추천 성능 향상을 위한 사용자별 가중치 자동 설정 기법*

이성진°, 이연정, 이수원
숭실대학교 컴퓨터 학과
ptnrev93@mining.ssu.ac.kr°, zzeong81@hanmail.net, swlee@ssu.ac.kr

An Autonomic User-Dependent Weighting Method to Improve Efficiency of Recommendation

SeongJin Lee°, YounJeong Lee, Soowon Lee
Dept. of Computing, Graduate, School, Soongsil University

요 약

추천 기술이란 과도하게 제공되는 정보를 여과하여 사용자에게 필요한 정보만을 제공해 주는 것으로 대표적으로는 협력적 여과가 있다. 그러나 협력적 여과는 희소성 문제와 확장성에 취약점을 보이고 있어 최근 이를 극복하기 위한 내용 기반 추천 기법에 관한 연구가 활발히 이루어지고 있다.

내용 기반의 추천 기법에서 효율적인 추천이 이루어지기 위해서는 각 요소별 가중치를 어떻게 설정할 것인가가 매우 중요하다. 기존의 연구에서는 요소별 가중치를 다양한 실험에 의해 결정하고 이를 모든 사용자에게 동일하게 적용하는 방식을 취하고 있다. 그러나 사용자마다 콘텐츠 선택 기준과 요인이 다를 수 밖에 없으므로 이러한 방식은 사용자의 선호 정보를 효과적으로 반영할 수 없다. 따라서 본 논문에서는 사용자의 선호 정보 분석과 함께 각 요소별 가중치를 사용자별로 자동으로 설정하여 보다 효과적인 추천이 이루어질 수 있는 기법을 제안한다.

1. 서 론

인터넷과 네트워크의 발전과 함께 일상적으로 접할 수 있는 정보의 양도 기하급수적으로 증가하면서 사용자들에게는 정보 선택의 문제가 대두되었다. 즉, 자신에게 필요한 정보를 검색하기 위해서는 더 많은 노력과 시간이 요구되고 있다. 따라서 사용자에게 필요한 정보를 제공해 주는 추천 기술에 관한 다양한 연구가 이루어지고 있다.

추천 기술이란 과도하게 제공되는 정보를 여과하여 사용자에게 필요한 정보만을 선별하여 제공하여 주는 것으로 IBM Watson 연구소를 중심으로 마이닝 분야에서 활발히 연구되어 왔다. 특히, 웹 환경에서의 추천 기술은 사용자의 웹 브라우징 히스토리, 웹 검색 히스토리, 웹 컨텐츠 히스토리 등으로부터 사용자의 취향을 분석하여 사용자가 관심을 가질만한 내용, 혹은 사용자에게 필요한 정보나 서비스를 제공하는 기술로 아마존(amzon.com) 등에서 상용화하여 사용되고 있다.

현재 상용화되어 사용되고 있는 대표적인 추천 기술로는 협력적 추천(Collaborative Recommendation)을 들 수 있으나 희소성과 확장성에 취약점을 가지고 있기 때문에[1] 이를 극복하기 위한 방안으로 내용기반 추천(Content-Based Recommendation)에 관한 연구가 활발히 진행되고 있다.

내용 기반 추천은 추천 대상이 되는 아이템 구조와 사용자 선호도 정보의 유사도를 계산하고 각 내용에 적절한 가중치를 적용하여 선호도를 계산한다. 내용 기반 추천에서는 각 내용에 적용하는 가중치에 따라 추천 결과가 달라지게 된다. 따라서 내용별 가중치를 어떻게 설정할 것인가가 중요한 문제가 된다[2][3].

기존의 연구 방식에서는 이를 위해 반복적인 실험을 통해 최적의 가중치를 찾고 이를 모든 사용자에게 적용하는 방식을 이용하고 있다. 그러나 이러한 방식은 사용자별 선택 기준을 효과적으로 반영하지 못한다. 예를 들어 일반적으로 배우나 감독보다는 장르가 영화 선택에 중요한 영향을 미친다는 것이 알려져 있으므로 장르에 대한 가중치를 가장 높게 설정했다고 하자. 그러나 장르보다는 자신이 좋아하는 배우가 출연하는 영화를 선호하는 다수의 사용자에 대해서는 효과적인 추천이 이루어질 수 없다.

따라서 본 논문에서는 사용자의 선호 정보 분석과 함께 각 내용에 대한 가중치를 자동으로 설정하여 각 사용자가 중요하게 생각하는 요소에 대한 가중치를 높여 줌으로써 보다 효과적인 추천이 이루어질 수 있는 기법을 제안한다.

본 논문의 2장에서는 추천과 관련된 대표적인 기술에 대해 알아보고 3장에서는 추천 성능 향상을 위한 사용자별 가중치 자동 설정 기법에 대해 설명한다. 4장에서는 본 논문에서 제안한 기법의 효율성을 검증하기 위한 실험 결과를 제시하며 5장에서는 결론 및 향후 연구 과제를 서술한다.

2. 관련 연구

2.1 협력적 추천(Collaborative Recommendation)

협력적 추천은 가장 대표적인 개인화 추천 기술로 일반적으로는 GroupLens[4]에서 처음 제안된 최근접 이웃 방법을 사용한다. 최근접 이웃 방식은 추천의 대상이 되는 목표 사용자에 대하여 가장 유사한 기호를 가지는 k명의 사용자를 선택하여 이들의 선호도를 고려하여 아이템을 추천하는 방식이다.

협력적 추천은 사용자간의 유사도를 계산하는 과정과 사용자에게 추천할 아이템에 대한 선호도를 예측하는 과정으로 나눌 수 있다. 일반적으로 사용자간의 유사도 계산은 코사인 유사도나 피어슨 상관 계수를 이용하며 다음의 [식 1]과 [식 2]는 각각 코사인 유사도와 피어슨 상관 계수를 구하는 식이다.

* 본 연구는 과학기술부 특정기초사업의 연구비 지원으로 이루어졌습니다.

$$sim(a,u) = cosine(a,u) = \frac{\vec{r_a} \cdot \vec{r_u}}{|\vec{r_a}||\vec{r_u}|}$$

[식 1] 코사인 유사도 계산식

$$sim(a,u) = corr(a,u) = \frac{\sum_{j \in Items}(R_{a,j} - \overline{R_a}) \times (R_{u,j} - \overline{R_u})}{\sqrt{\sum_{j \in Items}(R_{a,j} - \overline{R_a})^2 (R_{u,j} - \overline{R_u})^2}}$$

[식 2] 피어슨 상관 계수에 의한 유사도 계산식

위의 [식 1]에서 $\vec{r_a}$ 와 $\vec{r_u}$ 는 각각 사용자 a와 사용자 u가 각 아이템에 부여한 선호도를 벡터로 표현한 것이며 [식 2]에서 $R_{a,j}$ 는 사용자 u가 아이템 j에 부여한 선호도, $\overline{R_a}$, $\overline{R_u}$ 는 각각 사용자 a와 u가 부여한 선호도 값의 평균을 의미한다.

[식 1]과 [식 2]에 의해 목표 사용자와의 유사도가 가장 높은 k명의 사용자를 선택한 후 다음의 [식 3]을 이용하여 목표 아이템 i에 대한 선호도를 예측 할 수 있다.

$$P_{a,i} = \overline{R_a} + \frac{\sum_{u=1}^{k}(R_{u,i} - \overline{R_u}) \times sim(a,u)}{\sum_{u=1}^{k} sim(a,u)}$$

[식 3] 선호도 예측식

협력적 추천의 대표적인 문제는 희소성 문제와 확장성 문제로 희소성 문제란 사용자들의 각 아이템에 대한 선호 정보가 희소할 경우 추천의 정확도가 떨어진다는 것이며 확장성 문제란 사용자의 수가 많아질수록 유사도 계산하는 시간이 오래 걸린다는 것이다. 또한 선호 정보가 전혀 없는 새로운 사용자나 아이템에 대한 추천은 이루어질 수 없다.

2.2 내용기반 추천(Content-Based Recommendation)

내용 기반 추천은 사용자 입력한 내용이나 이전에 선호했던 것을 추천의 대상이 되는 아이템을 비교하여 유사도가 높은 아이템을 추천하여 주는 방식이다. 내용 기반 추천은 정보 검색(Information Retrieval)에 기반을 두고 있으며 일반적으로 정보 검색에서 사용하는 가중치 기법을 이용한다. 가중치 기법이란 추천 대상이 되는 아이템 구조와 사용자 선호도 정보의 구조의 유사도를 계산하고 각 내용에 적절한 가중치를 적용하는 것을 말한다. 따라서 내용 기반 추천에서는 추천의 대상이 되는 아이템의 구조와 사용자 선호도 정보의 구조가 일치해야한다. 예를 들어 영화 추천의 경우 사용자는 장르, 감독, 배우, 줄거리 등에 대한 선호 정보를 가지고 있어야 하며 마찬가지로 영화도 장르, 감독, 배우, 줄거리에 관한 정보를 가지고 있어야 상호간에 유사도 측정이 가능하다. 즉, 장르, 감독, 배우, 줄거리 간의 유사도를 계산한 후 각각의 유사도에 적절한 가중치를 곱한 값의 합으로 유사도를 계산하는 방식이다.

내용기반 추천은 각 아이템이 충분한 정보를 포함해야 하며 사용자에 의한 FeedBack이 명확할수록 높은 추천 효과를 기대할 수 있지만 사용자에게 불편함을 주는 단점이 있다. 또한 각 내용에 대한 적절한 가중치 설정에 따라 추천의 결과 및 효과가 달라지므로 사용자별로 적절한 가중치를 찾는 것이 문제가 된다.

3. 사용자별 가중치 자동 설정 기법

내용 기반 추천에서는 사용자로부터 FeedBack을 받는 것을 가정하는 경우가 일반적이다. 그러나 실제 상황에서 사용자로부터 명확한 FeedBack을 받기가 쉽지 않기 때문에 본 논문에서 제안하는 사용자별 가중치 자동 설정은 사용자에 의한 FeedBack없이 아이템 선택 히스토리만을 분석하여 각 내용별로 사용자의 아이템 선택에 영향을 미치는 중요도를 계산하는 방법을 이용한다.

사용자별 아이템 선택 히스토리에서 각 내용별로 가장 많이 출현하는 항목의 출현 확률을 비교해 본다면 아이템 선택에 중요한 영향을 미치는 내용의 경우 특정 항목의 출현 확률이 대단히 높게 나타날 것이다. 다시 말해 각 내용별로 가장 많이 출현한 항목의 확률이 높을수록 그 항목은 사용자의 아이템 선택에 중요한 영향을 미친다고 할 수 있다. 예를 들어 사용자 A는 장르의 최대 항목 출현 빈도가 0.8이고 감독이나 배우의 경우 0.2, 0.3값을 갖는다면 장르가 사용자 A의 영화 선택에 있어 가장 중요한 내용이라는 것이다. 따라서 사용자가 이용한 아이템들 중에서 각 내용별로 가장 많이 출현한 항목의 확률값의 정규화를 통해 중요도를 계산하고 이를 각 내용별 가중치로 이용한다.

3.1 내용별 중요도 계산

사용자별 가중치 자동 설정을 위해 우선 각 내용에 대한 중요도를 계산한다. 위에서 설명한 바와 같이 각 내용에 대한 중요도는 각 내용의 최대 출현 항목의 확률값을 이용한다. 그러나 같은 확률값을 가지더라도 몇 개의 항목을 가지느냐에 따라 중요도는 달라진다. 즉, 확률값이 똑같이 0.6이라 하더라도 항목 수가 3개인 경우보다는 항목 수가 6개인 경우에 그 중요도가 높다고 할 수 있다. 따라서 다음의 [식 4]와 같이 최대 출현 항목의 확률값과 항목 수를 곱한 값들의 정규화 결과를 중요도로 계산한다.

$$Importance_{u,i} = \frac{MaxP_{k=1}^{Num(i)}(C_{u,i,k}) \times Num(i)}{\sum_{j=1}^{n} MaxP_{k=1}^{Num(j)}(C_{u,j,k}) \times Num(j)}$$

[식 4] 중요도 계산식

[식 4]에서 u는 각 사용자, i와 j는 아이템을 구성하는 내용, n은 내용의 수를 의미한다. 또한 Num(i)는 i에서 나타날 수 있는 항목의 수를 나타내며, $C_{u,i,k}$는 사용자 u의 선택 아이템 중에서 내용 i를 구성하는 항목을 의미한다.

3.2 가중치 조정

협력적 추천이나 내용 기반 추천의 공통적인 문제는 사용자의 선호 정보가 충분하지 못한 추천 초기에는 충분한 효과를 기대하기 어렵다는 점이다. 이는 본 논문에서 제안하는 사용자별 가중치 자동 설정 기법에도 해당되는 문제이다. 따라서 히스토리가 충분하지 못한 초기에는 모든 사용자에게 동일한 가중치를 적용하고 히스토리가 갱신될 때 마다 가중치를 조정하는 방식을 사용한다. 이러한 동적 갱신의 장점은 사용자별로 특색있는 선택 패턴과 일반적인 사용자들의 선택 패턴을 함께 반영할 수 있다는 장점이 있다. 예를 들어 영화의 경우 장르가 영화 선택에 중요한 영향을 미치는 다는 것이 알려져 있으므로 초기 설정에서 장르에 대한 가중치를

보다 높게 설정하고 사용자별 히스토리 분석을 통해 이를 갱신하는 것이다. 이때 가중치 조정값은 각 내용별 중요도의 평균값과의 차이를 이용한다.

다음의 [식 5]는 가중치 조정값을 계산하는 과정을 보여주며 [식 6]을 통해 각 사용자별 가중치를 갱신한다.

$$UpdateWeight_{u,i} = Importance_{u,i} - Mean\_Importance_{u}$$

[식 5] 가중치 조정값 계산식

$$Weight_{u,i} = Weight_{u,i} + UpdateWeight_{u,i}$$

[식 6] 가중치 갱신 계산식

[식 5]와 [식 6]에서 $UpdateWeight_{u,i}$는 사용자 u에 대한 내용 i의 가중치 조정값을 의미하며 $Mean_Importance_{u}$는 사용자 u에 대한 Importance값의 평균이다.

4. 실험 결과

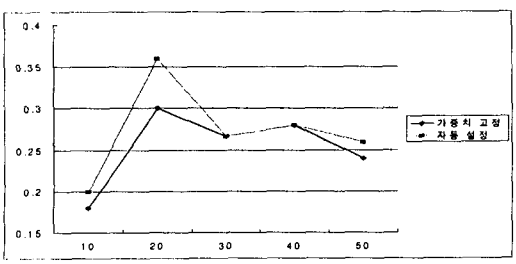
본 논문에서는 H사의 실제 영화 예매 데이터를 이용하여 실험을 진행하였다. 영화 예매 횟수가 높은 상위 10명의 사용자에 대해 예매일이 가장 늦은 영화 10개를 테스트 데이터로 활용하였다. 선호도 정보를 계산하기 위한 훈련 예제의 수를 10개, 20개, 30개, 40개, 50개로 구분하며 각각의 경우에 사용자에게 5개의 영화를 추천하였으며 타당성을 확인하기 위한 척도로는 추천한 아이템 중에서 실제 사용자가 선택한 아이템의 비율을 나타내는 Precision을 이용하였다. 실험을 위해 사용한 영화의 정보는 제작국가, 장르, 감독, 배우, 키워드의 벡터로 표현된 줄거리이다. 이때 줄거리의 중요도는 코사인 유사도의 평균을 이용하였다. 본 논문에서 제안한 기법은 초기 가중치 설정에 별다른 영향을 받지 않는다는 것을 확인하기 위해 다음의 [표 1]에서와 같이 세 가지 경우로 초기 가중치를 설정하여 실험을 진행하였다.

[표 1] 초기 가중치 설정 값

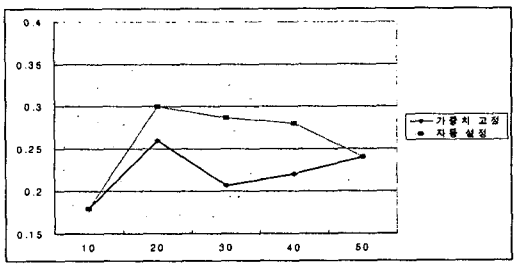
	국가	장르	감독	배우	줄거리
경우1	0.2	0.2	0.2	0.2	0.2
경우2	0.05	0.3	0.15	0.25	0.25
경우3	0.05	0.45	0.1	0.2	0.2

본 논문에서 각 항목에 대한 선호도는 항목별 출현 빈도의 최대값을 이용한 정규화를 통해 계산하였다.

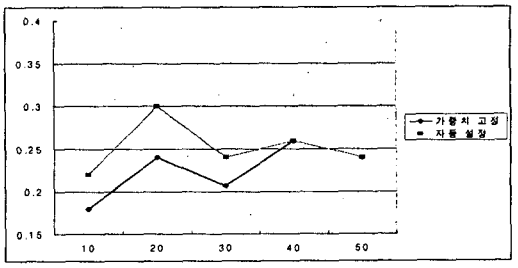
다음의 [그림 1], [그림 2], [그림 3]은 초기 가중치 설정의 세 가지 경우에 대한 실험 결과로 대부분의 경우 사용자별 가중치를 다르게 설정하는 것이 보다 좋은 추천 효과를 나타내고 있다. 또한 초기 가중치 설정과 관계없이 본 논문에서 제안하는 방식의 효과가 더 좋게 나타나고 있음을 확인 할 수 있다.



[그림 1] 초기 가중치 설정 1의 실험 결과



[그림 2] 초기 가중치 설정 2의 실험 결과



[그림 3] 초기 가중치 설정3의 실험 결과

5. 결론 및 향후 연구

본 논문에서는 내용 기반 추천에서 실험을 통해 찾아진 가중치를 모든 사용자에게 동일하게 적용함으로써 사용자의 선호 정보가 효과적으로 반영되지 못한다는 점을 극복하기 위해 아이템 선택 히스토리 분석을 통해 사용자별 가중치를 자동으로 설정하는 방식을 제안하였다. 사용자별 가중치의 자동 설정은 각 내용별로 가장 많이 출현하는 항목의 확률값의 정규화를 이용하였으며 영화 추천의 실험을 통해 타당성을 증명하였다.

그러나 만약 내용간의 연관성을 이용한다면 좀 더 나은 효과를 기대할 수 있을 것이다. 따라서 가중치 설정에 있어 내용간의 연관성을 활용할 수 있는 연구와 사용자 히스토리가 충분하지 못한 초기 추천의 어려움을 해결할 수 있는 연구가 필요하다.

참고문헌

[1] Sarwar, B., Karypis, G., Konstan, J., Riedl J., "Item-based Collaborative Filtering Recommender Algorithms", Accepted for publication at the WWW10 Conference, 2001

[2] Marko Balabnovic and Yova Shoham, "Fab:Content-Based, Collaborative Recommendation", CACM 40(3) p66-72, 1997

[3] 유상원, "내용 기반 추천 기법의 TV 환경 적용에 관한 연구", 한국정보과학회학술발표논문집, 2483호, p797-799, 2003

[4] Resnick,P., Iacovou,N., Suchak,M., Bergstrom,P., and Riedl,J., "GroupLens:An Open Architecture for Collaborative Filtering of Netnews", In proceddings of CSCS '94, Chapel Hill, NC, 1994

[5] 윤현호,이성진,강영길,이수원,김현, "사용자 선호도 기반의 TV 프로그램 추천 기법", KCC2005학술발표논문집,VOL. 32 NO. 01 pp.0730 ~ 0732, 2005