협력적 여과(Collaborative Filtering)에서 결측치(Missing Value) 예측에 관한 연구

황 철 현* 박 영 길** 박 용 준***

The Research for Prediction of Missing Value in Collaborative Filtering

요 약

성공적인 사이트를 위한 필수적인 요소로 각광 받고 있는 collaborative filtering 기술은 정보의 과부하를 줄일 수 있고 고객에 대한 충성 도를 높여주는 효과로 인해 많은 사이트에 적용되어 운용되고 있다. 이 논문에서는 collaborative filtering 적용 초기에 발생하는 정보의 부족으로 인한 정확도 저하를 막기 위해 상품간 연관성을 이용한 결측치 예측 방안을 제안한다.

Keyword : Collaborative filtering, Recommender systems

1. 서 론

인터넷이 점차 생활의 일부분으로 자리 잡으면서 정보의 과부하를 줄일 수 있는 collaborative filtering 기술이 빠르게 대중화 되어 가고 있는 기술로서 amazon.com 이나 cdnow.com 같은 site 에 적용되면서 성공적인 site 구축을 위한 중요한 요소로 부각되기 시작했다.[2]

collaborative filtering 은 동일한 정보를 필요로 하거나 동일한 성향을 가진 사람들을 연결 시켜 주고 같은 domain 내에 있는 아이템에 대해 사용자의 평가(Rating 같은)를 수집하여 동작한다. 또한, collaborative filter system 에 속한 사항자들은 자신들의 성향과 각 아이템에 대한 평가 결과를 공유함으로써 사용자들이 아이템에 대해 좀더 나은 결정을 하도록 도와 준다.

Collaborative filtering 을 이용한 대표적인 예는 추천 시스템이다. 추천 시스템이란 사용자로부터 명시적이거나 묵시적인 정보를 획득하여 학습하여 사용자가 선호할 만한 아이템을 추천하여 주는 시스템을 말한다.[5,6]

상업적인 사이트에 적용되는 대부분의 추천 시스템은 사용자로부터 명시적인 평가를 입력 받아 그 정보를 학습하지만 명시적인 평가 방법은 사용자들이 평가에 대한 거부감으로 인해 평가 행렬에 대한 희소성이 발생시키며 이는 사용자 성향 정보의 부족 문제를 야기 한다.[4,7]

이 같은 성향 정보부족 문제 중 사용자들 간의 상호 상관 관계를 구하기 위해 필요한 평가

* ㈜ 온빛 시스템 정보기술연구소 책임연구원
** ㈜ 온빛 시스템 정보기술연구소 연구원
*** ㈜온빛 시스템 정보기술연구소장

정보가 사용자에 의해 입력 되지 않은 경우 이를 결측치(Missing Data Value)라 하며, 이는 부정확한 사용자의 성향을 반영하거나 추천의 정확도를 저하시키는 주요한 원인이 된다.

본 논문에서는 Collaborative Filtering 의 방법 중 Jester 2.0 에 대한 결측치 문제점을 제시하고 이를 해결하기 위해 상품에 대한 clustering 방법을 전처리 과정으로 하는 결측치 예측 방안을 제안한다.

2. 관련 연구

2.1 Jester 2.0

Jester 2.0 은 사용자간의 평가 결과 차이를 이용하여 사용자간의 성향을 예측하는 기존의 Distance 방법이 가지는 수행 속도 저하의 문제점을 보완하기 위해 제안되었으며, 평가 행렬을 구성할 때 손쉽고도 빠르게 만들 수 있다는 것과 활성화된 사용자에 대한 선호도를 조사할 때 중복적인 계산을 피하게 함으로써 온라인에서도 적합하다는 장점이 있다.[1]

Jester 2.0 은 correlation 을 이용하여 사용자를 Clustering 하고 이를 바탕으로 Collaborative Filtering 을 수행하는 방법이다. Jester 2.0 의 처리 절차는 아래 그림 2-1 과 같다.

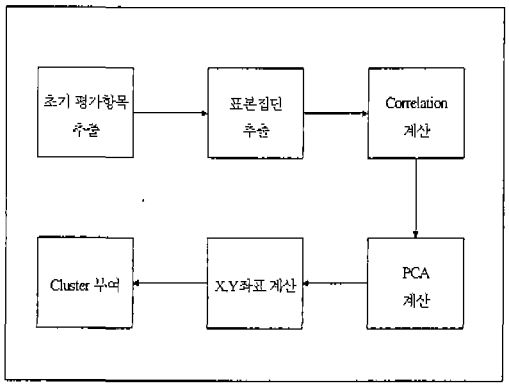

그림 2-1. Jester 2.0 처리 절차

아래의 그림 2-2 는 Jester 2.0 으로 Clustering 이 완료된 결과를 나타낸다.

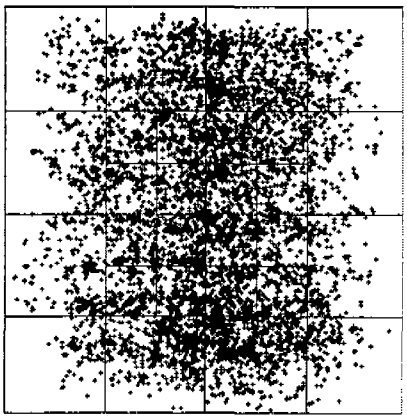

그림 2-2. Jester 2.0 처리 결과

2.2 적용 상 문제점

Jester 2.0 의 방법은 표본으로 추출된 사용자의 집단에 대해 완전한 평가 행렬을 요구하며 이 표본 집단의 Clustering 결과는 전체 사용자의 선호도를 반영하는 기준 자료로 활용된다. 이 과정에서 평가 행렬을 완전히 채우기 위해 결측치를 예측하게 되며 예측된 결측치는 사용자의 성향 정보와 동일하게 취급된다. 따라서 결측치가 개별 사용자의 성향을 제대로 추측해 내지 못한다면 잘못된 성향을 나타낼 수 있게 된다.

Jester 2.0 이 제안될 때 적용된 domain 은 유머 사이트 였다. 우선 선호도를 분류할 수 있는 공통된 평가 아이템을 선정하기 위한 적극적이며 충성도가 높은 집단으로부터 초기의 평가 행렬을 모두 채운 다음 이중 성향을 대표할 수 있는 몇 가지의 상품을 이용하여 아래 식 2-1 과 같이 correlation 을 생성한다.[1]

$$C_{jk} = \frac{\sum_{i=1}^{n_{i_k}} S_{ijk}}{\sqrt{\sum_{i}^{n_{i_j}} R_{ij}^2} * \sqrt{\sum_{i}^{n_{i_k}} R_{ik}^2}} \qquad \forall j,k = 1..n_J$$

....(식 2-1) Global Correlation Matrix 의 구성

그리고 표본 사용자 집단에 참여하지 못했던 신규 사용자는 공통된 평가 항목의 일정 범위를

평가 해야지만 자신의 선호도를 측정할 수 있고 '이웃'이라 불리는 공통된 선호도 집단을 확정 지을 수 있다.

그러나 상업적인 목적의 사이트에서는 명시적인 평가를 받는 것은 굉장히 힘들며 시간이 많이 소요된다. 국내의 한 상업적인 사이트에서 평가를 받은 결과 전체 상품의 약 1%정도를 받는데 걸리는 기간은 약 5 개월 정도가 소요되었으며 이때 기준이 되는 공통적인 아이템을 일정 수준 이상 평가한 사용자는 1% 미만으로 집계 되었다.

또한 상품의 순환 주기가 빠른 상업적 사이트에서는 공통된 아이템을 계속 우선적으로 평가하도록 유도 할 수 없다. 특히 신상품 위주의 선호도를 가진 국내의 소비 성향은 이러한 처리 절차를 적용하기 힘들게 한다.

이와 같이 Jester 2.0 의 처리 절차대로 상업적인 목적의 사이트에 적용하기에는 많은 문제점을 포함하고 있으므로 이를 개선하기 위한 방법으로 본 논문에서는 상품 연관성을 이용한 결측치 추론 방법을 제안한다.

3. 결측치 추론

3.1 상품 연관성

앞 장에서 알아본 것과 같이 명시적인 사용자의 평가를 받는 것은 상당히 힘들다. 또한 명시적인 평가 중 우선 특정 상품을 평가하게 하는 것은 더욱 힘들다.

따라서 사용자에 대한 선호도를 분류하기 전 우선 사용자가 평가하지 않은 아이템에 대한 평가 값을 예측할 수 있는 또 다른 예측 시스템이 필요하다는 가설을 세웠다.

이 가설을 기반으로 사용자를 분류하는 방법 이외에 별도로 상품(아이템)에 대한 분류 방법을 전(前) 처리과정으로 활용 하였다.

3.2 결측치 예측 방법

사용자의 선호도를 파악하기 위해 사용되는 correlation 에서 사용자와 상품에 대한 평가 값이 필수적으로 요구된다.

상품간 연관성을 부여하는 방법은 Jester 2.0 과 동일하게 처리 하였다. 다만 상품을 처리하는 부분과 사용자를 처리하는 부분을 모두 서로 바꾸어서 처리함으로써 최종 결과로는 선호도를 기반으로 한 각 상품의 X 와 Y 좌표가 산출되도록 하였다. 이 좌표를 활용하면 상품에 대한 cluster number 를 부여할 수 있다.

상품간 연관성을 기준으로 결측치를 예측 할 수 있는 방법은 아래의 두 가지 방법이 있다.

첫번째는 상품에 대한 연관성을 이용해서 상품간 선호도 차이를 구하는 방법으로 각 상품별 좌표가 인접하게 구성된 '이웃'의 상품에 평가한 경우 이웃의 상품에 대한 평가 값을 활용하는 방법이다. 이 방법을 이용하면 각 상품간의 선호도 차이별 weight 를 산출할 수 있어 좀 더 정밀하게 예측할 수 있는 장점이 있지만 수행 속도에 악영향을 미칠 수 있다.

각각의 상품간 선호도의 차이를 산출하는 방법은 아래의 식 3-1 과 같다.

$$Dist(x,y)= \sqrt{(|X_i-X_j|^2 + |Y_i-Y_j|^2)} \ldots \text{(식 3-1)}$$

X_i : I 번째 상품의 x 좌표

Y_i : I 번째 상품의 y 좌표

각 상품간 선호도의 차이가 산출되면 그 차이를 가중치로 이용하여 평가 값을 예측 한다. 평가 값을 예측하는 방법은 아래의 식 3-2 와 같다.

$$P = \frac{|W_1 \times R_1| + \cdots + |W_n \times R_n|}{n} \quad \ldots \text{(식 3-2)}$$

W_n : n 번째 상품과 예측 하고자 하는 상품의 선호도 차이를 나타내는 가중치

R_n : n 번째 상품에 대한 실제 평가 값

두 번째 방법으로는 cluster 를 이용하는 방법이다. 동일한 cluster 내에 분포된 상품은 동일한 선호도를 보이므로 [1] 결측치가 발생한 상품과 동일한 cluster 내에 활성화된 사용자가 평가한 상품이 있는 경우 실제 평가한 상품의 평가 값에 근거하여 결측치가 발생한 상품의 평가 값을 예측 하는 방법이다.

Cluster 를 이용한 평가 값 예측 방법은 아래의 식 3-3 과 같으며 F_n 는 식 3-2 와는 달리 동일한 cluster 에 속하는 지 많은 판단하는 bit 형으로 0 과 1 의 값 만을 가진다.

$$P = \frac{|F_1 \times R_1| + \cdots + |F_n \times R_n|}{n} \quad \cdots (식\ 3\text{-}3)$$

F_n : n 번째 상품과 예측 하고자 하는 상품이 동일한 cluster 내에 존재 하는가를 나타내는 Flag

R_n : n 번째 상품에 대한 실제 평가 값

4. 실험

4.1 실험 자료

상품의 연관성을 이용하여 결측치를 추론한 결과를 기존의 처리 방법과 비교하기 위해 본 논문에서는 Group Lens 에서 제공하는 943 명의 사용자가 1682 개의 아이템을 평가한 80,000 개의 평가 자료를 사용하였다.

자료의 구성은 사용자와 아이템에 대해 개인 정보 등의 중요 정보를 제거 한 후 제공 되었으며 각 평가 값의 범위는 1.0 에서 5.0 까지 5 단계의 평가 값으로 구성 되어있다.

자료의 구성 상 한명의 사용자는 평균 85 개 정도의 아이템을 평가 하였으며 이것은 전체 아이템의 약 5%에 달하는 비율이다. 이 자료 중에서 사용자가 평가한 자료의 80%는 학습자료로, 20%는 실험 자료로 활용하였다. 또한 Group Lens 사에서 임의로 선택된 평가자료와 학습자료의 다섯 가지의 데이터 유형을 이용하여 정확도를 측정 한 후 그 평균 값을 이용하여 종합적인 예측의 정확도로 제시하였다.

4.2 정확도 측정 기준

Collaborative Filtering System 평가의 정확도를 검증하기 위한 많은 측정 기준들이 제안되어 왔으며 각 제안들은 통계적인 정확도 측정 방법과 의사 결정 지원 정확도 측정 방법의 두 가지 범주에 포함된다.[2]

본 논문에서는 통계적인 정확도 측정 방법을 사용하였으며 이 방법은 예측 값과 평가 값을 동시에 가지는 경우에 시스템에서 예측한 값과 사용자의 평가 값을 비교함으로써 정확도를 측정하는 방법이다.[2]

4.3 실험 방법

본 논문에서는 앞에서 설명한 바와 같이 Group Lens 사에서 제공되는 다섯 가지의 평가 집합을 이용하여 Jester 2.0 의 처리 방법에 근거한 정확도 측정 자료와 제안된 상품 Clustering 에 의한 전처리 과정을 거친 처리 방법의 두 가지 방법에 대한 정확도를 측정하여 비교한다.

두 처리 방법에 대한 정확도 산정을 위해 아래 식 4-1 과 같이 오차율을 산정하며, mean absolute error 방법을 사용한다. 이 방법은 예측

$$MAE = \frac{|p_1 - a_1| + \cdots + |p_n - a_n|}{n \times R_T} \quad \cdots (식\ 4\text{-}1)$$

값과 평가 값의 차이에 대한 평균값을 산출하는 방법이다.

p_n : n 번째 상품에 대한 시스템 예측 값

a_n : n 번째 상품에 대한 실제 사용자 평가 값

4.4 실험 결과

아래 표 4-1 는 Jester 2.0 방법에 근거해서 예측을 수행한 후 MAE 를 백분율로 산출한 결과이다. U1..U5 는 Group Lens 에서 임의로 분류한 다섯 가지의 학습-검증 데이터 집합이다.

U1	U2	U3	U4	U5	평균
20.1	**18.6**	**19.1**	**19.2**	**18.9**	**19.1**

표 4-1. Jester 2.0 에 의한 예측의 MAE

Jester 2.0 을 기준으로 처리할 경우 Correlation 시에 발생되는 결측치는 평가 값의 범위의 중간 값인 3.0 을 적용하여 예측을 수행하였다.[1]

상품간 연관성을 이용해서 전처리 과정을 거친 경우의 정확도 산출 결과는 표 4-2 와 같다.

U1	U2	U3	U4	U5	평균
17.6	**18.0**	**18.3**	**19.1**	**19.0**	**18.4**

표 4-2. 상품간 연관성을 고려한 예측의 MAE

표 4-1 과 4-2 에서 보듯이 U5 외 경우만을 제외하고 전반적으로 정확도가 향상 된 것을 볼 수 있다. 또한 Jester 2.0 의 처리 방법 의 평균 에러 율 이상의 값은 발생 되질 않았다.

Group Lens 의 자료는 사용자-상품 대 평가 비율이 상당히 높으므로 결측치가 많이 발생되는 국내의 상업적인 사이트의 경우 MAE 의 차이는 더 많아 지게 되고 효과적인 활용 방안 되리라 예상된다.

5. 향후 연구 및 결론

본 논문에서는 correlation 을 계산 하기 위한 공통 적인 평가 항목 부족 시 발생하는 결측치를 보완하는 방법으로 상품간의 연관성을 이용하여 개개인의 평가 정보로부터 공통적인 평가 값을 예측하는 방법을 사용하였다.

이 방법은 실험 결과에서 알 수 있듯이 기존의 방법보다 정확도를 향상 시킬 수 있었다.

따라서, 향후의 연구 방향은 상품 연관성을 이용한 활용 방법에 대한 연구와 정확도를 좀 더 향상시킬 수 있는 상품간의 연관성 추출 방법에 대한 연구를 진행할 것이다.

6. 참고 문헌

[1] Dhruv Gupta, Mark Digiovanni, Hiro Narita, Ken Goldberg., Jester : Efficient Rating Prediction For Statical Joke Retrival, SIGIR, 1999

[2] Johnathan L. Herlocker, Joseph A. Konstan Al Borchers and John Riedl , An Algorithmic Framework for Performing Collaborative Filtering, Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 1999.

[3] 정준, 협력적 여과에서 평가 행렬의 희소성 문제를 해결하기 위한 SVD 의 적용방법에 관한 연구, 한국 지능정보시스템학회 2000 년 춘계학술대회 논문집 pages 317-322.

[4] Daniel Billsus and Michel J. Pazzani., Learnibg Collaborative Information Filters., Proceedings of the 15th International Conference on Machain Learning, pages 46-54.

[5] Pattie Maes. Agents that reduce work and information overload. Communications of the ACM, 37(7):30-40, July 1994.

[6] David Maltz and Kate Ehrlich. Pointing the way : Active collaborative filtering. In Proceedings of ACM CHI'95 Conference on Human Factors in Computing Systems, pages 202-209, 1995.

[7] Paul Resnick, Neophytos Iacovou, Mitesh Suchak, Peter Bergstorm, and John Riedl. Grouplens:An open architecture for collaborative filtering of netnews. In Proceeding of ACM CSCW'94 Conference on Computer Supported Cooperative Work, pages 175-186, 1994