잠재 요인 모델의 원리를 이용한 협업 태그 기반 추천 방법

Collaborative Tag-Based Recommendation Methods Using the Principle of Latent Factor Models

김형도(Hyoung-Do Kim)*

초 록

협업에 의한 태그 작성 시스템은 소셜 네트워크에서 다양한 공유 콘텐츠에 사용자가 태그를 부착할 수 있도록 허용하는데, 이러한 태그들은 본인뿐만 아니라 모든 커뮤니티 사용자들이 콘텐츠를 이용하는데 유용함을 준다. 협업 태그 기반의 추천에서는 사용자와 항목, 그리고 태그로 이루어진 3차원 데이터를 이용하는데, 이 데이터는 일반적으로 사용자와 항목으로 이루어진 2차원 데이터에 비하여 더 방대한 반면, 희소성(Sparsity)이 더 높다. 따라서 기존의 협업 필터링 기법을 바로 적용하는데 어려움이 많다. 잠재 요인 모델(Latent Factor Model)은 관찰된 값을 설명하는 잠재된 특징(요인)들을 밝히고, 이를 이용해서 문제를 해결하기 위한 모델로서 최근 협업 필터링에서도 성공적으로 적용되고 있으나, 모델을 학습하거나 개선하는 단계에서는 많은 시간과 노력이 필요하다는 단점이 있다. 이러한 잠재 요인 모델을 3차원 협업 태그 데이터에 적용하기 위해서는, 계산이 복잡한 협업 필터링 모델 수립의 어려움을 극복해야 한다. 이 논문에서는 사용자가 항목에 대해 사용한 태그들을 사용자 및 항목에 대한 잠재요인으로 간주하여 직관적인 모델을 수립하고, 사용자의 아이템에 대한 선호도를 결정하는 여러 가지 방법들을 제안하고, 실제 협업 태그 데이터를 이용하여 이들을 비교 평가한다.

ABSTRACT

Collaborative tagging systems allow users to attach tags to diverse sharable contents in social networks. These tags provide usefulness in reusing the contents for all community members as well as their creators. Three-dimensional data composed of users, items, and tags are used in the collaborative tag-based recommendation. They are generally more voluminous and sparse than two-dimensional data composed of users and items. Therefore, there are many difficulties in applying existing collaborative filtering methods directly to them. Latent factor models, which are also successful in the area of collaborative filtering recently, discover latent features(factors) for explaining observed values and solve problems based on the features. However, establishing the models require much time and efforts. In order to apply the latent factor models to three-dimensional collaborative filtering data, we have to overcome the difficulty of establishing them. This paper proposes various methods for determining preferences of users to items via establishing an intuitive model by assuming tags used for items as latent factors to users and items respectively. They are compared using real data for concluding desirable directions.

키워드 : 협업 태그, 잠재 요인 모델, 추천, 협업 필터링
Collaborative Tag, Latent Factor Model, Recommendation, Collaborative Filtering

---

* 한양사이버대학교 경영학부
2009년 09월 19일 접수, 2009년 10월 05일 심사완료 후 2009년 11월 06일 게재확정.

1. 서 론

협업 필터링(Collaborative Filtering)[1, 4, 5, 9, 10, 16, 17]은 사용자들이 항목(Item)들에 대해서 평가한 값(평가치)들을 기반으로 추천을 해주는 기법으로, 메모리 기반과 모델 기반의 협업 필터링으로 구분할 수 있다. 전자는 두 사용자 간의 또는 항목 간의 유사도(Similarity)를 측정하여 가장 유사한 사용자들 또는 항목들의 평가치를 반영하여 추천하며[2], 후자는 선형대수(Linear Algebra), 신경망(Neural Network), 군집화(Clustering) 등을 기반으로 사전에 모델을 수립하여 추천한다[8]. 메모리 기반의 협업 필터링에서 유사도 측정 방법은 피어슨의 상관계수(Pearson's Correlation), 코사인 유사도(Cosine Similarity) 등 다양하며, 예측의 정확성에 많은 영향을 미치게 된다[2]. 두 사용자 간의 또는 두 항목 간의 유사도는 사전에 계산될 수 있으나, 사용자의 수 또는 항목의 수가 늘어남에 따라서 저장해야할 유사도의 수가 기하급수적으로 증가하므로, 일반적으로 매 추천시마다 유사도를 계산함에 따라서, 문제의 크기가 확장되는데 대한 대응이 어렵다. 또한 평가치 데이터가 희소할 경우, 유사도 계산이 불가능하거나, 낮은 유사도의 이웃을 선택하게 되어, 추천 품질의 수준이 나빠진다. 모델 기반의 협업 필터링은 신속한 추천이 가능하지만, 모델을 학습하거나 개선하는 단계에서는 많은 노력이 수반된다는 단점이 있다.

온라인 소셜 네트워킹에서 협업에 의한 태그 작성 시스템은 콘텐츠를 공유하는데 있어서 핵심적인 역할을 수행한다[6]. 이러한 시스템은 e-비즈니스를 비롯한 다양한 분야의 공유 콘텐츠에 사용자가 태그를 부착할 수 있도록 허용하는데, 이러한 태그들은 본인뿐만 아니라 모든 커뮤니티 사용자들이 콘텐츠를 이용하는데 유용함을 준다. 이러한 태그 데이터는 검색 메카니즘을 개선하고, 콘텐츠를 보다 적절하게 구조화하거나, 사용자의 관심사항에 맞게 개인화된 추천을 제공하는데 이용될 수 있다[19]. 협업 태그 기반의 추천에서는 사용자와 항목, 그리고 태그로 이루어진 3차원 데이터를 이용하는데, 이 데이터는 일반적으로 사용자와 항목으로 이루어진 2차원 데이터에 비하여 더 방대한 반면, 희소성(Sparsity)이 더 높다. 따라서 기존의 협업 필터링 기법을 바로 적용하는데 어려움이 많다.

잠재 요인 모델(Latent Factor Model)은 관찰된 값을 설명하는 잠재된 특징(요인)들을 밝히고, 이를 이용해서 문제를 해결하기 위한 모델로서 최근 협업 필터링에서도 성공적으로 적용되고 있다[3, 13, 14]. 잠재 요인 모델도 모델 기반의 협업 필터링의 일종이므로 모델을 구축하는 단계에서 많은 시간과 노력이 필요하나, 일단 구축되고 나면 신속한 추천이 가능한 장점을 가지고 있다. 이러한 잠재 요인 모델을 협업 태그 데이터에 적용하기 위해서는 모델 수립의 어려움을 극복해야 한다. 이 논문에서는 잠재 요인 모델의 원리를 쉽게 이용할 수 있도록 사용자와 항목에 대해 사용된 태그들을 잠재요인으로 간주하여 직관적인 모델을 수립하고, 이들로부터 사용자의 아이템에 대한 선호도를 결정하는 6가지 방법을 구체적으로 제안하고, 실제 협업 태그 데이터를 이용하여 이들을 비교 평가하여 바람직한 대안을 제시한다.

이 논문의 구성은 다음과 같다. 제 2장에서

는 잠재 요인 모델을 이용한 협업 필터링과 협업 태그에 대한 관련 연구들을 정리하고, 제 3장에서는 협업 태그를 잠재 요인으로 보는 새로운 관점을 바탕으로 잠재 요인 모델의 원리를 이용하는 협업 필터링 방법들을 제시한다. 제 4장에서는 실제 협업 태그 데이터를 이용하여 실험 결과를 제시하고, 그 의미를 논한다. 마지막으로 제 5장에서는 결론을 맺고, 앞으로의 연구방향을 제시한다.

2. 관련 연구

2.1 잠재 요인 모델

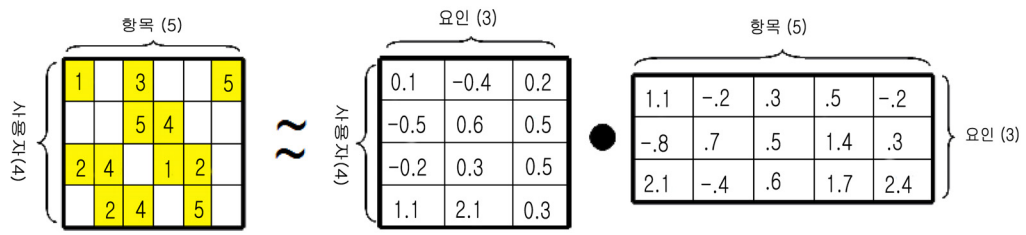
잠재 요인 모델은 관찰된 값을 설명하는 잠재된 특징(요인)들을 밝히고, 이를 이용해서 문제를 해결하기 위한 모델이다. SVD(Singular Vector Decomposition)나 PCA(Principal Component Analysis)와 같은 매트릭스 인수분해 알고리즘(Matrix Factorization Algorithm)이 가장 대중적이고, 협업 필터링 분야에서도 성공적으로 적용되고 있다[3, 14]. SVD의 경우 각각의 사용자 u와 사용자 요인 벡터 x_u를 연결하고, 마찬가지로 각각의 항목 i과 항목 요인 벡터 y_i를 연결하는 모델을 구축하게 되는데, 여기서 평가치가 존재하는 모든 경우에 대하여 $r_{ui} = x_u \cdot y_i$를 만족하도록 x_u와 y_i는 결정되어야 한다. 이후에 사용자 u의 항목 i에 대한 평가치 예측은 $p_{ui} = x_u \cdot y_i$로 구하게 된다. 벡터 x_u와 y_i의 차원 k는 고려하는 잠재 요인의 수로서 데이터의 가장 중요한 특징들을 고려하여 결정하게 된다.

예를 들어서 4명의 사용자가 5개의 항목을 평가한 데이터를 가지고 구축한 x와 y가 <그림 1>과 같다고 하자. 여기서는 3개의 요인을 고려하여 모델을 결정한 것이다. 두 번째 사용자의 첫 번째 항목에 대한 예측치는 x_2와 y_1을 곱하여 구하면 된다. 즉 $p_{21} = x_2 \cdot y_1 = (-0.5) * 1.1 + 0.6 * (-0.8) + 0.5 * 2.1 = -0.55 - 0.48 + 1.05 = 0.02$가 된다.

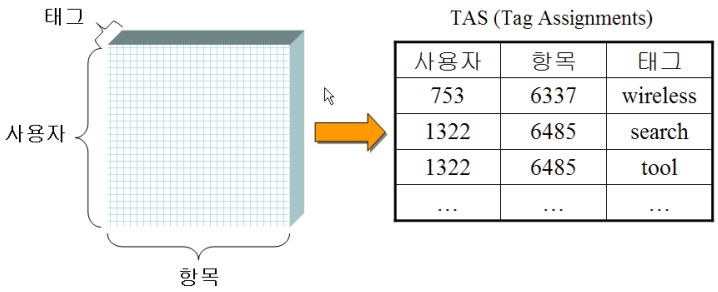
잠재 요인 모델을 이용한 추천은 모델기반 협업 필터링의 일종으로서 모델 구축에 많은 노력이 필요하지만, 일단 구축하게 되면 신속하게 추천을 해줄 수 있는 장점이 있다.

2.2 소셜 태그 작성

소셜 태그 작성(Social Tagging) 또는 협업 태그 작성(Collaborative Tagging)은 여러 사용자들이 협력하여 태그(Tag)라 불리는 단어들을 자원에 배정하여 주석을 붙이고 분류하는



〈그림 1〉 SVD 모델 사례



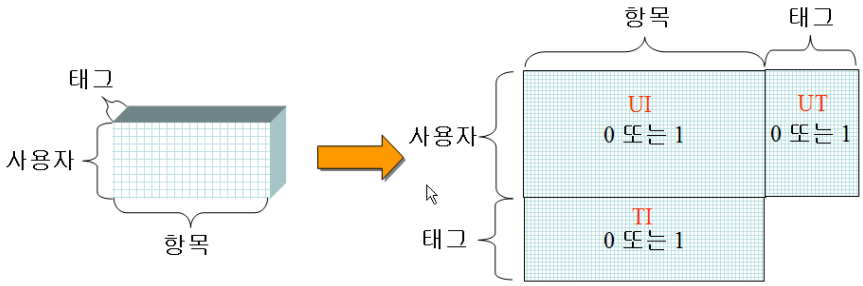
〈그림 2〉 협업 태그 데이터의 구조

것을 말한다[6, 7, 12, 18, 19]. 이러한 태그 작성은 전형적인 계층구조의 택소노미(Taxonomy)와 대비하여 폭소노미(Folksnomy)라고 부르기도 한다. 이렇게 작성된 태그들은 소셜 네트워크에서 자원을 공유하는데 있어서 핵심적인 역할을 하게 된다.

협업 태그 작성 시스템은 여러 가지 측면에서 분류될 수 있는데, 태그 작성에 있어서의 사용자 권한에 의하면 자원 생성자만의 태그 작성(사례 : YouTube), 허가 기반의 태그 작성(사례 : Flickr), 모두에게 자유로운 태그 작성(사례 : del.icio.us)으로 구분되며, 태그 모음에 의하면 중복된 태그를 허용하지 않는 집합형(사례 : YouTube와 Flickr)과 중복된 태그를 허용하는 가방형(사례 : del.icio.us)으로 구분할 수 있다. 이 논문에서는 모두에게 자유로운 가방형의 태그 작성을 대상으로 협업 필터링 기법을 논한다. 태그 데이터는 <그림 2>의 왼쪽과 같이 사용자, 항목, 태그로 이루어진 3차원 구조를 갖게 되며, 일반적으로 오른쪽과 같은 형태의 TAS(Tag Assignments)로 관리된다.

2.3 태그 기반의 협업 필터링

Tso-Sutter et al.[19]은 3차원인 협업 태그 데이터를 <그림 3>과 같이 2차원으로 펼쳐서 일반적인 협업 필터링 알고리즘을 적용할 수 있도록 제안하고 있다. 즉, 사용자와 항목으로 구성된 UI에다 사용자에 대한 태그 데이터인 UT를 결합해서 식 (1)과 같이 사용자 기반 협업 필터링을 적용하여 p^{ucf}를 구하고,



〈그림 3〉 협업적 태그 데이터의 재구성

UI에다 항목에 대한 태그 데이터인 TI를 결합해서 식 (2)와 같이 항목 기반 협업 필터링을 적용하여 p^{icf}를 구한 뒤, 이 두 값의 단위 차이를 없애고 결합하여 식 (3)과 같이 예측치를 얻는다. 여기서 N_u는 사용자 u의 이웃(Neighbor) 집합을 의미하며, 마찬가지로 N_i는 항목 i의 이웃(Neighbor) 집합을 나타낸다. β는 사용자 기반 예측치를 항목 기반 예측치와 결합하는 비율로서 0과 1사이의 값이다.

$$p^{ucf}(UI_{ui}=1)=\frac{\left|\{v\in N_u \mid UI_{vi}=1\}\right|}{\left|N_u\right|} \quad (1)$$

$$p^{icf}(UI_{ui}=1)=\sum_{j\in N_i, UI_{uj}=1} sim(i,\ j) \quad (2)$$

$$p^{iucf}(UI_{ui}=1)=\beta\frac{p^{ucf}(UI_{ui}=1)}{\sum_{v} p^{ucf}(UI_{vi}=1)} +(1-\beta)\frac{p^{icf}(UI_{ui}=1)}{\sum_{j} p^{icf}(UI_{uj}=1)} \quad (3)$$

Ji et al.[7]은 <그림 4>와 같이 사용자가 사용한 태그의 횟수를 이용하여 UT를 구성하고, 여기에서 사용자간 유사도를 구한 뒤, 식 (4)와 같이 사용자별로 우선순위가 높은 태그들의 집합인 후보 태그 집합(Candidate Tag Set)을 결정하고, 여기에 속한 태그들을 대상으로 식 (5)와 같이 나이브 베이즈(Naive Bayes)를 적용해서 예측치를 구하는 방법을 제시하고 있다. 단, UT의 값은 사용자가 태그를 사용한 횟수를, TI의 값은 항목에 태그가 사용된 횟수를 나타낸다.

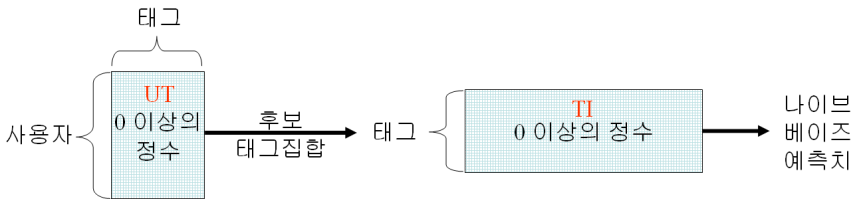
$$pref(u,\ t)=\sum_{v\in N_u} UT_{vt}\times sim(UT_{u.},\ UT_{v.}) \quad (4)$$

$$p_{ui}=P(I=i)\prod_{t} P(t\mid I=i) =\frac{\sum_{v} UI_{vi}}{\sum_{v,j} UI_{vj}}=\prod_{t}\frac{1+TI_{ti}}{m+\sum_{k} TI_{ki}} \quad (5)$$

3. 협업 태그와 잠재 요인

SVD 등 잠재 요인 모델의 우수한 추천 능력은 2006년부터 현재까지 진행 중인 Netflix Prize 대회[13]의 경쟁 결과에서도 쉽게 확인할 수 있다. 그러나 이러한 잠재 요인 모델을 협력적 태그 작성에 적용하고자 할 경우 심각한 문제들과 마주치게 된다. 무엇보다도 3차원으로 구성된 협업 태그 데이터는 매우 방대하나, 대부분의 데이터 값이 비어 있어서 고도로 희박한 형태를 가지고 있다. Ji et al.[7]의 논문에서 사용된 데이터는 1544명의 사용자가 17,390개의 항목을 10,077개의 태그로 평가한 것으로, 값이 0인 요소의 비율인 희소성(Sparsity)은 UI의 경우 99.89%이다.

이런 상황에서 잠재 요인 모델을 수립하기



〈그림 4〉 후보태그집합을 이용한 예측

위한 알고리즘은 매우 높은 계산 복잡성을 가지며, 실제로 많은 시간을 요구한다. 이러한 문제점을 개선하기 위하여 이 논문에서는 각각의 태그를 잠재 요인으로 보는 직관적인 아이디어를 근간으로 여러 가지 방법들을 제안하고, 이들을 비교 검토하여 바람직한 방안을 제시하고자 한다.

가장 기본적인 방법은 각각의 태그를 잠재 요인으로 직접 간주하는 것이다. 이 방법에서는 사용자 u가 태그 t를 사용한 횟수 UT_{ut}를 이 사용자가 태그를 사용한 모든 횟수로 나누어 0에서 1사이의 값으로 조정하며, TI에 대해서도 이와 동일하게 값을 조정한 뒤, 사용자 u의 항목 i에 대한 선호도를 식 (6)과 같이 구한다. 이 식에서 UT′는 <그림 1>에서의 사용자-요인 매트릭스에, TI′는 <그림 1>에서의 요인-항목 매트릭스에 해당되는 것으로, 선호도 계산 방식도 정확히 SVD와 일치한다.

$$pref(u,\ i) = \sum_t \left(\frac{UT_{ut}}{\sum_k UT_{uk}} \times \frac{TI_{ti}}{\sum_k TI_{ki}}\right) = \sum_t UT'_{ut} \times TI'_{ti} = UT'_{u.} \bullet TI'_{.i} \quad (6)$$

두 번째 방법은 정보 검색이나 추천에서 두 벡터를 비교하기 위하여 많이 사용되는 코사인 유사도를 이용하는 것이다. 이 방법에서는 두 태그 벡터 $UT_{u.}$과 $TI_{.i}$사이의 각도를 나타내는 코사인 유사도를 식 (7)과 같이 구하여 이를 사용자 u가 항목 i를 선호하는 정도로 사용한다. 그런데, 식 (7)의 코사인 유사도는 각각의 UT_{ut}를 $UT_{u.}$의 제곱근(Square Root)으로 나누어 UT''_{ut}로 조정하고, 이와 동일하게 각각의 TI_{ti}를 $TI_{.i}$의 제곱근으로 나누어 TI''_{ti}로 조정한 뒤, 이들을 곱해서 얻은 것과 같다. 즉, UT″는 <그림 1>에서의 사용자-요인 매트릭스에, TI″는 <그림 1>에서의 요인-항목 매트릭스에 해당된다.

$$pref(u,\ i) = \cos(UT_{u.},\ TI_{.i}) = \frac{\sum_t UT_{ut} \times TI_{ti}}{\sqrt{\sum_t UT_{ut}^2}\sqrt{\sum_t TI_{ti}^2}} = \sum_t \left(\frac{UT_{ut}}{\sqrt{\sum_k UT_{uk}^2}} \times \frac{TI_{ti}}{\sqrt{\sum_k TI_{ki}^2}}\right) = \sum_t (UT''_{ut} \times TI''_{ti}) = UT''_{u.} \bullet TI''_{.i} \quad (7)$$

사용자가 아이템을 평가했는지 여부를 나타내는 UI도 선호도를 계산하는데 도움이 될 수 있다. 여기서는 단순하게 식 (8)과 같이 특정 아이템에 대한 평가 횟수를 전체 평가 횟수로 나눈 값을 두 번째 방법에서 구한 코사인 유사도에 곱하여 선호도를 구하며, 이것이 세 번째 방법이다. 여기서 UT″는 <그림 1>에서의 사용자-요인 매트릭스에, TI‴는 <그림 1>에서의 요인-항목 매트릭스에 해당된다.

$$pref(u,i) = \cos(UT_{u.},\ TI_{.i}) \times \frac{\sum_v UI_{vi}}{\sum_v \sum_j UI_{vj}} = \sum_t \left(UT''_{ut} \times TI''_{ti} \times \frac{\sum_v UI_{vi}}{\sum_v \sum_j UI_{vj}}\right) = \sum_t \left\{UT''_{ut} \times \left(TI''_{ti} \times \frac{\sum_v UI_{vi}}{\sum_v \sum_j UI_{vj}}\right)\right\} = \sum_t (UT''_{ut} \times TI'''_{ti}) = UT''_{u.} \bullet TI'''_{.i} \quad (8)$$

사용자 태그 데이터의 희박한 특성을 보완하기 위해서, 식 (9)와 같이 사용자간 유사도를 이용해서 특정 사용자 u의 이웃들을 구하고, 이 이웃들이 태그를 사용한 빈도를 이용하여 이 사용자의 빈도를 수정한 뒤, 이렇게 수정된 UT^o와 TI간의 코사인 유사도로 선호도를 구하는 것이 네 번째 방법이다. 여기서 UT‴는 <그림 1>에서의 사용자-요인 매트릭스에, TI″는 <그림 1>에서의 요인-항목 매트릭스에 해당된다. 그리고 이웃은 피어슨의 상관계수(Pearson's Correlation)[2]를 이용하여 구하며, 그 값은 식 (9)에서 sim(u, v)로 사용된다.

$$
\begin{aligned}
pref(u,\ i) &= \cos(UT^o_{u.},\ TI_{.i}) \qquad (9)\\
&= \frac{\sum_t UT^o_{ut} \times TI_{ti}}{\sqrt{\sum_t UT^{o2}_{ut}}\sqrt{\sum_t TI^2_{ti}}}\\
&= \sum_t \left(\frac{UT^o_{ut}}{\sqrt{\sum_k UT^{o2}_{uk}}} \times \frac{TI_{ti}}{\sqrt{\sum_k TI^2_{ki}}}\right)\\
&= \sum_t UT'''_{ut} \times TI''_{ti} = UT'''_{u.} \bullet TI''_{.i}
\end{aligned}
$$

$$
where \quad UT^o_{ut} = \sum_{v \in N_u} UT_{vt} \times sim(u,v) \quad if \quad UT_{ut} = 0,
$$

$$
where \quad UT^o_{ut} = \sum_{v \in N_u} UT_{vt} \times sim(u,v) \quad if \quad UT_{ut} = 0,
$$

네 번째 방법에다 추가적으로 세 번째 방법과 같이 UI를 고려한 것이 식 (10)과 같이 선호도가 정의되는 다섯 번째 방법이다. 여기서 UT‴는 <그림 1>에서의 사용자-요인 매트릭스에, TI‴는 <그림 1>에서의 요인-항목 매트릭스에 해당된다.

$$
pref(u,\ i) = \cos(UT^o_{u.},\ TI_{.i}) \times \frac{\sum_v UI_{vi}}{\sum_v \sum_j UI_{vj}} \qquad (10)
$$

$$
\begin{aligned}
&= \sum_t (UT'''_{ut} \times TI''_{ti} \times \frac{\sum_v UI_{vi}}{\sum_v \sum_j UI_{vj}})\\
&= \sum_t \{UT'''_{ut} \times (TI''_{ti} \times \frac{\sum_v UI_{vi}}{\sum_v \sum_j UI_{vj}})\}\\
&= \sum_t (UT'''_{ut} \times TI''''_{ti}) = UT'''_{u.} \bullet TI''''_{.i}
\end{aligned}
$$

세 번째와 다섯 번째 방법에서는 단순하게 항목별 태그 사용 비율을 선호도 계산에 이용하는데, 이 단계에 항목 기반 협업 필터링과 같은 방법을 이용할 수도 있다. 식 (11)과 같이 사용자 u와 항목 i간의 코사인 유사도가 최소값 mincos보다 클 경우에, 항목 기반 협업 필터링의 방법을 적용하여 항목 i의 이웃들과의 유사도를 더한 값으로 선호도를 계산하며, 그렇지 않을 경우에는 선호도는 0으로 처리하는 것이 여섯 번째 방법이다.

$$
\begin{aligned}
pref(u,\ i) &= \sum_{j \in N_i, UI_{uj}=1} sim(UI_{.i},\ UI_{.j}) \qquad (11)\\
&\quad if \quad \cos(UT_{u.},\ TI_{.i}) > \min\cos
\end{aligned}
$$

4. 실험결과

이 논문에서 사용한 데이터 집합은 잘 알려진 소셜 북마크(Bookmark) 서비스인 del.icio.us에서 수집한 자료로, 2,000명의 사용자가 15,000 항목을 9,364개의 태그를 사용하여 평가한 것이다. 실제 작성된 태그 배정은 98,042개이며, 북마킹(Bookmarking)으로 보면 41,338개로서 희소성(Sparsity)은 99.862%이다. 예측치의 품질을 평가하기 위하여, 이들 중에서 약 80%인 33,104개를 훈련용 데이터로 사용

〈표 1〉 실험결과

방법	매개변수	Total Recall	Average Recall
1. Tags as Latent Factors I	nUsers = 2000	0.007894	0.006974
2. Tags as Latent Factors II	nUsers = 2000	0.0196745	0.020652
3. 방법2 + Item Frequency	nUsers = 2000	0.047122	0.066042
4. Users' Tag Extension	minUserSim = 0.1 nUsers = 2000	0.025868	0.031465
5. 방법4 + Item Frequency	minUserSim = 0.1 nUsers = 2000	0.044814	0.054780
6. 방법2 + Item-Based CF	mincos = 0.1 nUsers = 2000	0.038499	0.041317
7. Ji 등[7]	mincos = 0.1 w = 50 nUsers = 2000	0.034491	0.039992

하였고, 나머지 20%를 테스트 데이터로 사용하였다.

예측의 정확성을 측정하고 평가하기 위해서 이 논문에서는 회상도(Recall)[5]를 사용하는데, 여기서 회상도란 사용자에게 추천된 항목들 중에서 테스트 데이터에 포함된 것들(즉, 실제로 사용자가 이용한 항목들)이 테스트 데이터에서 차지하는 비율이다. 식 (12)와 같은 총 회상도(Total Recall)와 식 (13)과 같은 평균 회상도(Average Recall), 이렇게 두 가지를 사용하여 계산한다. 이 식에서 Test(u)는 테스트 데이터에서 사용자 u에 의해서 태그가 작성된 항목들의 집합이며, TopN(u)는 사용자 u에게 추천된 10개의 항목들의 집합이다. 따라서 총 회상도는 각 사용자 u에게 추천된 항목들 TopN(u) 중에서 테스트 데이터에 포함된 항목들의 개수를 모두 더한 뒤, 이를 모든 사용자가 이용한 항목들의 개수의 합으로 나눈 비율이다. 평균 회상도는 각 사용자별 회상도를 먼저 구한 뒤, 이를 평균한 것이다.

$$\sum |Test(u) \cap TopN(u)| / \sum |Test(u)| \quad (12)$$

$$\sum \{ |Test(u) \cap TopN(u)| / |Test(u)| \} / nTestUsers \quad (13)$$

<표 1>은 이 논문에서 제안한 여섯 가지 방법과 관련 연구에서 제시한 Ji et al.[7]의 방법을 테스트 데이터에 적용하여 얻은 결과이다. 가장 좋은 결과는 세 번째 방법에 의한 것으로, 테스트 데이터에 대하여 총 388번을 맞추어 총 회상도(Recall)는 약 4.7%, 평균 회상도는 약 6.6%이다. 방법 1과 같이 각각의 태그를 잠재 요인으로 간주하여 사용자와 항목에 대한 태그 벡터를 직접 곱하는 것보다는, 이들 벡터간의 코사인 유사도를 이용하는 방법 2가 좋은 결과를 보여주었다. 방법 2와 방법 4에 대해서 추가적으로 항목 빈도수를 고려한 방법 3과 방법 5가 각각 원래 방법보다 우수한 결과를 보여주었는데, 이것은 항목별 태그 사용 비율이 데이터의 희소성을 보

완하는 것으로 판단된다. 방법2에 항목기반 협업 필터링을 추가한 방법 6도 방법 2보다 우수한 결과를 보여 주었지만, 방법 3이나 방법 5보다는 우수하지 못하였다. Ji et al.[7]의 방법은 테스트 데이터에 대하여 총 284번을 맞추어 약 3.4%의 총 회상도를 보여주었다. 이것은 Ji et al.[7]의 방법이 이 논문에서 제안한 방법 1, 2, 4보다는 우수하지만, 방법 3, 5, 6보다는 우수하지 못함을 보여준다.

5. 결 론

구축하는 단계에서 많은 시간과 노력이 필요하나, 일단 구축되고 나면 신속하고 좋은 품질의 추천이 가능한 장점을 가지고 있는 잠재 요인 모델을 협업 태그 데이터에 적용하기 위해서는 모델 수립의 어려움을 극복해야 한다. 이 논문에서는 태그를 잠재 요인으로 보는 자연스럽고 합리적인 관점을 기본으로 잠재요인모델을 협업 태그 데이터에 적용하기 위한 구체적인 방법들을 제안하고, 이를 실재의 소셜 네트워크 사이트의 자료에 적용하여 테스트하였다. 사용자와 항목의 태그 벡터 간 유사도를 가지고 추천하는 방법 2에다 항목의 빈도수를 추가적으로 고려한 방법 3이 가장 우수하였고, 이것은 Ji et al.[7]의 기존 연구보다 우수한 결과를 보여주었다. 소셜 네트워크가 활성화됨에 따라서 태그를 이용한 평가와 활용이 e-비즈니스를 비롯한 다양한 분야에서 점증할 것으로 예상되며, 이런 환경에서 태그를 잠재 요인으로 사용하는 모델이 효율적이고 효과적으로 활용될 수 있을 것이다.

향후 연구 방향으로는 태그의 의미를 보다 정확하게 해석하기 위하여 세분화된 태그들을 그룹화하여 정제함으로써, 태그 데이터 량을 줄이고 예측 품질을 높이는 것이다. 이 과정에서 온톨로지(Ontology)[11]를 적용하는 것을 우선적으로 고려할 계획이다. 항목의 콘텐츠 메타 데이터를 태그의 일종으로 추가하여 고려하는 것[15]도 추천 품질을 개선할 수 있을 것으로 기대되고 있다.

참 고 문 헌

[1] 김형도, “일관성 기반의 신뢰도 정의에 의한 협업 필터링”, 한국전자거래학회지, 제14권, 제1호, 2009, pp. 1-11.

[2] Ahn, H. J., “A New Similarity Measure for Collaborative Filtering to Alleviate the New User Cold-Starting Problem,” Information Sciences, Vol. 128, 2008, pp. 37-51.

[3] Bell, R., Koren, Y., and Volinsky, C., “Chasing $1,000,000 : How We Won the Netflix Progress Prize,” Vol. 18, No. 2, December, 2007, pp. 4-12.

[4] Herlocker, J. L., et al., “An Algorithmic Framework for Performing Collaborative Filtering,” Proceedings of the 22nd Annual International ACM SIGIR Conf. on Research and Development in Information Retrieval, Berkeley, USA, 1999, pp. 230-237.

[5] Herlocker, J. L., et al., "Evaluating Collaborative Filtering Recommender Systems," ACM Transactions on Information Systems, Vol. 22, No. 1, 2004, pp. 5-53.

[6] Jaschke, R., et al., "Tag Recommendations in Folksonomies," Proceedings of the PKDD 2007 (LNAI 4702), 2007, pp. 506-514.

[7] Ji, A.-T., et al., "Collaborative Tagging in Recommender Systems," Proceedings of AI 2007 (LNAI 4830), 2007, pp. 377-386.

[8] Kim, H. -N., et al., "Error-Based Collaboration Filtering Algorithm for Top-N Recommendation," Proceedings of APWeb 2007 and WAIM 2007 (LNCS 4505), pp. 594-605, Huang Shan, China, June, 2007, pp. 16-18.

[9] Konstan, J., et al., "GroupLens : Applying Collaborative Filtering to Usenet News," Communications of the ACM, Vol. 40, No. 3, 1997, pp. 77-87.

[10] Lee, H. C., Lee, S. J., and Chung Y. J., "A Study on the Improved Collaborative Filtering Algorithm for Recommender System," Proceedings of the 5th International Conference on Software Engineering Research, Management and Applications (SERA2007), 2007, pp. 297-304.

[11] Mika, P., "Ontologies Are Us : A Unified Model of Social Networks and Semantics," Proceedings of the ISWC 2005 (LNCS 3729), 2005, pp. 522-536.

[12] Nakamoto, R., et al., "Tag-Based Contextual Collaborative Filtering," Proceedings of DEWS2007, Hiroshima, Japan, 2007, pp. 25-30.

[13] Netflix, "Netflix Prize," http://www.netflixprize.com//index, 2006.

[14] Paterek, A., "Improving Regularized Singular Value Decomposition for Collaborative Filtering," Proceedings of the KDD Cup and Workshop, San Jose, California, USA, August 12, 2007.

[15] Rhie, B. W., Kim, J. W., and Lee H. J., "Methods of User-Created Content Recommendation with Content Metadata," Proceedings of the Asian e-Biz Workshop, 2008.

[16] Sarwar B., et al., "Analysis of Recommendation Algorithms for e-Commerce," Proceedings of the 2nd ACM Conf. on Electronic Commerce, Minneapolis, USA, 2000, pp. 158-167.

[17] Sarwar, B., Karypis, G., Konstan, J., and Reidl, J., "Item-Based Collaborative Filtering Recommendation Algorithms," Proceedings of the 10th Int'l WWW Conf., Hong Kong, May 1-5, 2001, pp. 285-295.

[18] Shardanand, U., and Maes, P., "Social Information Filtering : Algorithms for Automating 'Word of Mouth'," Proceedings of the ACM CHI Conf. on Human Factors in Computing Systems, Denver, USA, May 1995, pp. 210-217.

[19] Tso-Sutter, K. H. L., Marinho, L. B., and Schmidt-Thieme, L., "Tag-aware Recommender Systems by Fusion of Collaborative Filtering Algorithms," Proceedings of the SAC'08, Brazil, March 2008.

저 자 소 개

김형도	(E-mail : hdkim@hycu.ac.kr)
1985년	서울대학교 산업공학과 (학사)
1987년	한국과학기술원 경영과학과 (석사)
1992년	한국과학기술원 경영과학과 (박사)
1993년~1999년	㈜데이콤 EC인터넷연구/기술 팀장
2000년~2002년	아주대학교 정보통신전문대학원 교수
2002년~현재	전자상거래표준화통합포럼 전자문서기술위원회 부위원장
2003년~현재	한양사이버대학교 경영학부 교수
2004년~2006년	ebXML 전문위원회 위원장
관심분야	전자상거래, 비즈니스 프로세스, 디지털 워터마킹, e-러닝, 데이터 마이닝 등