Singular Value Decomposition을 이용한 협력적 여과를 위한 임계값

정 준[O] 김용환 이필규
인하대학교 전자계산공학과
{g1991263, g1991260}@inhavision.inha.ac.kr

Thresholds for Collaborative Filtering using Singular Value Decomposition

Jun Jeong[O] Yong-Han Kim Phill-Kyu Lee
Dept. of Computer Science & Engineering, Inha University

요 약

협력적 여과는 사용자의 아이템에 대한 단계적 평가에 기초하여 그 평가 패턴이 유사한 사용자를 찾아 그 사용자들이 선호한 아이템을 상대방에게 교차 추천을 해주는 방법이다. 따라서, 유사한 사용자를 찾는 방법이 중요한 문제가 되며, 현재까지 여러 가지 방법들이 제안되어 왔다. 순수한 협력적 여과 방법은 n차원 공간에서 사용자를 모델링하여 가장 유사한 이웃을 찾는다. 이러한 모델링의 문제점은 사용자가 평가한 아이템의 집합은 전체 아이템의 집합에 비해서 극히 작으므로 유사한 사용자를 찾기 위해서는 충분한 수의 아이템에 대해서 평가해야 한다는 것이다. 따라서, 본 논문에서는 유사한 사용자를 찾기 위해서 충분한 수의 평가를 요구하는 명백하게 사용자의 평가를 비교하는 것 대신에 특징 가중치에 기초하여 사용자를 비교하는 방법을 사용하고 사용하는 방법의 정확성을 높일 수 있는 임계값을 제안하고자 한다.

1. 서론

협력적 여과는 사용자의 아이템에 대한 단계적 평가에 기초하여 그 평가 패턴이 유사한 사용자를 찾아 그 사용자들이 선호한 아이템을 상대방에게 교차 추천을 해주는 방법이다[6].

따라서, 유사한 사용자를 찾는 방법이 중요한 문제가 되며, 현재까지 여러 가지 방법들이 제안되어 왔다[1]. 순수한 협력적 여과 방법은 n차원 공간에서 사용자를 모델링하여 가장 유사한 이웃을 찾는다[3]. 이러한 모델링의 문제점은 사용자가 평가한 아이템의 집합은 전체 아이템의 집합에 비해서 극히 작으므로 유사한 사용자를 찾기 위해서는 충분한 수의 아이템에 대해서 평가해야 한다는 것이다[3]. 따라서, 본 논문에서는 유사한 사용자를 찾기 위해서 충분한 수의 평가를 요구하는 명백하게 사용자의 평가를 비교하는 것 대신에 특징 가중치에 기초하여 사용자를 비교하는 방법을 사용하고 사용하는 방법의 정확성을 높일 수 있는 임계값을 제안하고자 한다.

2. 관련연구

협력적 여과에 대한 연구는 [1]에서 자세히 살펴볼 수 있다. 본 논문에서 대표적인 몇가지 시스템들에 대해서 간략하게 살펴 보겠다.

Tapestry는 아이템에 대한 rating 혹은 주석을 입력받는 정보여과 시스템이며, 협력적 여과라는 개념을 최초로 제시하였다[4]. 그러나, 자신과 유사한 사용자들을 찾아주는 기능은 지원하지 않고 단지 "구전"이라는 과정에 대한 온라인 구조만을 제공하였다.

GroupLens는 협력적 여과를 뉴스그룹에 대한 개인화된 선택에 적용한 시스템이다[3]. GroupLens는 사용자들이 뉴스를 평가하고, 평가(rating)가 네트워크를 통해서 사용자 에이젼트에 분배된다. 그리고 유사한 사용자들을 찾기 위하여 Pearson r 상관계수를 적용하였다.

Ringo는 음악가에 대한 명백한 평가를 웹 혹은 전자우편을 통해서 받아들여 사용자 프로파일을 구성하는 음악 추천 시스템이다[6]. Ringo는 GroupLens에서 사용한 Pearson r 상관계수를 수정한 Contrainted Pearson r 방법을 소개하였다.

3. Singular Value Decomposition을 이용한 협력적 여과

3.1. Singular Value Decomposition

우선, A라는 행렬이 존재하면 행렬 A의 분해는 식(1)와 같은 형식의 m× n 대각 행렬 Σ 를 포함한다.

여기서, D는 m과 n에서 작은 값을 넘지 않는 어떤 r에 대해서 r× r 대각 행렬이다.

는 것은 $\hat{R}$ 가 자료의 중요하게 관련된 구조를 생성하고 대부분의 노이즈를 제거하는 효과를 가진다. 감

$$\Sigma = \begin{bmatrix} D & 0 \\ 0 & 0 \end{bmatrix}$$ ← (m−r)개의 열 / ↑ (n−r)개의 열 (1)

A는 계수가 r인 m× n 행렬이라고 하면, 식(1)와 같이 m × n 행렬인 ∑ 가 존재하며, D의 대각 성분은 행렬 A의 첫번째의 r singular value이고, δ 1 ≥ δ 2 ≥ ...≥ δ r > 0 이 되고, 식(2)과 같은 m× m 직교 행렬인 U와 n× n 직교 행렬 V가 존재한다.

$$A = U \sum V^{t} \quad (2)$$

3.2. Singular Value Decomposition을 협력적 여과

잠재적인 의미 구조(latent semantic structure)는 문서를 검색하기 위한 방법으로 사용되어 왔다[5]. 주로 용어(term)와 문서(document)의 행렬을 특별한 잠재적인 의미 구조를 파생하기 위하여 singular value decomposition(SVD)에 의해서 분석이 되었다.

SVD에서는 행과 열에 다른 존재를 가지는 임의의 직각 행렬을 가지고 시작한다. 예를 들면, 행은 사용자와 열은 아이템이 될 수 있다. 이런 직각 행렬은 "singular value decomposition"이라는 과정에 의해서 특별한 형식의 다른 세가지 행렬로 분해될 수 있다. 선형적으로 대부분의 독립적인 요소들은 매우 적을 것이고 혹은 무시될 수 있으며, 더 작은 차원을 가지는 근사 모델을 만들 수 있다. 이러한 감소된 모델에서 사용자와 사용자, 아이템과 아이템, 사용자와 아이템의 유사성은 축소된 차원에서의 값으로 어림잡아 진다. 결과적으로 두 벡터의 내적 혹은 코사인 거리는 측정된 유사성에 상응하는 지리적으로 공간 구성에서 표현된다.

구체적으로 설명하며, u명의 사용자와 i개의 아이템을 나타내는 $u \times i$ 직각 행렬인 R는 3가지의 다른 행렬로 분해될 수 있다.

$$R = USV^{T} \quad (3)$$

여기서, U와 V는 직교 행렬이고, S는 대각 행렬이다. 이러한 형태의 분해를 R 행렬의 singular value decomposition이라고 한다. U와 V는 각각 좌측,우측 singular vector라고 하며, S는 singular value의 대각 행렬이다.

일반적으로, SVD는 더 작은 행렬을 이용하여 최적의 근사형을 위한 단순한 방법을 제공한다. S의 singular value가 크기에 의해서 정렬된다면, k개의 가장 큰 값들은 유지되고, 나머지 작은 값들은 0으로 설정된다. 결과적으로 행렬의 곱은 R에 근사적으로 동일한 행렬인 $\hat{R}$ 이다.

$$\hat{R} = U_k S_k V_k^{T} \approx R \quad (4)$$

R의 k 개의 가장 큰 독립적인 선형 요소를 포함하고 있다. 사용자의 평가 정보는 {0.2,0.4,0.6,0.8,1.0} 과 같이 5단계로 이루어져 있다.

EachMovie data set의 방대한 크기 때문에 본 논문에서 소된 차원, k를 선택함에 있어서 자료의 실제 구조에 맞게 충분히 크게 선택하고 표본 추출의 오류와 중요하지 않는 상세함을 나타내지 않게 충분히 작게 선택해야 한다. 이러한 선택에서 있어서 본 논문은 실험을 통해서 협력적 여과에 적절한 k의 값을 제안한다.

일반적으로, SVD 의 분해에 위해서 사용자와 사용자, 사용자와 아이템, 아이템과 아이템의 유사성 비교가 가능하다. 그러나, 협력적 여과에서는 사용자들 사이의 유사성이 가장 중요한 요소이다. 사용자의 유사성의 비교는 다음과 같이 이루어 진다.

$\hat{R}$의 두 벡터의 내적은 두 사용자가 아이템에 대해서 유사한 패턴으로 평가한 정도를 반영한다. 모든 사용자 대 사용자의 내적을 포함하는 $\hat{R}\hat{R}^{T}$ 는 대칭 정방 행렬이다. S는 대각 행렬이고 U는 직교 행렬이므로 $\hat{R}\hat{R}^{T}$ 은 다음과 같이 계산될 수 있다.

$$\hat{R}\hat{R}^{T} = US^{2}U^{T} \quad (5)$$

$\hat{R}\hat{R}^{T}$ 안에서 (i,j) 점은 행렬 US 의 i와 j 행의 내적을 취해서 얻어 질 수 있다. 즉, 사용자에 대한 좌표로서 US 의 열을 고려한다면, 두 점의 내적은 사용자들의 비교를 의미한다.

사용자들의 유사성을 나타내는 기준으로 두 점의 내적을 사용할 수도 있고, 또한 $\hat{R}$의 두 벡터에 대한 코사인 거리를 이용할 수도 있다. US 의 행렬 곱은 두 사용자의 평가 패턴의 유사한 정도를 나타내는 벡터를 구성한다. 즉, 두 벡터의 코사인 거리가 작을수록 두 사용자의 유사한 정도는 크다고 할 수 있다.

$$\hat{R} = US \quad (6)$$

여기서, 중요한 점은 유사한 사용자로서 구분되기 위하여 두 사용자의 코사인 거리가 어느 정도가 되야 하는지의 선택하는 문제이다.

4.실험

본 논문에서 제안된 방법을 실험하기 위한 실험 자료는 DEC Systems Research Center에서 제공하는 EachMovie collaborative filtering data set을 사용하였다[]. DEC는 18개월 동안 협력적 여과 알고리즘을 실험하기 위하여 EachMovie 추천 서비스를 실행하였다. 그 결과로 수집된 자료가 EachMovie data set이다. 72,916명의 사용자들이 1628개의 영화와 비디오에 대해서 2,811,983개의 평가값을 가지고 있고, 사용자의 중요한 정보가 제거되고 협력적 여과 알고리즘에 쉽게 적용될 수 있도록 가공하여 제공되어지

즉, rank k는 6과 코사인 거리는 0.8일 때에 알고리즘의 오차가 최소가 된다. 오차가 최소화된다는 것은 효과적으로 유사한 사용자를 구분하는 임계값으로 볼 수

는 전체적인 자료에서 3000명의 사용자가 1628개의 영화와 비디오에 대한 125510개의 rating만을 임의적으로 추출하여 사용하였다.

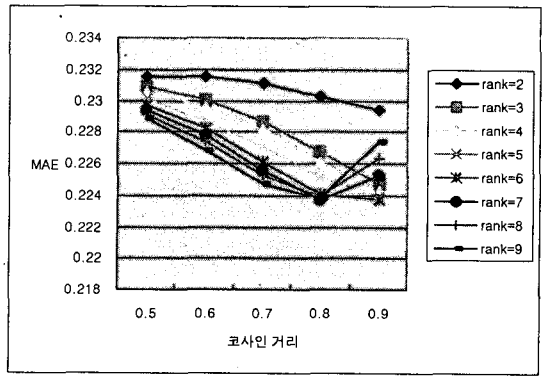
알고리즘에 대한 평가 방법의 기준은 여러 가지가 있을 수 있다. 그러나, 본 논문에서 제안하고 있는 방법은 사용자의 유사성을 측정하기 위한 방법이므로, 직접적인 성능 평가는 수행하기 어렵다. 그러나, 유사한 사용자들을 기반으로 사용자들의 선호도 값을 예측함으로 제안하는 방법의 성능에 대한 평가를 수행할 수 있다.

$$R_i = \frac{w_1 r_1 + w_2 r_2 + \ldots + w_n r_n}{w_1 + w_2 + \ldots + w_n} \qquad (7)$$

예상 값과 실제 값의 비교는 mean absolute error 방법을 사용하였다. mean absolute error 방법은 부호와 관계없이 각각 오차 크기의 평균이며, mean-squared error의 값이 작을수록 더 정확한 방법이다. mean absolute error는 다른 오차보다 더 큰 예측 오차의 결과를 강조하는 mean-squared error의 특성에 영향을 받지 않으며, 오차의 모든 크기는 그들의 오차량에 따라서 동일하게 취급되어진다.

$$\frac{|p_1 - a_1| + \ldots + |p_n - a_n|}{n} \qquad (8)$$

여기서, p_n은 선호도에 대한 예상된 값이고, a_n는 실제 사용자가 평가한 값이 된다. n은 a,p의 총개수이다.



[그림1] 코사인 거리와 계수 k에 따른 실험 결과

[그림1]는 0.5, 0.6, 0.8, 0.9의 코사인 거와 2~9사이의 rank k 값을 이용한 실험 결과이다. 그림에서 알 수 있듯이 k 값이 커질수록 오차가 작아 지는 것을 알 수 있다. 그러나 k값 6보다 커지면서 오차는 거의 작아지지 않고 있다. 한편, 코사인 거리는 k가 5일 때까지는 가까울수록 오차가 작아지고 k가 5보다 커지면 코사인 거리 0.9에서는 오차가 다시 증가 하는 것을 알 수 있다.

있다.

5.결론 및 향후 연구 과제

본 논문에서는 유사한 사용자를 찾기 위해서 충분한 수의 평가를 요구하는 명백하게 사용자의 평가를 비교하는 것 대신에 특징 가중치에 기초하여 사용자를 비교하는 방법을 사용하며, 사용하는 방법의 정확성을 높일 수 있는 임계값을 제안한다.

사용자를 비교하는 방법에서 중요한 임계값는 singular value의 개수 k와 사용자들의 유사성을 측정하는 기준의 코사인 거리이다. 본 논문에서는 실험에 의해서 k는 6과 코사인 거리는 0.8일 때 알고리즘의 성능이 최고가 되는 것을 알 수가 있었다.

본 논문에서는 적절한 임계값을 결정하기 위하여 실험적인 방법을 선택하였다. 따라서, 향후 연구 과제로서 다양한 실험 자료를 통해서 제안하고 있는 임계값의 유효성을 측정해보자 한다.

6. 참고 문헌

[1] John s. Breese, David Hecherman, and Carl Kadie., Empirical Analysis of Predictive Algorithms for Collaborative Filtering , IN Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence(UAI-98), pages 43-52, San Francisco, July 24-26 1998.

[2] Resnick, P. and Varian, H. R. , Recommender systems, CACM. 40(3), pp. 56-58, March 1997.

[3] Paul, R, Neophytos, I, Mitesh, S. Peter, B, John, R, GroupLens : an open architecture for collaborative filtering of netnews, In Proceedings of ACM CSCW'94 Conferece on Computer Supported Cooperative Work, pages 175-186, 1994.

[4] David Goldberg, David Nichols, Brian M. Oki, and Douglas Terry , Using collaborative filtering to weave an information tapestry , Communication of the ACM, 35(12) :61-70, December 1992.

[5] Deerwester, S., Dumais, S. T., Landauer, T. K., Furnas, G. W. and Harshman, R. A. , Indexing by Latent Semantic Analysis , Journal of the American Society for Information Science, 1990, 41(6), 391-407.

[6] Shardanand, Upendra., Social Information Filtering for Music Recommendation , S.M. Thesis, Program in Media Arts and Sciences, Massachusetts Institute of Technology, 1994.

[7] McJones, P.(1997) EachMovie collaborative filtering data set. DEC Systems Research Center. http://www.research.digital.com/SRC/eachmovie