Journal of KIIT. Vol. 16, No. 12, pp. 109-114, Dec. 31, 2018. pISSN 1598-8619, eISSN 2093-7571 109
http://dx.doi.org/10.14801/jkiit.2018.16.12.109

협업필터링의 희박 행렬 문제를 위한 이행적 유사도 평가 모델

배은영*, 유석종**

Transitive Similarity Evaluation Model for Improving Sparsity in Collaborative Filtering

Eun-Young Bae*, Seok-Jong Yu**

요 약

협업 필터링은 사회적 추천 방식으로서 뛰어난 성능을 제공하는 대표적인 추천 시스템 알고리즘으로 폭넓게 사용되어 오고 있다. 협업 필터링은 구조적으로 아이템 평가 데이터에 의존하고 있기 때문에 평가 행렬의 희박도는 추천 성능에 직접적으로 영향을 미친다. 평가 행렬의 희박성 문제 개선을 위해 협업 필터링과 내용 기반 방법을 결합하는 복합형 추천 방법에 대한 연구는 꾸준하게 이루어져 왔으며, 본 연구에서는 협업 필터링의 희소 평가 행렬(sparse rating matrix) 문제 개선 방안의 하나로 공통 평가 아이템이 누락되어 유사도 측정이 불가능한 상황에 대처하기 위한 방법을 제안한다. 이를 위하여 사용자간 이행적 관계 그래프에 기반하는 유사도 평가 모델을 설계하고 오픈 데이터셋인 Movielens에 적용하여 추천 정확도를 측정 비교하였다.

Abstract

Collaborative filtering has been widely utilized in recommender systems as typical algorithm for outstanding performance. Since it depends on item rating history structurally, The more sparse rating matrix is, the lower its recommendation accuracy is, and sometimes it is totally useless. Variety of hybrid approaches have tried to combine collaborative filtering and content-based method for improving the sparsity issue in rating matrix. In this study, a new method is suggested for the same purpose, but with different perspective, it deals with no-match situation in person-person similarity evaluation. This method is called the transitive similarity model because it is based on relation graph of people, and it compares recommendation accuracy by applying to Movielens open dataset.

Keywords
similarity evaluation model, collaborative filtering, recommender system, sparsity

* 숙명여자대학교 소프트웨어학부
- ORCID: https://orcid.org/0000-0002-7253-0778
** 숙명여자대학교 소프트웨어학부 교수(교신저자)
- ORCID: https://orcid.org/0000-0002-1631-4034

· Received: Sep. 07, 2018, Revised: Oct. 31, 2018, Accepted: Nov. 03, 2018
· Corresponding Author: Seok-Jong Yu
Department of Software, Sookmyung Women's University, Korea
Tel.: +82-2-710-9831, Email: sjyu@sm.ac.kr

Ⅰ. 서 론

협업 필터링은 사회적 방법을 통하여 뛰어난 성능을 제공해주는 대표적인 추천 알고리즘으로 널리 사용되고 있다. 그러나 구조적으로 아이템 평가 기록에 의존하기 때문에 평가 데이터가 부족하거나 없을 경우 선호도 예측 정확도가 떨어지거나 불가능해진다. 이를 보완하고자 초기 사용자나 아이템은 속성 정보를 이용하는 내용 기반 방식을 활용한다[1]-[3].

최근에는 사용자가 선호도를 직접 표현한 명시적 피드백(Explicit Feedback) 방식의 한계를 극복하고자 사용자의 행위를 관찰하고 정량적으로 분석하는 묵시적 피드백(Implicit Feedback) 기반 협업 필터링 연구가 이루어지고 있다. 명시적 피드백은 사용자의 긍정적, 부정적 선호도를 비교적 정확하게 전달해주는 반면 초기 추천 문제(Cold-start)가 발생하며, 평가 행렬의 희박성(Sparsity)으로 인해 이웃 탐색에 어려움을 겪는다. 한편, 묵시적 피드백은 데이터의 양은 풍부한 반면 긍정적 피드백만으로 구성되어 있고 부정적 피드백은 받기 어려운 단점이 있다. 또한 묵시적 피드백 데이터의 분석 결과가 대상에 대한 사용자의 선호도라고 명확히 단정할 수 없는 한계가 있다.

본 연구에서는 명시적 피드백 환경에서 평가 행렬의 희박성 문제를 개선하기 위한 방법에 초점을 맞추고자 한다. 협업 필터링은 목표 대상자에게 적절히 추천해주기 위해서는 유사 선호도를 갖는 이웃이 충분히 확보되어야 하고 이웃 탐색에는 공통 평가 아이템이 필요하다.

본 연구의 사전 연구로 추천 시스템 분야에서 가장 많이 이용되는 Movielens 데이터셋에서 이웃 탐색에 사용된 공통 아이템 수(Item Corating)의 비율과 분포를 분석하는 실험을 수행하였다. 이 실험에서 사용자간 공통 평가 아이템이 전혀 없어 유사도 계산이 불가능한 경우가 발생되는 것을 확인하였으며, 이를 활용할 경우 추천 정확도 향상에 얼마나 기여할 수 있는지 확인하고자 본 연구를 진행하였다. 사전 실험 결과에 대한 자세한 내용은 3장에서 기술한다.

관련 연구로 평가 행렬에서 누락된 아이템의 평가값을 주변 데이터를 보간하여 예측하는 방법에 대한 연구를 들 수 있다[1]-[3]. 본 연구는 평가 행렬의 희박성 문제를 개선한다는 점에서 기존 연구와 추구하는 목표가 동일하지만, 이웃 탐색에 사용되는 공통 아이템에 초점을 맞춘다는 점에서 접근 방법이 상이하다고 할 수 있다.

제안 연구는 사용자간 공통 아이템이 없는 경우에도 공통 이웃이 존재한다면 유사도 계산이 가능하다는 장점이 있다.

본 논문의 구성은 다음과 같다. 2장은 관련 연구로 추천 시스템의 종류를 구분하고, 대표적 추천 기법인 협업 필터링에 대해 기술한다. 3장에서는 본 논문에서 제안하는 이행적 유사도 평가 모델에 대해 자세하게 서술하고, 4장에서는 제안 방법의 성능 평가를 위한 실험 결과를 제시한다. 마지막으로 5장에서는 결론과 향후 연구 방향을 기술한다.

Ⅱ. 관련 연구

2.1 추천 시스템

추천 시스템은 사용자 및 상품의 속성 정보(Attribute)를 활용하는 내용 기반(Content-based) 추천과 과거의 상품 평가 이력을 이용하는 협업 필터링(Collaborative Filtering) 추천 방법이 있으며, 두 방식을 결합하는 하이브리드(Hybrid) 방식이 존재한다. 내용 기반 추천은 사용자를 나이, 성별 등 속성 정보로 분류하여 동일 그룹 내 사용자들이 많이 구매한 아이템을 추천해주는 방법과 과거 구매 이력을 분석하여 연관 상품을 추천해주는 방법으로 나뉜다[3].

내용 기반 추천은 구조적으로 사용자가 이미 경험한 것과 유사한 아이템이 주로 추천되는 포트폴리오(Portfolio) 현상이 불가피하게 발생되는 단점이 있다[4]. 이로 인해 추천 정확도와 함께 다양성에 관심이 많아지면서 평가 횟수가 적은 롱테일(Long-tail) 상품의 추천에 대한 연구가 진행되고 있다[2][4]-[6].

2.2 협업 필터링

협업 필터링은 사용자의 과거 상품 평가 기록을 분석하여 선호도가 유사한 사용자 집합(이웃)을 탐색하고 미경험 상품에 대한 선호도를 예측하는 사회적 추천 방법이다. 따라서 협업 필터링의 추천의 질은 탐색된 이웃에 달려 있으며, 뛰어난 추천 성능에도 불구하고 다음의 구조적 문제점이 발생한다. 첫째, 평가 이력이 없는 초기 사용자나 신규 아이템은 추천이 어려운 초기 평가 문제가 발생하며 개인화된 추천이 불가능하다. 두번째, 사용자의 평가 행렬이 희박할 경우 연관된 이웃 탐색이 어려우며 추천의 질도 낮아진다. 끝으로 확장성(Scalability)은 사용자와 아이템의 수가 증가할수록 연산 시간이 비례하여 증가하며 실시간 추천을 어렵게 한다[1][3]. 또한, 시간에 따라 변화하는 사용자의 선호도 문제를 위한 추천 다양성과 참신성(Novelty)에 대한 연구[6]-[9]가 진행되고 있다.

유사도 평가 방법에는 피어슨 상관계수(Pearson's Correlation), Cosine similarity 외에 Adjusted Cosine similarity, Constrained Pearson's correlation 등의 방법이 있으며[8], 개별 평가값의 차이에 의존하는 기존 방법과는 달리, 평가값들의 상호 관계를 분석하여 반영하는 PIP(Proximity - Impact - Popularity) 유사도 평가 모델에 대한 연구도 수행되고 있다[10].

III. 이행적 유사도 평가 모델

3.1 Movielens 데이터셋의 희박성

그림 1은 Movielens 데이터셋에서 사용자간 유사도 계산에 사용된 공통 평가 아이템 수의 분포를 분석한 그래프이고, 그림 2는 공통 평가 아이템 수(k)를 비율로 표시한 것이다. 두 사용자간 공통 평가 아이템 수가 하나도 없는 경우(k=0)가 전체의 12%이고 1회인 경우는 4%로 나타났다. 그리고 2~6회가 28%이었으며 대부분에 해당하는 k가 6이하인 경우가 전체 공통 평가 횟수의 44%에 해당한다. 공통 평가 횟수(k)가 0 또는 1인 경우는 사용자 유사도 계산이 사용될 수 없으며 그대로 버려지는 데이터이다.

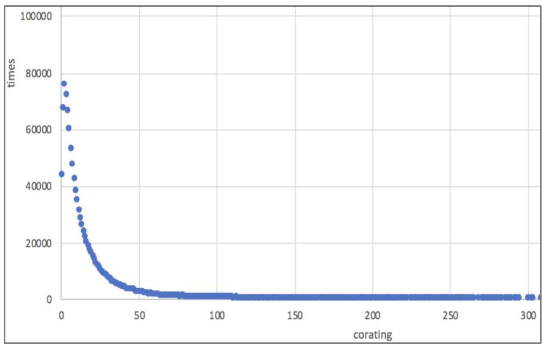

그림 1. Movielens 공통 평가 아이템 수 분포
Fig. 1. Distribution of Movielens item corating

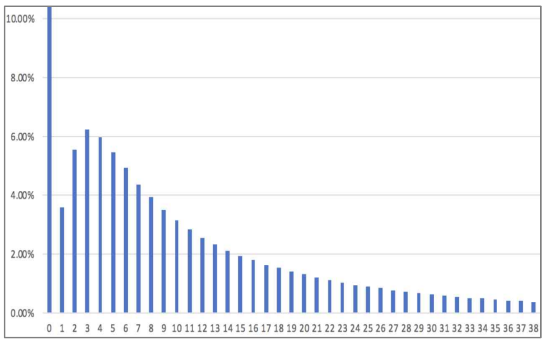

그림 2. 공통 아이템 수의 비율
Fig. 2. Ratio of item corating

Movielens 데이터셋의 희박도는 96%로 기본적으로 평가 데이터가 부족함에도 불구하고 약 16%의 데이터를 사용하지 못하고 있다.

본 연구의 목표는 사용자간 공통 평가 아이템이 없는 경우에도 적용할 수 있는 유사도 평가 방법을 설계하고 이를 통하여 협업 필터링의 추천 성능 개선에 도움이 되는지를 확인하는 것이다. 본 장에서는 이러한 무평가 데이터를 활용하기 위한 방법과 알고리즘을 제시한다.

3.2 이행적 유사도 모델

이행적 관계는 A=>B이고 B=>C이면 A=>C라고 하는 연관 관계를 의미하며 데이터베이스, 경제학 등 여러 분야에서 사용되는 개념이다. 이 개념을 사용자 유사도에 적용하면, 사용자 A와 B가 유사하고 사용자 B가 C와 유사하면 사용자 A는 C와 유사하다는 의미로 확장할 수 있다. 제안하는 이행적 유

사도 평가 모델(Transitive Similarity Model)은 공통 평가 아이템이 없는 사용자 A와 B간에 유사도 계산이 불가능한 경우 그림 3과 같이 사용자 U_a 의 이웃 그룹(C)과 사용자 U_b 의 이웃 그룹(D)사이의 공통 이웃(C∩D=$\{U_{c1}, U_{c2}, ..., U_{cp}\}$)을 탐색하여 아래의 식을 이용하여 사용자 U_a 와 U_b 간의 유사도를 근사적으로 계산하는(Approximation) 방법이다. 식과 같이 두 사용자의 공통 이웃과 각 사용자와의 유사도 행렬의 곱을 평균하여 계산한다.

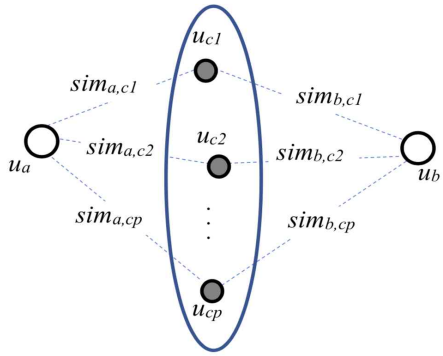

그림 3. 이행적 유사도 모델
Fig. 3. Transitive similarity model

$$sim(u_a, u_b) = \frac{\begin{bmatrix} sim_{a,c1} & sim_{a,c2} & \cdots & sim_{a,cp} \end{bmatrix} \begin{bmatrix} sim_{b,c1} \\ sim_{b,c2} \\ \vdots \\ sim_{b,cp} \end{bmatrix}}{p} \quad (1)$$

3.3 알고리즘

이행적 유사도 모델을 위한 알고리즘은 다음과 같다. U, R_{u_a}, N_{u_a}는 각각 사용자 집합과 사용자 u_a가 평가한 아이템의 선호도 집합, 사용자 u_a와 유사한 이웃의 집합을 의미한다. 두 사용자간에 공통 평가 아이템이 있는 경우와 없는 경우로 구분하여 수행된다. 공통 아이템이 있는 경우는 기존의 피어슨 상관계수를 통해 유사도를 계산하고 공통 아이템이 없는 경우에는 각 사용자의 공통 이웃을 추출하여 3.2절의 식을 통해 이행적 유사도를 계산한다. 유사도 계산 후에는 기존과 같은 예측 선호도를 계산한다.

$u_a, u_b \in U$
$R_{u_a} = \{r_{a,1}, r_{a,2}, \ldots, r_{a,i}\}$
$R_{u_b} = \{r_{b,1}, r_{b,2}, \ldots, r_{b,j}\}$
$N_{u_a} = \{u_{a,1}, u_{a,2}, \ldots, u_{a,s}\}$
$N_{u_b} = \{u_{b,1}, u_{b,2}, \ldots, u_{b,t}\}$

if $R_{u_a} \cap R_{u_b} = \emptyset$ *then*
find common neighbors.
$N_{u_a} \cap N_{u_b} = \{u_{c,1}, r_{c,2}, \ldots, r_{c,p}\}$
if $N_{u_a} \cap N_{u_b} \neq \emptyset$ *then*
calculate transitive similarity.

$$sim(u_a, u_b) = \frac{\sum_{i=1}^{p} sim(u_a, u_{c,i}) \cdot sim(u_b, u_{c,i})}{|N_{u_a} \cap N_{u_b}|}$$

else // $N_{u_a} \cap N_{u_b} = \emptyset$
user b is not registered as neighbor and just return.
go to ***(step-p)***

else // $R_{u_a} \cap R_{u_b} \neq \emptyset$
calculate regular similarity between user a and user b.

$$sim(u_a, u_b) = \frac{\sum_{k=1}^{m}(r_{u_a,i_k} - r_{u_a})(r_{u_b,i_k} - \overline{r_{u_b}})}{\sqrt{\sum_{k=1}^{m}(r_{u_a,i_k} - \overline{r_{u_a}})^2}\sqrt{\sum_{k=1}^{m}(r_{u_b,i_k} - \overline{r_{u_b}})^2}}$$

(step-p)
predict the preference of a user a over item x.

$$p(u_a, i_x) = \overline{r_{u_a}} + \frac{\sum_{k=1}^{m} sim(u_a, u_k)(r_{u_k,i_x} - \overline{r_{u_k}})}{\sum_{k=1}^{m} |sim(u_a, u_k)|}$$

Ⅳ. 실험 및 평가

4.1 실험 환경

성능 평가 실험은 MacOS환경에서 파이썬(3.6.0)으로 구현된 시스템에서 Movielens 데이터셋(사용자 수: 6040, 아이템 수: 3883, 평가횟수: 800,168)을 사용하여 수행되었다. 순수 협업 필터링(PCF, Pure CF)과 제안하는 이행적 협업 필터링(TCF, Transitive CF)을 사용하여 각 지표를 측정 비교하였다. 평균 절대 오차(MAE, Mean Absolute Error)는 예측 선호도와 실제 선호도 값의 차이의 평균을 의미한다[10].

$$MAE = \frac{\sum_{i=1}^{n} |r_{a,i} - p_{a,i}|}{n} \quad (2)$$

정확도는 추천된 아이템 중에 선호했던 아이템의 비율을 의미한다.

$$precision = \frac{|\{items\ preferred\} \cap \{items\ recommended\}|}{|\{items\ recommended\}|} \quad (3)$$

재현율은 선호했던 아이템 중에 추천된 아이템의 비율을 말한다.

$$recall = \frac{|\{items\ preferred\} \cap \{items\ recommended\}|}{|\{items\ preferred\}|} \quad (4)$$

4.2 성능 평가

그림 4의 그래프는 두 사용자간 공통 평가 아이템의 수가 k개 미만인 상황에서 유사도를 평가하여 측정한 MAE 결과이다. X축은 유사도 계산 시 공통 평가 횟수이다. Y축은 피어슨 상관계수에 의한 예측 선호도의 MAE 값이다. 표 1의 실험 결과와 같이 k가 2~5미만인 경우 제안한 TCF 방법의 MAE가 낮은 것으로 나타났다. 그러나 k가 6이상인 경우에는 PCF가 근소하게 우수한 것으로 나타났다. 반면, 공통 평가 아이템이 없는(k=0) 경우 PCF는 수행 불가능하지만 TCF는 정상적으로 수행되어 MAE 측정이 가능하였다. 전체적인 평균 MAE는 TCF가 낮아 정확도가 높은 것으로 확인되었다.

표 1. MAE 측정 결과
Table 1. MAE measurement results

item corating (k)	PCF	TCF
0	Unable	0.9037
2~5	0.9022	0.8836
six over	0.8245	0.8291
Average	0.8439	0.8327

이를 통해 최초의 가정과 같이 제안 방법은 특히 k값이 0이거나 적은 희소 행렬에 대해서 효과적이라고 할 수 있다. 그림 5와 그림 6과 같이 정확도에서는 PCF가 0.012, TCF가 0.0113으로 측정되어 PCF가 관련 아이템 추천에 근소하게 우수한 것으로 나타났으나, 재현율에서는 PCF가 0.8687, TCF가 0.8969로 TCF가 보다 우수한 것으로 확인되었다.

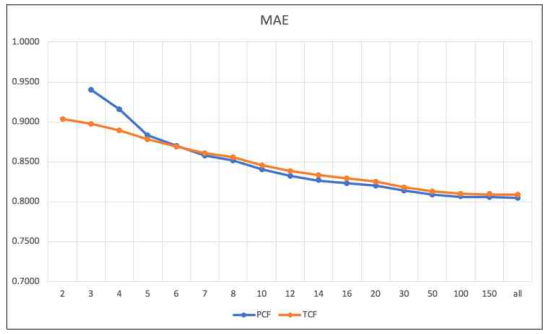

그림 4. 평균 절대 오차
Fig. 4. MAE

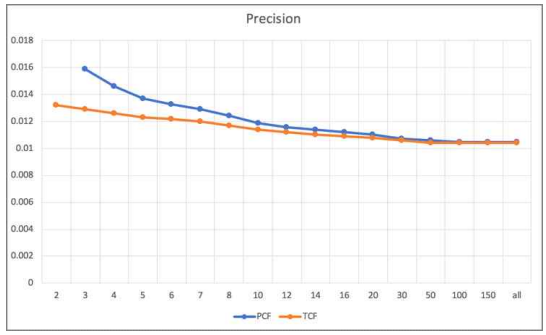

그림 5. 정확도
Fig. 5. Precision

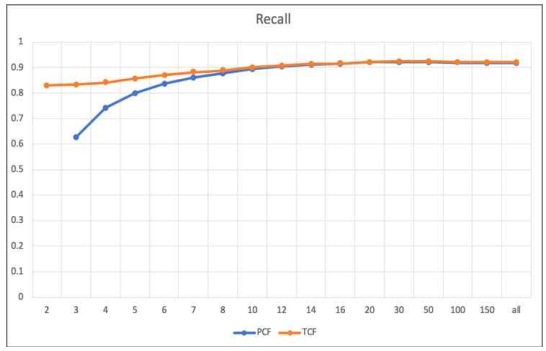

그림 6. 재현율
Fig. 6. Recall

V. 결론 및 향후 과제

본 연구는 협업 필터링의 구조적인 문제점인 평가 행렬의 희박성 문제를 개선하고자 하였다. 희박성이 96%에 이르는 Movielens 데이터셋에서 공통 평가 아이템의 누락으로 사용이 불가능한 상황에서는 실험을 통하여 제안 방법이 효과적임을 확인하였다. 실험에서 발생된 문제점은 제안 방법이 이행적 유사도 계산을 위해 많은 반복 연산을 필요로 한다는 점이다. 중복 계산을 하지 않도록 코드를 개선하였으나 향후 연구에서 이에 대한 추가적인 보완이 필요하다. 또한 묵시적 피드백 데이터에 대해서도 제안 방법이 적용 가능한지 후속 연구로 진행할 계획이다.

References

[1] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl, "Item-based collaborative filtering recommendation algorithms", in Proceedings of World Wide Web 10 Conference, pp. 285-295, May 2001.

[2] J. Herlocker, J. Konstan, A. Borchers, and J. Riedl, "An Algorithmic Framework for Performing Collaborative Filtering", Proc. of 22nd Annual International ACM SIGIR Conference, Research and Development in Information Retrieval, pp. 230-237, Aug. 1999.

[3] Z. Huang, H. Chen, and D. Zeng, "Applying Associative Retrieval Techniques to Alleviate the Sparsity Problem in Collaborative Filtering", ACM Trans. Information Systems, Vol. 22, No. 1, pp. 116-142, Jan. 2004.

[4] G. Groh and C. Ehmig, "Recommendations in Taste Related Domains: Collaborative filtering vs. Social filtering", Proc. of GROUP'07, pp. 127-136, Nov. 2007.

[5] C. Ziegler, S. McNee, J. Konstan, and G. Lausen, "Improving recommendation lists through topic diversification", WWW 2005, Chiba, Japan, pp. 22-32, Jan. 2005.

[6] G. Adomavicius and Y. Kwon, "Improving Aggregate Recommendation Diversity Using Ranking-based Techniques", IEEE Transactions on Knowledge and Data Engineering, Vol. 24, No. 5, pp. 896-911, May 2012.

[7] N. Lathia, S. Hailes, L. Capra, and X. Amatriain, "Temporal Diversity in Recommendation Systems", SIGIR, Geneva, pp. 210-217, Jan. 2010.

[8] H. Ahn, "A new similarity measure for collaborative filtering to alleviate the new user cold-starting problem", Information Sciences, Vol. 178, No. 1, pp. 37-51, Jan. 2008.

[9] S. Yu, "Frequency-sensitive diversification in collaborative filtering", Journal of KIIT, Vol. 13, No. 7, pp. 93-98, Jul. 2015.

[10] B. Patra, R. Launonen, V. Ollikainen, and S. Nandi, "A new similarity measure using Bhattacharyya coefficient for collaborative filtering in sparse data", Knowledge-Based Systems, Vol. 82, pp. 163-177, Jul. 2015.

저자소개

배 은 영 (Eun-Young Bae)

1993년 2월 : 숙명여자대학교 이과대학 통계학과(이학사)
2003년 2월 : 서강대학교 정보통신대학원 정보처리전공(공학사)
2014년 3월 ~ 현재 : 숙명여자대학교 컴퓨터과학과 일반대학원 박사과정
관심분야 : 추천시스템, 정보시각화, 협업필터링, 컴퓨터교육

유 석 종 (Seok-Jong Yu)

1994년 2월 : 연세대학교 컴퓨터과학과(이학사)
1996년 2월 : 연세대학교 컴퓨터과학과(이학석사)
2001년 2월 : 연세대학교 컴퓨터과학과(공학박사)
2005년 ~ 현재 : 숙명여자대학교 소프트웨어학부 교수
관심분야 : 추천시스템, 협업필터링, 정보시각화