협력적 여과 시스템의 예측 정확도 향상을 위한 전처리 방법

김교창[0], 전종훈
명지대학교 컴퓨터공학과
{kgchang[0], jchun}@mju.ac.kr

A Preprocessing Method for Improving Prediction Accuracy in Collaborative Filtering

Kyochang Kim[0], Jonghoon Chun
Department of Computer Science and Engineering
Myongji University

요 약

본 논문에서는 협력적 여과방식에서 고객의 특정 상품에 대한 선호도 예측의 정확도를 향상하기 위해 상품의 선호도 값에 가중치를 반영하는 전처리 방법을 제안한다. 이를 위해 고객별 상품의 선호도 값에 정보검색 분야에서 사용되고 있는 벡터 공간 모델을 이용하여 가중치를 부여하며, 이를 통하여 특정 상품을 선호하는 고객과 전체 상품을 고루 선호하는 고객간의 차별화 값을 반영하여 보다 정확한 선호도를 예측할 수 있게 된다. 전처리 과정을 수행하지 않은 기존의 협력적 여과 방식과의 실험을 통한 비교 분석을 통하여 본 논문이 제안하는 전처리 과정의 타당성과 비교우위를 검증한다.

1. 서론

인터넷 환경의 급속한 발달과 보급으로 인하여 이를 이용한 전자상거래 시장이 빠르게 증가하면서 개인화(Personalization)는 전자상거래의 중요한 단어로 떠올랐다. 현재 인터넷 상의 수많은 전자 상거래 업체들의 주요 관심사는 고객의 과거 구매 이력을 조사 분석하여 고객이 흥미를 가질만한 상품을 알아내고 이러한 상품을 추천함으로써 기업의 이익을 극대화 하는 것이다. 따라서 고객에게 어느 정도의 개인화 서비스를 제공하느냐는 앞으로 전자상거래의 성패를 좌우하게 될 것이다. 이러한 개인화 서비스를 위해 다양한 알고리즘들이 연구 되어왔다. 이들 알고리즘 중에 가장 대표적인 것은 협력적 여과(Collaborative Filtering)알고리즘 [1][2][3][4] 이다.

근래에 들어서 협력적 추천 알고리즘의 성능향상을 위한 많은 연구들이 진행되고 있으며, 연구결과로서 다양한 협력적 추천 알고리즘들이 제안되고 있다. 협력적 여과 방법은 TAPESTRY 에서 유래 되었으며[5] TAPESTRY의 문제점을 해결하고 상당한 수준의 성능향상을 가져온 협력적 여과 시스템으로 GroupLens가 있다[2].

GroupLens는 고객 기반의 추천시스템(user-based recommendation system)으로 고객의 상품에 대한 선호도를 평가할 때, 유사 성향의 고객들이 유사 상품들에 대해 유사한 선호도를 가질 것이라는 전제를 두고 있다. GroupLens의 추천방식은 고객 들이 상품에 대해 평가한 선호도를 이용하여 성향이 비슷한 고객들끼리 분류한 후, 상품을 추천할 때 그 고객과 가장 유사한 그룹을 찾아, 그룹에 속한 고객들의 상품에 대한 선호도로부터 특정 고객의 상품에 대한 선호도를 예측한다. 이러한 과정을 통해 예측된 선호도 값이 높다면 이 고객은 그 상품에 대해 흥미를 가질 확률이 높다고 예상할 수 있다. 그러나 고객 기반의 추천 시스템은 다량의 상품에 대해 일정한 선호도를 갖는 고객들과 극소수의 특정 상품에 대해서만 월등한 선호도를 갖는 고객들에 대한 차별화가 이루어 지지 않기 때문에 정확한 추천이 어렵다는 단점이 있다.

협동적 여과 시스템의 다른 방식은 상품 기반의 추천시스템(item-based recommendation system)으로 고객이 상품에 대한 선호도를 예측할 때 비슷한 상품들끼리 그룹으로 분류한 후에 예측하고자 하는 상품과 비슷한 그룹에 속한 상품간의 유사도를 측정하여 선호도를 예측하는 방식이다[6]. 이 상품 기반의 추천 시스템은 고객 기반의 추천 시스템보다 빠른 추천이 이루어 진다는 장점이 있지만 고객 기반의 추천 시스템과 마찬가지로 고객의 상품에 대한 선호도를 차별화 하지 않고 유사도를 계산하기 때문에 역시 정확한 선호도 예측이 어렵다는 단점이 있다.

결국 두 가지 방법 모두 고객의 상품에 대한 선호도 값을 가공하지 않고 그대로 사용하여 상품을 추천하기 때문에 정확한 상품 추천이 이루어 지지 않는다. 본 논문에서는 고객 기반의 추천 시스템에서 고객에게 상품을 추천할 때 정확한 추천이 이루어 지도록 고객의 각 상품에 대한 선호도 값에 가중치를 부여하는 방안을 전처리 단계로 제안한다. 이는 곧 상품 추천의 정확성 향상으로 이어질 것으로 사료되며 이를 실험을 통하여 검증한다.

본 논문은 다음과 같이 구성된다. 2장에서는 본 논문에서 제안하는 상품 추천의 전처리 단계인 가중치 계산 방법에 대해 기술한다. 3장에서는 실험을 통해 이를 검증 분석하고 4장에서는 결론 및 향후 계획에 관해서 논한다.

2. 전처리 방법

2.1. 가중치 계산

기존의 협력적 여과 알고리즘은 고객의 특정 상품에 대한 선호도를 예측하기 위해 각 상품에 대한 선호도 값을 아무런 가공 없이 그대로 이용한다. 그러나 고객이 상품에 대해 직접 평가한 선호도를 그대로 이용하여 특정 상품에 대한 선호도를 예측하는 경우 다수의 상품에 대해 선호도를 갖는 고객과 소수의 상품에 대해서 선호도를 갖는 고객들에 대한 차별화가

이루어 지지 않기 때문에 정확한 선호도 예측이 어렵다는 단점이 있다. 이러한 단점을 보완하기 위해서 본 논문에서는 Pearson 알고리즘을 사용하여 각 고객의 각 상품에 대한 선호도를 보다 정확하게 예측하기 위해 주어진 상품 선호도에 가중치를 반영 하는 방법을 제안 한다. 가중치를 구하는 방법으로는 정보검색에서 사용되는 벡터 공간 모델을 이용하며[7][8], 이 모델을 사용 함으로써 고객과 상품간의 연관성을 효과 적으로 수치화 하여 행렬로 표현할 수 있게 된다.

고객의 상품에 대한 선호도를 상품의 구매횟수, 상품의 클릭횟수, 상품 정보를 검색한 시간을 반영한 값이라고 가정하자. 아래 행렬은 고객의 상품에 대한 선호도를 행렬 $C(m \times n)$로 표현한 것이다.

$$C = \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1(n-1)} & c_{1n} \\ c_{21} & c_{22} & \cdots & c_{2(n-1)} & c_{2n} \\ \vdots & \vdots & c_{ij} & \vdots & \vdots \\ c_{(m-1)1} & c_{(m-1)1} & \cdots & c_{(m-1)(n-1)} & c_{(m-1)n} \\ c_{m1} & c_{m2} & \cdots & c_{m(n-1)} & c_{mn} \end{bmatrix}$$

- m : 고객
- n : 상품
- c_{ij} : 고객 i 의 상품 j 에 대한 선호도

기존의 협력적 추천 알고리즘의 문제점을 설명하기 위해 고객의 상품에 대한 선호도와 가중치를 반영한 선호도를 예로 들어 설명 하겠다.

$$C = \begin{bmatrix} 0 & 0 & 0 & 4 & 5 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 4 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 3 & 2 & 4 & 5 & 3 & 2 & 3 & 3 & 0 \\ 0 & 0 & 0 & 0 & 0 & 4 & 5 & 0 & 0 & 0 \\ 0 & 0 & 5 & 4 & 5 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

행렬 $C(5\times10)$ 은 5명의 고객이 10개의 상품에 대하여 직접 평가한 선호도 값을 나타낸 것이다. 행렬 C 에서 10개의 상품 중에서 고객1(1행)은 2개의 상품, 고객2(2행)는 1개의 상품, 고객3(3행)은 8개의 상품, 고객4(4행)는 2개의 상품, 고객5(5행)는 3개의 상품에 대해서 각각 평가하였다. 고객1과 고객3은 상품4, 5에 대해서 동일한 선호도 값으로 평가하였으나 유사도나 선호도 예측 값을 계산 할 때 그 값을 여과 없이 그대로 사용하면 정확한 예측이 어렵게 된다. 직관적으로 고객1은 상품4, 5에 대해서만 선호도를 평가하였기 때문에 상품4, 5는 고객1과 연관성이 높다고 할 수 있지만 고객3은 8개의 상품에 대해서 골고루 선호도를 평가하였기 때문에 상품4, 5가 고객3과 특별히 연관성이 높다고 할 수 없다. 즉 상품4, 5는 고객1을 대표하는 상품이지만 고객3을 대표하는 상품이라고 말할 수 없다.

그러므로 전체 상품을 골고루 구매하는 고객과 특정 상품만을 집중적 반복적으로 구매하는 고객이 같은 상품에 대해서 동일한 구매성향을 보일지라도 서로 다른 가중치를 반영하여 상품에 대한 선호도를 예측하는 것이 타당하다고 볼 수 있다.

기존의 협력적 여과 방법은 위의 예에서 상품4, 5에 대해서 고객1, 3, 5의 선호도를 모두 같은 값으로 취급하여 계산하기 때문에 정확한 선호도 예측이 어려운 반면, 본 논문에서는 전처리 과정을 통해 고객의 상품에 대한 선호도에 가중치를 차별화 하여 반영 함으로써 고객의 상품에 대한 선호도를 정확히 예측할 수 있다. 위의 내용을 수식으로 표현하면 식(1)(2)(3)로 나타낼 수 있다.

$U = \{u_1, u_2, \ldots, u_i, \ldots, u_m\}$ 는 모든 고객의 집합이고 m 는 모든 고객 수 이다. $P = \{p_1, p_2, \ldots, p_j, \ldots, p_n\}$ 는 모든 상품의 집합이고 n 은 총 상품의 개수 이다. 고객 u_i 의 상품 p_j 에 대한 선호도를 행렬 $C(m\times n)$ 로 표현할 수 있다. 행렬 C 의 원소 $c_{i,j}$ 는 상품 p_j 에 대한 고객 u_i 의 선호도 값이고 $\max[c_{i1}, c_{i2}, \ldots, c_{ij}, \ldots c_{in}]$ 는 상품의 집합 P 에 대한 고객 u_i 의 선호도중 가장 큰 값을 나타낸다. 수식(1)에서 $ufi_{i,j}$ 값은 상품 p_j 에 대한 고객 u_i 의 선호도를 0에서 1사이의 값으로 정규화 시킨 것이다. $ufi_{i,j}$ 값이 크면 고객 u_i 가 상품 p_j 에 대한 선호도 값이 크다는 것을 의미한다.

$$ufi_{i,j} = \frac{c_{i,j}}{\max[c_{i1}, c_{i2}, \ldots, c_{ij}, \ldots c_{in}]} \quad (1)$$

전체 상품의 개수를 N 이라고 하고, 전체 상품 P 에 대해서 고객 u_i 의 선호도가 0이 아닌 상품의 개수를 n_i 라 하자. 만약 고객 u_i 가 전체 상품 P 에 대한 선호도가 모두 0이 아니라면 $iufi$ 는 0이 되고 상품의 선호도가 0인 것이 많을수록 $iufi$ 의 값은 증가한다. 예를 들어 고객 u_i 와 u_j 가 있고 고객 u_i 는 다수의 상품에 대해서 0이 아닌 선호도 값을 갖고 고객 u_j 는 일부의 상품에 대해서 0이 아닌 선호도 값을 갖는다고 가정하자. 즉 고객 u_i 는 다수의 상품을 골고루 구매한 고객이고 고객 u_j 는 소수의 특정한 상품만 구매한 고객이라고 볼 수 있다. 수식(2)에서 고객 u_i 의 $iufi$ 값은 고객 u_j 의 $iufj$ 값보다 적은 수치 값을 보인다. 따라서 수식(2)는 다수의 상품을 골고루 구매한 고객보다 소수의 특정 상품을 지속적으로 구매한 고객에게 높은 수치를 부여함으로써 고객 u_i 와 고객 u_j 의 상품에 대한 선호도를 차별화 해준다.

$$iufi = \left(\log \frac{N}{n_i}\right) \quad (2)$$

행렬 $C(m\times n)$ 의 가중치는 아래와 같이 계산한다.

$$w_{ij} = ufi_{i,j} \times iufi \quad (3)$$

행렬 C 에 수식(3)의 가중치 공식을 이용하여 행렬 $W(m\times n)$ 로 표현할 수 있다. 행렬 W 의 원소 w_{ij} 는 고객 u_i 가 상품 p_j 에 대해서 평가한 선호도에 가중치를 반영한 값이다.

$$W = \begin{bmatrix} 0 & 0 & 0 & 0.56 & 0.70 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.80 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.06 & 0.04 & 0.08 & 0.10 & 0.06 & 0.04 & 0.06 & 0.06 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.56 & 0.70 & 0 & 0 & 0 \\ 0 & 0 & 0.52 & 0.42 & 0.52 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

행렬 W 은 행렬 C 의 모든 원소에 수식(3)을 이용하여 가중치를 부여한 것이다. 상품4에 대한 선호도가 행렬 C 에서는 모두 똑 같은 수치로 나타났으나 행렬 W 에서는 서로 다른 수치로 표현된다는 것을 알 수 있다. 고객1과 고객3를 비교해보도록 하겠다. 상품4에 대한 고객들의 선호도가{0.56, 0.08, 0.08, 0, 0.42}로 나타난다. 고객1은 상품4와 상품5에 대해서만 선호도를 갖기 때문에 8개의 상품에 대해 0이 아닌 선호도를 갖는 고객3보다 $iufi$ 의 값이 더 높게 나타난다. 즉, 전체 상품에서 소수의 특정 상품만을 지속적으로 구매한 고객은 구매 상품에 대해 선호도가 높다고 할 수 있지만 다수의 상품을 골고루 구매한 고객은 구매 상품 중에서 어느 특정 상품에 선호도가 높다고 단정하기가 어렵다. 따라서 행렬 C 보다 행렬 W 가 고객의 상품에 대한 선호도를

정확히 반영한다. 본 논문에서는 예측의 오차를 줄이기 위해서 전처리 방법을 사용하여 고객의 상품에 대한 선호도에 가중치를 부여하여 상품에 대한 선호도를 차별화 하였다.

2.2 상품에 대한 선호도 예측 식

$W(5\times10)$ 행렬을 이용하여 식(4)과 식(5)를 이용하여 고객의 상품에 대한 예측 값을 구한다[2].

$$w_{i,j}=\frac{\sum_{k}(r_{i,k}-\bar{r}_i)(r_{j,k}-\bar{r}_j)}{\sum_{k}(r_{i,k}-\bar{r}_i)^2\times\sum_{k}(r_{j,k}-\bar{r}_j)^2} \quad (4)$$

$$p_{i,p}=\bar{r}_i+\frac{\sum_{j}w_{i,j}\times(r_{j,p}-\bar{r}_j)}{\sum_{j}w_{i,j}} \quad (5)$$

식(4)의 Pearson 상관관계를 통하여 고객 i 와 j 사이의 유사도를 결정하고 하고 이렇게 계산된 유사도를 이용하여 식(5)와 같이 고객 i 의 상품 p 에 대한 선호도를 예측하게 된다. 여기서 k 는 고객 i 와 고객 j 가 모두 평가한 상품을 의미하고 $r_{i,k}$ 는 고객 i 가 상품 k 에 대한 평가한 값을 나타내며 $\bar{r}_i$ 는 고객 i 의 전체 상품에 대해 평가한 평균값을 나타낸다

3. 실험 및 성능 평가

3.1 실험 환경

실험에 사용한 데이터는 GroupLens Research Project에서 제공한 MovieLens 데이터를 사용하였다[9]. 이 데이터는 943명의 고객이 1682편의 영화에 대해서 선호도 값을 1~5점 사이의 수치로 평가한 10만개의 데이터로 구성되어 있다. 이 데이터는 최소 20편 이상의 영화에 대해서 평가한 고객 데이터로 이루어 졌으며 영화 데이터는 19개의 장르로 구분되어 하나 이상의 장르에 속 할 수 있도록 하였다. 본 실험에서는 총 100,000개의 데이터 중에서 임의로 80,000개를 뽑아 학습 데이터로 사용하고 20,000개를 실험 데이터로 사용 하였다.

3.2 평가기준

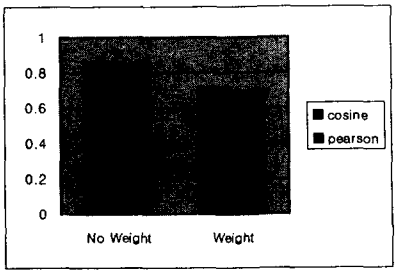
예측의 정확성을 평가하기 위해서 MAE(Mean Absolute Error)를 사용한다. 식(6)은 MAE를 수식으로 표현한 것이며 N은 총 예측횟수, εi 는 예측된 평가값과 실제 평가한 값 사이의 오차를 나타낸다.

$$|E|=\frac{\sum_{i=0}^{N}|\varepsilon i|}{N} \quad (6)$$

3.3 실험 결과 및 분석

본 논문에서는 고객의 상품에 대한 선호도와 선호도에 가중치를 부여한 데이터를 이용하여 기존의 협력적 여과 방법에서 사용하는 코사인(Cosine)을 이용한 유사도 측정 방식과 Pearson 상관관계를 이용하여 상품에 대한 선호도 예측 값을 측정하였다. 그림 1 은 전처리 단계를 수행하지 않은 선호도보다 전처리 단계를 수행한 선호도를 이용하여 고객의 상품에 대한 선호도를 예측하는 것이 평균적으로 15% 정확한 성능을 보인다는 것을 알 수 있다.

그림 1. 예측 오차 비교 결과



4. 결론

기존의 협력적 여과 방식은 고객의 구매 데이터를 그대로 사용하여 추천하기 때문에 고객의 특정 상품에 대한 선호도 예측에 있어 정확하지 못한 결과를 얻었다. 본 논문에서는 고객에게 상품을 추천함에 있어 협력적 여과 알고리즘을 기반으로 추천하되, 보다 정확한 예측을 가능하게 하기 위하여 선호도 값을 조정하는 전처리 방법을 제안하였다. 실험 결과, 고객마다 구매한 상품에 대해 차별적인 가중치를 반영함으로써 상품에 대한 선호도 예측 오차를 줄일 수 있음을 보였다. 그러나 본 논문에서 제시한 전처리 방법을 이용하여 전자상거래 시스템을 설계하고 구현하기 위해서는, 동일한 상품을 반복적으로 추천하는 단점을 보완하고, 고객간의 유사도 뿐만이 아니라 상품간의 유사도도 적극 고려하는 방식으로 발전시켜나가야 한다.

5. 참고문헌

[1] Konstan, J., Miller, B., Maltz, D., Herlocker, J., Gordon, L., and Riedl, J. (1997). GroupLens: Applying Collaborative Filtering to Usenet News.Communications of the ACM.

[2] Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., and Riedl, J, (1994). GroupLens: An Open Architecture for Collaborative Filtering of Netnews. In Proceedings of CSCW '94 Chapel Hill, NC

[3] Shardanand, U., and Maes, P. (1995). Social Information Filtering: Algorithms for Automating 'Word of Mouth'. In Proceedings of CHI '95. Denver, CO.

[4] Sarwar, B., Karypis, G., Konstan, J., Riedl, J.,(2001) Item-Based Collaborative Filtering Recommendation Algorithms

[5] D. Goldberg, D. Nichols, B. M. Oki, and D. Terry., Wsing Collaborative Filtering to Weaves and Information TAPESTRY, CACM, 1992.

[6] S. Badrul, K Georoge, K Joseph, and R. John. Item-Based Collaborative Filtering Recommendation Algorithms.

[7] R. Baeza-Yayes, and B. Ribeiro_Neto . Modern Information Retrieval. Addison Wesley 1998.

[8] G. Salton and M. J. McGill. Introduction to Modern Information Retrieval. McGraw-Hill, 1983

[9] http://www.grouplens.org