PCA 및 적응형 k-NN을 이용한 유머문서의 추천

Humor Document Recommendation using Adaptive k-NN with PCA

이종우, 장병탁
서울대학교 컴퓨터공학부

Jongwoo Lee, Byung-Tak Zhang
School of Computer Science and Engineering, Seoul National University
{jwlee, btzhang}@scai.snu.ac.kr

요 약

우리는 인터넷을 통한 사용자의 선호도(preference)를 분석하고 협력적 여과 기술을 학습하여 유머문서를 추천하는 MrHumor 시스템을 구축하였다. MrHumor에서는 사용자집합이 유머문서 집합에 대하여 보여준 등급매김값을 토대로 사용집합의 벡터공간(vector space)를 설정하고 노이즈에 강하면서 효율적인 학습을 위해 선형 PCA를 이용하여 축소된 2차원 공간상에서 유머문서의 통계적 특성을 반영하여 적응형 k-NN으로 지엽성을 적절히 조절하여 새로운 문서에 대한 선호도를 추정하게 된다.

ABSTRACT

We constructed *MrHumor* System which analyzes the preferences of users on internet, learning collaborative filtering technique, and recommends unrated humor documents to the users. In *MuHumor*, the vector space of rating vectors of users on humor documents is defined and transformed to 2-dimensional space by PCA to be immune to noises and to operate effectively, where the preferences of new documents are predicted using adaptive k-NN with statistical characteristics of humor documents.

1. 서론

정보추천(information recommendation), 혹은 정보 여과(information filtering)란 특정 정보 수요자에게 높은 선호도를 보일 만한 정보를 가려서 능동적으로 제공하여 주는 기술이다. 정보 여과에서 쓰이는 학습방법은 추천하고자 하는 데이터의 특성에 따라 인구통계적(demographic)방법, 내용적 여과(contents filtering), 협력적 여과(collaborative filtering) 등이 있다.

인구통계적(demographic) 방법은 개인의 보편적 프로파일(profile)에 대한 정보와 추천하는 아이템(item)의 특성과의 관계를 학습하는데 프로파일이 표현하지 못하는 사용자의 다양성을 반영하지 못한다는 단점을 지닌다. 내용적 여과 방식은 사용자가의 아이템에 대한 선호도로부터 선호하는 아이템의 특성을 학습하는 방법으로 개인의 다양한 선호도를 반영할 수 있지만 아이템의 특성이 복잡하여 학습하기 어려운 경우가 많다. 이에 비해 협력적 여과 방식은 다수의 사용자 집단으로부터 비슷하거나 다른 선호성향을 가진 타사용자의 선호정보를 이용하여 추천한다. 이 방법은 아이템의 내용이나 특성을 전혀 고려하지 않아 내용적 여과의 단점을 극복하는 반면 공통 등급매김 정보가 부족한 경우거나 사용자들의 선호도가 서로 관련이 적을 경우 좋은 성능을 보이지 않는다.

*MrHumor*는 협력적 여과 방법을 적용한 유머문서 추천시스템이다. 협력적 여과 방법은 학습 알고리즘와 추천에 이용하는 정보의 범위 등에 따라 상관계수를 이용하거나[Resnick, 1994] [Shardanand, 1995], 사용자집단을 클러스터링하거나[Gupta, 1998], 베이지안학습[Chien, 1998], 신경망[Pazzani, 1999] 학습, 기호학습[Lin, 2000]등의 방법이 있다. 이러한 방법들은 아이템의 특성에 따라 정보이용의 지협성(locality)을 변화시키지 않고 추천에 이용해서 아이템 의존적이지 않고 노이즈에 약해서 성능의 기복이 심하다.

이 논문에서는 유사성 측정의 노이즈를 줄이고 효율성을 높이기 위하여 등급매김에 의한 사용자 특성공간을 PCA를 통하여 축소하고 아이템의 통계적 특성에 따라 이용하는 정보의 지협성을 변화시키는 알고리즘을 제시하고 타방법과 비교실험을 하고자 한다.

II. 본론

협력적 여과 방법에서 이용하는 데이터는 다음의 공통 등급매김 벡터, $\mathbf{R}_u$ 들의 집합, R로 표시될 수 있다.

$$\mathbf{R}_u = (r_{u1}, r_{u2}, \mathrm{K}, r_{uN})$$

$$\mathbf{R} = \{\mathbf{R}_1, \mathbf{R}_2, \mathrm{K}, \mathbf{R}_U\}$$

윗 식에서 r_{ui}는 사용자 u의 문서 i에대한 등급매김이다.

협력적 여과가 좋은 추천 성능을 보이기 위해서는 R_i와 R_j의 임의의 k번째 요소, r_{ik}와 r_{jk} 없앤 두 개의 $N-1$ 차원 벡터간의 거리가 가까울 때 $|r_{ik} - r_{jk}|$도 작아야 한다.

이러한 조건을 표 표현하는 값 중 하나는 두 확률변수의 상관성을 측정하는 피어슨 상관계수(Pearson correlation)이다. 이 값들을 정규화시킨 후 확률로 이용하여 모든 사용자들의 등급매김의 기대값을 이용하는 방법은 실제로는 상관이 별로 없는 새로운 아이템에 대해서도 이전에 학습된 상관계수 값에 의해 과하게 가중되거나 그 반대의 경우가 발생할 수 있게 된다. 상관계수를 이외에도 유사성향을 가진 타사용자의 등급매

김만을 반영하는 방법으로 $|R_i - R_j|$의 값이 일정값이하인 사용자나 최소인 사용자의 등급매김만을 반영할 수도 있다. 이 방법 반대로 다수의 사람들이 보편적인 선호도를 보이는 문서에 대해 특정 타사용자의 선호도만을 반영하여 추정하므로 확률적으로 오류를 포함할 가능성이 높다. 또한 공통 등급매김된 모든 문서에 대해 공간상에서의 거리는 차원이 커질수록 작은 노이즈에도 공간상에서의 거리가 크게 변하여 유사도가 높은 사용자의 등급매김이 반영되지 않을 확률이 높다.

MrHumor에서는 이러한 문제점들을 극복하기 위해서 R_U공간을 PCA(Principal Component Analysis)를 통하여 2차원으로 줄이고 r_{ui}의 u에 대한 분산값을 k-NN에서 k값을 정하는데 이용하여 등급매김할 사용자의 크기를 조절하였다. 알고리즘 1.은 사용자 등급매김 벡터의 2차원으로의 변환과 k값을 조절하는 방법을 보여주고 있다.

Ⅲ. 실험

실험에서는 47명의 사용자와 80개의 유머문서에 대하여 등급매김되어진 3760개의 데이타를 가지고 다음의 네가지 방법을 이용하여 학습을 하고 동일한 47명의 사용자에 와 20개의 새로운 테스트 유머문서로 실제 사용자의 94개의 등급매김과 여러 학습을 통해 추정한 등급매김과의 MSE을 성능평가에 이용하였다.

1) 피어슨 상관계수를 이용한 기법

모든 사용자간의 상관계수를 계산하여 이를 정규화시켜서 확률로 이용하여 모든 타사용자의 등급매김값의 기대값으로 특정 사용자에 대해 선호도를 추정한다.

2) 축소 공간에서의 NN기법

알고리즘 1. 적응형 k-NN 알고리즘

입력: $\mathbf{r}_u = (r_{u1}, r_{u2}, \mathrm{K}, r_{uN})$

$\mathbf{R} = \{\mathbf{r}_1, \mathbf{r}_2, \mathrm{K}, \mathbf{r}_U\}$

1. 등급매김 벡터의 표준정규화

$$\mathbf{z}_u = \left(\frac{r_{u_1} - \mathrm{Exp}(r_u)}{\sigma(r_u)}, \mathrm{K}, \frac{r_{uN} - \mathrm{Exp}(r_u)}{\sigma(r_u)} \right)$$

2. Z에 대해서 SVD(Singular Value Decomposition)을 행한다.

$$\mathbf{Z} = \mathbf{U}\Sigma\mathbf{V}^T$$

3. 가장 큰 고유값을 가지는 고유벡터 2개로 이루어진 V' 을 구한다.

4. 2차원 공간상으로의 선형변환을 한다.

$Y = \{y \mid y = zV', z \in Z\}$

5. 아랫식을 만족하는 K_{iu} 를 이용하여 모든 u, i에 대하여 집합 $K=\{(K_{iu}, \sigma_i^2)\}$를 구한다. ($KNN(K,u)$는 y의 공간상에서 사용자 u와 가장 가까운 K개의 사용자 집합이다.)

$$K_{iu} = \arg\min_K \left(\frac{\sum_{j \in KNN(K,u)} r_{ji}}{K} - r_{ui} \right)$$

$$\sigma_i^2 = \underset{\forall u}{VAR}(r_{ui})$$

6. 5단계에서 구한 K로 선형회귀를 통해 K_{iu}와 σ_i^2의 선형관계를 구한다.

PCA를 통해 등급매김으로 구성되는 사용자의 공간을 2차원으로 축소한 후 특정 사용자와 가장 작은 거리를 가지는 사용자의 등급매김값만을 추정에 이용한다.

3) 축소 공간에서의 사용자 Cluster구성

축소 2차원 공간내에서 8개의 계층적 클러스터를 구성한다. 선호도 추정은 같은 클러스터내의 멤버의 등급매김의 평균을 이용한다.

4) 축소 공간에서 문서의 등급매김 분산에 따른 적응형 k-NN기법

축소 2차원 공간상에서 가장 가까운 K개의 타사용자의 등급매김값의 평균으로 추정

을 한다. K는 알고리즘 1.에서처럼 테스트 데이타의 벡터집합 K를 통해 선형회귀로 얻어진 1차 함수를 통해 얻는다.

그림 1. 은 2), 3), 4) 방법에서 공통적으로 이용하는 PCA에 의해 47명의 사용자들의 등급매김 벡터들을 2차원으로 변환시킨 결과를 보여준다.

그림 1. PCA을 통해 2차원으로의 등급매김벡터의 변환

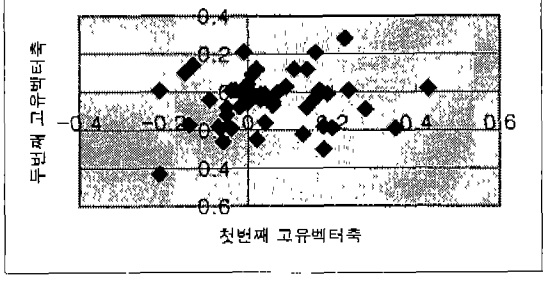
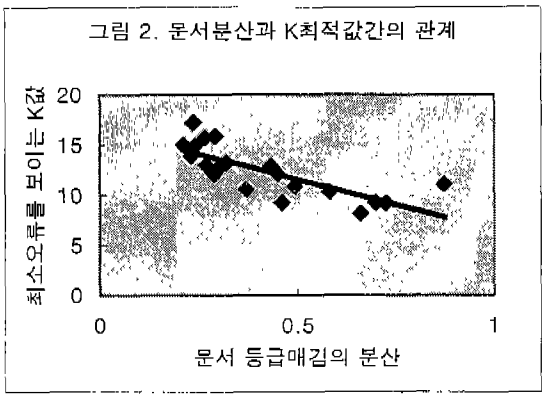


그림 2. 문서분산과 K최적값간의 관계



알고리즘 1.의 5, 6 단계를 통해 얻은 K와 선형회기의 결과는 그림 2.에서 보이고 있으며 등급매김 분산 V와 K와의 관계는 다음의 식으로 주어진다.

$$K = -10.1V + 16.6 \qquad (1)$$

표 1.에서는 테스트 데이터에 대해 4가지 학습방법에 의한 추정 등급매김과 실제 등급매김와의 MSE값을 나열한다.

학습 방법	상관 계수	NN	클러스터링	k-NN
MSE	0.583	0.899	0.539	0.421

표 1. 학습방법에 따른 추천 오류

. 결과

본 논문에서는 유머문서 추천을 4가지 협력적 여과 학습방법을 통해 실험하였다. 기존에 이용된 상관계수, NN, 클러스터링을 통해 추천하는 방법은 추천하고자하는 문서의 등급매김 분산값을 고려하여 사용자의 크기를 변화시켜 등급매김값을 이용하는 적응형 k-NN학습에 비해 높은 오류를 보인다.

감사의 글

본 연구는 첨단정보기술 연구센터(AITRC)를 통하여 과학재단의 지원을 받았음

V. 참고문헌

[Resnick, 1994] P. Resnick and N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl, "GroupLens: An open architecture for collaborative filtering," *Proc. ACM 1994 Conf. on Computer Supported Cooperative Work*, pp. 175-180.

[Gupta et al., 1999] D. Gupta, M. DiGiovanni, H. Narita, and K. Goldberg, "Jester2.0:Evaluation of a new linear time collaborative filtering algorithm applied to jokes," *Poster Session and Demonstration, 22nd Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pp. 191-192.

[Chien, 1998] Y. Chien, "A Bayesian model for collaborative filtering," *Proc. of Uncertainty in Artificail Intelligence*, Morgan Kaufmann.

[Billsus and Pazzani, 1999] D. Billsus and Michael J. Pazzani, "Learning Collaborative Information Filters," *In Shavlik, J., ed., Machine Learning: Proc. of the Fifteenth International Conference, Morgan Kaufmann Publishers*,

[Lin,2000] W. Lin, S. A. Alvarez, C. Ruiz, "Collaborative Recommendation via Adaptive Association Rule Mining," *Work shop on Web Mining for E-Commerce*, pp. 35-41.

[Shardanand, 1995] U. Shardanand and P. Maes, "Social Information Filtering: *Algorithms for Automating Word of Mouth", Proc. ACM CHI95*, pp.210-217.