사용자 클러스터링을 통한 개선된 협력적 정보여과

김학균, 조성배
연세대학교 컴퓨터과학과

Improved Collaborative Information Filtering with User Clustering

Hak-Gyoon Kim and Sung-Bae Cho
Computer Science Department, Yonsei University

요 약

정보추천 시스템은 사용자가 어떤 정보를 선호하는지를 식별함으로써 산재한 정보 중에서 적절한 정보만을 제공하는 것을 목표로 한다. 이러한 정보추천 시스템에서 사용되는 정보여과 기술에는 내용기반 여과와 협력적 여과가 있다. 기존의 협력적 정보여과 기술은 선호도를 적게 제시한 사용자에게 정보를 추천하기 어렵고, 동일한 상품 정보에 대해서 사용자의 평가가 없을 경우 사용자간의 유사성을 판단하기 어려운 단점이 있다. 본 논문은 SVD (Singular Value Decomposition)를 통해 사용자 프로파일을 정량화함으로써 사용자 선호도 행렬로부터 숨어있는 의미정보를 추출하여 동일한 정보에 대해 선호도를 평가해야 한다는 단점을 극복한다. 이 때, 사용자 프로파일 벡터를 비감독 학습 알고리즘인 SOM (Self-Organizing Map)으로 클러스터링하여 사용자를 분류하고, 정보추천은 사용자 그룹간에서 이루어지며 Pearson correlation 알고리즘을 이용한다. 기존의 방법과 비교한 결과, 제안한 방법이 새로운 사용자에 대해서도 적절한 정보를 추천할 수 있음을 볼 수 있었다.

1. 서론

기하급수적으로 증가하는 상품 정보 및 서비스는 개인의 취향이나 수준 등에 관계없이 모든 사용자들에게 일률적으로 제공된다. 따라서 개인의 축적되는 정보를 활용하지 못한 채, 사용자 스스로 정보를 선별적으로 획득하여야 한다. 이러한 과정은 사용자에게 많은 시간과 노력을 요구하기 때문에, 정보를 자동적으로 여과하여 사용자에게 유용하고 적절한 형태로 제공하는 정보 추천 시스템의 필요성이 제기되었다[1].

이러한 정보추천 시스템에서 사용되는 기술은 내용기반 여과(Content-based Filtering)와 협력적 여과(Collaborative Filtering)가 있다. 내용기반 여과는 정보의 특징을 기술하는 정보와 사용자의 기호를 담고 있는 프로파일을 비교하여 사용자가 필요한 정보들을 추천한다. 반면에 협력적 여과는 유사한 기호를 가진 사용자 사이의 선호도에 기반하여 정보를 추천한다.

하지만, 기존의 개인화된 정보추천 시스템은 새로운 사용자에게 정보를 추천하기가 어렵고, 새로운 정보에 대해 이를 추천할 방법이 없다. 또한 동일한 정보에 선호도를 입력해야만 사용자간의 유사성을 계산할 수 있다는 단점을 가지고 있다. 이를 해결하기 위해 본 논문에서는 사용자의 선호도에 기반하여 개인화된 정보를 생성하여 그 유사성에 따라 사용자를 분류하여 새로운 사용자나 상품 정보에 대하여 추천이 가능하도록 한다. 또한 SVD를 통해 사용자의 선호도간의 숨어있는 의미정보를 추출하여 동일한 정보에 대해서 선호도를 평가해야 한다는 단점을 극복한다.

2. 협력적 정보여과

협력적 정보여과는 사용자와 유사한 기호를 가진 사용자 그룹의 선호도에 따라 정보를 제공한다. 기호를 반영하기 위해서 각 사용자는 정보에 대한 선호도를 제공하여, 이로부터 사용자간의 상관관계를 계산하여 유사 기호 사용자 그룹을 형성하고 이들의 선호도를 종합하여 정보를 추천한다. 서로 다른 사용자의 프로파일을 비교하여 기호의 유사성을 평가하여 분류하는 방법으로는 크게 통계적 방법과 의사결정 방법이 있다. 통계적 방법에는 Mean Squared Differences, Pearson, Constrained Pearson, Artist-Artist 알고리즘 등이 있다[1]. 의사 결정 방법은 사용자 프로파일을 입력하여 적절한 출력을 얻기 위한 것으로서, 다층 퍼셉트론이나 SOM 등의 신경망, Bayesian Classification, Nearest Neighbor Classification 알고리즘 등이 있다.

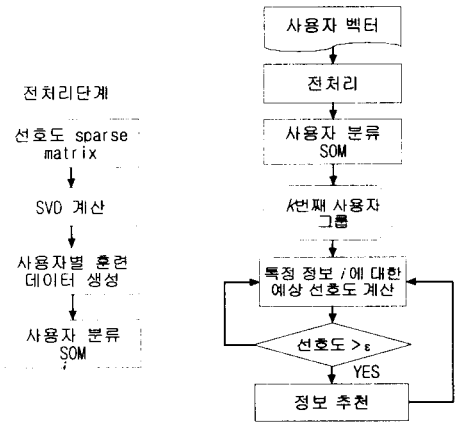
협력적 정보여과는 정보의 내용을 직접 분석할 필요없이 사용자들의 관계만을 이용하며, 정보추천의 범위를 넓혀 뜻하지 않은 것들을 추천할 수 있다. 또한 정보의 내용뿐만 아니라 정보의 우수성에 따라 정보를 추천할 수 있다[1]. Ringo 및 GroupLens는 pearson correlation을 통해 사용자들간의 정보에 대한 선호도를 비교하여 상관관계를 계산하여 유사 기호 사용자를 얻는다[1, 4]. 그래서 유사 기호 사용자들의 선호도에 기반하여 정보를 추천한다. 또한 사용자의

선호도를 SVD로 전화하여 다층 퍼셉트론의 입력으로 하여 정보에 대해 선호하는지를 결정하는 분류시스템이 있다[2].

그러나, 협력적 정보여과는 다음과 같은 단점을 가지고 있다. 첫째, 새로운 정보가 추가되었을 때, 다른 사용자들의 선호도 평가가 이루어지거나 다른 정보와의 유사성이 판명될 때까지 이를 사용자에게 제공할 방법이 없다[2]. 둘째, 기호가 특이한 사용자는 유사 기호의 사용자를 얻기 어렵기 때문에 정보를 추천하기 어렵다. 셋째, 새로운 사용자에게 정보를 추천할 방법이 없다[3].

3. 정보추천 시스템

전체적인 시스템의 구성은 그림 1과 같다. 초기 데이터는 사용자에 해당하는 행과 정보에 해당하는 열로 구성된다. 행렬에 채워진 값은 사용자의 선호도를 나타내며, 대부분의 값이 비어있는 sparse matrix 형태이다. 데이터가 2개 미만으로 있는 행을 삭제하고, 각 행에 대해 SVD를 계산하여 사용자별로 k차원의 벡터를 생성한다. 이 벡터는 각 사용자의 특징을 나타내는 것으로써, 분류 시스템의 입력으로 사용되어 사용자들을 선호도에 따라 분류한다. 정보추천은 동일 분류 그룹에 속하는 사용자들의 선호도에 따라 통계적 방법을 이용한다.



(a) 사용자 분류 (b) 정보 추천

그림 1 시스템 구성

3.1 전처리

사용자 분류를 위한 전처리 단계는 사용자들의 선호도 행렬로부터 의미 있는 숨겨진 정보를 얻고, 사용자별 벡터의 차원을 정량화하기 위해 SVD를 사용한다. m명의 사용자와 n개의 정보에 대한 선호도 값으로 채워진 sparse matrix를 A라 하고 Rank(A) = r이라 했을 때 SVD 식은 다음과 같다[5].

$$A = U\Sigma V^T$$

여기에서 orthogonal matrix인 U와 V는 각각 AA^T와 A^TA의 r개의 0이 아닌 eigenvalue와 연관된 orthonormal eigenvector를 정의하며, Σ는 singular values를 가지고 있는 diagonal 행렬이다.

singular value와 이에 해당하는 singular vector를 이용하여 행렬 A의 숨어 있는 의미 구조를 얻는다. singular values는 0으로 수렴하기 때문에 큰 값을 가지고 있는 것이 의미있는 정보이다. 벡터의 차원을 줄이기 위해 가장 큰 k개의 singular vector만을 사용한다. 변환된 행렬의 각 행은 SVD를 이용하여 k차원 벡터로 매핑된다. 즉, singular value로 가중된 행렬 V의 singular vector는 다음과 같이 k차원 사용자의 기호 벡터로 표현된다.

$$v_k = v^T V_k \Sigma_k^{-1}$$

여기에서 v는 사용자의 선호도를 담고 있는 특징 벡터이고 V_k는 k개의 원소를 가지고 있는 singular vector의 행렬이며, Σ_k^{-1}는 k개의 가장 큰 singular value를 가지고 있는 diagonal matrix이다.

3.2 사용자 분류를 위한 학습

전처리 과정을 통해 생성된 v_k는 사용자의 선호도를 다른 사용자의 선호도를 고려하여 k차원으로 변환된다. 이러한 사용자별 입력 벡터를 분류하는 알고리즘으로 본 논문에서는 SOM을 사용한다. SOM은 비감독 학습 알고리즘으로 벡터간의 유사성에 기반하여 입력 벡터를 분류한다[6].

w_i는 노드 i의 가중치 벡터일 때 입력벡터 x와 가장 근접한 노드는 다음과 같이 계산한다.

$$\| x - w_c \| = \min_i \| x - w_i \|$$

선택된 노드 i의 갱신은 다음과 같으며, 이와 같은 수식을 모든 입력 데이터에 대해 계산하여, 각 노드의 w_i를 구한다. 분류는 입력 벡터와 계산된 w_i에 중에 가장 근접한 노드를 구한다.

$$w_i(t_{k+1}) = \begin{cases} w_i(t_k) + \alpha(t_k)[x(t_k) - w_i(t_k)] & \text{if } i \in c \\ w_i(t_k) & \text{otherwise} \end{cases}$$

여기에서 w_i는 노드 i의 가중치, α는 학습률, x는 학습하고자 하는 입력 벡터이다.

3.3 정보추천

기존의 협력적 정보여과는 동일 정보에 선호도를 줄 경우에만 사용자의 유사성을 결정할 수 있고 이에 따라 유사 기호 사용자들을 결정하였다. 그러나 본 논문에서는 SVD를 이용하여 얻어진 사용자의 선호도 분포에 따라 SOM으로 사용자 그룹을 결정하여, 동일 분류 사용자 그룹간의 사용자의 유사성은 Pearson Correlation 알고리즘을 사용한다. 사용자 a와 사용자 u사이의 유사도는 다음과 같이 계산되며, 비슷한 기호의 사용자일수록 큰 값이 되고 반대의 경우는 작은 값이 된다.

$$w_{a,u} = \frac{\sum_{i=0}^{m} (r_{a,i} - r_a) * (r_{u,i} - r_u)}{\sqrt{\sum (r_{a,i} - r_a)^2 * \sum (r_{u,i} - r_u)^2}}$$

여기에서 $r_{a,i}$는 사용자 a가 정보 i에 준 선호도이고, r_a는 사용자 a의 평균 선호도이다.

정보추천은 동일 사용자 그룹의 평가 자료를 기반하여 예상 선호도를 계산하고, 이 선호도가 일정 임계값 0.6 이상이 될 경우 추천한다. 사용자 a의 특정 정보 i에 대한 선호도, 예상 추천값($p_{a,i}$)는 다음과 같다.

$$p_{a,i} = r_a + \frac{\sum_{u=0}^{n} (r_{u,i} - r_u) * w_{a,u}}{\sum_{u=0}^{n} w_{a,u}}$$

4. 실험

4.1 평가 기준

성능을 평가하는 기준으로 정확도(Accuracy)와 F-measure[2]를 이용한다. 정확도는 시스템이 예상한 선호도와 사용자가 실제 준 선호도가 얼마나 유사한지를 측정한다. 정확도를 측정하는 방법으로 Mean Absolute Error(MAE), Root Mean Squared Error 및 확률분포를 이용한 방법 등이 있다. 정확도는 사용자의 프로파일 즉, 선호도가 얼마나 잘 학습되어졌는지를 나타내는 척도이다. F-measure는 다음과 같이 정확도와 사용자가 선호하는 데이터 중에 실제로 선호한다고 평가된 양을 나타내는 리콜(Recall)의 가중치 조합을 나타내며, 그 범위는 0과 1 사이이다.

$$F = \frac{2 \times 정확도 \times 리콜}{정확도 + 리콜}$$

4.2 실험 데이터 및 환경

실험 데이터로 Eachmovie라는 영화 데이터베이스를 사용하였다[7]. 사용자 선호도는 6개 단위(0.0, 0.2, 0.4, 0.6, 0.8, 1.0)로 되어 있으며, 72916 명의 사용자가 1628개의 영화에 대해 준 2811983개의 선호도로 구성되어 있다. 본 실험에서는 그 중 나이 및 성별을 제공한 상위 2000명의 사용자 데이터를 사용하였다. 이들 2000명의 사용자는 1278개의 영화에 대하여 선호도를 입력하였다. 훈련 데이터로는 무작위로 추출된 50개의 영화에 대한 선호도를 사용하여 초기 선호도 행렬은 2000×50으로 구성된다. 이로부터 사용자를 32개의 유사 기호 사용자 그룹으로 분류하였다.

4.3 결과

제안된 협력적 정보여과의 성능을 보여주기 위하여 기존의 Pearson Correlation 방법과 비교하였다. 2000명중 20명의 사용자에 대하여 30개의 영화에 대한 선호도를 테스트 데이터로 사용한 결과는 그림 2와 같다.

	제안한 방법	Pearson Correlation
정확도	65.6%	64.4%
F-measure	59.5%	54.2%

그림 2 기존 사용자에 대한 실험

제안된 협력적 정보여과는 기존의 Correlation에 기반한 협력적 정보여과에 비해 좋은 성능을 보여줌을 볼 수 있다. 또한 제안된 방법은 새로운 사용자의 선호도 벡터를 만들어 분류 신경망에 입력함으로써 사용자의 유사 기호 그룹을 얻어낼 수 있다는 장점이 있다. 그림 3은 훈련 데이터에 포함되지 않은 10명의 사용자에 대하여 위 실험과 동일한 30개의 영화에 대한 선호도의 실험결과이다.

	제안한 방법	Pearson Correlation
정확도	61.0%	59.2%
F-measure	56.7%	51.2%

그림 3 새로운 사용자에 대한 실험

5. 결론

기존의 협력적 정보여과는 동일 정보에 대해 선호도를 입력한 사용자들간의 유사성만을 판단할 수 있다는 단점을 가지고 있다. 따라서 선호도 입력이 별로 없는 사용자에게는 추천할 수 없다. 본 논문은 이러한 단점을 보완한 개선된 협력적 정보여과 방법을 제안하였다. SVD를 이용하여 사용자 벡터를 정량화하여 이를 SOM의 입력으로 사용함으로써 사용자와 선호도 행렬에 숨어 있는 의미 정보를 추출하여 사용자를 분류하였다. 실험을 통해 기존의 협력적 정보여과에 비해 그 우수성을 입증하였다. 향후 몇몇의 선호도를 입력한 새로운 사용자에 대한 주관적 평가 실험을 할 것이다. 또한 현재의 시스템은 추천 대상 데이터에 대한 분류 작업이 이루어져 있지 않다. 따라서 대상 데이터의 분류 및 사용자 그룹과 대상 데이터 그룹간의 연관관계를 정의해야 할 것이다.

6. 참고 문헌

[1] U. Shardanand and P. Maes, "Social information filtering: Algorithms for automating 'Word of Mouth'," *Proc. of the Computer-Human Interaction Conference*, 1995.

[2] D. Billsus and M.J. Pazzani, "Learning collaborative information filters," *Proc. of Workshop on Recommender Systems*, 1998.

[3] M. Balabanovi and Y. Shoham, "Fab: Content-based, collaborative recommendation," *Comm. of the ACM*, 40(3), pp.66-72, 1997.

[4] J.L. Herlocker, J.A. Konstan, A. Borchers, and J. Riedl, "An algorithmic framework for performing collaborative filtering," *Proc. of Conf. on Research and Development in Information Retrieval*, 1999.

[5] M. W. Berry, "Large scale singular value computations," *International Journal of Supercomputer Application*, 6(1), pp. 13-49, 1992.

[6] T. Kohonen, *Self-Organizing Maps*, Springer, Berlin, 1995.

[7] P. McJones, EachMovie collaborative filtering data set, URL:http://www.research.digital.com/SRC/eachmovie/, 1997.