2020. 03. 31, Revised: 2020. 05. 12, Accepted: 2020. 05. 12.

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

I. Introduction

오늘날의 정보 홍수 시대에 인터넷 사용자들은 원하는 정보를 획득하는데 많은 시간과 노력을 소비하게 된다. 이러한 어려움을 해결하기 위한 가장 널리 알려진 방법이 추천 시스템(recommender system)이다[1][2]. 추천 시스템은 다양한 분야에서 실제로 상업용 목적 등으로 구현되어 서비스 중인데, 도서, 음악, 여행, 의복 등의 예가 있다.

추천 시스템의 구체적인 구현 방법으로서 내용 기반 필터링(content-based filtering, CBF)과 협력 필터링(collaborative filtering, CF)을 기본으로, 이들의 장단점을 고려한 하이브리드(hybrid), 인구통계 기반(demographic-based filtering)[3], 신뢰 기반(trust-based), 그리고 사회 연결망 기반(social network-based) 등이 개발되었다[2][4].

본 연구에서는 현재 상거래 시스템에서 성공적으로 서비스되고 있는 협력 필터링에 초점을 둔다. 이 방법의 근본 아이디어는 시스템에서 제공하는 항목들에 대해서 현 사용자와 유사한 선호 기록을 가진 다른 사용자들로부터 그들이 선호했던 항목들을 추천하는 것이다. 따라서, 대개 수집에 어려움이 있는 사용자 프로필이나 항목 특성 등의 정보를 보관하지 않아도 된다는 장점이 있다. 반면에, 시스템 사용자들이 매우 적은 경우, 유사한 타 사용자들의 선호 항목을 구하기 어렵게 되고, 반대의 경우에는 유사한 사용자들을 알아내기 위한 비용이 크다는 단점이 있다. 전자의 경우를 데이터 희소성 문제(data sparsity problem), 후자를 확장성 문제(scalability problem)이라고 일컫는다[2]. 이 두 문제는 협력 필터링 시스템의 본질적인 문제이므로 다양한 해결책이 연구되어졌다. 예를 들어 특이값 분해(Singular value decomposition, SVD), 주성분 분석(Principle component analysis, PCA), 베이지안 CF 알고리즘, 클러스터링 CF 등이 있다[1].

클러스터링 CF는 SVD나 PCA가 요구하는 막대한 매트릭스 계산이나 정보 손실을 초래하지 않으므로, 매우 유용한 기술 중 하나이다. 반면에 같은 클러스터 내의 사용자들만으로부터 추천 항목을 결정하므로, 예측 성능을 저하시킬 우려가 있다[1]. 본 연구에서는 성능 저하를 발생시키지 않으면서도 확장성 문제를 해결할 수 있는 클러스터링 CF를 제안한다. 이에 더하여 제안 방법은 퍼지로직을 활용하여 사용자 평가치에 내재한 주관성을 반영하였다. 본 연구는 영화 장르 정보를 보유한 시스템을 가정하였으며, MovieLens, CiaoDVD, Yahoo!Movie 등의 데이터셋이 그러한 시스템의 유명한 예이다. 제안 방법은 평가치 자체 뿐만 아니라 그 문맥을 파악할 수 있도록 사용자의 장르 선호도 측정 방법을 소개하고 클러스터링에 활용한다. 한 영화는 대개 여러 장르에 속하므로, 이는 데이터 희소성을 극복하는데 도움이 될 수 있다. 제안 방법의 성능 평가를 위해 다양한 실험을 실시하였고, 그 결과 기존 방법들에 비해 여러 성능 기준 측면에서 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

사용자들의 클러스터링 작업을 통하여 유사 사용자들은 한 클러스터 내에, 그렇지 않은 사용자들은 다른 클러스터에 속하게 된다. 클러스터링 CF 알고리즘은 현 사용자가 속한 클러스터 내의 다른 사용자들만을 참조하고, 대개 하나의 클러스터 크기는 전체 사용자 수보다 훨씬 작기 때문에, CF 시스템의 확장성 문제를 해결하는 방법이 된다.

K-means는 대표적인 클러스터링 알고리즘으로서 CF 시스템의 성능을 개선하기 위하여 여러 연구에서 활용되어 왔다. Tsai와 Hung은 K-means와 SOM(self-organizing maps)의 앙상블 방법이 단일 클러스터링 기법보다 좋은 추천 성능을 보임을 밝혔다[5]. Nilashi 외 3인은 멀티 기준의 사용자 평가 환경에서 K-means 실행 결과의 클러스터마다 회귀 기술을 이용하여 항목의 평가치를 구하는 방법을 제시하였다[6]. SOM은 또다른 클러스터링 알고리즘으로서 널리 알려졌는데, 신경망 기반의 비감독 학습을 통해 클러스터링을 수행한다. Ye의 연구에서는 연관규칙 마이닝으로써 미평가치를 채운 후 SOM 클러스터링으로 평가치를 예측하였다[7].

K-means와 SOM은 다양하게 활용되고 있지만, 초기 입력값으로의 클러스터 개수를 결정하기 어렵다는 단점이 있다. 또한, K-means는 간결하고 구현이 용이하다는 장점이 있지만 초기 중심값에 의해 성능이 크게 좌우되고 국부적 최적값에 빠질 수 있다는 단점이 있다[1]. 반면에 SOM은 초기 가중치와 종료 조건을 결정하기 어렵다. 이러한 단점을 개선하기 위하여 Liao와 Lee는 클러스터를 자동으로 구성하는 방법을 제안하였다[8].

퍼지 클러스터링은 K-means의 단점을 개선하기 위한 기법으로서, 사용자는 하나 이상의 클러스터에 서로 다른 퍼지 소속값으로서 구성원이 될 수 있다[9][10]. Koohi and Kiani는 피어슨 상관을 활용한 퍼지 C-means(FCM)

를 제안하였으며, 이는 사용자 기반의 협력 필터링을 위한 추천 결과의 성능을 향상시키는 것으로 보고하였다[9]. 사용자 기반이 아닌 항목들을 중심으로 한 퍼지 클러스터링 기법도 발표되었으며[11], Shivhare 외 2인은 유전자 알고리즘을 활용하여 최적의 유사도 척도를 개발하였다[10]. 이밖에 Fremal과 Lecron은 항목 메타데이타 정보를 활용하여 클러스터링을 실시하였고[12], Najafabadi 외 3인은 사용자의 간접 상호작용 정보와 연관규칙 마이닝으로써 노래 데이터를 클러스터링하였다[13].

[1]에서 언급한대로 클러스터링은 CF의 성능을 저하시킬 수 있기 때문에, 이에 대비하기 위해 본 연구에서는 사용자들의 평가치 활용 뿐만 아니라 평가치들로부터 문맥정보를 추출 및 반영한다. 즉, 영화 장르에 대한 선호도를 측정 후 이를 기준으로 클러스터링을 실시한다. 관련 연구로서, Al-Shamri와 M., Bharadwaj는 장르에 속한 영화들에 부여한 높은 평가치의 빈도수와 그 절대수치의 조화평균으로서 장르 선호도를 정의하였다[14]. Lee 외 2인은 장르 선호도를 평균 평가치로 정의하였고 [15], Zhang 외 2인은 각 영화 범주마다 두 사용자의 평가치 차이를 사용자의 선호도라고 정의하였다[16].

III. The Proposed Scheme

1. Estimation of Genre Preference

본 연구에서는 영화에 대한 사용자의 평가치를 유지관리하는 시스템을 가정한다. 이러한 시스템은 학계 및 상업계에서 접할 수 있는데, MovieLens, EachMovie, FilmTrust, Yahoo!Movie 등이 그 예이다. 협력 필터링 시스템은 사용자가 부여한 평가치를 기반으로 하여 영화를 추천하는데, 만약에 평가치 개수가 충분하지 않을 경우 추천의 질이 저하될 수 있다. 이러한 데이터 희소성 문제를 해결하기 위하여 본 연구에서는 영화의 장르 정보를 활용한다.

일반적으로 영화는 하나 이상의 장르에 속한다. 예를 들어, 전형적인 디즈니 영화는 애니메이션, 판타지 또는 모험, 아동용 등의 장르에 포함된다. MovieLens와 같은 데이터셋은 총 18개의 장르를 구성하였는데, 시스템이 제공하는 각 영화는 한 개 이상의 장르에 속하는 것으로 정의하였다. 본 연구에서는 사용자들 간의 장르 선호도가 얼마나 유사한지를 우선 파악한 후, 이를 기준으로 클러스터링을 구축한다. 기존 방법에서처럼 사용자의 항목 평가치를 기준으로 하지 않고 장르 선호도를 기준으로 하면, 평가치 개수가 적을 경우에도 사용자의 선호 맥락을 알 수 있으므로 사용자 간의 유사도를 파악하는데 유리하다고 판단된다. 사용자의 평가치는 클러스터링 작업이 완료된 후에 같은 클러스터 내에서 인접 이웃을 결정하기 위해 활용된다.

제안하는 장르 선호도의 정의를 예를 통하여 설명하기로 한다. 표 1에 제시한 두 사용자 u와 v의 사용자-항목 평가 매트릭스와 표 2의 항목-장르 매트릭스를 바탕으로 각 사용자의 장르 선호도를 추출한다. 표 1에서 7개의 항목을 가정하였고, 사용자들은 1부터 5까지의 평가치를 부여하였으며, - 표시는 평가를 하지 않았음, 즉, NULL을 의미한다. 표 2는 각 항목이 8개의 장르들 중에서 어느 장르에 속하는지를 v 표시로써 나타낸다. 예로서, 항목 i1은 장르 g1에 속하고, i2는 장르 g3과 g4에 속한다.

앞에서 언급한 두 개의 매트릭스로부터 사용자의 각 장르에 대한 선호도는 다음과 같이 산출한다. 사용자 u가 평가한 네 개의 항목들 중에서 g1에 속한 항목 수는 하나이고, g4에 속한 항목 수는 셋이므로, 이 사용자는 g1보다 g4 장르를 더 선호한다고 판단한다. 전체 네 개 평가항목 중 세 개가 g4에 속하므로, g4에 대한 선호도는 3/4이다.

이와 같이 산출한 두 사용자의 장르 선호도를 표 3에 정리하였다. 구체적으로, 전체 항목들의 집합을 I, 사용자 u의 i 항목에 대한 평가치를 $r_{u,i}$ 로 표기할 때, 사용자 u의 장르 g에 대한 선호도, $f_{u,g}$는 아래 식(1)과 같이 계산한다.

$$f_{u,g} = \frac{|\{i \in I | r_{u,i} \neq - \text{이고 } i\text{는 장르 } g\text{에 속함.}\}|}{|\{i \in I | r_{u,i} \neq NULL\}|} \quad \text{.....(1)}$$

Table 1. User-item rating matrix

item / user	i1	i2	i3	i4	i5	i6	i7
u	3	5	1	5	-	-	-
v	-	-	3	4	2	4	2

Table 2. Item-genre matrix

genre / item	g1	g2	g3	g4	g5	g6	g7	g8
i1	v	-	-	-	-	-	-	-
i2	-	-	v	v	-	-	-	-
i3	-	v	v	v	-	-	-	-
i4	-	-	-	v	-	v	v	-
i5	-	-	-	v	v	-	-	v
i6	-	-	-	-	-	v	-	-
i7	-	-	-	-	-	-	-	v

Table 3. User-genre preference matrix

genre / user	g1	g2	g3	g4	g5	g6	g7	g8
u	1/4	1/4	2/4	3/4	0	1/4	1/4	0
v	0	1/5	1/5	3/5	1/5	2/5	1/5	2/5

위와 같이 정의한 장르 선호도는 낮은 항목 평가치를 낮은 장르 선호도라고 판단한 기존의 연구[14][15][16]와는 다르게, 낮은 평가치는 영화에 부여한 것이고, 장르에 부여한 것이 아니라는 가정을 근거로 하여, 장르에 속한 영화 평가치 도수를 계상하였다.

2. Genre Preference Similarity

장르 선호도가 유사한 사용자들을 클러스터로 구축하기 위하여 피어슨 상관도를 활용하여 장르 선호도 간에 유사성을 측정하기로 한다. 따라서 두 사용자 u와 v 간에 장르 선호도에 대한 피어슨 상관계수는 다음 식(2)와 같이 산출한다.

$$GCOR(u,v)=\frac{\sum_{g\in G_{u,v}}(f_{u,g}-\bar{f}_u)(f_{v,g}-\bar{f}_v)}{\sqrt{\sum_{g\in G_{u,v}}(f_{u,g}-\bar{f}_u)^2}\sqrt{\sum_{g\in G_{u,v}}(f_{v,g}-\bar{f}_v)^2}} \quad \text{.....(2)}$$

위 식에서 $G_{u,v}$는 두 사용자의 공통 평가 장르 집합이며 $\bar{f}_u$는 이 집합에 속한 장르들에 대한 u의 선호도 평균값으로서 아래 식(3)과 같이 산출한다.

$$\bar{f}_u=\frac{1}{|G_{u,v}|}\sum_{g\in G_{u,v}}f_{u,g} \quad \text{.....(3)}$$

최종적으로 장르 선호도 기반의 유사도를 산출하기 위해서 앞에서 계산한 GCOR 값에 대해 기하급수 함수를 적용하기로 한다. 이는 피어슨 상관이 높은 사용자들 간에는 더욱 높은 유사도를 갖게 하고, 그 반대의 경우 더욱 낮은 유사도를 갖게 하려는 의도를 반영한 것이다. 그러므로, 최종 유사도 GSIM은 식 (4)와 같고, 값의 범위는 (0,1]이다. α0(0<α0≤1)와 α(α≥1)는 유사도값을 조정하는 파라미터이다.

$$GSIM(u,v)=\alpha_0 \cdot \exp(\alpha(GCOR(u,v)-1)) \quad \text{...(4)}$$

3. Fuzzy Clustering Algorithm

사용자들의 클러스터를 구축하기 위하여, 본 연구에서는 fuzzy C-means 알고리즘을 채택하였다. 이 방법은 기존의 또다른 대표적인 클러스터링 알고리즘인 K-means와 달리, 사용자가 하나 이상의 클러스터에 포함될 수 있도록 한다. 단, 퍼지 개념을 이용하여 각 클러스터마다의 소속값(membership degree)은 서로 다를 수 있다. 소속값은 클러스터의 센터와 사용자의 유사도값으로 결정한다.

Fuzzy C-means 알고리즘을 사용함으로써 시스템 확장성 문제를 극복함과 동시에, 사용자 평가치에 내재한 주관성 또는 모호성까지 고려하는 잇점이 있다. 본 연구에서는 사용자의 장르 선호도를 기준으로 클러스터링 작업을 실시하므로, 유사한 장르 선호도를 가진 사용자들이 하나의 클러스터에 포함된다. 장르 선호도에 대한 유사성은 앞절에서 기술한 GSIM 공식을 활용하여 산출한다. 제안된 클러스터링 알고리즘은 기존의 전형적인 Fuzzy C-means 알고리즘을 토대로 하지만, 클러스터의 센터를 해당 클러스터에 속한 사용자들의 각 장르별 선호도값들로부터 구하며, 장르 선호도값을 기준으로 하여 클러스터링한다는 점에서 차이가 있다. 따라서, 클러스터의 센터는 장르선호도 값들의 집합으로 구성된다. 상세한 클러스터링 알고리즘은 다음과 같으며 표 4에 기호들의 의미를 제시하였다.

1. 각 클러스터 k에 대하여, 각 사용자 u의 소속값 w_{uk}에 0과 1 사이의 난수를 부여한다.
2. while $iter$≤N and 연속된 두 반복 회차 간에 임의의 w_{uk}값의 변화 >ε begin
   2.1 각 클러스터의 센터를 장르 g에 대하여 다음 식(5)으로 계산한다.

$$f_{c_k,g}=\frac{\sum_u (w_{uk})^m f_{u,g}}{\sum_u (w_{uk})^m},\ m\geq 1 \quad \text{.....(5)}$$

   2.2 각 사용자 u에 대하여, 클러스터 k에의 소속값을 다음의 식(6)과 같이 계산한다.

$$w_{uk}=\begin{cases}\dfrac{1}{\sum_{j=1}^{C}\left(\dfrac{\|u-c_k\|}{\|u-c_j\|}\right)^{\frac{2}{m-1}}}, & \text{if } |G_{u,c_k}|>1\\ 0, & otherwise\end{cases} \quad \text{.....(6)}$$

   위 식에서

$$\|u-c_j\|=\begin{cases}1-\dfrac{\sum_{g\in G_{u,c_j}}(f_{u,g}-\bar{f}_u)(f_{c_j,g}-\bar{f}_{c_j})}{\sqrt{\sum_{g\in G_{u,c_j}}(f_{u,g}-\bar{f}_u)^2}\sqrt{\sum_{g\in G_{u,c_j}}(f_{c_j,g}-\bar{f}_{c_j})^2}}, & \text{if } |G_{u,c_j}|>1\\ 1, & otherwise.\end{cases} \quad \text{.....(7)}$$

$$\bar{f}_u=\frac{1}{|G_{u,c_j}|}\sum_{g\in G_{uc_j}}f_{u,g}, \quad \text{.....(8)}$$

$$\bar{f}_{c_j}=\frac{1}{|G_{u,c_j}|}\sum_{g\in G_{uc_j}}f_{c_j,g}. \quad \text{..... (9)}$$

   2.3 $iter$++
3. end while

Table 4. Notations used by our fuzzy clustering algorithm

Notation	Description
N	number of iterations
ε	lower bound of differences between membership degrees over two consecutive iterations
C	total number of clusters
w_{uk}	membership degree of user u of cluster k
c_k	center of cluster k
$f_{c_k,g}$	center of cluster k for genre g with respect to relative rating frequency
G_{u,c_k}	set of common genres rated by user u and ck
$\overline{f}_u$	mean of relative genre rating frequencies by u
$\overline{f}_{c_j}$	mean of relative genre rating frequencies of center of cluster j
m	hyper-parameter to control the fuzziness of the cluster

IV. Experiments

1. Experimental Background

성능 실험을 위하여 관련 연구에서 널리 활용되는 MovieLens 1M(http://www.grouplens.org)을 선정하였다. 이 데이터셋은 3952개의 영화와 영화의 장르를 18 종류로 구분하여 관리한다. 장르의 예로서, 액션, 코미디, 판타지, 모험, 만화 등이 있다. 사용자는 1~5 까지의 정수 평가등급을 부여한다. 데이터셋의 구체적인 특성은 표 5에 기술하였다. 희소성 수준(sparsity level)은 평가치 데이터 개수가 얼마나 적은지를 가늠하는 수치로서 1-T/(U*I)로서 산출하며, 값이 클수록 데이터셋은 희소하다.

일반적으로 협력필터링 시스템의 성능 평가는 시스템에서 채택한 유사도 척도에 따라 선정된 인접이웃 사용자(nearest neighbors)들의 평가치를 참조하여 획득한 다양한 측면에서의 성능 결과를 평가함으로써 이루어진다. 따라서, 유사도 척도의 선정이 성능에 핵심적 영향을 미친다. 현 사용자의 인접이웃은 현 사용자와 유사도가 높은 순으로 선정된다. 현 사용자가 아직 미평가한 항목에 대한 평가 예측치를 구하기 위하여 인접이웃 사용자들의 평가치를 참조하는데, 가장 널리 활용되는 weighted sum 방법인 Resnick's formula[1]를 활용하였다. 이 방법은 이웃과의 유사도가 클수록 해당 이웃의 평가치에 더 높은 가중치를 부여한다. 결과적으로 평가예측치가 정확할수록, 즉, 실제 평가치에 근접할수록, 선정한 유사도 척도의 성능이 우수한 것으로 판단한다. 본 연구에서는 전체 데이타를 통상적인 8:2의 비율을 적용하여 훈련데이타 집합과 시험데이타 집합으로 나누어 실험하였다.

성능을 비교할 방법들은 본 연구와의 관련성을 고려하여 다음과 같이 선정하였다: 평균자승차이(Mean Squared Differences, MSD), JMSD[17], K-means(KM), Self-Organizing Map(SOM), Fuzzy C-means(FCM). 본 연구 방법은 FCM-G로 표기한다. JMSD는 실험에 사용된 MovieLens와 같은 희소 데이터셋에서 이를 극복하기 위한 단순하지만 효율적인 유사도 척도로서 자카드 계수[18]를 MSD와 결합한 방식이며 성능 개선이 입증되었다.

성능 평가를 위하여 세 가지 일반적인 기준을 사용하는데, 예측 성능, 추천 성능, 그리고 순위 성능이다. 예측 성능의 대표적인 척도로서, MAE(Mean Absolute Error), 추천 성능을 위해 정밀도(precision)와 재현율(recall)의 조화평균인 F1을, 순위 성능을 위해 nDCG(normalized Discounted Cumulative Gain)을 선택하였다. 현 사용자의 미평가항목의 평가예측치가 추천 기준값 보다 크면 이 항목을 추천한다. MovieLens 데이터셋에서는 평가값의 범위가 1부터 5이므로, 본 실험에서는 추천 기준값을 4로 정하였다. 이밖에 III.3절의 알고리즘에서 사용한 파라미터 값 선정에 대한 기존 문헌[9][10][11]에 따르면, 하이퍼 파라미터 m의 전형적인 값으로 2.0을 제시하였고, N은 클수록, 또한 ε은 작을수록 바람직하므로 다음과 같이 정하였다: N=30, ε=0.001, m=2.0.

Table 5. Dataset Description

Feature	Value
Number of users (U)	6040
Number of items (I)	3952
Number of ratings per user	≥20
Total number of ratings (T)	1,000,209
Sparsity level	0.9581
Recommendation threshold	4

2. Results

2.1 Effect of Number of Clusters

본 연구의 제안 방법의 주요 비교 대상인 기존의 클러스터링 방법들은 클러스터 개수에 따라 성능에 큰 영향을 받는 것으로 알려져 있다[1]. 실제로 본 논문에서 제시한 실험 환경에서 KM과 SOM의 클러스터 개수에 따른 예측 정확도를

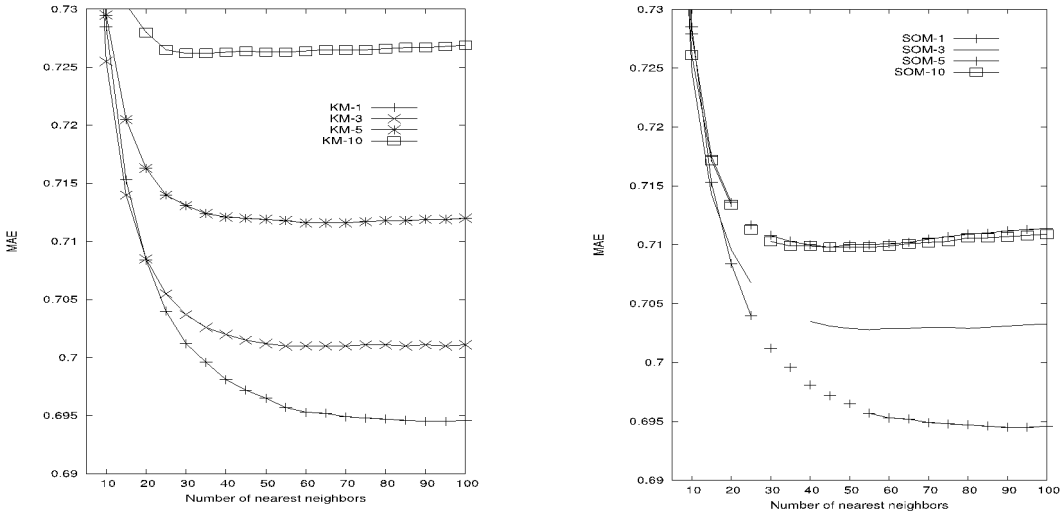

Fig. 1. MAE performance of KM and SOM for varying number of nearest neighbors

측정하였다. 그림 1에서 KM-n과 SOM-n은 n 개의 클러스터를 활용한 실험 결과를 나타낸다. KM의 실험 결과, 평가치를 참조할 인접 이웃수가 증가함에 따라 MAE 값은 점차 감소하여 안정화됨을 알 수 있다. 예측한대로 클러스터 개수가 적을수록 예측 정확도는 향상되는데, 이는 현 사용자가 속한 클러스터 내에서만 인접이웃을 참조할 수 있기 때문에 클러스터 크기가 클수록 성능에 유리하기 때문이다. 그림에서 SOM 결과도 KM과 유사한 행태를 보이는데, 다만 클러스터 개수가 5와 10일 때의 예측 정확도는 거의 차이가 없음을 확인할 수 있다. 이러한 결과에 따라, 다음 절에 제시될 실험 결과에서는 데이터 확장성 문제를 해결할 적정한 수치로서 모든 실험 대상 알고리즘을 위해 공통적으로 다섯 개의 클러스터를 선택하여 성능 평가를 진행하였다.

2.2 Performance Evaluation

그림 2는 실험 대상 방법들의 MAE 결과를 제시하였다. 인접 이웃수가 증가함에 따라 성능은 대체적으로 안정화되어 가는데, 다만 JMSD의 경우엔 오히려 저하된 성능을 보인다. 이는 MSD와 자카드 계수를 결합한 유사도값을 기준으로 인접 이웃을 선정하는 것이 이웃수를 크게 하면 할수록 예측 정확도를 저하시킴을 말한다. 이와는 대조적으로, 제안 방법은 참조 이웃수의 증가에 영향을 받지 않으며, 가장 우수한 예측 성능을 보였다.

KM의 성능은 실험 대상의 클러스터링 방법들을 포함한 모든 방법들 가운데 가장 저조한 것으로 나타났는데, 이는 KM 방법이 클러스터 개수 뿐만 아니라 초기 중앙값의 영향을 받기 때문인 것으로 보인다[8]. 반면에 그 밖의 클러

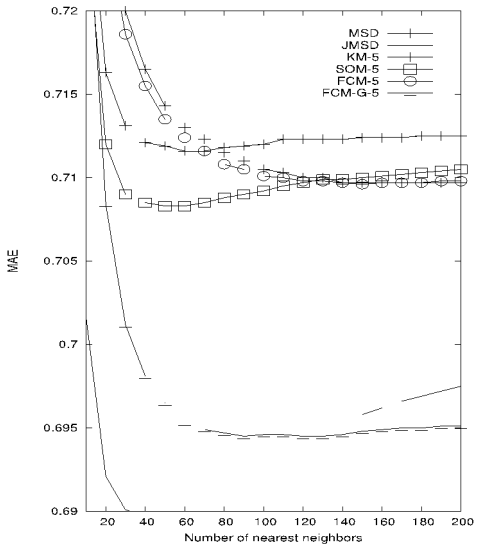

Fig. 2. Prediction accuracy of different methods

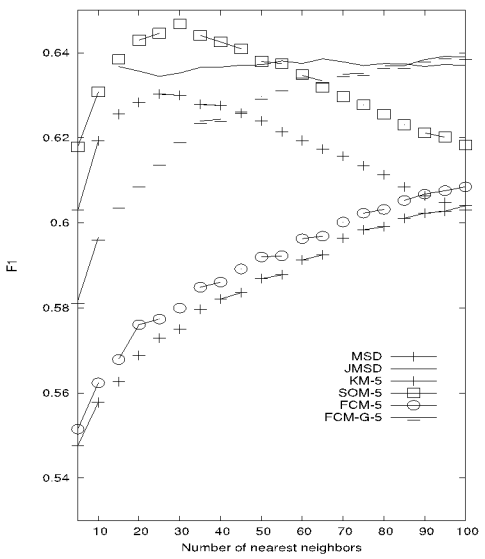

Fig. 3. Recommendation accuracy of different methods

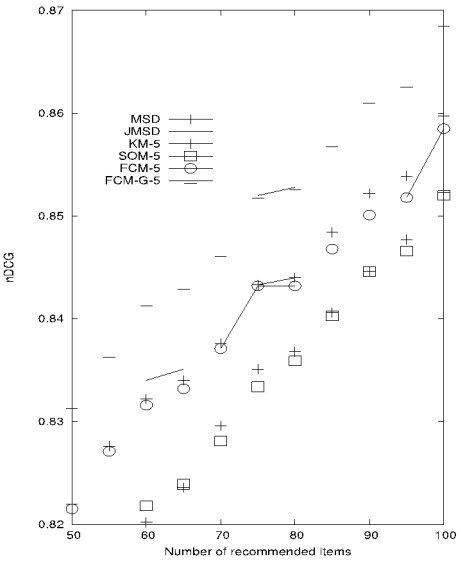

Fig. 4. Rank accuracy of different methods

스터링 방법인 SOM과 FCM은 비클러스터링 방법인 MSD와 거의 대등한 성능을 보이므로, KM과 비교할 때 상대적 우수성을 알 수 있다.

그림 3은 추천 성능을 F1 결과로 나타냈다. 인접 이웃수가 증가함에 따라 대부분의 방법들의 성능은 향상되는데 반해, 가장 저조한 MAE 결과를 보였던 KM과 SOM은 점차적으로 성능이 저하됨을 보였다. FCM과 MSD의 성능은 MAE 결과에서와 마찬가지로 저조함을 알 수 있다. 반면에 제안 방법은 충분한 수의 인접 이웃을 참조할 경우에 가장 우수한 성능을 보였다. 결론적으로 F1 측면에서, 실험 대상 방법들은 MAE 결과와 유사한 상대적 성능 결과를 나타냈다.

그림 4에서는 추천 항목 수에 따른 순위 성능을 제시하였다. 실험 방법들의 성능 결과는 대체로 세 그룹으로 나뉘어지는데, 이는 앞서의 성능 결과들과는 다소 다른 양상이다. 그러나 이전 결과들에서처럼 KM과 SOM은 가장 저조한 성능을 보였으며, 이는 사용자가 선호하는 항목들이 낮은 순위로 추천됨을 의미한다. 또한 FCM과 MSD는 이들보다 좋은 성능을 보였는데, 그 이유는 MSD는 비클러스터링 방법이기 때문이며, FCM은 퍼지 클러스터링의 특성상 클러스터 간의 경계가 모호하기 때문인 것으로 판단된다. F1 성능 결과에서는 JMSD가 제안 방법과 경쟁적으로 우수하였으나, 순위 성능에서는 제안 방법이 독자적으로 가장 우수한데, 이는 제안 방법은 예측 성능이 가장 월등하므로 정확한 예측치에 따른 순위 결과를 나타낼 수 있기 때문인 것으로 보인다.

V. Conclusions

사용자 기반의 협력 필터링 시스템은 사용자 간의 유사도를 파악하여 현 사용자와 유사한 선호를 가진 사용자들로부터 추천 항목들을 결정한다. 따라서 이러한 시스템에서 유사도 측정은 시스템의 성능에 매우 중요한 역할을 한다. 본 연구에서는 측정한 유사도의 신뢰성과 신속한 측정을 위해 데이터 희소성 문제와 확장성 문제를 고려한 새로운 협력 필터링 시스템을 제안하였다.

제안 방법은 퍼지 클러스터링을 기반으로 하되, 사용자 간의 영화 장르 선호도를 기준으로 클러스터링을 실시한다. 이를 위해 장르 선호도 측정 방법을 개발하였고, 다양한 기존의 유사도 척도 및 클러스터링을 적용한 시스템과의 성능 비교를 실시하였다. 실험 결과에 따르면, 제안 방법은 예측 성능과 순위 성능 측면에서 기존 방식들을 능가하였고, 추천 성능에 있어서는 가장 우수한 기존 방법과 대등한 결과를 보였다. 다만, 제안 방법은 영화 장르 정보를 활용하기 때문에, 일반화하기에 제약이 따를 수 있으므로, 장르 외에 다른 특성을 활용하는 방안의 개발이 사후 연구 주제로 적합하다고 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] 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

[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] C. F. Tsai and C. Hung, “Cluster Ensembles in Collaborative Filtering Recommendation,” Applied Soft Computing, Vol. 12, pp. 1417-1425, 2012. DOI: 10.1016/j.asoc.2011.11.016

[6] M. Nilashi, D. Jannach, O. Ibrahim, and N. Ithnin, “Clustering- and Regression-based Multi-criteria Collaborative Filtering with Incremental Updates,” Information Sciences, Vol. 293, pp. 235-250, 2015. DOI: 10.1016/j.ins.2014.09.012

[7] H. Ye, “A Personalized Collaborative Filtering Recommendation using Association Rules Mining and Self-organizing Map,” Journal of Software, Vol. 6, No. 4, 2011. DOI: 10.4304/jsw.6.4.732-739

[8] C. L. Liao and S. J. Lee, “A Clustering based Approach to Improving the Efficiency of Collaborative Filtering Recommendation,” Electronic Commerce Research and Applications, Vol. 18, pp. 1-9, 2016. DOI: 10.1016/j.elerap.2016.05.001

[9] H. Koohi and K. Kiani, “User based Collaborative Filtering using Fuzzy C-means,” Measurement, Vol. 91, pp. 134-139, 2016. DOI: 10.1016/j.measurement.2016.05.058

[10] H. Shivhare, A. Gupta, and S. Sharma, “Recommender System using Fuzzy C-means Clustering and Genetic Algorithm based Weighted Similarity Measure,” Proceedings of International Conference on Computer, Communication and Control, pp. 1-8, 2015. DOI: 10.1109/IC4.2015.7375707

[11] C. Birtolo and D. Ronca, “Advances in Clustering Collaborative Filtering by means of Fuzzy C-means and Trust,” Expert Systems with Applications, Vol. 40, No. 17, pp. 6997-7009, 2013. DOI: 10.1016/j.eswa.2013.06.022

[12] S. Fremal and F. Lecron, “Weighting Strategies for a Recommender System using Item Clustering based on Genres,” Expert Systems With Applications, Vol. 77, No. 1, pp. 105-113, 2017. DOI: 10.1016/j.eswa.2017.01.031

[13] M. K. Najafabadi, M. N.Mahrin, S. Chuprat, and H. M. Sarkan, “Improving the Accuracy of Collaborative Filtering Recommendations using Clustering and Association Rules Mining on Implicit Data,” Computers in Human Behavior, Vol. 67, pp. 113-128, 2017. DOI: 10.1016/j.chb.2016.11.010

[14] M. Al-Shamri and K. Bharadwaj, “Fuzzy-genetic Approach to Recommender Systems based on a Novel Hybrid User Model,” Expert Systems with Applications, Vol. 35, No. 3, pp. 1386-1399, 2008. DOI: 10.1016/j.eswa.2007.08.016

[15] M. Lee, P. Choi, and Y. Woo, “A Hybrid Recommender System Combining Collaborative Filtering with Neural Network,” Lecture Notes on Computer Sciences, Vol. 2347, pp. 531-534, 2002. DOI: 10.1007/3-540-47952-x_77

[16] L. Zhang, T. Qin, and P. Teng, “An Improved Collaborative Filtering Algorithm based on User Interest,” Journal of Software, Vol. 9, No. 4, 2014. DOI: 10.4304/jsw.9.4.999-1006

[17] 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

[18] S. Lee, “Improving Performance of Jaccard Coefficient for Collaborative Filtering,” Journal of The Korea Society of Computer and Information, Vol. 21, No. 11, pp. 121-126, 2016. DOI: 10.9708/jksci.2016.21.11.121

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.