2000년 한국정보처리학회 추계 학술발표논문집 제 7 권 제 2 호

동시출현 빈도에 기반한 협동추천시스템의 성능 향상

박지연, 박윤심, 유견아
덕성여자대학교 전자계산학과
e-mail : jypark@center.duksung.ac.kr

Improving performance of collaborative recommendation system based on co occurrence

Ji-Yeon Park, Yun-Shim Park, Kyeon-Ah Yu
Dept. of Computer Science, Duksung Womens University

요 약

인터넷이 발전하면서 인터넷을 이용한 여러 서비스들이 급속히 발달하고 있다. 이런 발전에 맞추어 사용자들은 적합한 상품을 선택하는 것이 점점 어려워지고 그에 따라 운영자들은 사용자들의 요구에 맞춰 원하는 상품을 쉽게 찾게 하여 매출을 올리는 노력을 하고 있다 이런 노력의 일환으로 기존의 사용자 데이터를 바탕으로 사용자의 선호도를 예측하고 사용자의 선호도에 따라 개인에게 적합한 상품을 추천하는 협동적 방식의 추천 시스템이 개발되어 많이 사용되는 추세이다. 본 논문에서는 현재 사용되고 있는 협동추천 시스템의 문제점을 보완할 수 있는 방법을 제시하며 실험을 통해 기존에 비해 성능이 향상되고 있음을 보인다.

1. 서론

인터넷이 급속히 발전하면서 인터넷을 이용한 여러 서비스들이 발달하고 있다. 이들 중 특히 전자상거래가 단연 돋보이는 발전을 하고 있다. 전자상거래는 규모 뿐만 아니라 상품의 다양성과 이를 사용하는 사용자들의 계층도 더욱 다양해지고 있는 실정이다. 상품의 다양성이 증가함에 따라 사용자들은 자신에게 보다 적합한 상품을 선택하는 것에 많은 어려움이 있으며 그에 따라 운영자들은 사용자들의 요구에 맞춰 원하는 상품을 쉽게 찾게 하여 매출을 올리는 노력을 하고 있다. 이러한 방식의 마케팅은 급속히 변동하는 인터넷 시장을 반영하기에 부족하여 이것을 자동화할 필요가 있다.

이런 이유로 기존의 사용자 데이터를 바탕으로 사용자의 선호도에 따라 개인에게 적합한 상품을 추천하는 협동적 방식의 추천 시스템이 개발되었다. 협동적 방식의 추천 시스템은 사용자들의 연관성을 기반으로 상품을 추천하는 방식이다. 즉 사용자 사이의 유사도를 측정하고 유사 사용자가 선호하는 상품을 현 사용자가 구매하지 않은 상품에 대한 선호도를 예측하는 것이다. 하지만 모든 사용자간의 유사도 계산을 기본으로 하는 이 방식은 전자상거래 사이트의 데이터 희박성(sparseness)로 인해 실제로 적용하는 데에 중요한 문제점이 있다. 즉, 실제 인터넷 상거래 시스템에서 사용자는 그 사이트에서 취급하는 항목의 적은 부분집합과 관련되어 있으며 사용자들 사이에 동시출현하는 빈도는 더 적기 때문에 모든 사용자간의 유사도를 계산하는 것은 시간적으로나 선호도 예측의 정확도 면에서 비효율적이라는 것이다.

본 논문에서는 동시출현 빈도를 이용해 공통으로 관심있는 항목이 많은 사용자들을 모아 그룹을 형성하고 그 그룹 안에서 우선적으로 유사도를 계산하여 계산 시간을 줄이는 동시에 선호도 예측의 정확도를 높이는 방법을 제안한다.

1.1 관련 연구

사용자간의 유사도를 기본으로 하는 협동추천 방식은 GroupLens 프로젝트에서 제안되었다[1]. GroupLens 는 인터넷 뉴스를 추천하는 시스템으로 소개된 이래 아마존, 리바이스, CDNOW 등과 같은 사이트들에서 여러

형태로 널리 사용되고 있다 [2]. GroupLens 는 정확도의 개선을 위해 유사도 가중치, 분산 가중치 등의 변형된 유사도 계산 공식을 제안했으며 [3], [4]에서는 없는 데이터에 대해 기본값을 산출하는 방식으로 예측 가능한 값의 범위를 확장시키고자 하였다. 계산량을 줄이고자 하는 시도로서는 [5]에서 SVD(singular value decomposition)을 통해 사용자-항목의 행렬 차원을 줄이는 방식이 제안되었다. [6]에서는 항목을 분류하여 상위 레벨의 카테고리를 형성하게 하여 사용자 사이에 공통항목이 없더라도 상위 레벨의 유사도를 계산하여 coverage 를 높이는 방법이 제시되었다.

2. 협동추천 시스템의 구현

2.1 피어슨 관계 계수를 이용한 기존의 방식

통계학 분야의 피어슨 관계 계수(Pearson correlation coefficient)는 사용자들간의 유사도 측정하는데에 많이 사용되는 방식이다.

[식 1]은 사용자 a 와 u 간의 피어슨 관계 계수를 이용해 유사도를 계산하는 식이다. [식 1]에서 $r_{a,i}$ 는 사용자 a 가 상품 i 를 택한 경우이며 σ_a 와 σ_u 는 각각 사용자 a 와 u 간의 편차이다.

$$w_{a,u} = \frac{\sum_{i=1}^{m}(r_{a,i} - \overline{r_a}) * (r_{ui} - \overline{r_u})}{\sigma_a * \sigma_u}$$

[식 1] 피어슨 관계 계수

피어슨 관계 계수는 각각의 공통된 상품에 대하여 현 사용자의 평가와 다른 사용자의 평가를 비교하는 방식으로 구성되어 있다. 사용자의 평가는 사용자가 평균적으로 평가한 점수와 상품의 평가 점수의 차를 구함으로써 상품에 대해 상대적 선호도를 가지고 있는지 판단한다.

$$\overline{r_a} + \frac{\sum_{u=1}^{n}(r_{u,i} - \overline{r_u}) * w_{a,u}}{\sum_{u=1}^{n} w_{a,u}}$$

[식 2] 선호도 예측을 위한 식

[식 2]는 위에서 구한 사용자들의 유사도를 이용하여 사용자 a 의 상품 i 에 대한 선호도를 예측하는 식으로 주로 아마존과 같은 대형 전자상거래 시스템에 사용되어지고 있는 방식이다. 기존의 방식보다 성능의 향상을 위해 기존 방식의 문제점과 그에 대한 대안을 제시한다.

2.2. 제안방식

모든 사용자 사이의 유사도를 계산하여 선호도 예측에 사용하는 기존의 방식은 몇 가지 문제점이 있으며 이를 지적함으로써 새로운 방식을 제안하려 한다.

첫째, 사용자가 몇 개의 공통항목에 대해 매긴 평가점수인지에 대한 정보를 내포하고 있지 않다. 즉, 100 개의 공통 항목을 가진 사용자를 사이에 계산된 유사도나 1 개 밖에 없는 사용자를 사이에 계산된 유사도가 같은 비중으로 취급되고 있지만 실제 공통 항목의 수가 많을수록 그 비중을 높이 하는 것이 합리적이다.

둘째, 유사도가 낮은 사용자들이 많을수록 실제 평가값이 극단으로 갈 때 선호도 예측값의 오차가 커진다. 모든 사용자들에 대한 유사도를 이용할 때 그 속에 상관관계가 없는 사용자들이 많아짐은 자명하다.

셋째, 상호 관계가 없는 것이 자명해 보이는 사용자간의 유사도 계산으로 불필요한 계산 시간이 소용된다.

기존방식을 적용하기에 앞서 간단한 전처리 과정을 통해 이와 같은 문제점들을 보완하는 방법을 제안한다. 기본개념은 항목들의 동시출현이 빈발하는 사용자들을 그룹화하여 같은 그룹내의 사용자 사이에서 우선적으로 추천하는 것이다. 이렇게 하면 공통항목의 수가 다른 사용자들을 차별화 할 수 있고 전체 사용자에 대한 유사도를 계산하지 않음으로 계산량도 크게 줄일 수 있다. 또한 다수의 같은 항목에 관심이 있는 사용자들이 서로 비슷한 취향을 가졌을 것이라는 가정 하에 같은 그룹내의 사용자들끼리의 유사도는 전체 사용자에 대한 것에 비해 높을 것이라는 것을 알 수 있다. 본 제안 방식을 정리하면 다음과 같다.

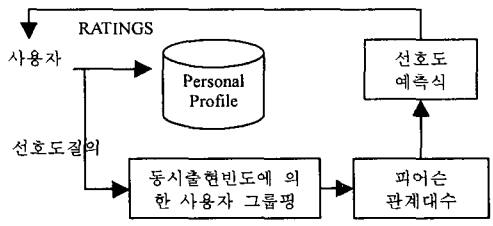
Step1. 동시출현 빈도를 이용해 선호도를 예측하고자 하는 사용자와 공통 항목이 가장 많은 그룹을 형성한다.

Step2. 그 그룹에 있는 사용자에 대해서 유사도를 계산하다.

Step3. 선호도 예측을 요구한 항목에 대해 그 그룹내의 사용자들의 데이터를 이용하여 [식 2]식으로 선호도를 계산한다.

Step4. Step3 이 실패한 경우 공통항목이 그 다음으로 많은 그룹을 만들어 Step2-3 을 반복한다.

공통항목이 많은 그룹을 만들기 위해서는 연관탐사에서 빈발 항목 집합을 찾아내는데 많이 이용되는 알고리즘인 Apriori 의 방식을 이용한다.[7]



[그림 1] 제안된 협동 추천 방식

3. 실험 및 분석

본 논문에서 제안한 추천 알고리즘은 자바로 구현되어 운영체제에 관계없이 수행 가능하며 실제 실험환경은 PentiumIII 550Hz, 128MB Ram 환경에서 수행되었다. 실험 데이터는 DEC 의 EachMovie data[8]를 이용

하였으며 자세한 실험 방법을 아래에서 설명한다.

3.1 실험방법

본 실험에서는 전자상거래 시스템에 이용되어지는 협동적 추천 시스템의 개발과 기존의 성능향상을 위한 효과를 보이고자 한다. 이를 위하여 실제 데이터에서 임의로 한 개씩의 데이터를 제거한 후 실험하였다. 실험 방법을 2 장에서 제안된 방식을 3 가지로 단계적으로 구분하였다.

제안 1. 기존의 방식에서 유사도가 낮은 것들을 필터링하여 일정값 이상의 유사도를 가진 사용자들의 데이터를 이용하였다. 좋은 상관관계만을 추출함으로써 기존방식의 오차율을 줄일 수 있다.

제안 2. 동시출현 빈도를 이용하여 최대 공통항목수의 75%까지의 공통항목을 공유한 그룹만 피어슨 관계 대수를 이용하여 계산하였다.

제안 3. 최대 50% 공통항목을 공유한 그룹 내에서도 유사도가 일정값 이상인 사용자들에 대해서만 피어슨 관계 대수로 계산하였다.

제안 2,3 의 경우 동시출현 빈도수가 높은 data set 의 순위를 정렬한 후 임의로 결정하였다. 제안 2 와 제안 3 의 차이를 보이는 이유는 동시출현 빈도가 낮은 경우와 높은 경우를 비교하고자 함이다.

3.2 Data Set

본 실험에 사용된 Data set 은 DEC 의 EachMovie data set 을 이용하고 있다. 이는 웹 상에 공개적으로 운영된 협동적 방식의 영화 추천 site 에서 수집되어진 것이다. 총 72916 명의 사용자가 0 부터 5 개 사이의 별의 개수로 평가를 하였다. 그러나 본 실험에서는 전체 데이터를 사용하지 않고 임의의 10, 100, 200 명의 사용자들을 증가, 선정해서 실험하였다. 영화의 수는 1628 개 모두 사용하였으며 10 회 반복한 결과의 평균을 취하였다.

3.3 평가 방법

시스템의 성능을 평가하는 여러 가지 방법 중에서 예측 값과의 실제 사용자 평가 값 사이의 차이를 표시하는 MAE(Mean Absolute Error) 방식[식 3]과 추천을 원하는 상품의 총수에 대한 선호도 값을 계산 가능한 상품의 수의 비를 나타내는 coverage 를 사용한다.

$$E = \frac{\sum |P - v|}{n}$$

P : 사용자 상품 선호도 예측값
v : 사용자 실제 평가 값

[식 3] Mean Absolute Error

Coverage 는 상품 추천 시스템에 있어 주요한 평가 요소 중 하나이다. 추천 시스템의 정확도는 일정 수준 이상의 Coverage 를 가질 경우에만 효용이 있는 것이다. 본 실험에는 백분율로 Coverage 를 표시한다.

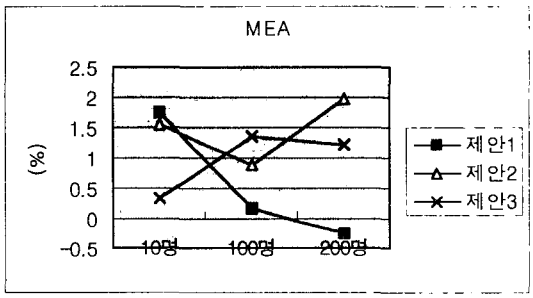
3.4 결과 분석

[표 1]는 규약에 대하여 제시한 방식을 10 회 반복하여 실험한 예측 값들의 MAE 를 평균한 값들이다.

단위(%)	10 명	100 명	200 명
기존방식	18.76	17.03	17.2
제안 1	18.43	17.0	17.244
제안 2	18.47	16.88	16.86
제안 3	18.7	16.8	16.99

[표 1] MAE

[그림 2]는 기존 방식 대비 백분율을 계산, 성능향상의 그래프를 보이고 있다.



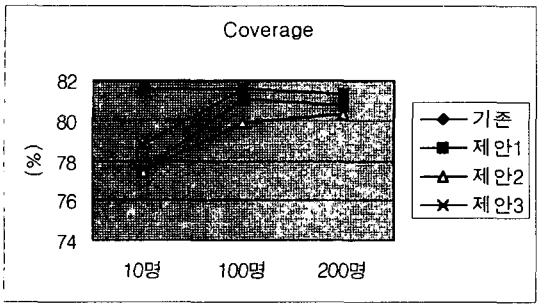
[그림 2] 성능향상을 보여주는 그래프

MAE 는 적을수록 성능이 우수하다. [표 1]과 [그림 2]를 보면 기존 방식과 비교해 볼 때 정확도가 향상됨을 볼 수 있다. 제안 1 은 상관관계가 비교적 좋은 것만을 추출, 연산하였기 때문에 기존보다 성능이 향상되었다. 제안 2 은 일반적으로 공통의 영화를 보는 것은 같은 취향의 사람이기에 유사도가 높은 것으로 보인다. 이에 기반을 두어 그룹간의 추천도를 연산한 결과, 좋은 결과가 나타남을 볼 수 있다. 제안 3 은 제안 2 를 보강하여 그룹 안에서도 가중치가 낮은 사용자를 제함에도 불구하고 제안 2 에 비해 정확도가 떨어지는 것은 추출된 데이터 집합의 크기가 작아 위험도가 커지기 때문이다.

제안 1 은 사용자가 증가할수록 기존에 비해 정확도가 떨어지는 것에 반면 제안 2 와 3 과 같이 그룹핑하는 방법에서는 정확도의 증가율이 사용자수에 따라 다른데 사용자수가 많아질수록 정확도가 증가하였다. 사용자 수가 적을 때에는 그룹핑의 의미가 크지 않으므로 이 결과는 예측과 부합한다고 할 수 있겠다.

단위(%)	10 명	100 명	200 명
기존방식	81.56	81.56	81.35
제안 1	77.48	81.11	80.71
제안 2	77.48	79.9	80.3
제안 3	78.83	81.56	81.34

[표 2] Coverage



[그림 3] Coverage 그래프

[표 2]와 [그림 3]은 실험결과의 coverage 를 나타낸다. 기존에 비해 coverage 가 전반적으로 떨어지는 것을 볼 수 있다. 이는 기존에 비해 자료를 많이 추출한 결과 상관관계가 적거나 오차율에 나쁜 결과를 미치는 자료를 제거했기에 coverage 가 약간 감소하는 것이다.

4. 결론 및 향후 연구과제

본 논문에서는 전자 상거래 시스템의 상품 추천 중 협동적 방식을 이용한 상품 추천 시스템을 구현하였다. 그리고 추천 시스템의 성능을 개선하기 위해 좋은 상관관계만을 고려한 피어슨 관계 계수 방식과 동시 출현 빈도를 응용으로 한 두 가지 방식은 모두 기존에 비해 coverage 는 떨어졌으나 오차율이 향상됨을 볼 수 있다. 또한 동시출현 빈도에 따른 그룹핑을 이용할 경우 연산할 수 있는 데이터 양이 줄어들기 때문에 연산 시간이 줄어드는 것을 알 수 있다. 참고로 본 논문에서는 EachMovie 를 사용하였지만 앞으로는 전자상거래가 더욱 활발해져 보다 많은 자료를 확보할 수 있기에 연관적인 패턴이 더 뚜렷하게 나타나고 우리가 제시한 방식 또한 coverage 가 향상될 수 있는 가능성을 보인다. 일반적으로 개성이 강한 경우를 제외한다면 일반적인 사용자들의 정확도가 개선되기 때문에 본 논문이 제시한 방식이 효과적임을 알 수 있다. Coverage 가 떨어지는 것은 Step4 에서 제안한 방식을 공통항목이 하나인 경우까지 확장하면 보완할 수 있으나 이렇게 하면 기존의 방식과 차별되지 않는다. 그러므로 앞으로는 이에 대한 연구와 함께 상관관계가 적거나 연관성이 적은 초기 전자상거래 자료를 보다 효율적으로 사용할 수 있는 연구가 병행되어야 할 것이다.

참고문헌

[1] Konstan J., Miller B., Maltz D., Herlocker J., Gordon L., and Riedl J., GroupLens: Applying collaborative filtering to Usenet news, *Communications of the ACM*, 40 (3), pp77-87, 1997.

[2] Schafer J., Konstan J., and Riedl J., Recommender Systems in E-Commerce, *Proceedings of the 22nd Conference on Research and Development in Information Retrieval*, 1999.

[3] Herlocker J., Konstan J., Borchers A., and Riedl J., An algorithmic framework for performing collaborative filtering, *Proceedings of the 22nd Conference on Research and Development in Information Retrieval*, 1999.

[4] Breese J., Heckerman D. and Kadie C., Empirical analysis of predictive algorithms for collaborative filtering", *Proceedings of the 14th Conference of Uncertainty in Artificial Intelligence*, 1998.

[5] Billsus, D. and Pazzani, M., Learning collaborative information filters. *Proceedings of the International Conference on Machine*, 1998.

[6] Kyeonah Yu, Sukmin Choi, and Juntae Kim, Improving the Performance of Collaborative Recommendation by Using Multi-Level Similarity Computation, IASTED *Proceedings on International Conference on Artificial Intelligence and Soft Computing*, pp241-245, 2000.

[7] 박종수, 유원경, 홍기형, 연관 규칙 탐사와 그 응용, *한국정보과학회지 제 16 권 제 9 호*,pp37-44,1998.

[8] EachMovie data set, http://www.research.digital.com/SRC/eachmovie/