은닉 변수 모델을 이용한 문서 추천

이종우 장병탁
서 울 대 학 교 컴 퓨 터 공 학 부
{jwlee, btzhang}@scai.snu.ac.kr

Learning Model for Recommendation of Humor Documents

Jongwoo Lee Byoung-Tak Zhang
School of Computer Science and Engineering, Seoul National University

요 약

우리는 유머문서의 추천을 위해서 문서 정보, 사용자 정보, 공통 등급매김 정보 등을 모두 이용하는 4 개의 관찰 변수와 이들간 관계의 학습을 위한 은닉변수를 사용한 확률모델을 구축하였다. 이 모델은 학습된 은닉 변수와 가시 변수 간의 관계를 통해 누락 관찰 데이터에 대해서도 추정값을 유도해 낼 수 있으므로 등급매김 정보가 부족하거나 새로운 사용자와 문서의 도입시에 안정적인 추천 성능을 보여 줄 수가 있다. 또한 확률 모델의 학습을 위해서 EM 알고리즘을 이용하였는데 저평가된 데이터의 이용도를 높이기 위해서 추천을 반대하는 확률 모델을 따로 두고 이들간에 분류모델(classification model)을 두어서 추정값을 분류해내는 방식을 취한다.

1. 서론

정보추천(information recommendation), 혹은 정보여과(information filtering)란 특정 정보 수요자에게 높은 선호도(preference)를 보일 만한 정보 혹은 아이템(item)을 가려서 능동적으로 제공하여 주는 기술이다. 정보 여과에서 쓰이는 학습방법은 추천하고자 하는 데이터의 특성과 응용 시스템의 종류에 따라 인구통계적(demographic) 방법, 내용기반 추천(contents-based recommendation), 협력적 추천(collaborative recommendation) 등이 있으며 이들을 부분적으로 결합하는 결합적 여과(hybrid recommendation) 방식이 있다.

내용적 여과 방식은 사용자가의 아이템(item)에 대한 선호도로부터 선호하는 아이템의 특성을 학습하는 방법으로 개인의 다양한 선호도를 반영할 수 있지만 아이템의 특성과 사용자간의 관계를 학습해야 하므로 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 추천방식의 일종이라고 할 수도 있다. 하지만 개인을 학습하는 것이 아니라 같은 부류의 보편적 프로파일을 가진 사람들의 집단과 아이템의 특징간의 관계를 학습한다는 점이 다르다. 이 방식이 필요한 경우는 프로파일이 표현하지 못하는 사용자의 다양성을 반영하지 못한다는 단점을 지니지만 새로운 사용자에 대하여 아무런 선호도 데이터가 마련되지 않을 때 이용할 수 있는 유일한 방법이기 때문이다.

이에 비해 협력적 추천 방식은 다수의 사용자 집단으로부터 비슷하거나 다른 선호 성향을 가진 타사용자의 선호정보를 추천에 이용한다. 이 방법은 아이템의 내용이나 특성을 전혀 고려하지 않아도 되기 때문에 아이템을 분석해야 하는 내용기반 여과의 단점을 극복하는 반면 공통 등급매김 정보가 부족한 경우거나 사용자들의 선호도가 서로 관련이 적을 경우 좋은 성능을 보이지 않는다[1].

추천에서 가장 어려운 문제 중 하나는 cold-start 상황이라고 알려진 것인데 실제 온라인 상에서 새로운 사용자가 추가된 경우 내용기반의 추천이, 새로운 문서가 추가된 경우 협력적 추천 방식이 이용할 정보가 없다는 것이다. 또한 가용 정보가 있다 하더라도 데이터의 희소성(sparse data)에 의해 모델의 신뢰성이 떨어지는 경우도 자주 등장하는 문제점이다.

cold-start 상황을 극복하기 위해 고안된 결합 추천 방식 혹은 통합 추천 방식은 위에서 언급된 각 방식의 단점을 보완하여 이들 방식의 일부 혹은 모두를 결합시킨 여과 방식이다. 추천 시스템의 모델은 단일 모델일 수도 있지만 여러 개의 모델을 두고 각 모델의 추천값을 가중치 합산(weighted

sum)시키거나, 환경에 따라 좋은 모델을 선택하거나 특정 모델의 출력이 다른 모델의 입력으로 주어지는 등의 형태가 있다.

본 논문에서는 3 가지 여과 방식의 부분적 단점을 보완하여 가용 정보를 모두 사용하여 단일 확률 모델을 구축하였다. 가용 정보로는 문서의 단어 벡터(term vector)에 문서의 길이 및 문단의 평균길이 등의 정보가 추가되었으며 사용자의 프로파일에 근거한 사용자 카테고리(category)값과 사용자가 유머문서를 보고 매긴 평가값(rating value) 등이 이용되었다. 가용 정보는 모델에서 입력 변수(observed variables)에 반영되며 이들간의 학습을 위해서 직접적인 조건부 확률값을 학습하는 대신 은닉변수(latent variable)을 중재자 역할로 사용하여 적은 데이터에 대해서 숨은 의미와 잠재적 클러스터의 생성을 꾀하였다.

2. 관련 연구

일반적인 인구통계적 방법은 CRM (customer relationship management) 분야에서 많이 활용되었다. 내용 기반의 방식과 유사하게 인구통계적 방법에서의 추천은 사용자의 프로파일과 추천 아이템의 특성간의 관계를 학습해야 한다.

내용 기반의 추천모델이 문서 추천이나 웹브라우징 시스템에 적용되는 경우 가장 자주 이용되는 문서의 특징은 단어 벡터와 문서의 길이, 저자, 장르 등이며 사용자의 피드백은 평가(rating) 데이터와 로그 파일, 이벤트 로깅(event logging), click stream 등이 된다. 만약 아이템이 상품이 주종인 전자 상거래 환경에서는 아이템의 주된 특징은 가격, 옵션, AS, 배송 등이 된다.

내용 기반의 추천 모델을 설계하는데 있어서 아이템의 특성간 유사도(similarity) 측정이 많이 이용되었는데 consine 유사도나 상관계수(correlation) 등이 아이템의 특징 벡터나 사용자의 프로파일 벡터간에 주로 사용되었으며 단어 벡터와 문서의 카테고리(category)간의 관계가 MDL (minimum description length)로 표현되기도 했다[2]. 사용자의 아이템 특성에 대한 선호도를 결정 트리(decision tree)로 표현하는 모델[3]도 있으며 비교적 간단한 확률모델로서 naïve Bayes classifier 를 이용하거나[4], SOM (self organizing map)으로 특성을 클러스터링(clustering)하기도 하였다[5].

협력적 여과 방식은 아이템의 특성을 전혀 고려하지 않는 방식이므로 사용자가 등급 매김한 정보만을 이용하여 그들간의 유사도를 측정하거나 비슷한 부류의 사용자들을 클러스터링하는 모델이 많이 이용되었다.

사용자의 아이템에 대한 직간접적 평가는 평가 벡터를 생성하게 되고 다른 사용자의 추천값을 이용하기 위해 k-NN 기법[6]이나 상관관계를 이용한 기법이 주로 쓰였으며 평가 벡터 공간의 크기와 데이터 부족 및 특징 요소 선별을 위해 PCA 등을 이용하여[7][8] 압축공간 상에서의 유사도 측정이 이루어지기도 하였다.

사용자와 아이템을 확률 변수로 정의하여 이들간의 선호도를 결합 확률 분포 혹은 조건부 확률로 정의하여 협력적 여과 모델을 제안하기도 하였으며[9][10], aspect model 을 이용하여 은닉변수가 사용자와 문서간의 관계를 확률 학습하기도 하였다[11]. 이 밖에도 신경망을 이용하거나[12], 회귀기반 선형모델(regression-base linear model)[13], 규칙 생성(rule generation)[14] 등이 기법이 쓰이기도 하였다.

cold-start 문제와 데이터 부족 현상을 극복하기 위한 결합 모델은 보통 내용 기반 추천 방식과 협력적 추천 방식을 결합하는데 제안되었으며 [15]에서는 추천을 위해서 협력적 정보와 내용적 정보를 모두 이용한 확률모델을 제안하였고 [16]에서는 협력적 추천방식을 내용적 요소가 협력적 모델 학습을 위해 데이터를 만들어주어 학습을 돕는 기법을 이용하였다. 결합모델은 대부분의 경우 사용자의 보편적 프로파일을 이용하지 않으므로 새로운 사용자가 들어왔을 때에는 무작위 추천을 할 수 밖에 없었다.

3. 본론

우리는 결합 추천 방식의 구현을 위하여 Probabilistic Latent Semantic Analysis (PLSA) 모델[2]을 통합 확률 모델로 이용하였다(그림 1.). 이러한 PLSA 모델은 [3]에서도 일부분 추천모델로 수용되었다.

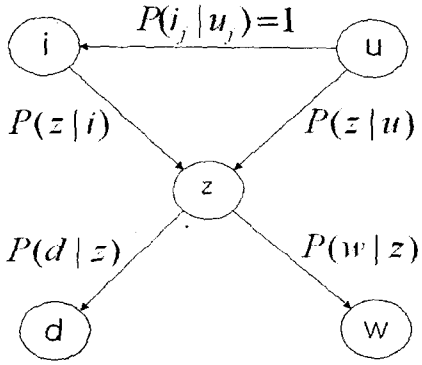

그림. 1 4 개의 관찰변수와 은닉변수를 가지는 PLSA 모델

그림 1. 에서 관찰 확률 변수로서 i 는 사용자의 인구통계적 카테고리(demographic category)를, u 는 사용자 번호를, d 는 문서번호를 w 는 문서의 단어 벡터를 나타내는 확률 변수이고 확률 변수 z 는 은닉 변수로서 입력 변수들간을 연결해서 z 가 주어진 상태에서는 d, w 와 u, i 두 그룹 사이에 확률적 독립성을 가정하게 해준다. i 는 u 에 함수 종속관계에 있고 w 는 d 에 함수 종속관계를 가진다는 것은 쉽게 알 수가 있다.

확률 변수들의 다음의 집합내의 값을 지닌다.

$$\mathbf{u} \in \ U = \{u_1, u_2, \mathrm{K}\ , u_N\}$$
$$\mathbf{d} \in \ D = \{d_1, d_2, \mathrm{K}\ , d_M\}$$
$$\mathbf{r} \in \ R = \{r_1, r_2, \mathrm{K}\ , r_{10}\}$$
$$\mathbf{z} \in \ Z = \{z_1, z_2, \mathrm{K}\ , z_k\}$$
$$\mathbf{i} \in \ I = \{i_1, i_2, \mathrm{K}\ , i_L\}$$
$$\mathbf{w} \in \ \mathrm{W} = \{\mathrm{w}_1, \mathrm{w}_2, \mathrm{K}\ , \mathrm{w}_V\}$$

학습에 쓰이는 데이터는 다음 집합으로 정의된다.

$$\{v \mid v = (u,i,d,w,r), u \in U, i \in I, d \in D, w \in W, r \in R\}$$

z 가 주어진 상태에서 다변량 결합확률은 독립가정에 의해 다음과 같이 표현된다.

$$P(z,u,i,d,w) =$$
$$P(z)P(u \mid z)P(i \mid z)P(d \mid z)P(w \mid z)$$

우리가 모델 학습을 위해 최대화시킬 목적함수는 다음의 로그 우도(log likelihood)로 설정할 수 있다.

$$L = \sum_{u,i,d,w} n(u,d,i,w) \log P(z,u,i,d,w)$$

입력 변수를 위한 주변 확률분포(marginal distribution)는 다음 식으로 얻을 수가 있다.

$$P(u,d,i,w) = \sum_{z} P(z)P(u \mid z)P(i \mid z)P(d \mid z)P(w \mid z)$$

입력 변수들에 대한 은닉 변수의 조건부 확률을 얻기 위해서는 다음의 E-step 과 M-step 을 반복하는 EM 알고리즘을 통해서 학습이 이루어진다.

E-step 에서는 아래의 식을 이용하여 입력 변수들에 대한 z 변수의 posterior 확률값을 재측정한다.

$$P(z \mid u,d,w,i) = \frac{P(z)P(u \mid z)P(i \mid z)P(d \mid z)P(w \mid z)}{\sum_{z'} P(z')P(u \mid z')P(i \mid z')P(d \mid z')P(w \mid z')}$$

M-step 에서는 E-step 에서 얻어진 posterior 확률값을 이용해서 다음의 매개변수값들을 재계산한다.

$$n(u,d,w,i,r) = f(r(u,d,w,i))$$
$$P(u \mid z) \propto \sum_{d,w,i,r} n(u,d,w,i,r)P(z \mid u,d,w,i)$$
$$P(d \mid z) \propto \sum_{u,w,i,r} n(u,d,w,i,r)P(z \mid u,d,w,i)$$
$$P(i \mid z) \propto \sum_{d,w,u,r} n(u,d,w,i,r)P(z \mid u,d,w,i)$$
$$P(w \mid z) \propto \sum_{d,u,i,r} n(u,d,w,i,r)P(z \mid u,d,w,i)$$
$$P(z) \propto \sum_{u,d,w,i,r} n(u,d,w,i,r)P(z \mid u,d,w,i)$$

이 과정은 사후 확률값(posterior probability)이 수렴할 때까지 반복 수행되어 진다. 윗 식에서 쓰인 샘플링 제어 함수, f()는 높이 평가된 문서에 대해서는 많은 학습데이터가 많이 쓰인 효과를, 낮게 평가된 문서는 적게 확률에 영향을 주도록 한다.

추천을 위해서 가용데이터의 종류에 따라 3 가지 경우로 다르게 평가값을 추정한다.

우선 이미 평가를 행해온 사용자가 이미 다른 사람들로부터 평가받은 적이 있는 문서를 추정할 때에는 다음 확률값으로 그 문서의 적합도를 계산한다.

$$P(d \mid u) \propto P(d,u) \propto \sum_{w,i} P(u,d,w,i) = \sum_{w} P(w,u,d,i)$$

한번도 평가를 하지 않은 새로운 사용자에게 특정문서의 평가값을 추정하기 위해서는 윗 식을 사용자 입력변수 u 에 대해 주변화(marginalizing)시킨 확률값을 이용한다.

$$\sum_{u} P(d,u) = \sum_{w,i,u} P(u,d,w,i)$$

새로운 사용자에 대한 평가와 마찬가지로 평가된 적이 없는 새로운 문서에 대한 사용자의 추정 평가값은 문서 입력변수 d 에 대해 주변화시킨 확률값을 이용한다.

$$\sum_{d} P(d,w) = \sum_{d,w,i} P(u,d,w,u)$$

확률 모델 학습을 위해 쓰인 EM 알고리즘은 샘플링 제어함수를 이용하게 되는데 7 등급(1, 2, 3, … , 7)으로 나뉜 평가치를 확률모델에 결합시키는 방법을 결정한다. 가장 쉽게 생각할 수 있는 제어함수는 f(r) = r 의 선형형태로 이용하는 것이다. 즉, 저평가된 데이터는 확률에 기여도가 떨어지는 효과를 가진다. 하지만 이렇게 샘플링을 제어하게 되면 저평가한 문서와 평가를 하지 않은 문서를 동일시하는 효과를 가져온다. 우리는 저평가된 데이터의 의미를 살리기 위해 데이터의 평균

평가값을 기준으로 고평가된 데이터는 선호 요소(preferred component) 모델을, 저평가된 데이터는 비선호 요소(dislike component) 모델을 각각 학습시키는데 이용시켰다. 즉, 그림 1. 의 모델을 2 개를 두고 학습을 시킨다.

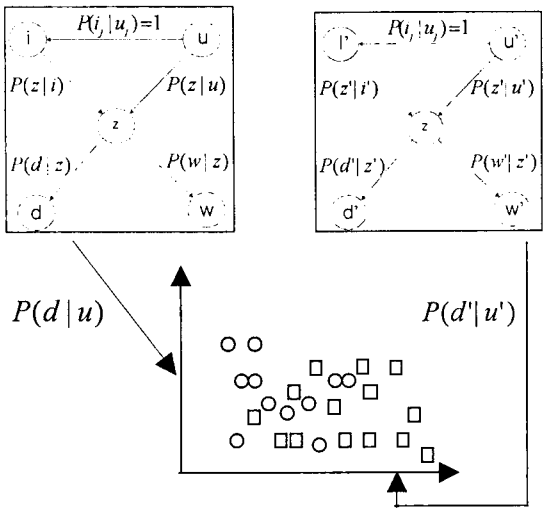

그림. 2 선호 및 비선호 요소의 추천확률에 대한 결정 평면

그림.2 가 보이는 평면에서 가로축은 특정 사용자가 특정 문서를 선호하지 않을 확률이고 세로축은 선호할 확률이며 작은 원으로 표시된 점은 실제로는 사용자가 선호했을 경우이고 작은 사각형으로 표시된 점은 실제로는 사용자가 선호하지 않았을 경우를 표현한다고 하면 이 평면상에서 추정모델의 학습은 2 진 분류문제가 된다.

만일 모델이 의미있게 학습된다면 선호 요소에서 높은 확률을 갖는 입력 확률 변수에 대해서 비선호 요소는 낮은 확률을 보일 것이다.

각 모델의 학습을 위한 샘플링 제어 함수는 그림.2 에서 보이는 함수를 이용하였다.

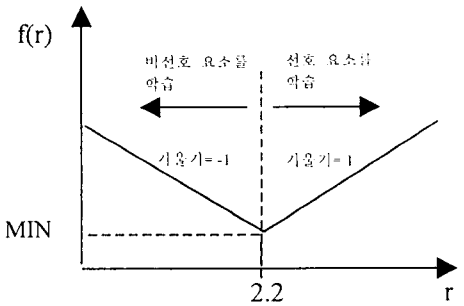

그림. 3 샘플링 제어 함수

샘플링 제어 함수의 선정은 결국 사용자의 지각적인 평가라는 값을 어떻게 확률변수의 빈도수로 변환시키는지에 관한 절차로서 실험을 통한 적절한 함수를 찾아내는 것이 중요하겠지만 이 논문에서는 아주 간단한 선형모델로서 구현하였다.

4. 실험

실험은 50 명의 사용자와 99 개의 유머문서에 대하여 실행되었다. 1(비선호)부터 7(선호)까지 등급매김되어진 4950 개의 데이터에 대하여 평균 평가값을 기준으로 저평가된 데이터는 비선호 요소를 고평가된 데이터는 선호 요소를 학습시키는데 사용하였다.

두개의 요소를 결합하여 유머 문서의 선호도를 추정하기 위해서 결정 평면상의 분류자로서 우리는 결정트리(decision tree)와 단일 perceptron 을 선형모델처럼 이용하여 2 가지로 실험하였다. 단말 노드에는 두개의 요소에서 추정된 확률값이 입력으로 들어가고 학습을 위한 출력값은 선호 및 비선호 정보로 표현된 이진값(binary value)을 이용하였다.

정확도 측정을 위해 우리는 정분류율(correct classification rate)으로서 평가평균값를 기준으로 추정값과 실제 평가값이 같은 편에 위치하면 정분류된 것이고 반대편에 위치하면 오분류된 것으로 판단하여 전체 실험데이터에 대해 그 비율을 계산한 값이다.

학습은 5-fold cross validation 방식을 이용하여 전체 데이터의 4/5 으로 학습을 시키고 1/5 로 테스트하여 정분류율을 측정하는 과정을 여러 번 수행하였다. 은닉변수 모델이 학습이 제대로 되었는지를 알아보기 위해서 학습데이터에 대해서도 정분류율을 측정해 보았다.

표 1.은 선호요소 모델과 비선호요소 모델을 은닉변수의 개수를 변화시켜가면서 결정트리를 이용하여 학습 시킨 후 학습 데이터와 테스트 데이터에 대해 각각 정분류율을 측정한 결과이다.

은닉변수	5 개	10 개	20 개	30 개
학습데이터	70.23%	74.19%	79.48%	82.77%
테스트데이터	62.00%	60.40%	59.30%	54.00%

표 1. 결정트리를 이용한 모델 결합 시 정분류율

표 2.는 두 요소의 결정 평면상의 분류자를 단일 perceptron 을 이용하여 결합 후 학습시킨 결과이다.

은닉변수	5 개	10 개	20 개	30 개
학습데이터	70.07%	74.95%	80.41%	83.78%
테스트데이터	60.70%	59.70%	57.30%	55.90%

표 2. 신경망을 이용한 모델 결합 시 정분류율

측정 결과 가장 높은 성능을 보이는 것은 5 개의 은닉변수를 가지는 확률모델을 결정트리를

이용하여 분류 추정한 것이었고 62%정도의 정확도를 보였다. 은닉변수를 늘려가면서는 학습데이터에 대한 정확도는 높아지는 반면 테스트 데이터에 대한 정확도는 줄어드는 것을 알 수가 있었다.

5. 결론

추천 아이템의 종류에 따라 추천성능이 변하겠지만 유머문서의 추천의 경우에서는 확률모델이 학습이 되긴 하였으나 테스트 데이터에 대해서 높은 정확도를 보이진 않고 있다. 이에 반해 학습 데이터에 대한 정확도가 높은 것은 적은 데이터량에 대해 과도학습(overfitting)한 것으로 볼 수가 있어서 좀더 많은 학습 데이터의 수집이 요구된다. 문서의 내용에 대한 정보는 주로 단어벡터에 의존하고 있는데 이 역시 문서의 개수가 부족하고 유머문서의 특징을 잘 반영시켜줄 수 있는 장르나 그 밖의 다른 부가정보의 추가가 필요하다.

추천 모델에서 통합된 확률 모델의 장점은 추천에서 자주 발생하는 누락된 정보에 대해 다른 정보를 이용하여 추천을 수행할 수 있다는 것이고 이는 은닉변수 모델에서 누락변수에 대한 주변화를 통해 이루어진다.

감사의 글

본 연구는 첨단정보기술 연구센터(AITRC)를 통하여 과학재단이 일부 지원하였으며 과학기술부 주관 뇌신경 정보학사업(BrainTech)에 의해 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] Hofmann, T. and Puzicha, J., "Latent class models for collaborative filtering," *Proc. Of the Sixteenth International Joint Conference on Artificial Intelligence*, pp. 688-693, 1999.

[3] B. Krulwich and C. Burkey, "The ContactFinder agent: Answering bulletin board questions with referrals," *Proc. of Thirteenth National Conf. on Artificial Intelligence* (AAAI-96), pp. 10-15, 1996.

[4] C. M. Kadie, C. Meek and D. HeckermanP. "CFW: A Collaborative Filtering System Using Posteriors Over Weights of Evidence", *Proc. of Uncertainty in Artificial Ingellegence*, pp. 242-250, 2002.

[5] S. J. Pelletier and J. F. Arcand, "STEALTH: A personal digital assistant for information filtering," *Proc. Practical Application of Intelligent Agents and Multi-Agent Technology*, pp. 455-474, 1996.

[6] Upendra Shardanand, Pattie Maes: "Social Information Filtering: Algorithms for Automating Word of Mouth", *CHI '95 Proceedings: Conference on Human Factors in Computing Systems : Mosaic of Creativity*, 1995.

[7] D. Gupta, M. DiGiovanni, H. Narita, and K. Goldberg, "Jester 2.0: Evaluation of a new linear time collaborative filtering algorithm applied to jokes," *Poster Session and Demonstration, 22nd International ACM SIGIR Conf. on Research and Development in Information Retrieval,* pp. 191-192, 1999.

[8] J. W. Lee, B. T. Zhang, "Humor Document Recommendation using User Clustering with PCA", *In Proc. Of Korea Fuzzy Logic and Intelligent Systems Society*, pp. 133-136, 2000.

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

[10] D. Billsus and Michael J. Pazzani, "Learning Collaborative Information Filters," *in Shavlik, J., ed., Machine Learning: Proc. of the Fifteenth International Conference, Morgan Kaufmann 0Publishers*, 1999.

[11] Lang, K., "NewsWeeder: Learning to filter netnews", *Proc. of the Twelfth International Conference on Machine Learning*, pp. 331--339 San Francisco, CA. Morgan Kaufman.

[12] D. Billsus and Michael J. Pazzani, "Learning Collaborative Information Filters," *in Shavlik, J., ed., Machine Learning: Proc. of the Fifteenth International Conference, Morgan Kaufmann 0Publishers*, 1999.

[13] Vucetic, S. and Obradovic, Z., "A Regression-Based Approach for Scaling-Up Personalized Recommender Systems in E-Commerce", *Workshop on Web Mining for E-Commerce, at the Sixth ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining*, Boston, MA, August 2000.

[14] Weiyang Lin, Sergio A. Alvarez, and Carolina Ruiz. "Collaborative Recommendation via Adaptive Association Rule Mining", *International Workshop on Web Mining for E-Commerce (WEBKDD'2000). held in conjunction with the Sixth International Conference on Knowledge Discovery and Data Mining (KDD2000)*

[15] A. Popescul and L. H. Ungar, "Probabilistic Models for Unified Collaborative and Content-Based Recommendation in Sparse-Data Environments" , *Proc. of the 17 Conference on Uncertainty in Artificial Intelligence*, pp. 437-444, 2001.

[16] P. Melville, R. J. Mooney and R. Nagarajan, "Content-Boosted Collaborative Filtering for Improved Recommendation", *Proc. of the 18 National Conference on Artificial Intelligence*, pp. 187-192, 2002.