JKIICE Journal of the Korea Institute of Information and Communication Engineering

한국정보통신학회논문지(J. Korea Inst. Inf. Commun. Eng.) Vol. 21, No. 2 : 277~284 Feb. 2017

사용자 선호도를 사용한 군집 기반 추천 시스템

김영현 · 신원용*

Clustering-Based Recommendation Using Users' Preference

Younghyun Kim · Won-Yong Shin*

Department of Computer Science and Engineering, Dankook University, Yongin 16890, Korea

요 약

사용자가 좋아할만한 콘텐츠를 정확하게 추천하는 것은 추천 시스템에서 매우 중요한 요소 중 하나이다. 원치 않는 콘텐츠를 추천하거나, 원하는 것을 추천하지 않는 것은 사용자 만족도 측면에서 안 좋은 영향을 끼친다. 본 연구에서는 콘텐츠의 정확한 추천을 위해 사용자 군집 기반 추천 시스템을 제안한다. 제안하는 알고리즘에서 사용자들의 실제 선호도 점수와 피어슨 상관 계수를 기반으로 사용자들을 여러 군집으로 나눈다. 이 후, 특정 사용자에게 어떤 콘텐츠의 추천 여부 결정은, 같은 군집 내에 있는 다른 사용자들의 해당 콘텐츠의 실제 선호도 점수를 근거로 정한다. 제안하는 알고리즘은 군집화를 사용하지 않는 아이템 기반 협력 필터링 알고리즘보다 정밀도, 재현율, F1 스코어와 같은 추천 정확도에 있어서 의미 있는 성능 향상을 보인다.

ABSTRACT

In a flood of information, most users will want to get a proper recommendation. If a recommender system fails to give appropriate contents, then quality of experience (QoE) will be drastically decreased. In this paper, we propose a recommender system based on the intra-cluster users' item preference for improving recommendation accuracy indices such as precision, recall, and F1 score. To this end, first, users are divided into several clusters based on the actual rating data and Pearson correlation coefficient (PCC). Afterwards, we give each item an advantage/disadvantage according to the preference tendency by users within the same cluster. Specifically, an item will be received an advantage/disadvantage when the item which has been averagely rated by other users within the same cluster is above/below a predefined threshold. The proposed algorithm shows a statistically significant performance improvement over the item-based collaborative filtering algorithm with no clustering in terms of recommendation accuracy indices such as precision, recall, and F1 score.

키워드 : 추천 시스템, 군집, 피어슨 상관 계수, 정밀도, 재현율, F1 스코어

Key word : Recommender system, Clustering, Pearson correlation coefficient, Precision, Recall, F1 score

Received 07 November 2016, **Revised** 08 November 2016, **Accepted** 18 November 2016

* **Corresponding Author** Won-Yong Shin(E-mail:wyshin@dankook.ac.kr, Tel:+82-31-8005-3253)
Department of Computer Science and Engineering, Dankook University, Yongin 16890, Korea

Open Access http://doi.org/10.6109/jkiice.2017.21.2.277 print ISSN: 2234-4772 online ISSN: 2288-4165

Ⅰ. 서 론

신문 기사나 도서, 예술작품 등과 같은 전통적인 오프라인 콘텐츠 뿐만 아니라 유튜브나 트위터, 페이스북, 온라인 커뮤니티, 포털 등과 같은 플랫폼 상에서의 온라인 콘텐츠 생성속도는 기하급수적으로 증가하고 있는 추세이다. 그러나 사람들이 접하고 즐길 수 있는 콘텐츠의 양은 제한적이다. 이런 정보의 홍수 속에서 사람들은 각자의 선호도에 따라 적절한 콘텐츠를 추천해 주는 서비스에 매력을 느낄 것이다. 이 때 콘텐츠는 모든 영역에서 통용될 수 있다. 예를 들어, 온라인 쇼핑몰에서는 사용자가 구입하고자 하는 물품들이 콘텐츠가 될 수 있고, 책이나 비디오와 같은 창작물도 포함되며, 매일 먹는 음식의 종류도 콘텐츠에 해당할 수 있다.

1.1. 관련 연구

추천시스템은 알고리즘에 따라 크게 사용자/아이템 기반의 협업 필터링 (user/item based collaborative filtering), 콘텐츠[1] 기반의 추천 (content-based recommendations) 으로 나눌 수 있다[1]. 협업 필터링은 나와 선호도가 유사한 사용자들을 기반으로 내가 접하지 않았던 아이템들에 대한 선호도를 예측하는 기법이다. 반면에 콘텐츠 기반 추천은 내가 평상시에 자주 접했던 아이템을 분석하여 이와 유사한 아이템들을 추천하는 방법이다[2, 3]. 각각의 추천 기법은 고유한 장점 및 단점을 가진다. 협업 필터링은 나와 선호도가 유사한 사용자들 기반으로 아이템을 추천해 주기 때문에, 평상시에 관심을 가지지 않았던 예상치 못한 아이템들을 추천받을 수 있는 장점이 있다. 반면에 커뮤니티 데이터를 활용해야하기 때문에 사용자들의 아이템들에 대한 선호도를 측정할 수 있는 데이터를 축적해야 하는 단점을 가진다. 또한 새로운 사용자나 새로운 콘텐츠에 대해서는 추천 결과에 반영시킬 수 없는 콜드 스타트 (cold start) 문제점 역시 지니고 있다. 콘텐츠 기반의 추천 기법은 전통적인 정보 검색에 근거를 두고 있는 방법이다[4]. 즉, 특정 콘텐츠와 유사한 성질을 가지는 콘텐츠를 검색하는 것이 콘텐츠 기반 추천 기법의 핵심 기술이다. 예를 들어, 평상시 대하소설을 선호하는 사용자에게 유사한 장르의 소설을 추천해 주기 위해서는 추천엔진이 소설의 내용을 보고 이것이 대하소설인지 아닌지를 판단해야 하는데, 이는 검색 분야의 핵심 기술과 유사하다. 콘텐츠 기반의 추천 기법은 사용자가 선호하는 아이템들을 선별하여 추천해 주기 때문에 일정수준 이상의 사용자 경험 만족도 (user experience: UX)를 보장할 수 있다. 그러나 포털 사이트 뉴스와 같은 환경에서는 항상 같은 주제를 가지는 뉴스만을 추천해 주기 때문에 사용자는 피로도와 지루함을 쉽게 느낄 수 있는 단점을 가진다.

협업 필터링의 목적은 사용자가 아직 접하지 않은 콘텐츠에 대해서 예상 선호도를 최대한 오차 없이 예측하는 데에 있다. 이는 넷플릭스 대회 [5]를 통해 비약적인 발전을 이뤄왔으며, 대상을 받은 팀의 알고리즘을 살펴보면 하나의 기법만을 가지고 목표를 이루기보다는 다양한 기법을 활용하여 오차를 줄여나가는 마치 딥 러닝 [6]에서의 딥 레이어와 같은 방법을 사용한다. 다른 한편으로는, 자체적으로 콘텐츠의 게놈 지도를 정의하여 사용자들이 좋아했던/싫어했던 콘텐츠와 유사한 게놈 지도를 가지는 것들을 추천/비추천함으로써 추천 정확도를 높이는데 집중하는 업체들이 등장한다. 넷플릭스, 아트시, 판도라뮤직이 콘텐츠의 게놈 지도를 개발하여 활용하는 대표적인 기업이다[7, 8].

또한, 추천 시스템에서 사용자들 혹은 아이템들의 군집을 활용하는 다양한 연구들이 제시되었지만, 이들의 목적은 예상선호도 값과 실제선호도 값의 차이를 줄이는 것에 초점을 맞추었다[9, 10]. 예상 선호도 오차와 상관없이 최종적으로 추천 정확도가 높아야 사용자들에게 만족도를 줄 수 있는데, 기존 연구에서는 아직 군집화를 통해 정밀도, 재현율, F1 스코어와 같은 추천 정확도를 개선하는 방안은 제시된 바가 없다.

1.2. 주요 제안 사항

본 논문에서는 같은 예상 선호도 기법 하에서 추천 정확도를 높이는 방법을 제안한다. 즉, 추천 시스템에서 예상 선호도와 실제 선호도의 오차를 줄이는 것보다는 정밀도, 재현율, F1 스코어와 같은 “추천 정확도”를 높이는 방안에 대해 초점을 맞춘다. 즉, 같은 예상 선호

1) 추천 알고리즘 방식에 따라 아이템, 콘텐츠 등과 같은 용어가 혼재되어 사용된다. 두 용어 간에는 미묘한 의미의 차이가 있으나, 본 논문에서는 아이템과 콘텐츠를 혼용하여 같은 의미로서 사용한다.

도를 가지는 아이템이더라도 사용자들의 아이템들에 대한 선호도 경향에 따라 추천을 하거나 하지 않는 방안을 제시한다. 이를 위해 다음과 같이 사용자들의 군집화에 기반한 알고리즘을 제시한다. 우선 사용자들의 실제 선호도 점수를 이용하여 사용자들 간의 아이템 선호도에 대한 유사성을 계산한다. 이에 따라 사용자들을 군집하면, 선호도 경향이 유사한 사용자들은 같은 군집 내에 묶이고, 선호도 경향이 서로 다른 사용자들은 서로 다른 군집으로 묶인다. 이후 어떤 사용자에게 특정 아이템의 추천 여부를 결정할 때, 같은 군집 내에 있는 다른 사용자들의 해당하는 아이템에 대한 실제 선호도 평균 점수가 높은 경우에는 가점 (advantage)을 주고, 그렇지 않은 경우에는 감점 (disadvantage)을 준다. 즉, 선호도 경향이 유사한 사용자들이 같은 군집 내에 존재하기 때문에 같은 군집 내의 다른 사용자들이 이미 접근하여 만족도가 높은 아이템에 대해 아직 접근하지 않았던 사용자 역시 만족도가 높을 것이라는 착안에서 개발한 알고리즘이다.

본 논문의 구성은 다음과 같다. 2장 시스템 모델 및 문제 정의에서는 사용자 군집 기반 추천 시스템을 위한 여러 가지 변수 및 데이터베이스 구조를 정의한다. 그리고 3장과 4장에서는 각각 제안하는 알고리즘과 추천 정확도에 대한 성능분석을 보여준다. 마지막으로 5장에서는 본 연구의 결과와 향후 연구 과제를 제시한다.

Ⅱ. 시스템 모델 및 문제 정의

본 연구에서 제안하는 알고리즘은 사용자들을 아이템에 대한 선호도 점수 기반으로 군집한 후, 같은 군집 내에 있는 다른 사용자들의 실제 선호도 값을 근거로 아이템들의 추천 여부를 결정하는 것이다. 그림 1은 제안하는 알고리즘의 예를 보여준다. 그림 1에서 사용자 u_1과 u_2, u_6, u_{17}은 군집 C_1에 있다고 가정한다. 그리고 적색으로 색칠된 원은 사용자들이 특정 아이템에 대해 실제로 선호도 점수를 매긴 것을 의미한다. 예를 들어 사용자 u_1과 u_2는 아이템 i_1에 대해 선호도 점수를 각각 5점을 준 것으로 가정한다. 그리고 사각형으로 표시된 것은 성능분석을 위한 테스트 데이터이다. 이 후, 사용자 u_6에게 i_1의 추천 여부는 u_6과 같은 군집 내에 존재하는 다른 사용자들의 i_1에 대한 실제 선호도 값들의 평균값 $\left(\overline{C}_1^{i_1} = 4.67\right)$을 근거로 결정한다. $\overline{C}_1^{i_1}$이 미리 정의한 파라미터 γ보다 클 때에는 예상 선호도 값 $\hat{r}_{u_6,i_1}$이 β보다 크면 u_6에게 i_1을 추천한다. 그러나 $\overline{C}_1^{i_1}$이 γ보다 작을 때에는, $\hat{r}_{u_6,i_1}$이 α보다 클 때에만 i_1을 추천한다. 제안하는 알고리즘에서 $\overline{C}_c^i$는 군집 c에 존재하는 사용자들의 아이템 i에 대한 실제 선호도 값들의 평균으로 정의한다. $\hat{r}_{u,i}$는 사용자 u의 아이템 i에 대한 예상 선호도 값으로 정의한다.

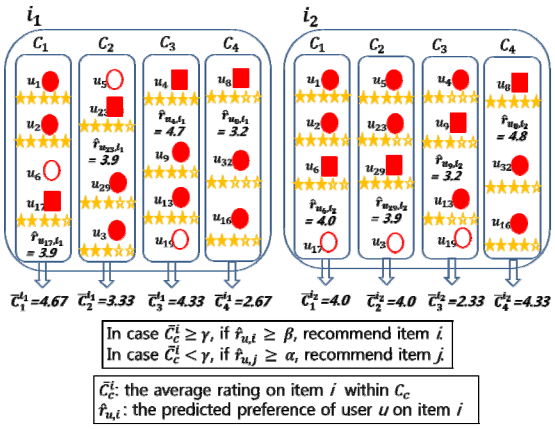

Fig. 1 Example of the proposed algorithm

사용자 군집을 위해 사용자들의 아이템에 대한 선호도 점수 기반 Pearson correlation coefficient (PCC)를 사용한다 [11]. PCC 기법은 두 오브젝트 사이의 상관관계를 계산하는 방법으로서 $[-1.0, 1.0]$ 사이의 값을 가진다. 두 오브젝트의 PCC 값이 1.0에 가까울수록 양적 상관관계가 높고, -1.0에 가까울수록 음적 상관관계가 높은 것으로 간주한다. 그리고 PCC 값이 0.0일 때, 두 오브젝트의 상관관계는 없다고 판단한다. PCC는 두 명의 사용자가 2개 이상의 선호도 점수를 매긴 공통된 아이템들만 존재하면 계산할 수 있기 때문에, 선호도 데이터가 많지 않은 상황에서의 군집에 적합하다.

위와 같이 사용자들 간의 PCC 유사도를 계산한 후 이를 기반으로 사용자들을 군집화한다. 데이터셋의 흩어진 모양과 형태에 따라 적절한 군집화 알고리즘을

선택할 수 있으나, 본 연구에서는 스펙트럴 (spectral) 군집 알고리즘을 택한다. 스펙트럴 군집은 그래프 분할 기반의 군집 방법으로써 다양한 형태의 군집에 잘 동작하는 것으로 알려져 있다 [12]. 또한 군집의 수를 정하는 것도 중요한 설정중의 하나이다. 군집의 수는 오브젝트의 수가 n개일 때, $\sqrt{n/2}$ 로 계산하는 방법이 널리 통용되지만 정확한 방법은 아니며 상황에 따라 적응적으로 정해야 한다. 따라서 군집의 수를 여러 가지로 정하여 실험을 진행할 필요가 있다.

사용자 군집 기반 추천 시스템을 위한 데이터베이스 구조는 표 1과 같다. 표 1에서 볼 수 있듯이, 제안하는 알고리즘의 데이터베이스는 user ID, item ID, ratings, 이렇게 3개의 필드로 이뤄진다.

Table. 1 Database structure

user ID	item ID	ratings
u_1	i_1	$r_{1,1}$
u_1	i_2	$r_{1,2}$
u_1	i_8	$r_{1,8}$
⋮	⋮	⋮
u_n	i_{m-4}	$r_{n,m-4}$
u_n	i_m	$r_{n,m}$

Ⅲ. 사용자 군집 기반 추천 알고리즘

본 장에서는 사용자 군집을 이용한 추천 기법에 대해 묘사한다. 그림 2는 사용자 군집 기반 추천 시스템의 의사 코드를 보여준다. 그림 2에서처럼 사용자들은 총 c개의 군집으로 나뉘고, 전체 사용자의 수와 아이템의 수는 각각 n과 m으로 둔다.

그림 2의 첫 번째 라인에서처럼 n명의 사용자들은 군집 C_1부터 군집 C_c까지 집합에 각각 할당된다. 군집은 사용자들의 실제 선호도 점수를 이용한 PCC 기반으로 수행했기 때문에 같은 군집 내에 존재하는 사용자들은 다른 군집 내의 사용자들보다 실제로 접근했던 아이템들에 대한 선호도 경향이 보다 더 유사하다고 말할 수 있다[2]. 그리고 표 1의 데이터베이스를 이용하여 그림 2의 두 번째 라인에서처럼, 선호도 행렬 R을 초기화한다. 이 후, 행렬 R을 예상 선호도 함수에 대입하여 나온 결과를 예상 선호도 식 (1)과 같이 행렬 $\hat{R}$에 저장한다.

식 (1)에서 $r_{a,b}\ (1 \le a \le n, 1 \le b \le m)$로 표시된 부분은 사용자 a의 아이템 b에 대한 실제 선호도 점수를 의미하고, $\hat{r}_{a,b}$는 예상 선호도 점수를 나타낸다. 그리고 그림 2의 네 번째 라인에서처럼 사용자의 예상 선호도 점수와 함께 추천 여부를 결정할 파라미터인 α, β, γ값을 설정한다. 세 개의 파라미터 값에 따라 추천 시스템의 성능은 달라지는데, 이 부분은 다음 장에서 자세히 서술한다.

$$\hat{R}=\begin{pmatrix} r_{1,1} & \hat{r}_{1,2} & r_{1,3} & \cdots & \hat{r}_{1,m} \\ \hat{r}_{2,1} & r_{2,2} & \hat{r}_{2,3} & \cdots & \hat{r}_{2,m} \\ r_{3,1} & \hat{r}_{3,2} & \hat{r}_{3,3} & \cdots & r_{3,m} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ r_{n,1} & r_{n,2} & \hat{r}_{n,3} & \cdots & \hat{r}_{n,m} \end{pmatrix} \quad (1)$$

다음 단계는 사용자들에게 추천할 아이템들을 결정하는 과정이다. 사용자들의 수를 n으로 가정했기 때문에, 모든 사용자들에 대한 추천 아이템을 결정하기 위해 그림 2의 다섯 번째 라인에서처럼 사용자 u_1에서 u_n까지 총 n번의 루프 반복을 수행한다. 루프 안에서 그림 2의 라인 6에서처럼 우선 사용자 u를 위한 top-N 아이템을 I_u에 저장한다. 이 때, top-N 아이템은 사용자 u가 접근하지 않은 아이템들 가운데 예상 선호도 점수가 가장 높은 N개의 아이템을 의미한다. 그리고 사용자 u의 top-N 아이템에 대한 예상 선호도 점수를 $\hat{r}_{u,I_u}$에 저장한다. 이 후, top-N 아이템들 가운데에 사용자 u에게 실제로 추천할 아이템들을 선별하기 위해 그림 2의 여덟 번째 라인에서처럼 총 N번의 루프 반복을 수행한다.

첫 째로, 그림 2의 아홉 번째 라인에서처럼 사용자 u가 속하는 군집을 검색하여 C_{tmp}에 할당한다. 그리고 군집 C_{tmp} 내에 존재하는 사용자들 중에 아이템 i

2) 사용자들이 실제로 접근했던 아이템들의 선호도 경향에 따라 사용자들을 군집했기 때문에, 같은 군집 내에 존재하는 사용자들의, 접근하지 않았던 아이템들에 대한 선호도 경향 역시 서로 유사할 것이라는 가정하에 본 알고리즘을 제안한다.

에 대한 실제 선호도 평균을 계산하여 $\overline{C}^{j}_{tmp}$에 저장한다. 예를 들어 그림 1에서 사용자 u_6의 아이템 i_1에 대한 $\overline{C}^{i_1}_{1}$ 값은 $4.67(=(5+5+4)/3)$이다. 이 후, 그림 2의 11번째 라인에서처럼 사용자 u의 아이템 i에 대한 예상 선호도 값 $\hat{r}_{u,i}$이 α보다 크면 해당 아이템은 사용자에게 추천할 것으로 결정한다. 반면에 $\hat{r}_{u,i}$ 값이 α보다는 작지만 β보다 클 때에는 $\overline{C}^{j}_{tmp}$ 값을 보고 추천 여부를 결정한다. 그림 2의 13번째 라인에서처럼 $\hat{r}_{u,i}$이 값이 β보다 크고 또한 $\overline{C}^{j}_{tmp}$ 값이 γ보다 크면 해당 아이템 역시 사용자에게 추천할 아이템으로 결정한다. 반면에 그림 2의 11번째 라인 그리고 13번째 라인의 조건을 충족시키지 않는 아이템은 사용자의 추천 리스트에서 제외한다.

```
1  Clusters C = C_1, C_2, ..., C_c;
2  Initiate the rating matrix R;
3  R̂ ← predicted rating matrix;
4  Initiate the threshold values α, β, and γ;
5  for u ← 1 to n do
6      I_u ← top-N items for user u;
7      r̂_{u,I_u} ← predicted rating values of I_u;
8      for i ← 1 to N do
9          C_tmp ← a cluster to which user u belongs;
10         C̄^i_tmp ← average rating on item i within C_tmp;
11         if r̂_{u,i} ≥ α then
12             Recommend item i to user u;
13         else if r̂_{u,i} ≥ β && C̄^i_tmp ≥ γ then
14             Recommend item i to user u;
15         else Drop item i;
16     end
17 end
```

Fig. 2 Pseudocode of the proposed algorithm

본 연구에서 제안하는 알고리즘의 아이디어는 다음과 같이 요약할 수 있다. 임의의 사용자 u에게 아이템 i의 추천 여부를 결정한다고 가정하자.

- 사용자 u의 아이템 i에 대한 예상 선호도 값이 아주 높으면 $\left(\widehat{r_{u,i}} \geq \alpha\right)$ 추천한다.
- 사용자 u의 아이템 i에 대한 예상 선호도 값이 아주 높지는 않지만 일정 값보다 크고 $\left(\widehat{r_{u,i}} \geq \beta\right)$, 사용자 u와 같은 군집 내에 존재하는 사용자들의 아이템 i에 대한 실제 선호도 평균 점수가 일정 값보다 높으면 $\left(\overline{C^{i}_{tmp}} \geq \gamma\right)$ 추천한다.

Ⅳ. 성능 평가

본 장에서는 사용자 군집 기반 추천 시스템에서의 정밀도, 재현율, 그리고 F1 score 측면에서 성능 평가를 한다. 정밀도, 재현율, F1 score는 추천 시스템의 정확성을 판단하기 위해 활용되는 대표적인 요소들이고, 이것은 표 2와 같은 에러 종류의 횟수에 의해 결정된다. True positive (TP)는 사용자에게 추천을 했고 실제로도 만족한 경우, true negative (TN)은 추천을 하지 않았고 실제로도 불만족한 경우, false positive (FP)는 추천은 했지만 실제로는 불만족한 경우, false negative (FN)은 추천은 하지 않았지만 실제로는 만족한 경우를 의미한다. 이와 같이 오류가 발생하는 경우의 수를 계산하여 정밀도, 재현율을 아래와 같이 구할 수 있다 [13]. F1 score는 정밀도와 재현율의 조화평균이다.

$$\text{정밀도} = TP/(TP+FP) \tag{2}$$

$$\text{재현율} = TP/(TP+FN) \tag{3}$$

Table. 2 Type of error

	Recommend (Positive)	Nonrecommend (Negative)
True	True Positive	True Negative
False	False Positive	False Negative

본 연구의 성능 분석을 위해, 여러 논문에서 활용되었던 MovieLens[3] 데이터셋을 이용한다[14, 15]. 구체적으로, MovieLens 100K 데이터셋를 이용하고 이는 1,682편의 영화에 대한 943명의 사용자의 100,000개의 평가가 포함된다. MovieLens 100K 데이터셋에서 각 사용자는 적어도 20개 이상의 영화에 대한 선호도 데이터를 가진다.

3) GroupLens Research, http://grouplens.org/datasets/movielens/

시뮬레이션은 Apache Mahout 오픈소스를 이용하고, 선호도 예측 함수는 item-based CF를 사용한다. 그리고 사용자들의 군집 개수는 10으로 설정한다. 5점 스케일의 경우 테스트 데이터 셋에서 사용자들에게 추천한 아이템들 가운데에 사용자들의 실제 선호도 점수가 4.0 이상일 때 올바른 추천이라고 가정 (10점 스케일의 경우 8.0 이상)한다.

그림 3은 $\gamma = 3.4$일 때, α와 β의 변화에 따르는 F1 score의 변화를 보여준다. α와 β는 각각 $[3.5, 4.5]$, $[2.5, 3.5]$ 사이의 값을 가진다. 제안하는 알고리즘에서 $\alpha = 3.7$, $\beta = 2.9$, $\gamma = 3.4$일 때, 최대의 F1 score 값을 (= 0.7451) 가진다. 그리고 그림 3에서 보듯이 α와 β 값이 커질수록 F1 score 값은 감소한다. 이는 α와 β 값이 증가할수록 제안하는 알고리즘의 정밀도 역시 증가하지만, 재현율의 감소하는 속도가 더 빠르기 때문에 나타나는 현상이다.

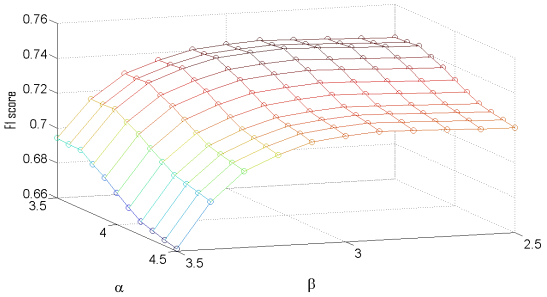

Fig. 3 F1 score when $\gamma = 3.4$

그림 4는 같은 환경에서의 item-based CF 사용 시 임계값 (threshold)에 따른 F1 값의 변화를 보여준다. Item-based CF에서의 예상 선호도가 임계값보다 높으면 사용자에게 추천하고, 그렇지 않으면 추천하지 않는다. 이 후, 추천한 아이템의 실제 선호도가 4.0이 넘으면 올바른 추천이고 그렇지 않으면 잘못된 추천으로 해석한다. 임계값이 3.1일 때 최대의 F1 score 값을 가지는 것을 볼 수 있고, 이때의 F1 score는 0.7282이다. 위의 그림 3과 유사하게, 임계값이 커질수록 F1 score 값이 떨어지는 것을 볼 수 있다. 임계값의 증가는 정밀도의 증가로 이어지지만 재현율의 감소하는 속도가 더 빠르기 때문에, 전체적으로 F1 score 값이 감소하는 것이다.

제안하는 알고리즘의 최대 F1 score는 0.7451, item-based CF의 최대 F1 score는 0.7282이다. 같은 예상 선호도 기법을 활용하고 사용자들의 군집 결과를 추천 시스템에 추가했을 때 대략 3%의 성능 향상이 있는 것을 확인할 수 있다.

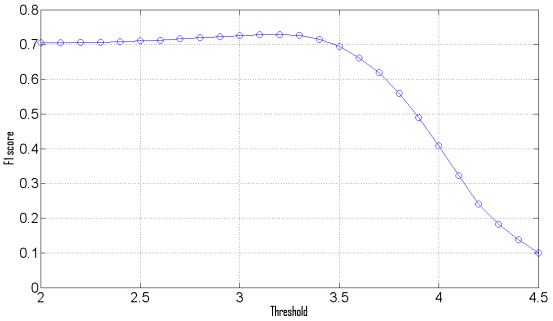

Fig. 4 F1 score of item-based CF

표 3은 item-based CF와 제안하는 알고리즘의 정밀도와 재현율의 비교를 보여준다. Item-based CF의 최대 정밀도 값은 0.7449이고[4], 이때의 임계값은 4.0이다. 같은 정밀도 값에서 제안하는 알고리즘은 최대의 재현율 값으로 0.4343을 가진다. 이때의 α, β, γ는 각각 3.9, 2.1, 4.2이다. 즉, 같은 정밀도를 가질 때, 재현율은 대략 50%의 성능향상을 보인다. 이와 유사하게 item-based CF와 비교하여, 같은 정밀도 값을 가질 때, 보다 향상된 재현율을 보이는 것을 표 3으로부터 알 수 있다.

Table. 3 Maximum recall for given precision

Precision	Item-based CF		Proposed scheme	
	Recall	F1 score	Recall	F1 score
0.7449	0.2815	0.4085	0.4343	0.5487
0.7201	0.4565	0.5588	0.5706	0.6367
0.7074	0.5499	0.6188	0.6842	0.6956
0.6519	0.7914	0.7149	0.825	0.7283
0.6036	0.9177	0.7282	0.9402	0.7352

4) 제안하는 알고리즘에서의 최대 정밀도 값은 0.8162이다.

Ⅴ. 결 론

본 논문에서는 추천 시스템의 추천 정확도 향상을 위해, 사용자들을 군집화한 후 다른 사용자들의 실제 선호도 값을 이용하여 아이템의 추천 여부를 결정하는 방안을 제시하였다. 사용자들의 군집을 위해 실제 선호도 데이터와 PCC 유사도를 활용한다. 그 결과, 같은 예상 선호도 기법 (item-based CF) 하에서 제안하는 알고리즘은 F1 측면에서 대략 3% 성능향상을 보였다. 또한 같은 정밀도를 가질 때 재현율을 비교한 결과, 최대 50%의 성능이 좋아지는 것 역시 확인하였다. 사용한 협업 필터링의 성격에 따라 사용자들의 군집을 좀 더 정확하게 한다면 더 나은 결과를 보일 것으로 예상한다. 따라서 향후 연구과제로 사용자들 혹은 아이템들을 군집화하는 새로운 방안과 이를 추천 시스템에 적용시킬 수 있는 방안을 제시한다.

ACKNOWLEDGMENTS

This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2014R1A1A2054577) and by the Ministry of Science, ICT & Future Planning (MSIP) (2015R1A2A1A15054248).

REFERENCES

[1] D. Jannach, M. Zanker, A. Felfernig, and G. Friedrich, “Recommender systems: An introduction,” *Cambridge University Press*, 2010.

[2] J. A. Konstan and J. Riedl, “Recommender systems: From algorithms to user experience,” *Springer User Modeling and User-Adapted Interaction*, vol. 22, no. 1, pp. 101-123, March 2012.

[3] X. Su and T. M. Khoshgoftaar, “A survey of collaborative filtering techniques,” *Advances in Artificial Intelligence*, vol. 2009, no. 421425, pp. 1-19, 2009.

[4] M. J. Pazzani and D. Billsus, “Content-based recommendation systems,” *Springer Lecture Notes in Computer Science*, vol. 4321, pp. 325-341, 2007.

[5] Netflix Prize, http://www.netflixprize.com/

[6] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature* 521, pp. 436-444, May 2015.

[7] P. M. Napol, “Special issue introduction: Big data and media management,” *International Journal on Media Management*, vol. 18, no. 1, pp. 1-7, June 2016.

[8] M. S. Berrie, “Curatorial compass: Organising meaning in institutional and online displays,” *Museological Review*, vol. 18, no. 1, pp. 61-68, 2014.

[9] C. Cheng, X. Wang, Z. Li, and Y. Lin, “A new TV recommendation algorithm based on interest quantification and item clustering,” in *Proceedings of the IEEE ICSESS*, Beijing, China, pp. 215-200, September 2015.

[10] X. Wang, X. Wang, Z. Ding, X. Nie, and L. Xiao, “A new algorithm based on item clustering and matrix factorization,” *International Journal of Engineering and Technology*, vol. 9, no. 2, pp. 160-165, January 2017.

[11] W. H. Gomaa and A. A. Fahmy, “A survey of text similarity approaches,” *International Journal of Computer Applications*, vol. 68, no. 13, pp. 13-18, April 2013.

[12] U. Luxburg, “A tutorial on spectral clustering,” *Springer Statistics and Computing*, vol. 17, no. 4, pp. 395-416, December 2007.

[13] J. Lee, D. Lee, Y.-C. Lee, W.-S. Hwang, and S.-W. Kim, “Improving the accuracy of top-*N* recommendation using a preference model,” *Information Sciences*, vol. 348, no. 20, pp. 290-304, June 2016.

[14] J. Schaffer, T. Hollerer, and J. O’Donovan, “Hypothetical recommendation: A study of interactive profile manipulation behavior for recommender systems,” in *Proceedings of the FLAIRS*, Hollywood, USA, pp. 507-512, May 2015.

[15] D. Song and D. A. Meyer, “Recommending positive links in signed social networks by optimizing a generalized AUC,” in *Proceedings of the AAAI*, Austin, USA, pp. 290-296, January 2015.

김영현(Younghyun Kim)

2005년 숭실대학교 컴퓨터학부 학사
2007년 숭실대학교 컴퓨터공학과 석사
2013년 고려대학교 전자전기컴퓨터공학부 박사
2016년 5월~현재 단국대학교 컴퓨터학과 통신및네트워킹연구실 연구교수
※관심분야 : 네트워크이론, 빅데이터 분석

신원용(Won-Yong Shin)

2002년 연세대학교 기계전자공학부 학사
2004년 KAIST 전자전산학과 석사
2008년 KAIST 전자전산학과 박사
2008년 9월~2009년 2월 KAIST BK 정보전자연구소 박사후연구원
2009년 3월~4월 KAIST 고성능집적시스템연구센터 선임급 위촉연구원
2009년 5월~2011년 10월 Harvard University Postdoctoral Fellow
2011년 10월~2012년 2월 Harvard University Research Associate
2012년 3월~현재 단국대학교 국제대학 모바일시스템공학과/대학원 컴퓨터학과 부교수
※관심분야 : 정보이론, 통신이론, 신호처리, 모바일 컴퓨팅, 빅데이터 분석, 온라인소셜네트워크 분석