유머문서 추천을 위한 기계학습 기법

이종우 장병탁

서 울 대 학 교 컴 퓨 터 공 학 부

{jwlee, btzhang}@scai.snu.ac.kr

A Learning Model for Recommendation of Humor Documents

Jongwoo Lee Byoung-Tak Zhang

School of Computer Science and Engineering, Seoul National University

요 약

인터넷을 통한 사용자의 선호도를 분석하고 협력적 여과 및 내용기반 여과 기술을 결합 이용하여 유머문서를 추천하는 *MrHumor* 시스템을 구축하였다. 유머문서 추천 기술은 다양한 아이템에 대한 여과 및 추천 기술로 확장되어 인터넷을 통한 과다 정보 시대에 필요한 소프트봇 혹은 지능형 에이전트 기술에 적용될 수 있다. *HrHumor* 추천시스템은 적응형 학습 시스템으로서 새로운 사용자의 선호도에 대한 학습량과 추천시기에 따라 이용할 추천방식이 다른 성능을 보이는데 여러 가지 상황에서도 적절한 동작을 보이기 위하여 *MrHumor*에서는 은닉변수 모델을 이용하여 사용자의 인구통계적 정보와 문서의 내용적 특징간의 관계를 학습하여 초기 추천을 행하고 SVM을 이용하여 개인의 선호도를 학습한 내용기반의 여과와 적응형 k-NN모델을 이용한 협력적 여과를 결합하여 추천을 수행한다. 제안된 방식에 의한 추천 성능은 3 방식이 각각 이용된 경우에 비해 안정적이고 높은 예측 정확도를 보인다.

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, 내용기반 여과 (contents-based filtering), 협력적 여과 (collaborative filtering) 등이 있다[1].

내용적 여과 방식은 사용자의 아이템(item)에 대해 제시하는 선호도로부터 선호하는 아이템의 특성을 학습하는 방법으로 개인의 다양한 선호도를 반영할 수 있지만 아이템의 특성이 복잡한 경우 학습하기 어려운 경우가 많다. 인구통계적(demographic) 방법은 개인의 보편적 프로파일(profile)에 대한 정보와 추천하는 아이템의 특성과의 관계를 학습하는데 아이템의 특성이 학습요소로 취급되므로 내용기반 여과방식의 일종이라고 할 수도 있다. 프로파일은 사용자의 다양성을 모두 반영하지는 못한다는 단점을 지니지만 새로운 사용자에 대하여 아무런 선호도 데이터가 마련되지 않을 때 이용할 수 있는 유일한 방법이다. 이에 비해 협력적 여과 방식은 다수의 사용자 집단으로부터 비슷하거나 다른 선호 성향을 가진 타사용자의 선호정보를 이용하여 추천한다. 협력적 여과 방법은 아이템의 내용이나 특성을 전혀 고려하지 않아도 되므로 내용적 여과의 단점을 극복하는 반면 공통 평가 정보가 부족한 경우거나 사용자들의 선호도가 서로 관련이 적을 경우 좋은 성능을 보이지 못한다[1].

MrHumor(http://MrHumor.snu.ac.kr)는 이러한 3가지 방식의 장단점을 고려하여 사용자에 대한 정보가 부족한 초기에는 인구통계적 방식을 학습할 데이터가 증가하면서 내용기반 여과방식과 협력적 여과 방법을 사용자 선호 정보의 크기에 따라 결합 적용한 유머문서 추천시스템이다. *MrHumor*는 인구통계적 정보를 학습하기 위해서 은닉변수모델(latent variable model)의 일종인 PLSA (probabilistic latent semantic analysis)를 통하여 사용자의 인구통계적 프로파일과 문서의 특징간의 관계를 학습하며, 내용적 여과 방식의 학습을 위해 각 사용자가 평가한 문서의 특성을 SVM (support vector machine)으로 선호 문서를 분류하였다. 또한 협력적 여과 방식의 학습을 위하여 각 모든 사용자의 평가벡터(rating vector)를 PCA (principal component analysis)를 통하여 2차원 공간상으로 변환시킨 후에 추천할 문서의 평가 분산값(variance)에 적절히 k값을 조절하여 k-NN 방식을 적용하였다.

2. 본론

PLSA방식을 이용한 사용자의 인구통계적 프로파일과 문서 특성과의 관계는 그림-1 에서 제시된 바와 같이 은닉변수(latent variable)에 의한 조건부 확률분포를 따른다고 가정한다. 사용자의 프로파일은 카테고리(category)정보로서 벡터값이 아닌 이산적인 값으로 변환하여 이용하며 문서 특성은 문서에 출현하는 단어(term)와 평균 문단 길이(sentence length or *sl*), 문서의 길이(document length or *dl*) 등으로 이루어진 벡터, $\mathbf{t}$로 정의된다.

학습은 EM (expectation maximization) 알고리즘을 통해 식(1)~식(4)로 주어지는 조건부 확률분포를 추정함으로써 이루어진다[3]. 식에서 $\mathbf{t}$는 문서의 특성을 나타내는 단어 벡터이고 $\mathbf{c}$는 사용자의 인구통계적 범주를 표현하는 기본벡터 (elementary vector)이며 z_k 들은 $\mathbf{t}$와 $\mathbf{c}$간의 관계를 학습하기 위한 은닉변수이며 $n(\cdot)$는 데이터의 크기를 나타낸다.

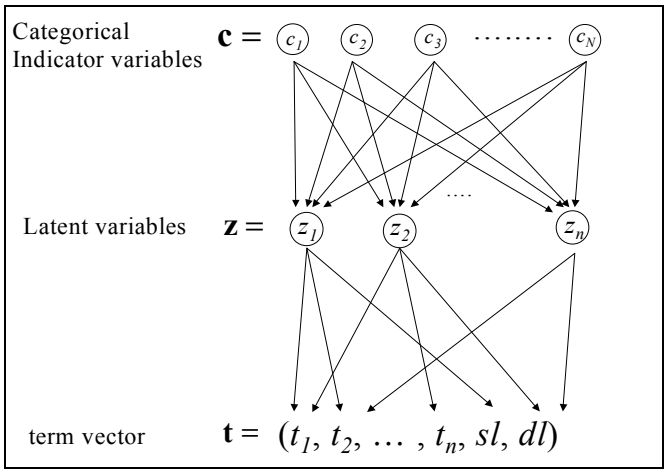


그림-1 인구통계적 여과에 적용된 PLSA 모델

$$P(z_k \mid \mathbf{t},\mathbf{c}) = \frac{P(z_k)P(\mathbf{c} \mid z_k)P(\mathbf{t} \mid z_k)}{\sum_{\mathbf{z}_k'} P(z_k')P(\mathbf{c} \mid z_k')P(\mathbf{w} \mid z_k')} \quad (1)$$

$$P(\mathbf{c} \mid z_k) = \frac{\sum_{\mathbf{t}} n(\mathbf{c},\mathbf{t})P(z_k \mid \mathbf{c},\mathbf{t})}{\sum_{\mathbf{c}',\mathbf{t}} n(\mathbf{c}',\mathbf{t})P(z_k \mid \mathbf{c}',\mathbf{t})} \quad (2)$$

$$P(\mathbf{t} \mid z_k) = \frac{\sum_{\mathbf{c}} n(\mathbf{c},\mathbf{t})P(z_k \mid \mathbf{c},\mathbf{t})}{\sum_{\mathbf{c},\mathbf{t}'} n(\mathbf{c},\mathbf{t}')P(z_k \mid \mathbf{c},\mathbf{t}')} \quad (3)$$

$$P(z_k) = \frac{1}{R}\sum_{\mathbf{c},\mathbf{t}} n(\mathbf{c},\mathbf{t})P(z_k \mid \mathbf{c},\mathbf{t}), R = \sum_{\mathbf{c},\mathbf{t}} n(\mathbf{c},\mathbf{t}) \quad (4)$$

내용적 여과 방식의 구현을 위해 사용된 SVM의 추천 모델은 그림-2에 나타나 있다. 이것은 SVM이 2진(선호함, 선호하지 않음) 분류(classification) 문제에 적용된 것이고 하나의 SVM이 사용자 한명의 프로파일을 구성하게 된다. 그림에서 x_i들은 단어벡터의 각 구성요소를 나타내며 y는 2진 분류의 결과값이 된다.

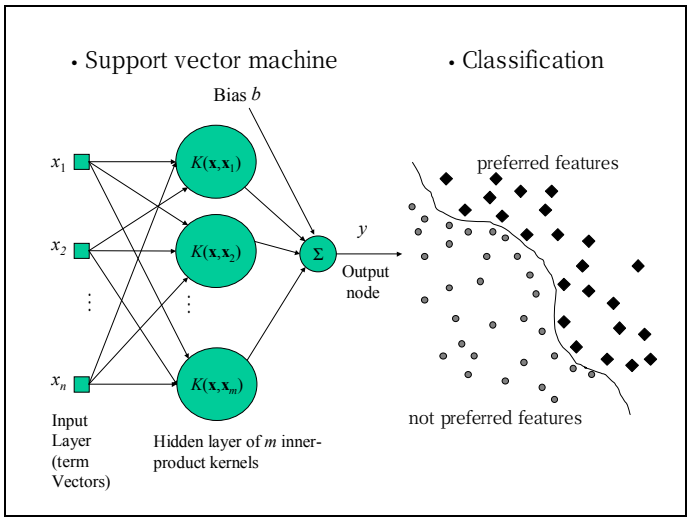


그림-2 특성벡터 분류에 이용된 SVM 모델

SVM 학습은 식(6)의 제약조건하에서 식(5)를 최대화시키는 α_i 값을 이용하여 식(7)로 SVM 의 기저함수(basis function)의 선형결합에 이용되는 가중치 벡터 $\mathbf{w}$ 를 결정한다[4].

$$Q(\alpha) = \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{n}\boldsymbol{\alpha}_i\boldsymbol{\alpha}_j\mathbf{y}_i\mathbf{y}_j H(\mathbf{x}_i,\mathbf{x}_j) \quad (5)$$

$$\sum_{i=1}^{n} y_i\alpha_i = 0, \qquad 0 \le \alpha_i \le \frac{C}{n}, i = 1,\ldots,n \quad (6)$$

$$\mathbf{w} = \sum_{i=1}^{n}\alpha_i y_i \mathbf{x}_i, \qquad \alpha_i \ge 0, i = 1,\ldots,n \quad (7)$$

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

$$\mathbf{r}_u = (r_{u1}, r_{u2}, \ldots, r_{uN})$$
$$\mathbf{R} = \{\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_U)$$

윗 식에서 r_{ui} 는 사용자 u 의 문서 i 에대한 평가이다. r_u 는 사용자 u 의 선호특성을 나타내는 벡터가 되고 $|r_u - r_v|$ 값이 작은 v 사용자와 선호도가 비슷하다고 할 수 있다. *MrHumor*에서는 오류에 민감한 다차원 $\mathbf{r}_u$ 공간상에서의 거리를 이용하지 않고 $\mathbf{R}$ 을 정규화시킨 후 PCA (principal component analysis)를 이용한 공간 축소를 통한 2차원 공간상에서의 거리를 사용자간의 유사도로 이용하였다. 축소된 2차원 공간상에서 가장 가까운 k명의 평균 평가값을 추천 예측값으로 이용하는데, k는 문서의 평가값에 대한 분산값에 따라 결정하였다. 이상의 적응형 k-NN방식을 이용한 협력적 여과방식은 표-1에 정리되어 있다.

1. 등급매김 벡터의 표준정규화
$$\mathbf{z}_u = \left(\frac{r_{u_1} - \mathrm{E}(r_u)}{\sigma(r_u)}, \ldots, \frac{r_{uN} - \mathrm{E}(r_u)}{\sigma(r_u)}\right)$$
2. $\mathbf{Z}$ 에 대해서 SVD(Singular Value Decomposition)을 행한다.
$$\mathbf{Z} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T$$
3. 가장 큰 고유값을 가지는 고유벡터 2 개로 이루어진 $\mathbf{V}'$을 구한다.
4. 2 차원 공간상으로의 선형변환을 한다.
$$\mathbf{Y} = \{\mathbf{y} \mid \mathbf{y} = \mathbf{z}\mathbf{V}', \mathbf{z} \in \mathbf{Z}\}$$
5. 아랫식을 만족하는 k_{iu} 를 이용하여 모든 u, i 에 대하여 집합 $K = \{(k_{iu}, \sigma_i^2)\}$ 를 구한다. ($KNN(k,u)$는 $\mathbf{y}$의 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. 5 단계에서 구한 K로 선형회귀(linear regression)를 통해 k_{iu}와 σ_i^2의 선형관계를 구한다.

표-1 협력적 여과에 이용된 적응형 k-NN

3. 실험

실험은 58명의 사용자와 100개의 유머문서에 대하여 평가된 5800개의 데이터에 대하여 행해졌다. 데이터 희소 문제로 인하여 인구통계적 여과모델의 학습을 위해서 5800개의 데이터가 모두 이용되었고 이는 또한 테스트 데이터로도 이용되었다. 하지만 다른 두 여과방식의 학습을 위해서는 추천해 나가는 과정에서 얻은 평가값만을 가지고 학습하였다. 100개의 문서는 차례로 추천되어졌고 추천된 직후 사용자는 이를 평가하도록 되어있다.

사용자의 평가는 1부터 10까지의 정수로 이루어지지만 협력적 여과를 제외한 나머지 학습방법은 2진 분류에 의한 학습이 이루어지므로 전체 평가 데이터의 평균값(mean value)을 기준으로 선호와 비선호를 양분하였다.

각 모델은 모든 사용자에 대한 옳은 예측평가를 취한 횟수의 비율(정분류율)로써 성능이 평가되었다. 인구통계적 여과 모델을 제외하고는 모델이 학습되지 않으므로 예측평가값이 없는 경우 0.5 의 가중치를 곱하여 횟수를 반영하였다. 그림-3 은 유머문서를 차례로 추천하면서 정분류율의 변화를 그래프로 표현한 것이다.

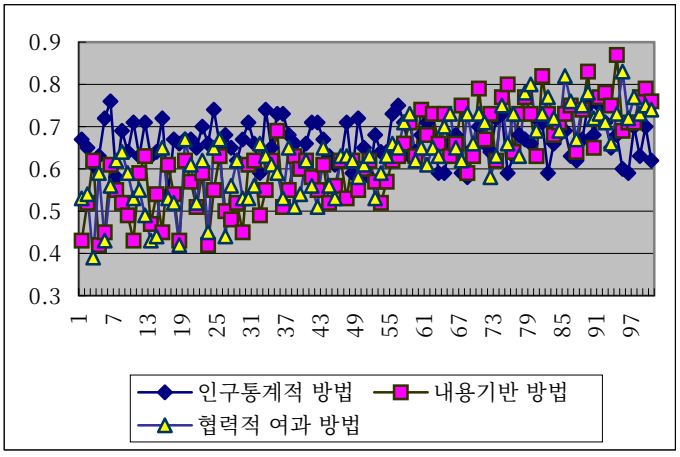
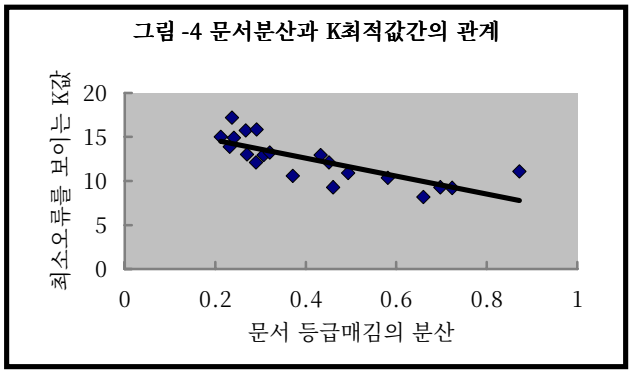


그림-3 학습방법에 따른 정분류율의 변화

그림-3 에서 보듯이 초기에는 내용기반의 여과 방식과 협력적 여과방식은 평가데이타의 부족으로 인해 임의추천 방식이 되어 좋은 성능을 기대하기 어렵지만 추천 문서량이 커지면서 인구통계적 방식에 의한 추천보다 좋은 성능을 보이게 된다.

그림-4 는 적응형 k-NN에 쓰이는 최적의 k값과 문서의 평가값의 분산과의 선형관계를 학습한 결과를 보여주고 있다.

결합방식은 학습된 데이터 및 각각의 정분류율 분석에



기반하여 다음의 식(8)에 의해 추천을 행한다.

$$D = \arg\max_{D}(\lambda_d f_d(D) + \lambda_s f_s(D) + \lambda_k f_k(D)) \qquad (8)$$

where, $\lambda_d(N) = 1 \ (N < 55), \ 0 \ (N \geq 55)$

$$\lambda_s(N) = 0 \ (N < 55), \ \frac{4N_k}{4N_k + N_s} \ (N \geq 55)$$

$$\lambda_k(N) = 0 \ (N < 55), \ \frac{4N_s}{4N_k + N_s} \ (N \geq 55)$$

윗 식에서 f_d, f_s, f_k 는 각각 인구통계적 방식, SVM에 의한 방식, k-NN에 의한 방식에 의한 정규화된 추정 선호도값이고 $N_s, N_k (N_s + N_k = N)$ 는 각각 특정 사용자에 대한 SVM에 학습한 데이터 크기와 공통 평가 데이터의 크기이다. 그림-4 는 식(8)에 의한 추천 정분류율의 변화를 보여준다.

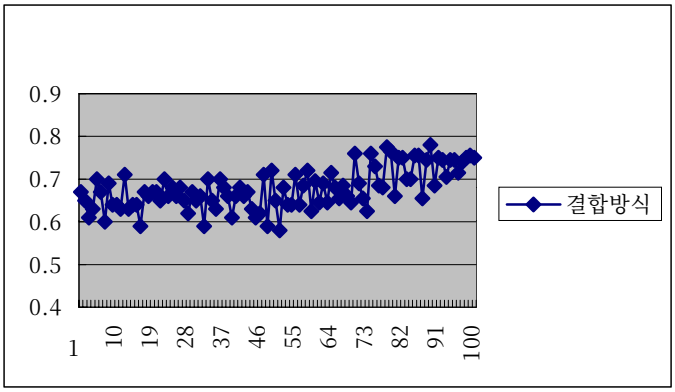


그림- 4 결합방식에 의한 정분류율의 변화

3. 결론

본 논문에서는 인구통계적 방식, 내용기반 방식, 협력적 방식에 의해 유머문서 추천을 행하였다. 새로운 사용자에 대해서는 인구통계적 방식이 좋은 성능을 내고 새로운 문서에 대해서는 내용기반 방식이, 평가값을 많이 가진 문서에 대해서는 협력적 방식이 좋은 성능을 내었으며 이들의 적절한 조합을 통해 추천의 최적성능을 낼 수가 있었다

감사의 글

본 연구는 첨단정보기술 연구센터(AITrc)를 통해 과학재단이 일부 지원하였고 2001 년도 두뇌한국 21 사업에 의하여 일부 지원되었음.

참고 문헌

[1] 이종우, 장병탁, "PCA 및 적응형 k-NN 을 이용한 유머문서의 추천", *한국 퍼지 및 지능 시스템 학회 2000 추계학술대회 학술발표 논문집*, pp. 133-136, 2000.

[2] T. Hofmann, Probabilistic Latent Semantic Analysis, In *Proc. Of the 15th Conference on Uncertainty in Artificial Intelligence (UAI'99),* 1999.

[3] Vapnik, V., S. Golowich, and A. Smola, Support Vector Method for Function Approximation, Regression Estimation, and Signal Processing, *Adv. Neur. Info. Process. Syss, 10,* 1996.