평가 점수와 평가 횟수를 활용하는 추천 방안

박정빈*, 박주안*, 황원석**, 김상욱**
*한양대학교 컴퓨터공학부
**한양대학교 전자컴퓨터통신공학과
e-mail:alien087@naver.com

A Recommendation Method Exploiting the Scores and Numbers of Ratings

Jeong-Bin Park*, Ju-An Parc*, Won-Seok Hwang**, Sang-Wook Kim**
*Dept. of Computer Science, Hanyang University
**Dept. of Electronics and Computer Engineering, Hanyang University

요 약

기존에 제안된 K-NN 기반 추천 방안은 타겟 유저와 유사한 유저들 중 소수만이 높은 평점을 부여한 아이템을 높게 평가하는 문제점이 존재한다. 따라서 본 논문에서는 유사한 유저들 중 일부만 평가한 아이템은 추천에서 제외하는 방안을 제안하고, 실험을 통해 제안하는 방안의 정확성을 검증하였다.

1. 서론

Amazon.com, Epinion.com, MovieLens와 같은 상품 판매 또는 상품 정보 제공 서비스가 온라인상에서 활성화되고 있다. 이와 같은 서비스에서 등록된 상품은 지속적으로 증가하고 있으며, 이로 인하여 자동으로 사용자가 좋아하는 상품을 추천해 주는 추천 시스템(recommendation system)에 대한 연구의 필요성이 대두되고 있다.

기존의 추천 시스템들은 크게 내용 기반 방안, 협업 필터링 방안, 통합 방안으로 구분할 수 있다[1]. 협업 필터링은 가장 널리 연구되고 있는 분야이며, 가장 대표적인 방법으로는 K-Nearest Neighbor(K-NN) 기반 추천 방안[2]이 있다. K-NN 기반 추천 방안은 각 아이템에 부여한 평점이 타겟 유저와 유사한 k명의 유저를 찾고, 타겟 유저가 평가하지 않은 아이템의 평점을 유사한 k명이 그 아이템에 대해 부여한 평점을 종합하여 예측한다. 이 때, 유사한 k명의 유저 중 일부만이 그 아이템에 대해 부여하였다면, 그 유저들의 평가만으로 점수를 예측하게 된다.

K-NN 기반 추천 방안은 평점을 예측하는 방법을 제안하였으나, 실제로는 예측된 평점이 높은 일부 아이템이 유저에게 추천된다. 이 경우, K-NN 기반 추천 방안은 타겟 유저와 유사한 유저들 중 몇 명이 해당 아이템을 평가하였는지를 무시하기 때문에 문제가 발생할 수 있다.

그림 1은 이러한 문제점을 그림으로 나타낸 것이다. 그림 1에서 타겟 유저와 유사한 대부분의 유저는 i_2를 평가하지 않았다. 그러나 유저 c_1이 높은 평점을 부여하였기 때문에 최종적으로는 i_2가 추천된다. 그러나 i_1은 다수의 유저가 관심을 가지고 평가하였으며, 꽤 높은 점수를 받았기 때문에 i_2보다 우선적으로 추천되는 것이 옳다. 이러한 문제를 해결하기 위하여 유사한 k명의 유저가 부여한 평가와 함께 몇 명의 유저가 평가를 부여했는지 또한 고려하여 추천을 해야 한다.

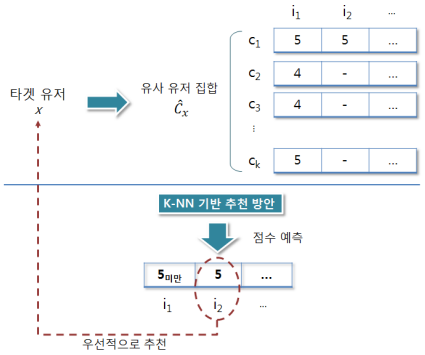

그림 1. KNN 기반 추천 방안의 문제.

2. 제안하는 방안

본 논문에서 제안하는 추천 방안에서는 타겟 유저와 유사한 유저들을 찾아낸다. 이를 위하여, 우선적으로 아이템에 부여한 평점에 대해 각 유저간의 유사도를 correlation coefficient를 이용하여 측정한다.

$$sim(x,y) = \frac{\sum_{s \in S_{xy}} (r_{x,s} - \overline{r_x})(r_{y,s} - \overline{r_y})}{\sqrt{\sum_{s \in S_{xy}} (r_{x,s} - \overline{r_x})^2 \sum_{s \in S_{xy}} (r_{y,s} - \overline{r_y})^2}} \quad (1)$$

수식1에서 $r_{x,s}$는 유저 x가 아이템 s에 부여한 평점을 의미하며, $\overline{r_x}$은 유저 x가 평가한 모든 아이템의 평점의 평균이다. 또한 S_{xy}는 유저 x와 y가 모두 평가한 아이템 집합을 나타낸다. 제안하는 방안에서는 타겟 유저와 타 유저와의 유사도를 계산하여 가장 유사도가 높은 k명의 유저를 찾아 유사 유저 집합 $\hat{C}_x$으로 정의한다.

제안하는 방법에서는 타겟 유저에게 $\hat{C}_x$에 포함된 유저들 중 다수가 평가한 아이템만을 추천한다. 이를 위해 $\hat{C}_x$의 유저들 중 아이템을 평가한 유저의 수와 함께 타겟 유저와 $\hat{C}_x$의 유저들의 유사도를 동시에 고려하여 수식2로 정의한다. 이는 동일한 수의 유저가 평가한 두 아이템이 있는 경우, 더 유사한 유저들이 평가한 아이템이 추천될

가능성이 높아지도록 하기 위함이다. 수식 2에서 $r'_{x,s}$는 유저 x가 아이템 s를 평가했다면 1, 그렇지 않은 경우 0의 값을 가진다. 제안하는 방법에서는 $p_{x,s}$의 값이 큰 일부 아이템만을 추천 대상으로 고려한다.

$$p_{x,s} = \frac{\sum_{y \in \hat{C}_x} sim(x,y) \times r'_{y,s}}{\sum_{y \in \hat{C}_x} sim(x,y)} \quad (2)$$

추천 대상으로 선별된 아이템들에 대해 제안하는 방법에서는 타겟 유저가 각 아이템에 부여할 평점을 예측한다. 예측 점수는 기존의 K-NN 기반 추천 방법을 통해 결정한다[2]. 타겟 유저 x의 아이템 s에 대한 평점은 $\hat{C}_x$의 유저가 아이템 s를 평가한 평점과 타겟 유저와 $\hat{C}_x$의 유저간의 유사도를 고려하여 수식3으로 정의된다.

$$r_{x,s} = \overline{r_x} + \frac{\sum_{y \in \hat{C}_x} sim(x,y) \times (r_{y,s} - \overline{r_y})}{\sum_{y \in \hat{C}_x} sim(x,y)} \quad (3)$$

점수 예측은 유사 유저 집합의 유저들 중 다수가 평가한 것으로 선별된 아이템들에 대해서만 수행되며, 이 결과로 각 아이템에 부여된 점수를 통해 가장 점수가 높은 아이템 일부가 추천된다.

제안하는 방법의 점수 예측 방법은 기존의 추천 방법과 동일하지만, 점수 예측 이전에 아이템을 선별하는 과정을 거침으로써 유사 유저들 중 다수가 평가하지 않은 아이템은 추천하지 않도록 한다. 그 결과 최종적으로 추천되는 소수의 아이템은 기존 방법과 제안하는 방법이 전혀 다른 결과가 되며, 더 나은 추천 결과를 제공할 수 있다.

3. 실험

제안하는 방법의 추천 결과가 유저에게 의미 있음을 보이기 위하여 MovieLens 데이터를 이용하였다. 수집된 데이터는 유저 943명의 영화 1,682편에 대한 100,000개의 평가(1-5의 정수)로 구성되어 있다.

실험의 평가를 위해 5-fold cross validation을 이용하였다. 또한, 평가를 위한 척도로는 *top-k hit ratio*[3]를 이용하였다. top-k hit ratio는 추천된 정답들 중 실제 정답이 얼마나 있는지 그 비율로 계산되며, 그 값이 높을수록 정확도가 높은 것을 나타낸다. 이 척도에서 정답은 테스트 세트에서 타겟 유저가 5점을 준 아이템들로 간주하였다.

본 실험에서는 제안하는 방법과 함께 K-NN 기반 추천 방안을 비교하였다. 제안하는 방법과 K-NN기반 추천 방안에서는 $\hat{C}_x$의 유저 수인 k를 결정해야 한다. 이를 위하여 k를 변화시켜가며 top-k hit ratio를 측정하였다. 그림 2는 그 실험 결과이다. 그림 2에서 x축은 유사 유저의 수인 k를 나타내며, y축은 top-k hit ratio를 나타낸다. 그림 2에서와 같이 k가 10일 때 가장 좋은 결과를 보였으며, 이후 실험에서 이용하는 모든 방안의 k는 10으로 수행을 한다.

제안하는 방안에서는 수식2를 통해 아이템을 미리 선별하며, 이 때 몇 퍼센트의 상위 아이템을 남길 것인지를 결정해야 한다. 이를 위하여 선별할 상위 아이템의 퍼센트를 조절하며 top-k hit ratio를 측정하였다. 그 결과 그림 3과 같은 결과를 얻었다. 그림 3에서 x축은 남긴 선별된 상위 아이템의 비율을 나타내며, y축은 top-k hit ratio를 나타낸다. 그림 3에서와 같이 10%일 때 가장 좋은 결과를 보였다. 따라서 이후 실험에서 제안하는 방안은 상위 10%의 아이템을 선별한다.

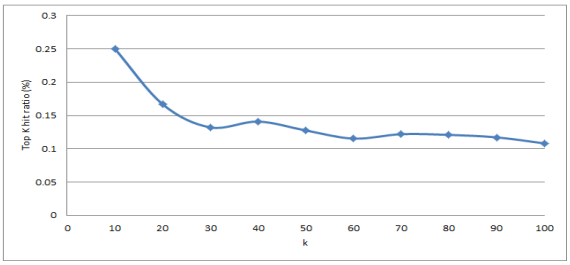

그림 2. 제안하는 방안/K-NN 기반 추천 방안의 k.

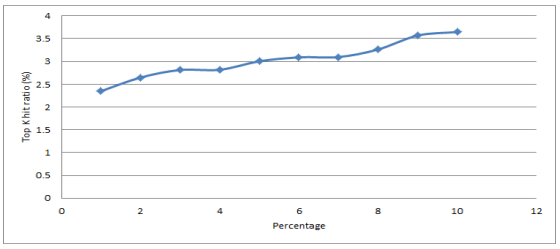

그림 3. 제안하는 방법에서 상위 아이템의 퍼센트.

표 1은 제안하는 방안과 K-NN 기반 추천 방안을 top-k hit ratio를 이용하여 비교한 결과이다. K-NN 기반 추천 방안에 비해 제안하는 방안이 더 정확한 것으로 나타났다. 이는 몇 명의 유저들이 평가했는지를 고려하는 것이 최종 아이템 추천에 큰 도움이 된다는 것을 의미한다.

표 1. 제안하는 방안과 기존 방안의 정확도 비교

척도 / 알고리즘	top-k hit ratio (%)
제안하는 방안	3.65
K-NN 기반 추천 방안	0.25

4. 결론

기존에 제안된 K-NN 기반 추천 방안은 타겟 유저와 유사한 유저들 중 일부만이 높은 평점을 부여하였더라도, 그 아이템이 우선적으로 추천되었다. 그러나 다수의 유저가 높게 평가한 아이템을 우선적으로 추천하는 것이 더 정확한 결과를 도출할 수 있다. 본 논문에서는 타겟 유저와 유사한 유저들 중 아이템을 평가한 유저의 수를 고려하는 추천 방안을 제안하였으며, 실험을 통해 제안하는 방안이 기존의 추천 방안보다 더 정확함을 검증하였다.

감사의 글

본 연구는 지식경제부 및 정보통신산업진흥원의 IT융합 고급인력과정지원사업의 연구결과로 수행되었음(NIPA-2011-C6150-1101-0001)

참고문헌

[1] G. Admavicius 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*, pp. 734-749, 2005.

[2] J. Herlocker, J. Konstan, A. Borchers, and J. Riedl, "An algorithmic framework for performing collaborative filtering," In *Proc. of the 22nd ACM Conf. on Research and Development in Information Retrieval,* SIGIR '99, pp. 230-237, 1999.

[3] H. Steck, "Training and Testing of Recommender Systems on Data Missing Not at Random," In *Proc. of the 16th ACM SIGKDD Int'l. Conf. on Knowledge Discovery and Data Mining*, pp. 713-722, 2012.