임의 사상을 이용한 저차원 공간에서의 협력적 여과

정 준[0] 이필규
인하대학교 전자계산공학과
{jjeong[0], pkhree}@im.inha.or.kr

A Collaborative Filtering in a Lower-Dimensional Subspace using Random Projection

Jun Jeong[0] Phill_Kyu Rhee
Dept. of Computer Science and Engineering, Inha University.

요 약

추천 시스템에서 사용되고 있는 중요한 방법인 협력적 여과는 유사한 사용자들에 기초하여 그 사용자들이 선호하는 아이템을 교차 추천을 해주는 방법이다. 사용자들에 대한 정보는 아이템을 평가한 등급에 기초하며, 그 평가 등급 패턴이 유사한 사용자를 찾게 된다. 협력적 여과는 사용자와 정보의 증가에 따라서 성능이 저하되는 문제점을 가지고 있다. 이러한 문제점을 해결하기 위하여 SVD, PCA, LSI와 같은 차원 감소 방법이 제시되어 왔으나, 이러한 방법은 계산 비용이 크다는 단점을 가지고 있다. 따라서, 계산 비용이 적고, 정확성에 있어서도 충분히 정확한 임시 사상이 최근에 주목을 받고 있다. 본 논문에서는 임의 사상을 이용한 차원 감소 방법이 협력적 여과에 미치는 효과를 실험을 통하여 제시한다. 실험적으로, 임의 사상 방법은 협력적 여과에서 충분히 정확한 성능을 보였다.

1. 서론

인터넷을 발달은 정보 과잉이라는 현상을 불러 일으켰으며, 이러한 문제를 해결하기 위한 여러 방법 중에서 협력적 여과는 대표적인 해결책으로 제시되어 왔다[3]. 추천 시스템에서 사용되고 있는 중요한 방법인 협력적 여과는 유사한 사용자들에 기초하여 그 사용자들이 선호하는 아이템을 교차 추천을 해주는 방법이다. 사용자들에 대한 정보는 아이템을 평가한 등급에 기초하며, 그 평가 등급 패턴이 유사한 사용자를 찾게 된다[6].

협력적 여과는 사용자와 정보의 증가에 따라서 성능이 저하되는 문제점을 가지고 있다. 이러한 문제점을 해결하기 위하여 SVD, PCA, LSI와 같은 차원 감소 방법이 제시되어 왔으나, 이러한 방법은 계산 비용이 크다는 단점을 가지고 있다[2]. 따라서, 계산 비용이 적고, 정확성에 있어서도 충분히 정확한 임시 사상이 최근에 주목을 받고 있다. 본 논문에서는 임의 사상을 이용한 차원 감소 방법이 협력적 여과에 미치는 효과를 실험을 통하여 제시한다.

2. 관련 연구

Dhruv Gupta [1]는 PCA와 군집화에 기초한 협력적 여과 시스템인 Jester 2.0를 소개하였다. 이 논문에서는 실험을 통하여 인접 이웃 방법보다 효과적임을 보였다.

Badrul M. Sarwar [7]는 SVD를 이용한 협력적 여과 방법을 제시하였다. 이 논문에서는 명백한 평가 등급에 기반한 자료에서 선호도 예측과 실제 전자상거래 사이트에서 수집된 자료를 이용한 Top-N 방법을 실험을 통하여 차원 감소 방법의 효과성을 보였다.

Michael H. Pryor [5]는 문서 자료에서 SVD를 이용한 대한 협력적 여과 방법을 제시하였다. 이 방법은 명백한 평가등급뿐만 아니라 웹사이트 방문 정보를 이용하였다.

3. 임의 사상을 이용한 협력적 여과

3.1 임의 사상(Random Projection)

임의 사상 [2]에서는 원래의 d-차원 자료가 행이 단위 길이를 가지는 임의 k×d 행렬인 R을 이용하여 k-차원의 하위 공간으로 사상된다(단, k << d). $X_{d\times N}$이 N개의 d-차원 관찰의 본래의 집합일 때, 행렬표기법을 사용하면, 다음과 같다.

$$X^{RP}_{k\times d} = R_{k\times d}\, X_{d\times N} \qquad (1)$$

식(1)은 자료를 더 낮은 k-차원 공간으로 사상한다. 임의 사상의 주요한 개념은 Johnson-Lindenstrauss lemma[2]에서 기인한다. 만약 한 벡터 공간에서 점이 적당히 높은 차원의 임의로 선택된 하위공간에 사상된다면, 점간의 거리는 근사적으로 유지된다.

임의 사상은 계산상으로 간단하다. 임의 사상 행렬 R을 구성하고 d×N 자료 행렬 X를 k 차원에 사상하는 것은 O(dkN)이다.

엄격히 말해서, R은 일반적으로 직교하지 않기 때문에, 식(1)은 사상이 아니다. 식(1)과 같은 선형 매핑은 R이 직교하지 않는다면 자료에 중대한 왜곡을 일으킬 수 있다. 그러나, R을 직교화하는 것은 계산상으로 아주 비용이 크다.

임의 사상의 성능을 다른 차원 감소 방법과 비교 할 때, 두 벡터의 유사성이 차원 감소에 의해서 왜곡되는

정도를 고려한다.

임의 행렬 R의 선택은 주요한 문제점이다. R의 요소 r_{ij}는 보통 Gaussain 분포를 따르나, 반드시 그럴 필요는 없다. 최근에 Gaussain 분포는 다음과 같은 더욱 간단한 분포로 대치될 수 있음을 보였다[2].

$$r_{ij} = \sqrt{3} \times \begin{cases} +1 & 1/6 \text{ 확률} \\ 0 & 2/3 \text{ 확률} \\ -1 & 1/6 \text{ 확률} \end{cases} \quad (2)$$

3.2 임의 사상을 이용한 협력적 여과

본 논문에서 제안하고자 하는 방법은 사용자의 평가 행렬 X를 임의 행렬 R에 의해서 임의 저 차원공간으로 사상하여 사용자간의 유사성을 계산한다. 3.1절에서 설명한 임의 사상을 이용하여 전체적인 과정을 표현하면 다음과 같다.

N명의 사용자가 d개의 아이템을 평가한 자료는 $U_{d\times N}$으로 표현한다. 사용자 평가 행렬 $U_{d\times N}$의 요소 u_{ij}는 j 사용자가 i 아이템에 평가한 값을 의미한다. $R_{k\times N}$은 N명의 사용자들을 k차원으로 사상하는 임의 행렬이다. $U^{RP}_{k\times N}$는 새롭게 생성된 저차원의 사용자 평가 행렬이 된다.

$$U^{RP}_{k\times N} = R_{k\times N}\, U_{d\times N} \quad (3)$$

임의 행렬 $R_{k\times N}$로서 Gaussain 분포와 식(2)에서와 같은 희소 임의 행렬을 사용하였다. Gaussain 분포를 이용한 것을 RP라고 하고, 희소 임의 행렬을 이용한 것을 SRP라고 하자.

$U^{RP}_{k\times N}$에 기초하여 사용자들의 유사성은 식(4)와 같이 코사인 거리로 구하여진다.

$$S_{ij} = \frac{u_{ki} \bullet u_{kj}}{\| u_{ki} \| \, u_{kj} \|} \quad (4)$$

임의의 임계값 T보다 큰 코사인 거리를 갖는 사용자들을 유사한 사용자 N_u로 간주하고, 유사한 사용자들간에 가중치를 구한다.

$$W_{iu} = \left(\frac{S_{iu} - T}{1 - T} \right)^2 \quad (5)$$

마지막으로, 사용자 j에 대한 아이템 i의 선호도는 모든 유사한 사용자들의 가중치가 부여된 평가 값의 평균으로 구해진다.

$$P_{ij} = \frac{\sum_{d}^{N_u} w_{jd} \times U_{id}}{\sum_{d}^{N_u} w_{jd}} \quad (6)$$

4. 실험

4.1 실험방법

차원 감소를 이용한 방법은 감소될 차원수가 중요하다. 따라서, 최적의 k를 결정하기 위하여 k의 값을 변화시켜 실험한다. 사용자의 유사성에 대한 임계값을 결정하기 위한 실험도 수행한다. 또한, Gaussain 분포와 식(2)에 따른 분포를 가지는 임의 행렬에 따른 영향에 대한 실험을 수행한다. 마지막으로 대표적인 협력적 여과 방법인 상관계수를 이용한 방법과 성능을 비교한다.

4.2 실험자료

본 논문에서 제안된 방법을 실험하기 위한 실험 자료는 DEC Systems Research Center에서 제공하는 EachMovie collaborative filtering data set을 사용하였다[4]. 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 제공되어 있다. 사용자의 평가 정보는 {0.0,0.2,0.4,0.6,0.8,1.0}과 같이 6단계로 이루어져 있다.

4.3 평가방법

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

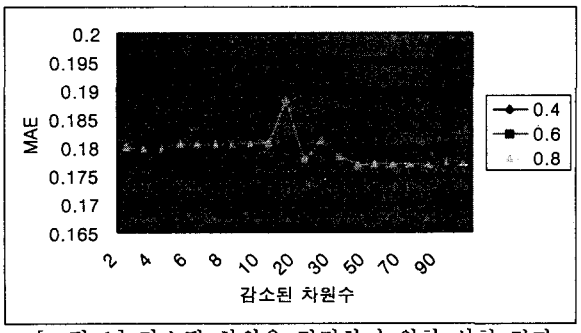
$$MAE = \frac{\sum |p - a|}{n} \quad (7)$$

p : 선호도 예측값
a : 실제 평가값

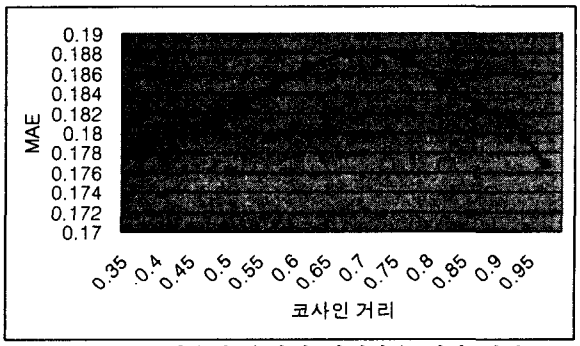
예측된 선호도 값과 실제 값의 비교는 Mean Absolute Error(MAE)방법을 사용하였다. MAE 방법은 부호와 관계없이 각각 오차 크기의 평균이며, MAE의 값이 작을수록 더 정확한 방법이다. MAE는 다른 오차보다 더 큰 예측 오차의 결과를 강조하는 MSE의 특성에 영향을 받지 않으며, 오차의 모든 크기는 그들의 오차량에 따라서 동일하게 취급되어진다.

4.4 실험 결과

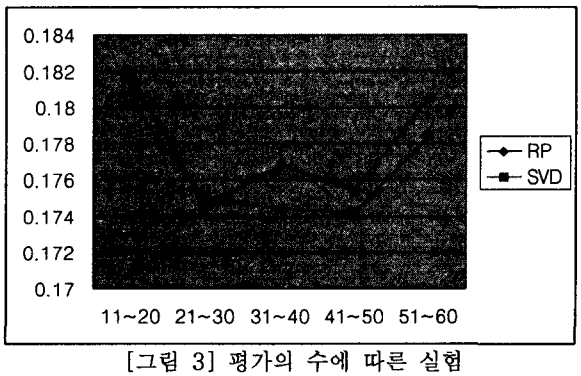
[그림 1]은 최적의 감소될 차원의 수를 구하기 위한 실험 결과이다. 그림에서 보여지듯이 10차원이하에서 좋은 성능을 보여주고 있다. [그림 2]는 감소되는 차원의 수를 8로 고정하고 코사인 거리를 0.4에서 0.95까지 변경하면서 성능을 측정한 실험결과이다. 코사인 거리가 0.95일 때 좋은 성능을 보여주고 있다. [그림 3]은 평가의 수에 따른 정확도를 SVD 방법과 비교한 실험 결과이다. 전체적으로 임의사상 방법이 SVD를 이용한 방법만큼 정확한 성능을 보여 주고 있다.



[그림 1] 감소될 차원을 결정하기 위한 실험 결과



[그림 2] 사용자 유사성 임계값을 위한 실험



[그림 3] 평가의 수에 따른 실험

5. 결론

협력적 여과는 사용자와 정보의 증가에 따라서 성능이 저하되는 문제점을 가지고 있다. 이러한 문제점을 해결하기 위하여 SVD, PCA, LSI와 같은 차원 감소 방법이 제시되어 왔으나, 이러한 방법은 계산 비용이 크다는 단점을 가지고 있다. 따라서, 계산 비용이 적고, 정확성에 있어서도 충분히 정확한 임시 사상이 최근에 주목을 받고 있다. 본 논문에서는 임의 사상을 이용한 차원 감소 방법이 협력적 여과에 미치는 효과를 실험을 통하여 제시하였다. 실험적으로, 임의 사상 방법은 협력적 여과에서 충분히 정확한 성능을 보였다.

향후 과제로 임의 사상에서 임의행렬의 분포에 따라 미치는 영향을 연구할 필요가 있다.

6. 참고문헌

[1] Dhruv Gupta, Mark Digiovanni, Hiro Narita, Ken Goldberg Jester 2.0 : Evaluation of a New linea Time Collaborative Filering Algorithm Proceedings on the 22nd annual international ACM SIGIR conference on Research and development in information retrieval August 15 - 19, 1999, Berkeley, CA USA.

[2] Ella Bingham and Heikki Mannila, Random projection in dimensionality reduction: applications to image and text data, Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2001), August 26-29, 2001, San Francisco, CA, USA, pp. 245-250.

[3] J. Ben Schafer, Recommender Systems in E-Commerce, Proceedings of the ACM Conference on Electronic Commerce, 1999, Pages 158 - 166.

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

[5] Michael H. Pryor, The Effects of Singular Value Decomposition on Collaborative Filtering, Dartmouth College Technical Report PCS-TR98-338, June 1998.

[6] 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.

[7] Sarwar, B. M., Karypis, G., Konstan, J. A., and Riedl, J. Application of Dimensionality Reduction in Recommender System--A Case Study. ACM WebKDD 2000 Web Mining for E-Commerce Workshop.

[8]정준, 이필규, Singular Value Decomposition을 이용한 협력적 여과의 임계값, 한국정보과학회 추계학술대회 2000년 10월.

[9]정준, 정대진, 김용환, 이필규, 협력적 여과에서 평가 행렬의 희소성 문제를 해결하기 위한 Singular Value Decomposition의 적용 방법에 관한 연구 , 한국지능정보시스템학회 춘계학술대회 2000년 6월.