협업 필터링을 위한 EMD 기반 유사 사용자 선별 기법

강윤석, 정세현, 이상철, 장민희, 김상욱
한양대학교 컴퓨터공학과
e-mail : poyche@gmail.com

A Method for Selecting Similar Users for Collaborative Filtering

Yoon-Suk Kang, Seihyun Jeong, Sang-Chul Lee, Min-Hee Jang and Sang-Wook Kim
Dept. of Computer Engineering, Hanyang University

요 약

협업 필터링은 유사한 사용자를 선별하여 아이템을 추천하는 대표적인 추천 방법이다. 협업 필터링을 이용한 추천에서 추천 품질은 유사 사용자를 선별하는 기법에 따라 크게 달라질 수 있다. 본 논문에서는 협업 필터링 추천의 품질을 크게 개선 시킬 수 있는 새로운 유사 사용자 선별 기법을 제안한다. 제안하는 기법에서는 Earth Mover's Distance (EMD)를 이용하여 사용자간의 유사도를 정의한다. EMD를 적용하기 위해서 각 사용자를 히스토그램으로 표현하며, 히스토그램 빈(bin)간의 거리를 정의한다. 이렇게 정의된 유사도를 이용하여 타깃 사용자와 유사한 사용자들을 선별하며, 이를 기반으로 타깃 사용자가 부여한 타깃 아이템에 대한 점수를 예측한다. 다양한 실험을 통하여, 제안된 기법이 기존 기법들과 비교하여 추천의 정확도를 최대 30%까지 향상시키는 것으로 나타났다.

1. 서론

다양한 아이템 중 사용자가 선호할 만한 아이템을 찾아주는 시스템을 추천 시스템이라 부른다[1][3][4]. 협업 필터링 기법은 가장 널리 사용되는 추천시스템의 한 기법으로써, 타깃 사용자와 성향이 유사한 사용자를 이용하여 아이템을 추천한다. 유사 사용자를 정확히 선별함으로써 협업 필터링의 추천의 질을 향상 시킬 수 있다.

본 논문에서는 유사 사용자 선별을 위해 Earth Mover's Distance (EMD)를 사용하여 사용자간 유사도를 계산하고자 한다. EMD 는 멀티미디어 데이터에서 널리 사용되는 유사도 측정 함수로써, 그 정확도가 매우 높다고 알려져 있다[2].

EMD 를 유사 사용자 선별에 적용하기 위해서는 다음과 같은 두 가지 문제를 해결해야 한다. 첫째, 사용자 데이터를 히스토그램 형태로 표현해야 한다. 사용자 데이터는 사용자 번호, 사용자가 구매한(혹은 이용한) 아이템 번호, 아이템 평가 점수로 구성되어 있다. 둘째, 히스토그램 빈(bin)간의 거리가 정의되어 있어야 한다.

본 논문에서는 이러한 문제를 해결할 수 있는 새로운 EMD 적용 기법을 제안한다. 제안하는 기법은 EMD 를 통해 정확한 유사 사용자 식별이 가능하기 때문에 기존 기법에 비해 높은 정확도를 보인다. 다양한 실험을 통하여 제안하는 기법이 기존 기법에 비해 정확도가 우수함을 검증한다.

2. Earth Mover's Distance

EMD 는 데이터간 거리 측정을 위해 비교하고자 하는 두 데이터를 히스토그램으로 표현한다, 두 히스토그램 $P=\{p_1, p_2, \ldots, p_n\}$, $Q=\{q_1, q_2, \ldots, q_n\}$ 가 있을 때 p_i, q_i는 각 히스토그램의 i 번째 bin 의 비중(weight)을 의미한다. P와 Q의 비율 총합은 동일하다고 가정한다. EMD 는 이 두 데이터간의 거리를 측정하기 위해 최소(minimum) WORK 를 계산한다. WORK 란 히스토그램 P의 분포를 Q의 분포로 옮기는데 들어가는 최소의 양을 의미한다. WORK 란 한 히스토그램에서 다른 히스토그램으로 옮겨진 빈들의 양 f 와 ground distance d의 곱으로 구할 수 있다. ground distance 란 히스토그램 각 빈간의 거리를 의미한다. 식 1 은 WORK 의 계산법을 나타낸다.

$$WORK(P,Q,F) = \sum_{i=1}^{n}\sum_{j=1}^{n} f_{ij}d_{ij} \quad (1)$$

여기서, $F=[f_{ij}]$는 p_i에서 q_i로 옮겨진 히스토그램 빈의 양을 의미하고 $D=[d_{ij}]$는 옮겨진 빈간의 거리를 의미한다. EMD 는 이처럼 ground distance 를 통해 히스토그램 내 다른 위치의 빈간에도 유사도를 측정할 수 있기 때문에 정확도가 매우 높은 것으로 알려져 있다[2].

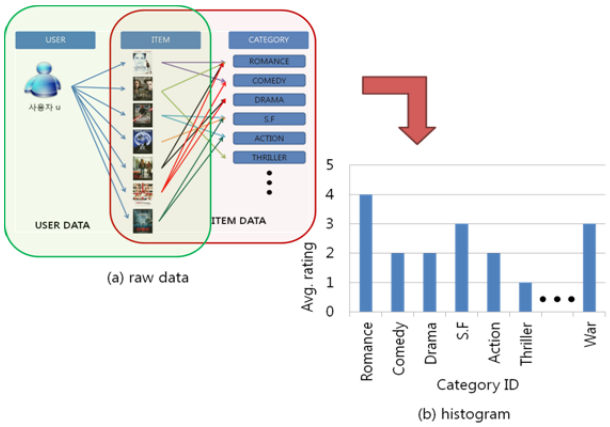
3. 제안하는 협업 필터링 기법

협업 필터링 기법은 우선 타깃 사용자와 유사한 사용자를 찾는다. 그 뒤, 유사 사용자들이 평가한 아이템들을 이용하여 타깃 사용자가 평가하지 않은 아이템의 평점을 예측한다. 예측된 평점을 기반으로 타깃 사용자에게 아이템을 추천한다. 협업 필터링 기법의 추천 품질을 향상시키기 위해서는 타깃 사용자와 유사한 사용자를 선별하는 것이 매우 중요하다.

본 장에서는 EMD 를 이용하여 유사 사용자 선별 기법을 제안한다. 유사 사용자 선별에 EMD 를 적용하기 위해서, 우선 사용자 데이터를 EMD 에 적용 가능한 히스토그램 형태로 표현해야 한다. EMD 는 유사도를 측정하기 위하여 사용자를 히스토그램의 형태로 표현해야 한다[2]. 임의의 사용자에 대한 히스토그램의 각 빈을 아이템의 카테고리로 정의하고, 해당 카테고리를 선호하는 정도를 값으로 표현한다. 이때, 선호하는 정도를 각 아이템의 평균 평가 점수을 이용하거나, 각 카테고리 내에서 평가한 아이템의 수를 이용하는 기법을 고려할 수 있다.

그림 1 은 사용자 u 를 히스토그램으로 표현한 예를 나타낸다. 그림 1 (a)는 영화 제공 사이트에서 사용자 u와 각 영화에 대한 정보와 그 관계를 나타낸다. 사용자 데이터는 사용자 ID, 사용자가 평가한 영화 ID, 해당 영화에 대한 평가 점수로 구성되어 있다. 아이템 데이터는 영화 ID, 영화가 속한 장르 ID 로 구성되어 있다. 하나의 영화는 여러 장르에 포함 되어 있다. 그림 1 (b)는 사용자 데이터와 아이템 데이터를 이용하여 사용자 u 를 히스토그램으로 표현한 예이다. 여기서 x축은 히스토그램의 각 빈을 나타내며, y축은 각 빈의 비중을 나타낸다. 여기서, 빈의 비중은 해당 빈에 대응하는 카테고리 안의 영화에 대하여 사용자 u 가 부여한 평

가 점수의 평균이다. 예를 들어, 사용자 u의 Romance에 해당하는 빈의 비중은 4이다.



<그림 1> 사용자 u를 히스토그램으로 표현하는 예.

유사 사용자 선별에 EMD를 적용하기 위해서는 각 빈간의 거리(ground distance)를 정의해야 한다. 대부분의 응용에서는 카테고리간 거리가 정의되어있지 않다. 또한, 카테고리는 서로 독립적이기 때문에 기존 EMD에서 일반적으로 사용하는 거리 함수 공식인 Euclidean 거리를 사용 할 수가 없다. 따라서, 본 논문에는 두 카테고리에서 동시에 존재하는 아이템의 비중을 이용하여 카테고리간 거리를 계산하는 기법을 제안하며, 그 기법은 식 2와 같다.

$$J(x, y) = \frac{|I_x \cap I_y|}{|I_x \cup I_y|} \quad (2)$$

여기서, x, y는 카테고리 ID, I_x는 카테고리 x에 속한 아이템 집합, I_y는 카테고리 y에 속한 아이템 집합이다.

앞서 제안한 기법을 협업 필터링 기법에 적용하여 k명의 유사 사용자를 선별하여 평가 점수를 예측하는 기법을 제안한다. 평가 점수를 예측하는 기법은 식 3과 같다[5].

$$r_{uq} = \bar{r_u} + \sum_{v \in N_{uq}} \frac{sim(u,v)}{\sum_{j \in N_{uq}} sim(u,j)} * (r_{vq} - \bar{r_v}) \quad (3)$$

여기서, r_{uq}는 사용자 u가 아이템 q에 대해 예측 할 평가 점수, $\bar{r_u}$는 사용자 u의 평균 평가 점수, N_{uq}는 사용자 u와 아이템 q를 기준으로 했을 때 유사 사용자의 집합, v와 j는 집합 N_{uq}에 속한 사용자, r_{vq}는 사용자 v가 아이템 q를 평가한 평가 점수, 그리고 $\bar{r_v}$는 사용자 v의 평균 평가 점수이다. $sim(x, y)$는 사용자 x와 y의 유사도이며, 본 논문에서는 EMD 유사도를 사용한다.

사전 실험을 통해 유사 사용자의 수를 10명으로 설정 했을 때 정확도가 가장 우수했으며, 유사 사용자의 수가 그 이상의 경우 정확도에는 큰 변화가 없었다. 따라서 유사 사용자의 수를 10명으로 하여 실험을 수행하였다.

4. 실험

4.1. 실험 환경

본 실험을 위해 MovieLens의 dataset을 사용하였다[7]. 실험 결과의 정확도의 척도로는 MAE와 RMSE를 사용하였다[6]. MAE(mean absolute error)은 실제값과 예측값의 오차에 대한 절대값들의 평균이며, RMSE(Root Mean Square Error)은 실제값과 예측값의 오차의 제곱에 대한 평균의 제곱근이다. 식 4와 5에서 p_i는 i번째 아이템에 대한 임의의 사용자 u가 평가 점수를 예측 한 값이며, r_i는 i번째 아이템의 실제 평가 점수이다.

$$MAE = \sum_{i=1}^{n} \frac{|p_i - r_i|}{n} \quad (4)$$

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(p_i - r_i)^2}{n}} \quad (5)$$

본 실험에서, test 사용자에 대해 평균 MAE와 RMSE를 통하여 각 기법의 정확도를 보인다. 평가한 사용자-아이템 쌍의 수는 118,860 쌍이며, 비교대상은 제안하는 기법과 Cosine 유사도, Pearson 상관계수를 이용한 협업 필터링 기법이다.

4.2. 실험 결과 및 분석

표 1은 사용자의 선호도를 평가한 영화의 수로 했을 때의 정확도를 나타낸다. 제안하는 기법이 Cosine 유사도에 비해 MAE는 11.33%, RMSE는 8.49% 향상 되었으며, Pearson 상관계수에 비해 MAE는 22.89%, RMSE는 17.03% 향상 되었다. 표 2는 사용자의 선호도를 장르별 평균 평점으로 했을 경우의 정확도를 나타낸다. 제안하는 기법이 Cosine 유사도에 비해 MAE는 13.10%, RMSE는 10.16% 향상 되었으며, Pearson 상관계수보다 MAE는 4.35%, RMSE는 29.66%향상 되었다. 제안하는 기법이 기존 기법에 비해 유사 사용자를 정확하게 선별하였기 때문에, 아이템 평가 점수 예측에서 우수한 결과를 나타내는 것으로 보인다.

<표 1> 선호도를 평가한 영화의 수로 했을 때 정확도 비교

평가 대상	MAE	RMSE
Count Cosine	0.90	1.76
Count Pearson	1.03	1.94
Count EMD	0.80	1.61

<표 2> 선호도를 장르의 평균 평점으로 했을 때 정확도 비교

평가 대상	MAE	RMSE
평점 Cosine	0.92	1.80
평점 Pearson	0.83	2.29
평점 EMD	0.80	1.61

5. 결론

본 논문에서는 EMD를 이용하여 유사 사용자를 선별하는 기법을 제안하였다. 유사 사용자를 선별하기 위해 각 사용자를 히스토그램의 형태로 표현하고, 히스토그램의 빈간 거리를 정의한다. 제안하는 유사 사용자 선별 기법에 기반한 협업 필터링을 통하여, 각 유사 사용자가 평가하지 않은 아이템의 평가점수를 예측하는 기법을 제안한다. 다양한 실험을 통하여 제안하는 기법이 기존 기법에 비해 MAE는 최대 23%, RMSE는 최대 30% 우수함을 검증하였다.

감사의 글

본 연구는 2012 년도 정부 (교육과학기술부)의 재원으로 한국연구재단의 지원 (No. 2012007817)과 지식경제부 및 정보통신산업진흥원의 IT 융합 고급 인력과정 지원사업 (NIPA-2012-H0401-12-1001)의 지원을 받아 수행되었습니다. 또한, 본 연구는 중소기업청의 재원으로 산학연공동기술개발사업 (No.C0006278)의 지원과 문화체육관광부 및 한국저작권위원회의 저작권기술개발사업의 지원으로 수행되었습니다.

참고 문헌

[1] G. Adomavicius, A. Tuzhilin, "Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions," *IEEE transactions on knowledge and data engineering*, Vol. 17, No.6, pp.734-749, 2005.

[2] Y. Rubner, C. Tomasi, and L. J. Guibas, "A metric for distribution with applications to image databases," *In Proc. of IEEE Int'l Conf. on Computer Vision*, pp.59-66, 1998.

[3] A. F. Smeaton and J. Callan, "Personalisation and recommender systems in digital libraries," *International Journal on Digital Libraries*, Vol.57, No.4, pp.299-308, 2005.

[4] W. Yeo, H. Park, Y. Kwon and Y. Park, "Application of Research Paper Recommender System to Digital Library," *Korea Contents Association*, Vol.10, No.11, pp.10-19, 2010. (In Korean)

[5] Y. Shi, M. Larson and A. Hanjalic, "Exploiting user similarity based on rated-item pools for improved user-based collaborative filtering," *In proc. of ACM conf. on Recommender Systems*, pp.125-132, 2009.

[6] J. Han, M. Kamber, *Data Mining*, Morgan Kaufmann, 2006.

[7] GroupLens, http://www.grouplens.org.