신뢰 네트워크에서 직접 연결된 이웃들을 활용한 추천을 위한 대치 방법

차정민*, 황원석**, 김상욱*[1]
*한양대학교 컴퓨터전공
**한양대학교 전자컴퓨터통신공학과
e-mail : {cjm9236, hws23, wook}@hanyang.ac.kr

An Imputation Method Using Directly Connected Neighbors in a Trust Network for Recommendation

Jeong-Min Cha*, Won-Seok Hwang**, Sang-Wook Kim*
*Dept. of Computer Science and Engineering, Hanyang University
**Dept. of Electronics and Computer Engineering, Hanyang University

요 약

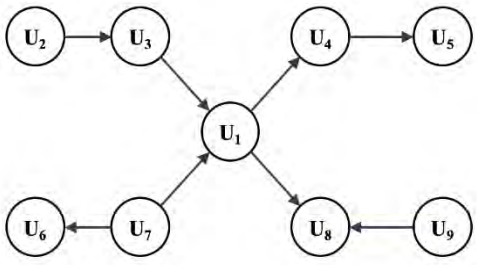
데이터 희소성 문제를 해결하기 위한 방법으로 신뢰 네트워크를 이용한 대치 방법이 제안되었다. 특정 유저로부터 신뢰 네트워크에서 직접 연결된 이웃들이 그 유저와 매우 유사한 취향을 지니고 있음에도 기존의 방법은 이를 간과하였다. 본 논문에서는 직접 연결된 이웃들이 부여한 평점을 통해 데이터 희소성 문제를 더욱 효과적으로 해결하는 방법을 제안한다.

1. 서론

추천 시스템은 사용자, 아이템의 프로필 정보와 함께 사용자가 아이템을 평가, 구매, 검색한 기록 등을 분석하여 사용자가 과거에 평가하지 않은 아이템에 대한 평점, 선호 또는 구매할 가능성을 예측한다. 추천 시스템에서는 사용자가 남긴 평점 정보를 *평점 행렬*로 구성한다. 평점 행렬에서 행은 사용자를, 열은 아이템을 각각 의미하고, 그에 대응하는 셀은 사용자가 아이템에 부여한 평점이 부여되거나 비어있다. 이 때, 평점 행렬이 적게 비어있을수록 추천 시스템은 더 정확한 결과를 도출할 수 있다. 그러나 대부분의 사용자는 다수의 아이템을 평가하지 않기 때문에 평점 행렬에는 비어있는 셀이 많이 존재하게 된다. 그 결과, 추천 시스템은 부정확한 추천 결과를 도출하게 되며, 이러한 문제를 *데이터 희소성 문제* (data sparsity problem) [1]라고 한다.

데이터 희소성 문제를 해결하기 위해서 *대치 방법* (imputation method) [2]이 제안되었다. 이 방법은 평점 행렬에서 비어 있는 셀을 적절한 값으로 채우는 방법이다. 이 때, 채우는 값 (*대치 값*, imputed value)이 사용자가 남길 평점과 유사할수록 추천 시스템이 더 정확한 추천 결과를 도출할 수 있다.

대치 값을 정확하게 계산하기 위하여 *신뢰 네트워크* (trust network)를 이용하는 대치 방법 [3]이 제안된 바 있다. 신뢰 네트워크는 소셜 네트워크의 일종으로, 한 사용자가 다른 사용자를 신뢰하는 신뢰 관계를 기반으로 만들어진 그래프이다. 그림 1 은 신뢰 네트워크의 한 예이다. 그래프에서 정점은 한 사용자를 의미하고, 간선은 한 사용자가 다른 사용자를 신뢰한다는 신뢰 관계를 나타낸다. 예를 들어, 간선 $U_3 \rightarrow U_1$ 은 사용자 U_3가 사용자 U_1을 신뢰함을 의미한다.



(그림 1) 신뢰 네트워크의 한 예시

기존의 대치 방법 [3]은 신뢰 네트워크에서 특정 사용자로부터 도달 가능한 사용자들이 그와 취향이 유사하다고 가정한다. 왜냐하면 신뢰 관계로 연결된 두 사용자는 취향이 유사한 것으로 알려져 있기 때문이다 [4, 5]. 기존 연구에서는 신뢰 네트워크에서 특정 사용자 U 로부터 일정한 거리 내에서 도달 가능한 사용자들을 U의 *신뢰할 수 있는 이웃* (reliable neighbors)이라고 정의하고, 신뢰할 수 있는 이웃들이 부여한 평점들을 종합하여 대치 값을 계산한다. 그림 1 에서 사용자 U_1의 신뢰할 수 있는 이웃들은 U_2, U_3, U_4, U_5, U_6, U_7, U_8, U_9이다.

기존 방법에서는 대치 값을 정확하게 계산하기 위하여 신뢰할 수 있는 이웃들 중 다수가 평가한 아이

[1] 교신 저자

템들에 대해서만 대치 값을 계산한다. 이는 소수 이웃의 의견만을 집계하면 부정확한 대치 값을 도출할 수 있기 때문이다. 즉, 평점 행렬에서 많은 셀을 채우기보다 정확한 값을 계산할 수 있는 셀만 채우려고 하였다.

특정 사용자와 직접 연결된 이웃들은 다른 이웃들에 비해 그 사용자와 매우 유사한 취향을 지닌 사용자들일 것이다. 왜냐하면 신뢰 네트워크에서 거리가 가까운 사용자일수록 더 유사한 취향을 가질 가능성이 높기 때문이다. 따라서 직접 연결된 이웃들 중 한 명만이 아이템을 평가한 경우에도 대치 값을 정확하게 계산할 수 있을 것이다. 이를 통해, 평점 행렬을 더 많이 채우게 되어 데이터 희소성 문제를 더 잘 해결할 수 있다.

2. 제안하는 방법

제안하는 대치 방법은 다음 과정을 통해 대치 값을 계산하여 평점 행렬을 채운다. 먼저, 각 사용자의 신뢰할 수 있는 이웃들을 신뢰 네트워크를 기반으로 찾는다. 그 사용자가 평가하지 않은 아이템들 중 신뢰할 수 있는 이웃들 2 명 이상이 평가하였거나, 직접 연결된 이웃 한 명만이 평가하였다면 이 아이템에 대한 대치 값을 계산한다. 대치 값은 이 아이템을 평가한 이웃들이 부여한 평점을 종합하여 계산된다.

제안하는 대치 방법에서는 사용자 u와 아이템 i에 대한 대치 값은 수식 1 과 같이 계산한다.

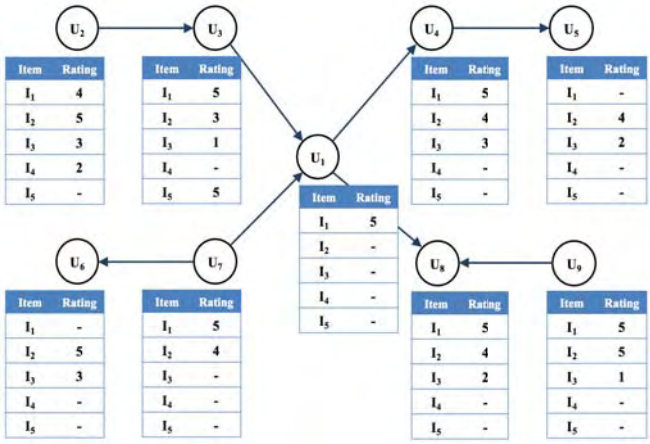
$$r'_{u,i\in\{i|i\in C(u)\}} = \bar{u} + \frac{1}{k}\Sigma_{v\in\{v|r_{v,i}\neq null, v\in N(u)\}}(r_{v,i}-\bar{v}) \qquad (1)$$

$r'_{u,i}$ 는 사용자 u 와 아이템 i 에 대한 대치 값을 나타내고, $r_{u,i}$ 는 u 가 i 에 부여한 평점이다. $C(u)$ 는 u 가 평가하지 않은 아이템 중 대치 값으로 채울 아이템들의 집합을 의미한다. $\bar{u}(\bar{v})$는 사용자 $u(v)$가 부여한 평점들의 평균을 나타내고, $r_{v,i} \neq null$ 는 사용자 v 가 아이템 i 에 대해서 평가하였음을 의미한다. $N(u)$ 는 사용자 u의 신뢰할 수 있는 이웃들의 집합을 의미하며, 본 논문에서는 신뢰 네트워크에서 거리 2에서 도달 가능한 사용자들의 집합으로 정의하였다. 분모 k 는 $\{v|r_{v,i} \neq null, v \in N(u)\}$에 속하는 사용자들의 수를 나타낸다.

기존 대치 방법 [3]에서는 신뢰할 수 있는 이웃들 중 다수가 평가한 아이템에 대해서만 대치 값을 계산하기 때문에, 대치 값을 계산할 때 신뢰할 수 있는 이웃들 중 2명 이상이 평가한 아이템들이 $C(u)$ 에 속한다. 이와 달리, 본 논문의 제안 방법에서는 직접 연결된 이웃 중 단 한 명만이 평가한 아이템에 대해서도 대치 값을 계산한다.

위의 차이점은 그림 2를 통해 자세히 설명 가능하다. 그림 2는 그림 1에서의 각 사용자들이 남긴 평점을 추가로 보이고 있다. 사용자가 아이템에 부여한 평점은 표에서 숫자로 표현되며, "-"는 사용자가 아이템을 평가하지 않은 경우이다.

그림 2에서 사용자 U_1이 평가한 아이템은 I_1뿐이며 아이템 I_2, I_3, I_4, I_5에 대한 평점은 부여되지 않았다. 이때, 제안하는 방법은 아이템 I_2, I_3, I_5에 대한 대치 값을 계산한다. 아이템 I_2, I_3는 신뢰할 수 있는 이웃 중 2명 이상이 평가하였고, 아이템 I_5는 직접 연결된 이웃 U_3가 평가했기 때문이다. 반면, 기존 방법 [3]에서는 아이템 I_2, I_3에 대한 대치 값만이 계산된다. 아이템 I_5는 신뢰할 수 있는 이웃 중 단 한 명만이 평가하였기 때문이다. 아이템 I_4는 2명 이상의 이웃에 의해 평가되지 않았으며, 이 아이템을 평가한 이웃인 U_2는 U_1과 직접 연결된 사용자가 아니기 때문에, 위의 두 방법은 아이템 I_4에 대한 대치 값을 계산하지 않는다.



(그림 2) 신뢰 네트워크와 각 사용자의 평점 정보

본 논문의 제안 방법은 기존의 대치 방법보다 평점 행렬의 비어 있는 셀들을 더 많이, 정확하게 채우기 때문에 데이터 희소성 문제를 더 잘 해결하고, 추천 시스템의 정확도를 향상시킬 수 있다. 또한, 본 방법은 기존 평점 행렬과 동일한 형태인 평점 행렬을 제공하기 때문에 행렬 인수 분해 (matrix factorization) 기법 [6]을 포함한 다양한 추천 시스템에 적용이 가능하다.

3. 평가

실험을 위해 현실 세계의 데이터 셋인 Ciao 를 사용하였다. 이 데이터 셋은 7,375 명의 사용자와 106,797 개의 아이템, 282,619 개의 평점, 그리고 111,781 개의 신뢰 관계를 포함하고 있다.

이 실험에서는 RMSE (평균 제곱근 오차, root mean square error)와 MAE (평균 절대 오차, mean absolute error)를 평가 측정의 기준으로 두었다. 5-fold 교차 검증 (5-fold cross validation) 기법을 사용하여 실험 결과를 측정하였다. 각 fold 에 대해서 80%를 학습 집합, 20%를 검증 집합으로 설정하였다.

기존의 신뢰 네트워크 기반 대치 방법을 적용시킨 행렬 인수분해 기법 [6]과 본 논문에서 제안하는 대치 방법을 적용시킨 경우에 대해서 비교하였다. 표 1 은 두 경우에 대해서 RMSE 를 측정하여 비교한 실험 결과를 보여주고 있다. 모든 fold 에 대해서 기존 방법에 비해서 제안 방법이 적용된 경우가 더 정확도가 높은 것을 확인 할 수 있으며, RMSE 값이 평균적으로 약 1.14% 정도 낮은 것을 확인할 수 있다.

<표 1> 기존 방법과 제안 방법이 적용된 행렬 인수분해 기법에서의 RMSE 측정 값 비교

Fold	RMSE	
	Existing Method	Proposed Method
1	0.263	**0.260**
2	0.262	**0.259**
3	0.264	**0.260**
4	0.263	**0.260**
5	0.263	**0.259**
Avg.	0.263	**0.260**

표 2는 MAE를 측정하여 비교한 실험 결과를 보여주고 있다. 모든 fold 에 대해서 기존 방법보다 제안 방법이 MAE 값이 평균적으로 2.58% 정도 더 낮게 측정되었다.

<표 2> 기존 방법과 제안 방법이 적용된 행렬 인수분해 기법에서의 MAE 측정 값 비교

Fold	MAE	
	Existing Method	Proposed Method
1	0.156	**0.151**
2	0.155	**0.151**
3	0.156	**0.151**
4	0.155	**0.151**
5	0.155	**0.151**
Avg.	0.155	**0.151**

표 1 과 표 2 의 실험 결과에서 모두 제안 방법이 적용된 경우의 오차 값이 더 낮게 나왔다. 이로써, 본 논문에서 제안한 대치 방법이 기존의 대치 방법보다 추천 시스템에 적용되었을 때 더 정확한 추천 결과를 도출할 수 있음을 알 수 있다.

4. 결론

본 논문은 신뢰 네트워크에서 직접 연결되어 있는 이웃의 평점을 통해 더 많은 셀에 대한 대치 값을 계산하는 방법을 제안하였다. 실험을 통해, 제안하는 방법을 통해 생성된 평점 행렬이 기존 대치 방법을 통해 생성된 행렬보다 추천 시스템에서 더 정확한 결과를 도출하는 것을 확인 하였다.

5. Acknowledgments

본 연구는 2015년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (NRF-2014R1A2A1A10054151, No. 2015R1A5A7037751). 또한, 정보통신기술진흥센터의 대학ICT연구센터육성지원사업 (IITP-2015-H8501-15-1013)과 미래 창조과학부 및 정보 통신 기술 진흥 센터의 서울 어코드 활성화 지원 사업 (IITP-2015-R0613-15-1149)의 연구결과로 수행되었음.

6. 참고 문헌

[1] G. Adomavicius and A. Tuzhilin, “Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-art and Possible Extensions,” *IEEE Transactions on Knowledge and Data Engineering*, Vol. 17, No. 6, pp. 734–749, 2005.

[2] H. Ma, I. King, and M. R. Lyu. “Effective Missing Data Prediction for Collaborative Filtering,” In *Proc. of the 30th Annual Int’l ACM SIGIR Conf. on Research and Development in Information Retrieval, SIGIR*, pp. 39–46, 2007.

[3] W. Hwang et al., “Data Imputation Using a Trust Network for Recommendation,” In *Proc. of the 23th ACM Int’l. Conf. on World Wide Web, WWW*, pp. 299-300, 2014.

[4] M. Jamali and M. Ester. “A Matrix Factorization Technique with Trust Propagation for Recommendation in Social Networks,” In *Proc. of the 4th ACM Conf. on Recommender Systems, RecSys*, pp. 135–142, 2010.

[5] P. Massa and P. Avesani. “Trust-Aware Recommender Systems.” In *Proc. of the 2007 ACM Conf. on Recommender Systems, RecSys*, pp. 17-24, 2007.

[6] Y. Koren, R. Bell, and C. Volinsky. “Matrix Factorization Techniques for Recommender Systems,” *IEEE Computer Society*, Vol. 42, No. 8, pp. 30–37, 2009.