2000 년 한국정보처리학회 추계 학술발표논문집 제 7 권 제 2 호

협동적 필터링을 이용한 전자상거래에서의 추천시스템

김영설*, 장수현, 윤병주
명지대학교 컴퓨터공학과
e-mail : {kys0716, shjang, yoonbj}@mju.ac.kr

Recommender Systems in E-Commerce using Collaborative Filtering

Youngseol Kim* , Suhyun Jang, Byungjoo Yoon
Dept. of Computer Engineering, Myongji University

요 약

인터넷이 생활의 일부분이 되어감에 따라 인터넷상에서 이루어지는 전자상거래는 빠르게 발전하고 있다. 지금까지의 전자상거래는 고객이 요구하는 제품을 판매하는 단순한 형태였다. 하지만 앞으로의 전자상거래에서는 고객이 선호할 만한 제품을 예상하여 고객에게 해당 제품을 추천해 줌으로서 양질의 서비스를 제공하고 더 많은 이익을 창출 할 수 있는 전자상거래 시스템이 요구되고 있다. 본 논문에서는 전자상거래시스템에서 이용될 수 있는 추천시스템을 개발하기 위하여 추천시스템의 핵심이 되는 사용자간 유사도에 기초한 GroupLens 의 협동적 필터링 알고리즘을 실제 Data Set 을 통해서 실험하였다. 또한 Data Set 을 분석하여 아이템을 대표할 수 있는 장르를 결정하여 전체 학습데이터로부터 대표장르에 속하는 데이터들만을 분리하여 학습데이터로 사용하는 추천시스템을 제안하였고, 실험을 통하여 제안한 추천시스템의 타당성을 보였다.

1. 서론

인터넷의 급속한 발달은 사람들의 생활 방식을 바꾸기에 이르렀다. 지금까지의 상거래가 실제 상점을 통해서 이루어졌다면 앞으로의 상거래는 인터넷을 통한 전자상거래(E-Commerce)형태의 상거래가 빠르게 확산될 것으로 예상되고 있다. 또한 현재의 전자상거래 시스템이 고객의 웹 사이트 방문을 기다려 고객이 요구하는 제품을 제공하는 단순한 형태였다면, 앞으로의 전자상거래 시스템은 고객이 선호할 만한 제품을 미리 예측하여 고객에게 제품 정보를 제공함으로서 고 부가가치를 창출 할 수 있는 수준을 요구하고 있다. 이와 같이 전자상거래 시스템에서 구매 촉진을 위하여 사용되는 핵심 기술은 전자상거래 시스템을 이용하는 사용자들로부터 얻어진 구매 정보를 기초로 고객이 좋아할 만한 제품을 예측하여 추천해주는 추천시스템(Recommender Systems)이다.

이에 따라 추천시스템에 대한 연구가 활발히 진행되고 있으며 대표적 전자상거래 시스템인 Amazon[9]에서는 'Who Bought' 서비스를 통해서 사용자가 특정 책을 조회하면 그 책을 구입한 사람들이 많이 구매한 책 리스트를 제시해주는데 이는 협동적 필터링이 사용된 대표적 추천시스템의 예가 된다[6].

이 논문은 전자상거래에서 추천시스템을 구현하기 위하여 협동적 필터링이 사용된 GroupLens[2]를 실제 Data Set 을 통해서 실험하고, 또한 Data Set 을 분석하여 아이템을 대표할 수 있는 대표장르를 결정하여 전체 학습데이터로부터 대표장르 데이터만을 분리하여 분리된 데이터를 학습데이터로 사용하는 추천시스템을 제안하였다. 제안한 방법은 한 아이템이 여러 장르에 속하는 경우 장르 정보를 이용하여 학습데이터를 줄임으로서 아이템에 대한 평가의 희소성을 감소시켜 추천시스템의 추천성능을 개선할 수 있음을 실험을 통하여 확인할 수 있었다.

2. 관련연구

인터넷의 정보 종류가 다양해지고 정보 양도 증가

함에 따라 사용자가 필요로 하는 정보를 찾기 위한 시간과 노력은 앞으로 계속 증가하여 심각한 문제가 될 것이다. 이러한 정보과잉(Information Overload)을 해결하는 방안으로 협동적 필터링(Collaborative Filtering)에 대한 연구가 시작되었다. 협동적 필터링은 사용자의 특정 아이템에 대한 선호도를 유사 사용자의 의견을 참고하여 예측하는 방법으로 추천시스템의 핵심 기술로 사용되고 있다[3].

협동적 필터링의 개념은 TAPESTRY[1]로부터 유래되었으며 대표적인 자동화된 협동적 필터링 시스템은 Netnews(Usenet News)에서 개별화된 추천을 제공하는 GroupLens[2]이다. GroupLens 에서는 사용자가 읽은 문서에 대한 평가를 하지 않은 경우 다른 사용자와의 유사성에 기초하여 평가값을 예측한다.

예로서 6 개의 문서에 대한 사용자의 평가가 <표 1>과 같다고 가정하자.

<표1> 사용자의 문서에 대한 평가표

문서번호	평가자			
	Ken	Lee	Meg	Nan
1	1	4	2	2
2	5	2	4	4
3			3	
4	2	5		5
5	4	1		1
6		2	5	?

<표 1>에서 빈칸으로 되어 있는 것은 아직 사용자가 그 문서를 읽지 않은 것을 의미한다. 그리고 물음표가 되어 있는 것은 문서를 읽었지만 문서에 대한 평가를 하지 않은 것을 의미한다. GroupLens 는 Ken 의 6 번째 문서에 대한 평가를 예측하기 위하여 상관계수(Correlation Coefficients)를 사용하는데 이 값은 –1 과 1 사이의 값을 가진다. Ken 과 Lee 사이의 상관계수는 다음과 같이 구해진다.

$$r_{KL} = \frac{Cov(K,L)}{\delta_K \delta_L} = \frac{\sum_i (K_i - \overline{K})(L_i - \overline{L})}{\sqrt{\sum_i (K_i - \overline{K})^2}\sqrt{\sum_i (L_i - \overline{L})^2}} = \frac{-2-2-2-2}{\sqrt{10}\sqrt{10}} = -0.8 \quad (2.1)$$

식(2.1)은 Ken 과 Lee 사이의 상관계수를 구하는 식으로 $\overline{K}$ 과 $\overline{L}$ 는 각각 Ken 과 Lee 의 문서에 대한 평가 평균값이다. 상관계수가 1 이면 Perfect Positive Relationship 이라고 하며 –1 이면 Perfect Negative Relationship 이라고 한다. 만약 0 이면 상관관계가 없는 경우이다. 위와 같은 방법으로 다른 사람들과의 상관관계를 구해보면 Meg 와의 상관계수는 1 이며, Nan 과의 상관계수는 0 이 된다. Ken 과 다른 사람과의 상관관계가 모두 구해졌으면 Ken 의 6 번째 문서에 대한 평가를 예측하게 된다. 평가의 예측에는 상관계수를 포함한 평균값을 이용한다.

$$K_{6\,pred} = \overline{K} + \frac{\sum_{J \in raters} (J_6 - \overline{J}) r_{KJ}}{\sum_J |r_{KJ}|} = 3 + \frac{2r_{KM} - r_{KL}}{|r_{KM}| + |r_{KL}|} = 3 + \frac{2 - (-0.8)}{|1| + |-0.8|} = 4.56 \quad (2.2)$$

Nan 과의 상관계수는 0 이므로 식(2.2)에서 Nan 과의 관계값은 제외되게 된다. 따라서 Ken 의 6 번째 문서에 대한 평가값은 4.56 으로 예측된다. 위와 같은 방법을 통해서 평가를 하지 않은 문서에 대한 예측을 수행한다.

이와 같이 한 사용자와 다른 사용자의 상관관계에 기초한 평가 예측을 일반화 시키면 식(2.3)과 같다.

$$p_{a,i} = \overline{e}_a + \frac{\sum_{u=1}^{n} (e_{u,i} - \overline{e}_u) * r_{a,u}}{\sum_{u=1}^{n} r_{a,u}} \quad (2.3)$$

식(2.3)의 $p_{a,i}$ 는 사용자 a 의 아이템 i 에 대한 예측값, $\overline{e}_a$ 는 사용자 a 의 평가 평균값, n 은 사용자의 수, $e_{u,i}$ 는 사용자 u 의 i 번째 아이템 평가값, $\overline{e}_u$ 는 사용자 u 의 평가 평균값, $r_{a,u}$ 는 사용자 a 와 사용자 u 사이의 상관계수를 나타내며 식(2.4)로 계산된다.

$$r_{a,u} = \frac{\sum_{i=1}^{m} (e_{a,i} - \overline{e}_a) * (e_{u,i} - \overline{e}_u)}{\sigma_a * \sigma_u} \quad (2.4)$$

GroupLens 의 예측 결과는 상당히 성공적이었다. 그러나, 1) 아이템에 대한 초기 평가가 없는 경우에는 추천이 불가능하며, 2) 아이템의 전체 수에 비해서 사용자들이 평가한 아이템 수가 매우 부족한 경우(희소성이 높음)는 추천의 정확성이 떨어지는 점 등이 문제점으로 지적되고 있다[3].

3. 추천시스템의 제안

GroupLens 는 전체 학습데이터를 사용하여 사용자간 상관계수를 계산하므로 추천까지 많은 시간이 소요 된다. 또한 아이템에 대한 평가의 희소성이 큰 경우는 추천의 정확성이 떨어지게 된다. 따라서 본 논문에서는 GroupLens 의 이런 단점을 극복할 수 있는 새로운 추천시스템을 제안한다. 제안 추천시스템에서는 한 아이템이 여러 장르에 속하는 경우 아이템의 대표 장르를 결정하여 전체 학습데이터를 줄임으로서 예측 시간과 데이터의 희소성을 감소시켜 추천성능을 향상시키고자 하였다.

다음은 아이템의 장르 중 아이템을 대표할 수 있는 대표장르를 이용하여 평가값을 예측하는 과정이다. 100 명의 사용자가 3 개의 영화(X-men, Patriot, Titan A.E.)에 대해 평가하였고 각 영화의 장르는 <표 2>와 같으며 장르별 영화는 <표 3>과 같다고 가정하자.

<표 2> 영화별 장르

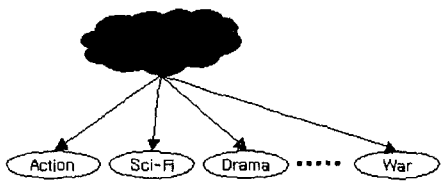
영화	장르
X-men	Action, Sci-Fi
Patriot	Action, Drama, War
Titan A.E.	Adventure, Animation, Sci-Fi

<표 3> 장르별 영화

장르	영화
Action	X-men, Patriot
Sci-Fi	X-men, Titan A.E.
Animation	Titan A.E.
War	Patriot

3-1. 장르별 데이터 분리

전체 데이터를 장르별 데이터로 분리하기 위하여 전체 데이터에서 해당 장르의 영화를 평가한 {사용자-영화} 데이터를 <그림 1>과 같이 장르별로 분리한다. <표 4>는 전체 데이터에서 Action 장르에 속하는 X-men, Patriot 영화를 평가한 데이터를 분리한 예이다.



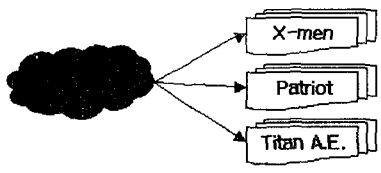
<그림 1> 장르별 데이터 분리

<표 4> Action 장르 데이터의 예

사용자	영화	평가값
Kim	X-men	4
Kim	Patriot	5
Park	X-men	1
Lee	X-men	3

3-2. 영화별 대표 '장르 k'결정

A. <그림 2>와 같이 전체 데이터로부터 해당 영화만을 평가한 데이터를 뽑는다. <표 5>는 전체 데이터에서 X-men 을 평가한 데이터를 뽑은 예이다.



<그림 2> 영화별 데이터 추출

<표 5> X-men 을 평가한 데이터의 예

사용자	영화	평가값
Kim	X-men	4
Lee	X-men	3
Park	X-men	1

B. X-men 이 속하는 Action, Sci-Fi 중 X-men 을 대표하는 장르를 구하기 위하여 [3-1]에서 구한 Action 장르 데이터를 사용자간 상관계수를 구하기 위한 학습데이터로 사용하여 X-men 을 평가한 <표 5>의 {사용자-영화}쌍의 평가값을 예측한다. 예측값과 실제값을 비교하여 X-men 의 Action 장르에 대한 평균에러를 구한다. 같은 방법으로 X-men 의 Sci-Fi 장르에 대한 평균에러도 구한다.

C. X-men 의 모든 장르에 대한 평균에러가 구해지면 평균에러가 최소가 되는 장르를 X-men 의 대표 '장르 k'로 선정한다. 같은 방법으로 <표 6>과 같이 영화 Patriot 와 Titan A.E.에 대한 대표 장르를 모두 구한다.

<표 6> 영화별 대표 '장르 k'의 예

영화	장르
X-men	Sci-Fi
Patriot	War
Titan A.E.	Animation

3-3. 평가값 예측

만약 사용자 a 가 Titan A.E.를 보지 않은 경우 사용자 a 의 Titan A.E.에 대한 선호도를 예측하기 위해서 Titan A.E.의 대표 장르 Animation 장르 데이터를 사용자 사이의 상관계수 계산에 사용하여 사용자 a 의 Titan A.E.의 평가값을 예측한다. 같은 방법으로 X-men 은 Sci-Fi 장르 데이터를, Patriot 는 War 장르 데이터를 사용하게 된다.

대표 장르를 이용하여 축소된 데이터를 선호도 예측에 사용하므로 학습데이터 양과 평가의 희소성을 줄일 수 있어서 예측시간과 예측성능이 개선될 것으로 기대된다.

4. 실험 및 토의

실험에 사용된 Data Set 은 GroupLens Research Project [8]에서 제공한 MovieLens Data Set 으로 그 구성은 다음과 같다. 1) 943 명의 사용자가 1682 영화에 대해서 1-5 점 사이의 점수로 평가한 100,000 개 데이터이다. 2) Data Set 은 최소한 20 개 이상의 영화에 대해서 평가한 사용자 데이터만으로 구성된다. 3) 영화는 19 개 장르로 구분되며 다양한 장르에 속할 수 있다. 실험을 위하여 총 100,000 개의 Data Set 중에서 80,000 개를 뽑아 Training Data 로 사용하고 20,000 개를 Test Data 로 사용하였다.

본 실험은 GroupLens 의 예측 알고리즘과 본 논문에서 제안한 대표 장르를 이용하여 학습데이터를 줄이는 알고리즘에 대해 예측의 평균에러(Mean Absolute Error), 최소에러, 최대에러 그리고 전체 영화 중 예측 가능한 영화의 수를 나타내는 예측율을 비교하여 제안한 방법의 타당성을 살피고자 한다.

4-1. GroupLens 에 의한 실험

GroupLens 의 예측성능을 알아보고자 Training Data 를 사용하여 사용자간 상관계수를 구한 후 Test Data 에서 무작위로 선택한 200 개의 {사용자-영화}쌍의 평가값을 예측한다. 그리고 사용자가 평가한 영화 수(희소성)가 예측에 미치는 영향을 알아보기 위하여 <표 7>과 같이 평가한 영화의 수에 따라 6 개 그룹으로 데이터를 재구성하였다. 평가한 영화의 수가 증가함에 따라 사용자의 수, 영화의 종류, 데이터의 희소성은

감소하였다.

<표 7> 평가한 영화 수에 따른 학습데이터 비교

x 개이상 영화평가	사용자 (명)	영화 종류	환경크기	평가 수	데이터 희소성
> 20	943	1650	1,555,950	80,000	94.86%
> 50	471	1642	773,382	66,810	91.37%
> 100	288	1627	468,576	53,576	88.57%
> 150	165	1604	264,660	38,154	85.59%
> 200	98	1586	155,428	27,175	82.52%
> 250	49	1472	72,128	16,280	77.43%

<표 8>은 GroupLens 알고리즘을 이용해 Test Data 에서 무작위로 선택한 200 개 {사용자-영화} 쌍에 대한 6 개 그룹별 평균에러, 최소에러, 최대에러, 예측율을 나타낸다. 실험결과 평가한 영화수가 증가할수록 평균에러가 감소하였다. 이는 많은 영화를 평가한 사용자 데이터를 사용할수록 정확한 예측이 가능하다는 것을 의미한다. 그러나 평가한 영화의 수가 증가함에 다라 예측율은 줄어들었다. 이는 평가한 영화의 수에 따라 학습데이터에 제약을 가하면 GroupLens 가 예측할 수 있는 영화 수가 줄어들게 되는 것을 의미한다.

<표 8> 평가한 영화 수에 따른 성능비교

x 개이상 영화평가	평균 에러	최소 에러	최대 에러	영화 예측율
> 20	0.7425	0.0082	2.6277	100%
> 50	0.7813	0.0009	2.6277	76%
> 100	0.7688	0.0082	2.8312	53%
> 150	0.7608	0.0027	2.8312	37%
> 200	0.6545	0.0351	1.9927	15%
> 250	0.5839	0.0227	2.1276	6%

4-2. 대표 장르를 이용한 실험

943 명의 사용자가 1650 개 영화에 대해서 평가한 80,000 개 Training Data 를 분석하여 <그림 1>과 같이 Training Data 로부터 19 개 장르에 대한 장르별 데이터를 분리한다. 다음으로 1650 개 영화에 대해서 [3-2]방법으로 영화를 대표할 수 있는 대표 '장르 k'를 구한다. [4-1]에서 사용한 Test Data 에서 무작위로 선택한 200 개 {사용자-영화} 쌍에 대한 평가값 예측을 [3-3] 방법에 따라 수행한다.

<표 9> 대표 '장르 k' 를 이용한 경우 환경

사용자 수	영화종류	환경크기	평가 수	희소성
894	341	304,854	18,300	94.00%

<표 9>은 Test Data 에서 무작위로 선택한 200 개 {사용자-영화}쌍을 제안된 방법에 적용한 경우의 사용자간 상관계수 계산에 사용된 데이터의 환경이다. 평균 894 명이 341 영화에 대해서 평가한 18,300 개의 데이터로 희소성은 94%였다.

<표 10> 대표 '장르 k' 를 사용한 경우 성능

평균에러	최소에러	최대에러	예측율
0.7748	0.0074	2.9762	99%

<표 10>은 대표 '장르 k'를 학습데이터로 사용하여 평가값을 예측한 결과이다. GroupLens 와 비교하여 평균에러와 예측율은 비슷하였다. 하지만 GroupLens 가 943 명이 1650 개 영화에 대해 평가한 80,000 개 데이터를 사용하는데 반하여 제안한 방법에서는 894 명이 341 개 영화에 대해 평가한 18,300 개의 데이터로 데이터를 축소하여 예측시간을 약 4 배 정도 단축시킬 수 있었다. 비슷한 예측시간이 보이는 150 개 이상의 영화를 평가한 38154 개 데이터가 37%의 예측율을 보이는데 반하여 제안된 방법은 99%의 예측율을 보였다.

5. 결론

실험을 통해 제안된 추천시스템에서는 GroupLens 의 평가값 예측을 수행할 때 아이템을 대표할 수 있는 대표 장르를 이용하여 학습데이터를 축소하여 평가값 예측까지의 시간을 단축할 수 있었으며, 같은 예측시간을 요구하는 학습데이터를 사용하는 GroupLens 에 비해서 높은 예측율을 보임을 확인 할 수 있었다.

참고문헌

[1] D. Goldberg, D. Nichols, B. M. Oki, and D. Terry., Using Collaborative Filtering to Weaves and Information TAPESTRY, *CACM, Vol.35, No.12, pp.61-70*, 1992.

[2] Resnick P., Iacovou N., Sushak M., Bergstrom P., and Riedl J., GroupLens : An open architecture for collaborative filtering of Netnews. *Proceedings of the 1994 Computer Supported Collaborative Work Conference*. 1994.

[3] Sarwar, B., Konstan, J., Borchers, A., Herlocker, J., Miller, B., and Riedl, J., Using Filtering Agents to Improve Prediction Quality in the GroupLens Research Collaborative Filtering System. *Proceedings of the 1998 Conference on Computer Supported Cooperative Work*. Nov. 1998.

[4] Herlocker, J., Konstan, J., Borchers, A., Riedl, J., An Algorithmic Framework for Performing Collaborative Filtering. *Proceedings of the 1999 Conference on Research and Development in Information Retrieval*. Aug. 1999.

[5] Good, N., Schafer, J.B., Konstan, J., Borchers, A., Sarwar, B., Herlocker, J., and Riedl, J., Combining Collaborative Filtering with Personal Agents for Better Recommendations. *AAAI-99*. pp 439-446

[6] Schafer, J.B., Konstan, J., and Riedl, J., Recommender Systems in E-Commerce. *Proceedings of the ACM Conference on Electronic Commerce*, November 3-5, 1999.

[7] Sarwar, B. M., Karypis, G., Konstan, J. A., and Riedl, J. T., Application of Dimensionality Reduction in Recommender System--A Case Study. *WebKDD 2000 Workshop*. August 20, 2000.

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

[9] http://www.amazon.com/