그래프 클러스터링을 이용한 추천 시스템 성능 개선 방안

홍동균*, 홍지원**, 이연창**, 김상욱**
*한양대학교 공과대학 컴퓨터전공
**한양대학교 컴퓨터 소프트웨어 학과
e-mail : ghdehdrbs798@naver.com, {nowiz, lyc0324, wook}@hanyang.ac.kr

A Method for Improving Recommender System using Graph Clustering

Dong-Gyun Hong*, Jiwon Hong**, Yeon-Chang Lee**, Sang-Wook Kim**
*Dept. of Computer Science & Engineering, Hanyang University
**Dept. of Computer and Software, Hanyang University

요 약

추천 시스템의 정확도를 향상시키기 위한 방법으로 그래프 클러스터링을 활용한다. 본 논문에서는 실험을 통하여 RWR 알고리즘을 사용하는 추천 시스템의 정확도를 Modularity 기반 클러스터링 알고리즘을 활용함으로써 개선하는 것을 보인다.

1. 서론

추천시스템에 대한 산업계와 학계의 관심은 1990년대부터 계속되고 있으며 많은 발전이 있어 왔다. 그러나 여전히 기존의 추천 시스템들의 정확도는 추천의 대상이 되는 아이템의 방대한 양과 종류에 따라 영향을 받는다. 이러한 문제를 완화하기 위해, 유저와 아이템에 관련한 정보를 더 많이 활용하거나 유저의 상황에 따른 문맥 정보를 통합하는 것, 또는 더욱 적합한 추천 모델링 기법을 사용하는 등의 많은 관련 연구들이 진행되어 왔다 [1].

본 논문에서는 먼저 추천을 위해 데이터를 그래프로 모델링한다. 다양한 모델링 방법 중, 특히 데이터 마이닝 응용 분야 중 하나인 장바구니 분석에서 고객과 구매한 아이템의 관계를 표현하기 위하여 주로 사용하는 이분 그래프(bipartite graph)를 사용하고자 한다 [2].

이와 같이 이분 그래프로 모델링된 데이터를 그래프 분석 알고리즘을 이용하여 각 유저에게 적절한 아이템들을 추천해 줄 수 있다. 그러나 이러한 방법 또한 추천의 대상이 되는 아이템의 수가 많아질수록 부정확한 분석 결과를 제공한다.

따라서, 본 논문에서는 그래프 클러스터링을 통해 얻어진 클러스터 정보를 추가로 활용고자 한다. 그래프 클러스터링 알고리즘을 통해 얻어진 각 클러스터에는 유사한 유저와 아이템으로 구성된다. 해당 클러스터들을 기반으로 각 유저에게 아이템을 추천할 때, (1) 해당 유저가 속하지 않은 클러스터들에 포함된 아이템들은 제외하고, (2) 해당 유저가 속해 있는 클러스터 내의 아이템만을 포함할 수 있다. 다시 말해, 각 유저와 매우 유사한 아이템들만을 추천의 후보로 바라보는 것이다.

본 논문에서는 그래프 클러스터링 기법을 활용하여 추천의 정확도가 향상됨을 실제 데이터를 기반으로 실험하여 확인한다.

2. 추천 시스템

2.1. 그래프 기반 추천 시스템

본 논문에서는 먼저 유저와 아이템을 각각 노드로 바라 보고, 유저가 평가하거나 경험한 적 있는 아이템과 해당 유저 간의 관계를 링크로 바라보는 이분 그래프로 모델링한다. 모델링된 데이터를 기반으로 추천을 하기 위해, 그래프 분석 알고리즘인 Random Walk with Restart (RWR) 알고리즘을 사용하여 각 유저와 가장 유사도가 높은 아이템들을 추천하는 방안을 사용한다.

RWR 은 random surfer 모델을 이용하여 그래프 상의 두 노드 간의 관련성 점수(relevance score)를 계산하는 알고리즘이다. 즉, 각 유저에게는 관련성 점수가 높은 아이템들이 추천할 아이템의 후보로 고려된다 [3].

그러나 서론에서 언급한 바와 같이, 이러한 방법은 추천의 대상이 되는 아이템의 수가 많아질수록 부정확한 분석 결과를 제공한다는 한계를 가진다.

2.2. 제안 방안

본 논문에서는 기존 추천 시스템의 정확도 개선을 위해, 그래프 클러스터링 알고리즘을 사용하여 얻어진 클러스터 정보를 추가로 활용하는 방안을 제안한다. 먼저 클러스터링을 위해, 그래프 클러스터링 알

고리즘인 Modularity 기반 클러스터링을 사용한다. Modularity 기반 클러스터링 알고리즘은 그래프를 계층적으로 클러스터링하는 방법으로, 클러스터 내부의 링크와 그 링크들 간의 밀집도를 기반으로 modularity 를 계산하여 가장 높은 modularity 를 갖는 클러스터들로 구성하기 위해 사용되는 방법이다 [5]. 기존 방법과 마찬가지로 추천을 위해 RWR 알고리즘을 사용한다. 다만, 최종적으로 추천 아이템을 선별할 때, 클러스터 정보를 확인하여 각 유저와 같은 클러스터에 포함되지 않은 아이템들은 제외한다.

본 논문에서는 클러스터링 알고리즘을 통해 얻은 클러스터 정보를 추가로 활용함으로써 기존 추천 시스템의 정확도를 개선할 수 있는지 확인하기 위하여 실험을 진행하였다.

3. 실험

3.1. 데이터

본 실험은 미네소타 대학 컴퓨터 공학과의 GroupLens 연구실에서 제공하는 MovieLens 데이터를 사용하여 수행하였다 [4].

데이터 셋으로는 약 1000 명의 유저가 약 1700 편의 영화에 평점(1~5)을 매긴 10 만개의 평가 데이터 셋과, 약 6000 명 정도의 유저가 약 4000 편의 영화에 평점을 매긴 100 만개의 평가 데이터 셋을 사용한다.

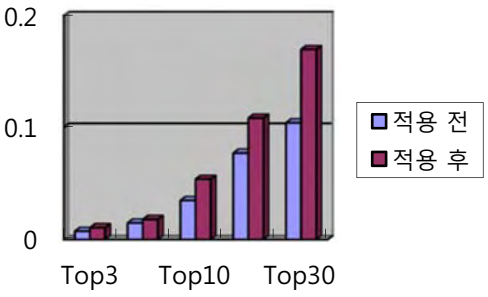
3.2. 실험 방법

MovieLens 데이터를 기반으로 추천을 수행하여 클러스터링 정보 적용 전과 후의 정확도를 비교한다. 이를 위해, MovieLens 100K, 1M 데이터 셋을 트레이닝 셋과 테스트 셋으로 구성하는데, 이 때 테스트 셋은 전체 데이터 셋에서 각 유저가 평점을 매긴 아이템들 중 하나씩 만을 임의로 선택하여 구성하였고, 트레이닝 셋은 테스트 셋에 포함되지 않은 데이터로 구성하였다.

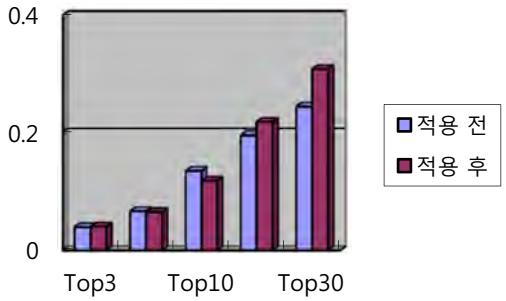
본 논문에서는 정확도 측정을 위해, 재현율(recall)을 사용하는데, 이는 추천된 아이템들 중에 정답 아이템이 포함된 경우를 전체 정답 아이템의 수로 나누어 계산된다.

3.3. 실험 결과

그림 1 과 2 는 각각 100K, 1M 크기의 데이터를 가지고 실험을 수행한 결과이다. X 축은 상위 N 개를 가지고 정확도를 측정했음을 의미하며, Y 축은 재현율을 의미한다.



(그림 1) 100K 데이터 셋 추천 정확도.



(그림 2) 1M 데이터 셋 추천 정확도

먼저, 100K 데이터를 이용한 실험 결과는 클러스터링을 적용한 후에 추천 정확도가 평균 43% 상승했음을 보여준다. 또한, 1M 데이터를 이용한 실험 결과는 상위 10 개의 경우에 정확도가 약간 감소하지만, 결과적으로 상위 30 개까지 추천하였을 때는 정확도가 상승하여 평균 5%의 정확도 향상을 가지는 것을 보여준다.

4. 결론

본 논문에서는 그래프 클러스터링 기법을 활용하여 추천 시스템의 추천 정확도를 향상시킬 수 있는지 실험을 통해 검증해 보았다. 결과로는 추천 아이템을 클러스터링의 결과를 활용해 제한함으로써 추천 정확도를 향상시킬 수 있음을 보였다.

Acknowledgements

본 연구는 (1) 미래창조과학부 및 정보통신기술진흥센터의 서울어코드활성화지원사업 (IITP-2015-R0613-15-1149)과 (2) 미래창조과학부 및 정보통신기술진흥센터의 대학 ICT 연구센터육성 지원사업 (IITP-2015-H8501-15-1013)의 연구결과로 수행되었음. 또한, (3) 미래창조과학부의 재원으로 한국연구재단 (NRF-2014R1A2A1A10054151)의 지원을 받아 수행된 연구임.

참고문헌

[1] Adomavicius, Gediminas, and Alexander Tuzhilin. "Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions." Knowledge and Data Engineering, IEEE Transactions on 17.6 (2005): 734-749.

[2] Zha, Hongyuan, et al. "Bipartite graph partitioning and data clustering." Proceedings of the tenth international conference on Information and knowledge management. ACM, 2001.

[3] Fouss, Francois, et al. "Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation." Knowledge and data engineering, ieee transactions on 19.3 (2007): 355-369.

[4] http://grouplens.org/about/what-is-grouplens/

[5] Shiokawa, Hiroaki, Yasuhiro Fujiwara, and Makoto Onizuka. "Fast Algorithm for Modularity-Based Graph Clustering." AAAI. 2013.