임계값이 표준편차에 미치는 영향에 관한 연구

김선옥*°,이석준**, 이희춘***

*한라대학교 정보통신공학부, **상지대학교 경영정보학과, ***상지대학교 컴퓨터데이터정보학회

A Study about the Impact of Standard Deviation for critical point

Kim, Sun-Ok, Lee, Seok-Jun, Lee, Hee-Choon

Halla University, sanggi University, sanggi University

E-mail : sokim@halla.ac.kr, digitaldesign@sangji.ac.kr, choolee@sangji.ac.kr

요 약

이웃기반 협력 필터링을 이용한 추천시스템은 적은 평가 자료로 인해 추천 성능에 문제가 생긴다. 이는 다른 고객의 정보도 추천에 사용하는 협력 필터링에서 이웃고객 선정에 문제가 생겨 추천시스템의 신뢰가 떨어진다. 본 논문은 추천시스템의 신뢰를 높이기 위한 방법으로 선호도 평가치가 적은 상품을 임계값을 이용하여 선별하고 이에 따라 고객의 표준편차를 조사하였다. 그리고 표준편차가 낮은 고객에 대한 MAE를 분석하여 예측의 정확도가 높아짐을 알 수 있었다.

1. 서론

전자상거래에서 사용되는 추천시스템은 고객에게 필요한 정보를 제공하여 고객이 편리하게 정보에 접근할 수 있는 시스템이다. 상품에 선호도를 표시한 고객에게 맞춤 정보와 고객이 원하는 정보를 미리 예측해서 상품을 추천하는 추천시스템은 인터넷에서 성공적인 기법으로 상용화되고 있다. 추천시스템은 연관규칙이나 사례기반추론 등 다양한 방법으로 구현될 수 있는데, 추천시스템의 기술 중 하나는 복잡한 정보 속에서 고객이 원하는 정보를 선택해 추천하는 정보 여과 기술이다(Kim, 2008). 정보여과 기술은 크게 두 가지 기술로 나누어진다. 이 기술 중 하나인 내용기반 필터링은 고객 자신이 과거에 선호했던 상품에 대한 정보만을 가지고 추천을 한다. 이에 반해 협력기반 필터링은 고객 자신의 정보와 이웃고객에 대한 정보를 함께 사용하므로 상품에 대한 추천의 범위가 매우 다양하다. 하지만 협력기반 필터링은 일정수준의 평가치가 있어야 하며 선호도 평가치가 적은 경우에는 고객에게 추천할 상품에 대한 추천 성능에 문제가 생긴다. 이러한 추천시스템의 성능을 개선하기 위한 연구가 꾸준히 진행되어 왔다.

본 논문은 추천시스템의 성능을 개선하기위해 적은 평가치를 갖는 상품에 대한 정보를 임계값을 이용하여 선정하고, 선정된 상품에 대한 선호도를 표시한 고객에 대한 표준편차를 조사하였다.

2. 연구배경

2.1 이웃기반 협력 필터링

협력 필터링은 1992년 Goldberg가 도입하여 문맥기반시스템에서 이메일을 필터링하기 위해 처음으로 사용하였다(Goldberg, 1992). 미네소타 대학의 GroupLens에서는 인터넷을 기반으로 한 토론 시스템인 유즈넷 뉴스(UseNet News) 그룹의 기사를 추천하기 위해 최초로 자동화된 협력기반 알고리즘인 이웃기반의 협력 필터링(Neighborhood-Based Collaborative Filtering) 알고리즘을 제시하였다. 이웃기반의 협력 필터링은

추천대상고객의 선호도 평가 자료와 이웃 고객의 선호도 평가 자료를 가지고 추천 상품에 대한 선호도를 예측하는 시스템이다. 이때 이웃 고객의 선호도 평가자료가 어느정도는 있어야 추천이 신뢰있게 이루어질 수 있다. 충분한 평가 자료를 이용하여 추천대상고객에게 특정 상품에 대한 선호도를 예측하기 위한 선호도 예측은 다음식을 이용한다(Resnick, 1994; Konstan, 1997).

$$\hat{U}_x = \overline{U} + \frac{\sum_{J \in Raters} (J_x - \overline{J}) r_{uj}}{\sum_{J \in Raters} | r_{uj}|} \quad (1)$$

식(1)에서 $\hat{U}_x$은 상품 x에 대한 추천 대상 고객 u의 선호도 예측 평가값이다. $\overline{U}$는 추천 대상 고객 u가 평가한 모든 상품에 대한 평균이다. J_x는 상품 x에 대한 이웃 고객 j의 선호도 평가값이고, $\overline{J}$는 이웃 고객 j가 평가한 모든 상품에 대한 선호도의 평균을 나타내고 아이템 x에 대한 평가치는 제외하며 다음식을 이용하여 계산한다.

$$\overline{J} = \frac{\sum_{i=1}^{n} J_i}{n}, \quad i \neq x \quad (2)$$

식(1)에서 r_{uj}는 추천 대상 고객 u와 추천 대상 고객의 이웃고객인 j의 선호 유사 정도를 나타내는 유사도 가중치이다.

2.2 유사도 가중치

이웃기반의 협력 필터링을 적용하기 위한 첫 번째 단계는 특정 상품에 대한 선호도 예측을 위해 이웃고객을 선정해야 한다. 다음 단계는 이렇게 선정된 이웃고객들과 추천 받을 고객간의 선호도 유사 정도를 나타내는 유사도 가중치를 구하는 것이다. 특정 상품을 추천 받고자 하는 고객은 이웃기반의 협력 필터링을 이용해서 이미 선호도를 표시한 이웃 고객들이 경험을 통해 얻어진 정보를 바탕으로 보다 정확한 추천을 얻을 수 있다. 이때 특정 상품을 평가한 이웃 고객과 추천을 받고자하는 고객과의 선호도 유사관계를 나타내는 정도를 값으로 표시할 수 있는데 이값을 유사도 가중치라 하며 GroupLens 에서는 피어슨의 상관계수를 사용하였다.(Resnick, 1994)

다음은 고객과 이웃고객의 선호 유사 정도를 나타내는 유사도 가중치인 피어슨의 상관계수이다.

$$r_{uj} = \frac{\sum_{i=1}^{m} (R_{u,i} - \overline{R_u})(R_{j,i} - \overline{R_j})}{\sqrt{\sum_{i=1}^{m} (R_{u,i} - \overline{R_u})^2 \cdot \sum_{i=1}^{m} (R_{j,i} - \overline{R_j})^2}} \quad (3)$$

식(3)에서, r_{uj}는 추천을 받고자하는 고객 u와 이웃고객 j가 선호도를 평가한 상품에 대한 피어슨의 상관계수이며, $R_{u,i}$는 추천 대상 고객 u가 평가한 상품 i에 대한 선호도 평가 치이고 $\overline{R_u}$는 추천 대상 고객 u가 평가한 상품들에 대한 평균이다. 그리고 $R_{j,i}$는 이웃고객 j가 선호도를 평가한 상품 i 에 대한 평가값이며 $\overline{R_j}$는 이웃고객 j 가 평가한 상품들의 선호도 평가치에 대한 평균이다.

또한 두 문서간 유사성을 계산하기 위해 각 문서에서 단어의 출현 빈도를 벡터로 처리하여 계산하는데 이 때 사용하는 코사인 벡터를 협력 필터링에 적용하여 이웃 고객과의 선호도 유사 정도인 유사도 가중치로 사용하며 이를 벡터 유사도라 한다(Breese, 1998). 피어슨의 상관계수와 벡터 유사도는 모두 고객 간의 선호도 유사 정도를 나타내며, 피어슨 상관계수는 두 고객의 유사 정도를 1에서 −1까지의 양과 음의 관계로 표현되고, 벡터 유사도의 경우는 음의 값이 존재하지 않으며 최대 1의 유사도 가중치 값으로 표현된다. 다음은 고객 u와 이웃 고객 j의 선호도 유사 정도를 나타내는 유사도 가중치인 벡터 유사도이다.

$$r_{uj} = \cos(\overrightarrow{R_u} \cdot \overrightarrow{R_j}) = \frac{R_u \cdot R_j}{|R_u| \cdot |R_j|} \quad (4)$$

Breese(1998)은 피어슨 상관계수와 벡터 유사도 외에 사용할 수 있는 유사도 가중치로 기본 선호도, 역사용자 빈도, 사례확대 등의 다양한 유사도 가중치를 소개하고 평가하였다.

본 논문은 피어슨 상관계수를 고객 간의 선호 정도를 나타내는 유사도 가중치로 이용하여 분석하였다.

2.3 희소성에 관한 연구.

이웃기반의 협력 필터링을 이용한 추천시스템은 적은 평가 자료를 사용할 경우 평가 자료의 희소성으로 인해 예측 정확도에 문제가 생긴다. 이것을

희소성 문제라 한다. 이러한 희소성의 문제를 해결하기 위하여 다양한 방법의 연구가 진행되고 있다. Pazzani(1999)는 희소성의 데이터를 분리하여 속성별로 데이터를 추출하여 선호도 예측을 향상시키는 연구를 하였다. 또한 Kim(2007)은 희소성이 높은 데이터를 희소하지 않는 상태로 변형하는 데이터 변형기법을 제안하였다. 이 논문에서 사용한 데이터 변형기법은 아이템의 추가 속성 정보에 대한 확률분포를 이용하여 희소성의 데이터를 변경하고, 변경된 선호도 데이터를 협력기반 필터링을 이용하여 추천의 성능을 향상 시키는 것이다. 여기서는 다양한 형태의 선호도 평가에 대한 데이터들의 특성을 무시하고 확률분포만을 사용하였으므로 각 데이터들에 대한 정보가 정확하게 반영되지 않았다. Melville(2002)는 희소성이 있는 사용자의 평가치를 행렬을 이용하여 내용기반 필터링을 통해 사용자 평가 행렬을 생성하고, 이를 기반으로 협력기반 필터링을 이용하여 추천에 사용하였다. 이 논문에서는 희소성의 문제가 조금 완화되었지만 추천의 정확도는 크게 향상되지 못하였다. 그리고 Soboroff(1999)는 행렬을 이용한 SVD (Singular Value Decomposition)를 계산하여 희소성의 문제에 접근하였으며 이는 계산속도 향상에는 기여하였으나 결과적으로 정확도는 크게 나아지지 않았다. Kim(2007)은 희소성의 수에 따라 집단을 분리하여 희소성이 MAE에 미치는 변화를 분석하였고, 분류된 집단에 따라 MAE의 유의적인 차이가 있음을 밝혔다.

본 논문은 희소성이 있는 데이터를 선별하는 방법을 임계값을 이용하였으며, 임계값에 의해 선별된 고객들의 선호도 표준편차를 분석하였다.

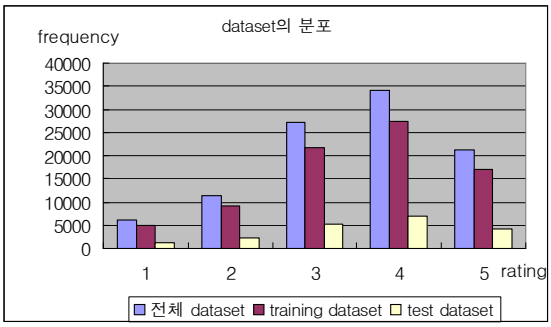
3. 연구방법

3.1 실험 데이터

GroupLens에서 제공되는 MovieLens Data set의 100k 데이터를 사용하였으며 MovieLens 100k 데이터는 943명의 사용자가 1682편의 영화에 대한 선호도를 평가한 평가 자료 100,000개로 구성되어 있다. 사용자는 20편 이상의 영화에 대해 선호도를 평가하였으며, 최소 1점에서 최대 5점까지 선호도를 표시하였다. 본 논문에서는 GroupLens에서 제공되는 MovieLens 100K dataset을 80%의 training dataset 과 20%의 test dataset 으로 랜덤하게 분할하여 test dataset을 사용하여 training dataset에 대한 선호도 평가치를 산출하였다.

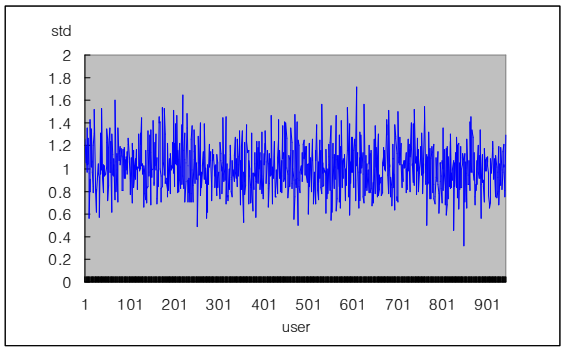
3.2 실험방법

GroupLens에서 제공되는 MovieLens 100K dataset을 80%의 training dataset 과 20%의 test dataset 으로 랜덤하게 분할한 dataset들에 대한 선호도 분포도를 살펴보면 전체자료에 대한 선호도 분포와 유사하여 실험 데이터로 적합함을 알 수 있다. 그림1은 MovieLens의 100K dataset 전체 자료에 대한 고객의 선호도 평가치 빈도 분포와 training dataset과 test dataset에 대한 선호도 평가치의 빈도 분포를 나타낸다.



<그림 1> dataset들의 선호도 빈도분포

임계값에 따라 표준편차의 변화를 조사하기 위해 실험에 사용되는 training dataset에 대한 고객별 표준편차의 분포는 아래와 같다.



<그림 2>training dataset의 고객별 표준편차 분포

본 논문의 연구를 위해 사용된 dataset들에 대한 희소성 아이템에 대한 희소율을 전체 dataset과 training dataset 으로 나누어 조사하였다. 표1에서 보는 바와 같이 모든 dataset이 높은 희소성을 가

지고 있음을 알 수 있다.

<표 1> 실험데이터의 희소성

Dataset	User:Item	Sparsity(%)
MovieLens 100K	943:1682	94.95
training dataset	943:1378	93.90

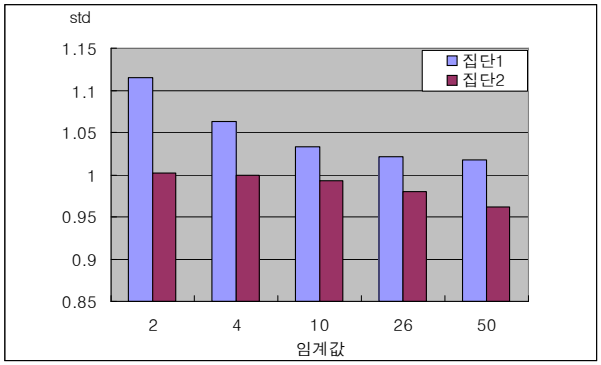
희소성이 있는 데이터는 추천의 정확도에 영향을 줄 수 있다(Kim, Lee, 2007). 그러므로 본 논문에서는 희소성 데이터를 분류하기 위하여 아래 식을 이용하여 training dataset에서 희소성이 있는 데이터를 분류한다.

$$T_k(j) = \sum_{i=1}^{user} X_k, \quad k = \{1,2,3,4,5\} \qquad (5)$$

여기서, $T_k(j)$는 선호도를 k값으로 평가한 상품 j에 대한 고객의 모든 평가 값이다. 이에 따라 희소성이 있는 데이터를 추출하기 위해 본 논문에서 사용된 식은 다음과 같다.

$$\sum_{k=1}^{5} T_k(j) \le s \qquad (6)$$

여기서, j는 고객이 선호도를 표시한 상품을 나타내고 s는 희소성을 추출하기 위한 임계값이다. 임계값 s에 따라 희소성 데이터가 선택되며 희소성 데이터를 포함하는 집단을 집단1, 나머지 데이터들을 포함하는 집단을 집단2 이라 구분하여 이들 집단 간의 training dataset들에 대한 표준편차의 변화를 살펴보면 아래와 같다.



<그림 3> 임계값에 따라 추출된 집단 간 표준편차 분포

3.3 연구결과

실험에서 training dataset들을 임계값에 의해 구분하여 임계값보다 작은 집단을 집단1라하고, 임계값보다 큰 집단을 집단2로 하고 이들 집단 간의 표준편차의 변화를 알아보기 위하여 t검정 결과를 조사하였다.

<표 2> 임계값에 따라 추출된 집단 간 표준편차의 변화에 대한 t검정결과

임계값	구분	N	std	t	유의확률
2	집단1	11	1.1153	1.803	0.072**
	집단2	932	1.0024		
4	집단1	56	1.0732	2.600	0.009*
	집단2	887	0.9993		
10	집단1	234	1.0375	2.892	0.004*
	집단2	709	0.9926		
26	집단1	526	1.0230	3.234	0.001*
	집단2	417	0.9794		
50	집단1	693	1.0188	3.762	0.000*
	집단2	250	0.9618		

*:p<0.05, **:p<0.01

t검정결과 희소성 데이터가 포함된 집단1과 임계값보다 큰 dataset집단인 집단2의 집단 간에는 유의적인 차이가 있음을 알 수 있다. 임계값이 50인 경우 집단1과 집단2의 모든 dataset에서 표준편차가 작아졌고, 집단2의 경우 표준편차의 값이 0.9618로 가장 작았다. 희소성 data가 가장 많은 임계값이 2인 집단1의 경우 표준편차의 평균값이 1.1153으로 가장 큰 값을 갖는다. 따라서 임계값이 커질수록 표준편차가 작아짐을 알 수 있다. 이는 data의 희소성을 보다 더 완화하였을 경우 표준편차가 작아짐을 의미한다.

4. 결론

본 논문은 이웃기반 협력 필터링을 이용하는 추천시스템에서 data의 희소성에 따른 선호도 예측 정확도의 신뢰를 높이기 위한 방법으로 임계값과 표준편차의 변화를 비교하였다. 우선 희소성이 있는 데이터는 추천시스템의 예측정확도가 떨어지므로 희소성이 있는 데이터를 선별하기 위해 임계값을 사용하였다. 이 임계값을 이용하여 희소성이 많은 집단과 희소성이 적은 집단으로 나누어 표준편차를 조사하였다. 연구 결과 임계값이 커짐에 따라 표준편차가 작아졌으며 희소성을 제거할수록 표준편차가 작아짐을 알게 되었다. 따라서 추천시스템의 예측 정확도를 높이기 위해서는 본 연구에서 제시한 임계값에 따라 선별된 표준편차가 큰 data들에 대한 연구가 필요하다. 이것은 이웃기반 협력 필터링을 이용하는 추천시스템의 희소성문제를 해결하는 한 방법이 될 것

이다. 따라서 향후 연구로 표준편차가 큰 data들의 예측 정확도를 높이는 방법에 대한 연구가 필요하다.

[참고문헌]

[1] 이희춘, 이석준 (2006). 사용자 기반 추천시스템에서 근접이웃 알고리즘과 수정알고리즘의 예측 정확도에 관한 연구, Journal of the Korean Data Analysis Society, Vol.8, No.5, 1893-1904.

[2] 강현철, 한상태, 정병주, 신연주 (2004). 개인화를 위한 추천시스템 알고리즘에 관한 연구, Journal of the Korean Data Analysis Society, Vol.6, No.4, 1043-1049.

[3] 강현철, 한상태 (2003). 웹유시지 패턴 분류를 위한 군집분석 알고리즘, Journal of the Korean Data Analysis Society, Vol.5, No.14, 537-334.

[4] 김재경, 오희영, 권오병 (2007), 유비쿼터스 환경에서 협업필터링을 이용한 상품그룹추천, 한국IT서비스학회지 Vol.6, No.2, 113-123.

[5] 이희춘 (2006). 추천시스템에서 Top-N 추천을 위한 순위적합에 관한 연구, Journal of the Korean Data Analysis Society, Vol.8, No.6, 2597-2607.

[6] Pazzani, M.J. (1999). Framework for Collaborative, Content_Based and Demograghic Filtering, Artificial Intelligent Review, 394-408,

[7] Kim Hyungil, Kim Juntae. (2005). Modifying Sparse Date for Collaborative Filtering, Journal of The Korean Society of Computer Information, Vol.32, No.1, 610-613.

[8] Melville. P., Mooney. R., Nagarajan. R, (2002). Content-Boosted Collaborative Filtering for Improved Recommendations, Proceedings of the eighteenth national Conference on Artificial Intelligence, 187-192.

[9] Soboroff. I., Nocholas. C.(1999). Combining content and collaborative in text filtering, Preceedings of the IJCAI Workshop on Machine Learning in Information Filtering, 86-92.

[10] Kim SunOk, Lee SeonJun. (2007). The Effect of Data Sparsity on Prediction Accuracy in Recommender System, Journal of the Korean Society for Internet Information, Vol.8, No.6, 95-102.

[11] Kim, S. O., Lee, S. J. and Lee, H. C. (2008). A Study on Improvement of Prediction Accuracy by Critical value, Journal of the Korean Data Analysis Society, Vol.10, No.1(B), 591-601.

[12] Resnick, P., N. Iacovou, M. Suchak, P. Bergstrom, J. Riedl, (1994). GroupLens: an open architecture for collaborative filtering of netnews, in Proceedings of the 1994 ACM conference on Computer supported cooperative work. ACM Press: Chapel Hill, North Carolina, United States. 175-186.

[13] J. Konstan, B. Miller, D.Maltz, J. Herlocker, L. Gordon, and J. Riedl,(1997). GroupLens: Applying Collaborative Filtering to Usenet News, Communications of the ACM, Vol.40, No.3, pp.77-87.