온라인 공군집화를 이용한 추천 및 콜드스타트 문제 해법 연구

김병희 · 장병탁
Byoung-Hee Kim and Byoung-Tak Zhang

서울대학교 컴퓨터공학부
E-mail: {bhkim, btzhang}@bi.snu.ac.kr

요 약

추천 시스템을 준비하는 과정에서 새로운 사용자 또는 추천 아이템을 고려해야 하는 콜드스타트 문제는 현재의 표준적인 협업 필터링 방식으로 처리하기 힘든 구조적 한계가 있다. 본 논문에서는 온라인 방식의 군집화를 적용한 협업 필터링 기법을 이용한 콜드스타트 문제의 해법을 제시한다. 특히, 최신 협업 필터링의 핵심 기법인 행렬 분해 방식에 기반한 사용자-아이템 공군집화를 통해 예측 정확도와 규모에 대한 적응을 지향한다. 영화 추천 데이터에 대한 온라인 방식의 추천 시스템 확장 사례에 제안한 기법을 적용하여 유용함을 보인다.

키워드 : 추천 시스템, 공군집화, 콜드스타트

1. 서 론

추천 시스템은 고객이 소비할 대상을 선택하는 과정을 도와주는 도우미로서 다양한 온라인 서비스의 필수 요소 기술로 자리잡고 있다. 서비스의 규모가 커짐에 따라 추천 시스템도 함께 변화할 필요가 있으며, 이를 위해 온라인 방식으로 추천 시스템을 업데이트하는 방법이 필수적이다. 즉, 데이터가 순차적으로 들어올 때 새로운 데이터를 고려하여 기존 모델을 순차적으로 업데이트해야 한다.

추천 시스템의 핵심 모델을 순차적으로 업데이트하기 위해서는 콜드스타트 문제에 대한 해법이 필요하다. 즉, 새로운 사용자, 또는 새로운 추천 대상 아이템을 고려한 추천 방법이 필요하다. 내용 기반 추천 방식(content-based recommendation)이나 인기도 기반의 협업적 필터링(collaborative filtering)이 기본적인 해법으로 사용 가능하지만, 온라인 환경에서의 개인화된 추천에는 적절하지 않다. 추천의 다양성을 위한 해법으로서 군집화(clustering)를 도입하는 개념이 여러 연구자들에 의해 검토된 바 있다. 특히, 사용자와 아이템을 함께 군집화하는 공군집화는 근접 이웃(neighborhood) 기반의 협업적 필터링에 비해 온라인, 증분적 확장이 용이한 장점이 있다.

본 논문에서는 근접 이웃 방식과 행렬 분해 방식(matrix factorization)이 결합된 협업 필터링 기법과 군집화 방식의 협업 필터링을 분석하여 온라인 방식의 추천이 가능하고 특히 콜드스타트를 해결할 수 있는 해법을 도출하는 과정을 논의한다.

2. 추천 시스템과 콜드스타트 문제

협업 필터링 방식은 아이템에 대한 명시적(explicit) 또는 비명시적인(implicit) 평점 정보를 재료로 추천을 수행한다. 명시적 평점 정보가 주어진 경우, 아이템-아이템 이웃 모델의 경우 새로운 사용자가 한 번이라도 피드백을 주는 경우 바로 추천 가능한 기본 콜스트타트 해법이다. 새로운 아이템에 대해서는 파라미터를 추가로 학습해야만 하기 때문에 콜스트타트 해법으로서는 적합하지 않다.

Netflix 대회 우승팀의 일원으로 유명한 Y. Koren의 SVD++ 계열 알고리즘은 현재 Netflix[1]를 비롯하여 여러 추천 시스템에 적용되고 있다. Koren이 [1]에서 제안한 알고리즘 중, 비대칭적 SVD(Asymm-SVD의 경우의 경우 아이템-아이템 이웃 방식과 마찬가지로 새로운 사용자가 단 한 건이라도 피드백을 주면 이에 기초한 추천이 가능하다.

코렌의 비대칭 SVD에서는 아이템 유사도 기준과 은닉요인 판별 기준을 결합하였으며, 사용자 u의 명시적 선호 정보와 비명시적 선호 정보를 기준으로 이웃 범위를 설정하고 다음과 같이 평점 예측에 반영한다[1].

$$\hat{r}_{ui} = b_{ui} + q_i^T\Big(|R(u)|^{-1/2}\sum_{j\in R(u)}(r_{uj}-b_{uj})x_j + |N(u)|^{-1/2}\sum_{j\in N(u)} y_j\Big), \quad (1)$$

$$b_{ui} = \mu + b_u + b_i. \text{ (baseline estimator)} \quad (2)$$

사용자 u의 아이템 i에 대한 평점에 대한 예측치 $\hat{r}_{ui}$에 대해, 데이터 전체의 평점 평균 μ 및 사용자와 아이템 각각에 대한 개별적 편향치 b_u, b_i의 합으로 기반 추정치 b_{ui}를 계산할 수 있다. 사용자와 아이템 간의 연관성을 추가로 고려하기 위해, 명시적 평점 데이터 행렬과 비명시적 평점 데이터 행렬을 SVD를 기반으로 각각 분해하여 사용자와 아이템 데이터를 공통으로 표현하는 저차원의 은닉 공간을 탐색하고, 이 은닉공간 상에서 표현되는 아이템 벡터 q_i와 사용자 벡터 p_u를 구하는 것이 행렬 분해의 기본 개념이다. Koren은 p_u를 식(1)에서와 같이 명시적 행렬 및 비명시적 행렬을 기준으로 한 사용자 u의 로컬 이웃 $R(u)$, $N(u)$ 간 유사도 x_j, y_j를 고려하여 평균을 취함으로써, 계산 속도와 정확도가 향상됨을 보였다.

이러한 Koren의 연구를 확장하여, [2]에서는 명시적 평점

1) 2014년 말 기준 SVD++과 RBM 기반 협업 필터링[3]의 앙상블 기법을 적용하는 것으로 알려져 있다.

정보에서 행렬분해로 얻은 은닉 공간을 기준으로 공군집화를 적용한 후, 군집 간 평균 선호도의 관계를 고려한 확장을 시도하였다.[2] 클러스터링 결과(C_i와 같은 아래 첨자가 포함된 벡터)와 기존 정보 간의 선택적 가중치 a를 도입하여 종합적으로 평점을 예측한다.

$$\hat{r}_{ui} = b_{ui} + \Big((1-\alpha)q_i + \alpha q^*_{C_i}\Big)^T (|R(u)|^{-1/2} \sum_{j\in R(u)} (r_{uj} - b_{uj})\Big((1-\alpha)x_j + \alpha x^*_{C_j}\Big) + |N(u)|^{-1/2} \sum_{j\in N(u)} \Big((1-\alpha)y_j + \alpha y^*_{C_j}\Big)\Big), \quad (3)$$

이 방식에서 사용자와 아이템의 연관 관계를 표현하는 은닉 공간을 온라인 방식으로 업데이트를 하고, 온라인 방식의 군집화를 적용하면, 비명시적 데이터까지 고려한 신규 사용자 고려 가능 추천 모델을 얻을 수 있다. 이 모델은 저자의 기존 공군집화 모델[4]에서 선호 행렬의 관측값만을 고려하되 차원 축소를 하지 않고, 보조 정보를 결합하기만 했던 연구가 콜드스타트 해법을 제시하지 않는 점을 보완한다.

3. 행렬 분해 기반 온라인 공군집화

이 장에서는 최근의 연구를 기초로 [1]과 [2]를 온라인 방식으로 확장하여 개인화하는 솔루션을 모색한다. 솔루션이 갖출 기본 절차는 다음과 같다.

Step 1. 대규모 공용 선호도 데이터에 '온라인' 방식의 행렬 분해 기법 적용하여 사용자와 아이템 간의 공통적 저차원 은닉공간 획득 및 추적

Step 2. 행렬 분해 결과 얻은 은닉 공간 상에서 '온라인' 방식의 공군집화 기법을 적용하여 사용자와 아이템으로 구성된 공군집 탐색 및 추적

Step 3. 공군집화 결과를 추가로 고려하여 식 (3)과 같은 방식으로 일반 사용자의 평점 예측.

Step 4. 개인 사용자의 비명시적 정보 위주의 추천 시스템에 공용의 명시적 평점 행렬에서 얻은 공군집화 단위의 평점의 요약 정보(평균, 분산 등)를 추가로 적용하여 개인 사용자의 선호도 예측

각 절차별로 적용 가능한 최근의 연구를 정리하면 다음과 같다. 행렬 분해에 대한 효율적인 증분적 해법을 구성하는 문제는 매우 도전적인 과제이다. Step 1에는 Luo 등이 제안한 증분적 선형 저차원 행렬 분해 기법[5]를 적용할 수 있다. 계산 시간이 오래 걸리는 것을 감수할 수 있다면 [6]과 같은 active Bayesian 방식의 해법을 고려할 수도 있다. Step 1에서 행렬 분해를 통해 얻은 저차원 공간에 대해 Step 2에 적용 가능한 온라인 공군집화 기법으로는 [7][8] 등이 있다. 이 단계에서 새로운 사용자에 대해 $\rho(u) = \arg\min_g \sum_h n_{uh}(\bar{\epsilon}_{uh} - \bar{\epsilon}_{gh})^2$와 같이 공군집 할당 및 할당된 공군집을 기초로 한 평점 예측이 가능하다. 공군집화 단계에서 사용자와 아이템이 하나의 군집이 아닌 다수의 군집에 부분적 멤버쉽을 가지도록 하는 경우는 비모수적 베이지안 기법을 기반의 다양한 해법을 적용할 수 있다. 예를 들면, PMF, Bayesian PMF, pLSI, LDA, HDP 등이 가능하다. 이러한 경우 Step 3에서는 다수의 평점 예측값을 얻게 되며, 이를 취합하기 위해, 단순한 가중치합, 또는 Bayesian 모델 평균(BMA) 방식의 종합을 적용할 수 있다.

2) 이 논문에서는 [1]과 달리 전체 평균을 빼는 'centering'이 적용되지 않았고, 비대칭 SVD 설정시 원본에서는 다른 요인으로 대체하여 생략하는 p_u를 추가로 고려하는 방식을 취한다.

4. 논의 및 결론

본 논문에서는 온라인 공군집화 기반의 기법을 제안하였다. 영화 추천을 위한 표준적인 데이터셋에 적용한 실험 결과를 후속 연구로 보고할 예정이다.

본 논문에서는 행렬 분해로 얻은 은닉공간에서의 사용자 및 아이템 군집화가 핵심 과정이다. 실제 사례에 적용하는 경우 [9]에서 지적하였듯이 전체 평점 행렬을 분해하는 것은 신뢰할 만한 은닉공간을 도출하는데 한계가 있을 수 있다.

추천시스템의 성능을 비명시적 피드백과 콜드스타트 위주의 환경에 적용하기 위한 확장 연구를 위해 다음과 같은 핵심질문을 도출할 수 있다. 명시적으로 표현된 선호도에서 얻은 군집화 또는 은닉변수 표현을 비명시적 선호도 표현 학습에 활용할 수 있는가? 전이학습 관점에서 강인한 방법론은 어떠한 접근법이 가능한가? 이러한 질문에 대한 답을 구하는 방향으로 향후 연구를 진행할 계획이다.

감사의 글

이 논문은 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구이며(NRF-2010-0017734, Videome), 정부(미래창조과학부 및 정보통신기술진흥센터)의 정보통신·방송 연구개발사업 지원(10035348- mLife, 10044009-HRI.MESSI)을 일부 받았음.

참고문헌

[1] Y. Koren, "Factorization Meets the Neighborhood: A Multifaceted Collaborative Filtering Model," in *Proc. KDD*, pp. 426 - 434, 2008.

[2] N. Mirbakhsh and C. X. Ling, "Clustering-based factorized collaborative filtering," in *Proc. RecSys*, pp. 315 - 318, 2013.

[3] R. Salakhutdinov, A. Mnih, and G. Hinton, "Restricted Boltzmann Machines for Collaborative Filtering," in *Proc. ICML*, pp. 791 - 798, 2007.

[4] 김병희, 장병탁, Infinite Relational Model 기반 Co-Clustering을 이용한 영화 추천, 한국지능시스템학회논문지, 24(4):443-449, 2014.

[5] X. Luo, Y. Xia, and Q. Zhu, "Incremental Collaborative Filtering Recommender Based on Regularized Matrix Factorization," *Know.-Based Syst.*, vol. 27, pp. 271-280, 2012.

[6] J. Silva and L. Carin, "Active learning for online Bayesian matrix factorization," in *Proc. KDD*, pp. 325 - 333, 2012.

[7] M. Khoshneshin and W. N. Street, "Incremental Collaborative Filtering via Evolutionary Co-clustering," in *Proc. RecSys*, pp. 325 - 328, 2010.

[8] J. Peltonen, J. Sinkkonen, S. Kaski, "Sequential information bottleneck for finite data," in *Proc. ICML,* 2004.

[9] J. Lee, S. Kim, G. Lebanon, and Y. Singer, "Local Low-Rank Matrix Approximation," in *Proc. ICML*, pp. 82 - 90, 2013.