DOI: 10.3745/KIPSTD.2009.16-D.2.273

신규 사용자 추천 성능 향상을 위한 가중치 기반 기법

조 성 훈[†] · 이 무 훈[††] · 김 정 석[††] · 김 봉 회[†††] · 최 의 인[††††]

요 약

오늘날 컴퓨팅 환경의 진보와 웹의 이용이 활발해짐에 따라 오프라인에서 이루어졌던 있었던 많은 서비스들과 상품의 제공이 웹에서 이루어지고 있다. 이러한 웹 기반 서비스 및 상품은 개인에 적합하게 취사선택되어 제공되는 추세이다. 이렇듯 개인에 적합한 서비스 및 상품의 선택과 제공을 위한 패러다임을 개인화(personalization)라 한다. 개인화된 서비스 및 상품의 제공을 위한 분야로서 연구된 것이 추천(recommendation)이다. 그러나 이러한 추천 기법들은 신규 사용자에게 적합한 추천을 제공하지 못하는 문제와 사용자의 상품에 대한 평점에만 의존하여 추천을 생성한다는 계산 공간에서의 제약 사항을 가지고 있다. 두 문제 모두 추천 분야에서 지속적인 관심을 보이는 분야로서 신규 사용자 추천 문제의 경우는 신규 사용자의 평점이 없기 때문에 유사 사용자들을 분류할 수 없음에 기인한다. 그리고 추천 공간 제약에 따른 문제는 추천 차원의 추가에 따른 처리 비용이 급격히 증가한다는 문제를 가지고 있기 때문에 쉽게 접근하기 어렵다. 따라서 본 논문에서는 신규 사용자 추천 향상을 위한 기법과 평점 예측 시 예측에 대한 가중치를 적용하는 기법을 제안한다.

키워드 : 추천, 콜드스타트, 협력 필터링, 하이브리드 필터링, 가중치

Weight Based Technique For Improvement Of New User Recommendation Performance

Sun-Hoon Cho[†] · Moo-Hun Lee[††] · Jeong-Seok Kim[††] · Bong-Hoi Kim[†††] · Eui-In Choi[††††]

ABSTRACT

Today, many services and products that used to be only provided on offline have been being provided on the web according to the improvement of computing environment and the activation of web usage. These web-based services and products tend to be provided to customer by customer's preferences. This paradigm that considers customer's opinions and features in selecting is called personalization. The related research field is a recommendation. And this recommendation is performed by recommender system. Generally the recommendation is made from the preferences and tastes of customers. And recommender system provides this recommendation to user. However, the recommendation techniques have a couple of problems; they do not provide suitable recommendation to new users and also are limited to computing space that they generate recommendations which is dependent on ratings of products by users. Those problems has gathered some continuous interest from the recommendation field. In the case of new users, so similar users can't be classified because in the case of new users there is no rating created by new users. The problem of the limitation of the recommendation space is not easy to access because it is related to moneywise that the cost will be increasing rapidly when there is an addition to the dimension of recommendation. Therefore, I propose the solution of the recommendation problem of new user and the usage of item quality as weight to improve the accuracy of recommendation in this paper.

Keywords : Recommendation, Cold Start, Collaborative Filtering, Hybrid Filtering, Weight

1. 서 론

추천 시스템(recommender system)은 사용자가 요구하는 것이 무엇인지 분석하여 이를 토대로 상품이나 서비스를 찾으며, 사용자의 요구를 만족시키는 상품을 제공하는 역할을 수행한다. 이를 위해 보다 정확하고 효과적으로 수행하기 위한 연구가 1990년대 중반부터 활발히 이루어지고 있으며, 추천 시스템에 적용하기 위한 연구는 협력적 필터링 기법에서부터 시작되었다[1].

협력적 필터링 기법(collaborative filtering approach)의 기본적인 아이디어는 사용자와 비슷한 취향을 가진 다른 사용자들이 선호하는 것은 사용자 역시 선호할 확률이 높은 점에 착안한 것이며, 이를 이용해서 사용자가 이전에 평가

※ 본 논문은 2008년도 한남대학교 교비학술연구비의 지원에 의한 연구결과임. 본 연구는 교육과학기술부와 한국산업기술재단의 지역혁신인력양성사업으로 수행된 연구결과임.
† 정 회 원 : 한국원자력연구소 하나로운영부 연구원
†† 준 회 원 : 한남대학교 컴퓨터공학과 박사과정
††† 정 회 원 : (주)유비엔씨 대표이사
†††† 종신회원 : 한남대학교 컴퓨터공학과 교수
논문접수 : 2009년 3월 6일
수 정 일 : 1차 2009년 3월 30일
심사완료 : 2009년 3월 31일

한 경험이 없는 상품에 대한 추천을 수행한다. 따라서 협력적 필터링 기법은 사용자들의 평가 내역을 통해서 사용자와 비슷한 관심사를 갖는 사용자(또는 이웃)들을 찾고, 유사 사용자들이 상품에 대해 평가한 정보를 토대로 상품을 필터링하는 과정이 필요하다.

이러한 협력적 필터링 기법을 적용한 기존 연구들로 Group Lens 시스템[2], Tapestry 시스템[3], Ringo 시스템[4], MovieLens 시스템[5], Jester 시스템[6] 등이 있다. 적극 활용되고 있다.

이들 기법이 추천을 위해 고려하는 사항은 사용자의 특징, 상품들의 특징, 사용자 간의 유사도, 사용자에 의한 상품의 평점 등이지만, 이들 기법은 상품 자체의 품질은 고려하지 않고 있다. 만약 상품 자체의 품질이 추천을 위한 기준으로서 제공될 수 있다면 사용자들의 요구에 보다 근접해질 수 있게 된다.

따라서 본 논문에서는 전술한 신규 사용자 추천 문제를 해결하기 위해 평가 항목을 선정하고 이를 피드백 받아 초기 평가 집합을 생성하는 기법과 정확한 추천을 제공하기 위해 상품 품질을 가중치로 가중하는 기법을 제안한다. 그리고 웹 환경에서 정확한 추천을 위해 상품 품질을 가중치로 적용하는 기법과 신규 사용자에 대한 추천을 위해 초기 평가 집합을 구성하기 위한 기법에 대한 연구를 병행하였으며, 이 연구를 통해 기존 웹 환경에서 운용되는 추천 시스템에서 추천을 위해 평가하는 2차원 추천 공간에 가중치를 적용함으로서 다차원적인 요소들을 고려한 추천이 가능하도록 하였다.

2. 추천 기법

2.1 추천 기법

추천 시스템은 다양한 데이터에 기반하여 사용자의 선호를 분석하고 이에 적합한 추천을 제공한다. 이러한 추천 시스템의 근원은 정보검색과 인공지능에 있다[7, 8]. 추천 시스템은 서비스 또는 상품의 특성 데이터, 사용자의 특성 데이터나 프로파일, 사용자와 서비스 간의 관계와 같은 정보에 기반하여 추천을 생성한다. 이런 추천의 생성을 위한 기법으로 내용기반 필터링, 협력적 필터링, 하이브리드 기법, 다차원 추천 기법 등이 있다[9, 10].

2.1.1 내용기반 추천기법

내용기반 추천기법은 단순한 키워드 비교를 통한 추천에서 진보하여 사용자의 취향과 선호를 포함하는 사용자 프로파일을 통한 추천에 이르게 되었다. 사용자 프로파일 정보는 사용자로부터 명시적으로나 묵시적으로 도출할 수 있는데, 이 프로파일은 사용자의 특성을 나타내는 키워드들로 구성된다. 따라서 내용기반 추천 기법은 이 키워드에 대한 가중치를 계산함으로서 추천을 생성하는데, 키워드 가중치 계산을 위해 사용되는 대표적인 기법이 TF/IDF(Term Frequency/Inverse Document Frequency)이다[11].

TF/IDF는 문서에 많이 출현하면서 관련성 있는 키워드의 가중치를 구하여 이를 통해 사용자에게 추천을 제공할 수 있도록 한다.

내용기반 추천 기법은 Bayesian classifiers를 비롯한 다양한 기계학습(machine-learning) 기술들도 사용된다[12, 13]. 그러나 내용기반 추천 기법에서 이 기법들은 휴리스틱 공식에 기초하지 않고 유용도 예측(utility prediction)을 계산하기 때문에 정보검색과는 차이가 있다. 이 기법들은 통계 학습과 기계 학습 기술을 이용하여 기반이 되는 데이터로부터 학습한 모델(model)에 기반한다고 볼 수 있다.

내용기반 추천기술은 여러 가지 문제를 포함하고 있다. 먼저 추천하는 상품과 명시적으로 관련된 특징들로 제한된다는 것이다. 따라서 이러한 특징 집합을 가지기 위해서는 컨텐츠가 컴퓨터에 의해 자동적으로 해석될 수 있는 형태이거나, 상품에 있는 특징을 직접 해석할 수 있어야 한다. 정보검색 기술들은 텍스트 문서들에서 특징들을 추출하는데 잘 동작하지만, 다른 도메인들은 자동적인 특징 추출이 가지는 근본적인 문제를 가지므로 자동 특징 추출 방법은 멀티미디어 데이터에 적용하기는 어렵다[17].

2.1.2 협력적 필터링 기법

협력적 필터링은 내용기반 추천기법과는 다르게 이전에 다른 사용자에 의해 상품 평가에 기초하여 특정 사용자의 상품 평가를 예측한다.

즉 사용자가 어떤 상품에 대한 평가(평점)를 예측하는데 있어서 다른 사용자들의 평가를 기반으로 예측한다는 것이다. 그러면 협력적 추천기법에서 모든 사용자들을 고려하여 상품에 대한 추천을 예측하는 것인지에 대한 의문이 생긴다. 만약 협력적 추천기법이 모든 사용자를 예측을 위해 고려한다면 그 정확성은 현저하게 낮아질 것이다. 이러한 이유로 협력적 추천기법은 특정 사용자와 유사 취향의 사용자들을 찾고, 이 사용자들의 평가에 기반하여 예측을 수행한다. 예를 들면, 어떤 사용자가 영화를 보려고 한다고 했을 때, 이 사용자가 이전에 보았던 영화 내역을 추천 시스템이 분석하고 유사한 취향을 가진 사용자들을 선정하고 이 유사 사용자들의 평가에 기초해서 아직 사용자가 보지 않은 영화에 대한 예측을 수행한 뒤에 전체 영화 리스트 중에서 가장 적합한 영화를 추천하는 것이다. 물론 이전에 사용자가 보았던 영화는 배제가 될 것이다. 이렇게 유사 사용자를 찾고 이 유사 사용자들의 평가에 기초한 예측의 수행이 협력적 추천기법의 기본적인 처리 방식이다.

이러한 협력적 추천 시스템은 Grundy 시스템Tapestry 시스템[14], GroupLens, Ringo, Video Recommender Amazon.com, PHOASK 시스템, Jester 시스템 등 많은 시스템에 적용되었다.

2.1.3 하이브리드(Hybrid Approach)

몇 몇 추천 시스템들은 협력적 필터링 기법과 내용 기반 필터링 기법을 결합함으로서 만들어진 하이브리드 방법을

사용한다. 이 방법은 협력적 필터링 기법과 내용기반 필터링 기법이 가지는 한계를 벗어나도록 도와주며, 협력적 필터링 기법과 내용 기반 필터링 기법을 통합하는 방법은 아래처럼 분류될 수 있다[15].

- 협력적 필터링 기법과 내용 기반 필터링을 각각 실행하고 예측 평점들을 결합
- 협력적 필터링 기법에 내용 기반 필터링 기법의 특성 중 일부를 통합
- 내용 기반 필터링 기법에 협력적 필터링 기법의 특성 중 일부를 통합
- 협력적 필터일 기법과 협력적 필터링 기법 모두의 특성을 통합하여 하나의 단일 모델로 구성

전통적인 추천 시스템의 한계를 다루고 추천 정확도를 개선하기 위해서 사례기반 추론 같은 지식기반 기술에 의한 하이브리드 추천 시스템들이 증가하고 있다.

2.1.4 다차원 추천 기법

추천 공간을 다차원으로 확장해야 한다는 주장이 제안되었다[16]. 이 논문에서 주장하는 것은 기존 2차원 추천 공간에 기초하여 평가되지 않은 상품에 대한 예측을 수행하는 것은 사용자에게 만족스럽지 못하기 때문에, 보다 적합한 예측을 수행하기 위해서는 사용자 × 아이템의 2차원 공간 이외의 추가적인 차원이 포함되어야 하며, 이러한 다차원 추천 공간을 통해 예측이 수행되어야 한다는 것이다. 이러한 요구에 의해서 Gediminas는 다차원 추천 모델을 제안하였다[16].

다차원 추천 시 어떤 차원을 추천에 이용할 것인가는 매우 중요한 문제이다. 이 문제는 데이터 마이닝(data mining)과 통계에서 다뤄졌던 것으로서 특성 선택 문제와 관련 있다. 문제를 이해하기 위해서, 두 개의 값 X=h, X=t를 가지는 하나의 속성 차원 X의 경우를 생각해보자. X=h와 X=t에 대한 평점의 분포(distribution)가 같다면, 차원 X는 추천을 위해 고려될 필요 없다. 실제로 추천에 영향을 주지 않는 차원들은 추천 공간에서 배제되어도 추천에 영향을 미치지 않는다. 그리고 이러한 차원은 추천 공간에서 없어져야 오히려 타당하다.

추천에서 근본적인 문제인 알지 못하는 상품의 평가에 대한 예측은 다차원 추천에서도 마찬가지이다. 전통적인 추천 시스템에서처럼, 다차원 시스템에서의 주요 문제는 평점에 대한 다차원 큐브에서 사용자가 지정한 평점으로 부터 평점을 추정하는 것이다. 그리고 다차원 중 추천에 사용할 차원을 선택하는 것이다. 바꾸어 말하면, 추천 공간에서 배제할 추천 공간을 찾아내서 예측 계산을 위한 추천 공간에서 배제하는 것으로, 이렇게 추천 공간을 축소하는 기법으로서 제거 기반(reduction based) 기법이 있다.

제거 기반 기법은 다차원 추천의 문제를 2차원 User × Item 추천 공간으로 감소시키므로, 2차원 추천 시스템에 대한 이전의 모든 연구가 직접적으로 다차원 추천에 적용할 수 있다는 것이다. 그러나 다차원 추천기법은 이러한 차원 감소를 위한 비용이 크다는 단점이 있다.

3. 상품 품질(item quality) 기반의 가중치

이 장에서는 상품 품질을 가중하는 추천 기법과 협력적 필터링 기법의 신규 사용자 추천 문제 해결 기법에 대해 제안과 상품 품질에 관련된 개념들에 대해 기술한다.

3.1 추천 가중치 상품 품질

3.1.1 상품 품질

기존 추천 기법들에서는 상품의 특징 집합을 평가 요소로서 이용하여 추천을 수행하는데 이러한 평가 요소에 상품 품질이 포함되지 않는 것은 상품에 대한 품질을 사용자로부터 일일이 피드백 받기 어렵기 때문이다.

상품 품질은 상품의 품질이 높다면 그 상품을 좋아하는 사용자가 많을 것이고, 그 상품이 빠르게 사용자들 사이에서 인지된다는 것을 개념화한 것이다. 따라서 본 논문에서 이러한 가능성을 추천 시스템에 가중치로 적용하여 그 효율을 높이고자 한 것이다.

상품 품질은 주어진 상품 i를 한 사용자가 좋아할 확률, $Q(i)$,이다. 일반적으로 인기가 높은 상품의 품질은 상품을 구매한 사용자에게 있어서 대체로 만족되기 때문에 상품에 대한 품질(Q)과 인기도(P) 사이에는 비례 관계가 있으며, 사용자들이 많이 찾는 상품일수록 그 상품의 품질도 좋다. 따라서 상품에 대한 인지도(A)와 품질 간에도 비례적 관계가 있다. 이와 같이 상품에 대한 품질, 인기도, 인지도 간에 존재하는 관련성을 토대로 상품 품질을 구할 수 있다. 즉 상품에 대한 사용자가 느끼는 품질은 얼마나 많은 사람들에게 알려졌으며, 얼마나 선호되고 있는가를 통해 계산될 수 있으며, 이를 식으로 표현하면 (식 1)과 같다.

$$Q(i) = P(i) \times A(i) \tag{1}$$

인기도는 상품 품질이 높으면 그것을 선택할 확률이 높아진다는 사실을 이용하기 때문에 상품이 빈번히 사용자들에 의해 선택된다면, 그 상품의 품질은 높다고 말할 수 있다. 이렇듯 상품 품질은 사용자들의 상품 선택이라는 묵시적인 피드백을 통해서 인기도를 측정하며, 얼마나 많은 사람들이 그 상품을 좋아하는지 평가할 수 있게 된다.

사용자 인지도는 얼마나 많은 사용자들이 상품을 알고 있는가이다. (식1)을 변경하면 상품 품질과 사용자 인지도의 관계를 다음처럼 알 수 있다.

$$A(i) = \frac{Q(i)}{P(i)} \tag{2}$$

(식2)는 개개의 사용자들이 상품 i를 알고 있을 확률이

얼마나 되는지를 인기도와 상품 품질을 통해 구할 수 있음을 보여준다.

3.1.2 인기도와 사용자 인지도

추천 시스템은 협력 기법이나 내용기반 필터링 기법 또는 두 기법 모두의 장점을 취한 혼합 추천 기법을 통해 사용자에게 적합한 추천 후보들을 생성하며, 사용자는 추천 시스템이 제시한 추천 후보들 중에서 원하는 것을 선택한다. 만약 여러 사용자들로부터 동일한 질의가 있을 때 상품 *a, b, c*가 추천 후보로 주어지고 사용자들에 의해서 상품이 선택되는 비율이 $a > b > c$라면, *a*가 가장 인기 있는 상품이라고 말한다. 많이 선택될수록 그 상품은 높은 인기도를 가진다고 볼 수 있으므로 인기도는 상품 *i*를 선택한 사용자의 비율, *P(i)*,로 정의할 수 있다.

사용자들은 인기 있는 상품을 빈번히 찾고 선택하는데, 인기 있는 상품은 현재 유행하는 상품이기 때문에 선택될 수도 있고, 객관적인 품질이 좋아서 선택될 수도 있고, 여러 사람들의 취향에 부합하기 때문에 선택된다. 이렇듯 높은 품질을 가진 상품은 많은 사용자들로 하여금 선택과 구매를 유발한다. 따라서 높은 품질의 상품은 사용자들 사이에서 보다 많이 검색됨을 쉽게 알 수 있다. 이런 사실을 통해 사용자 인지도는 상품 *i*를 인지하는 사용자의 비율, *A(i)*, 로 정의한다.

3.2 상품 품질 측정 기법

인기도와 사용자 인지도를 통해 상품 품질을 계산하는 기법에 대해 설명한다.

3.2.1 인기도 측정

상품의 인기도는 얼마나 많이 선택되었는지에 의해 계산되는데, 예를 들면 상품 *i*에 대해 백만 번 추천이 수행되었고 이 중 100,000명이 선택했다면, 그것의 인기도는 0.1이다. 즉, (식3)과 같이 된다.

$$P(i) = \frac{z_i}{y_i} \tag{3}$$

여기서 z_i는 상품 *i*를 선택한 횟수, y_i는 *i*가 전체 추천 중 포함된 횟수를 가리킨다. 논문에서 제안하는 상품 품질은 특정 상품 *i*를 아는 사람 중에 상품 *i*를 선호할 확률이다.

3.2.2 사용자 인지도 측정

사용자 인지도 계산에 앞서 무엇을 통해 사용자 인지도를 평가할 것인지 고려해야 한다. 추천 시스템에서 특정 시점에 상품을 인지한 사용자가 얼마나 되는지 아는 것은 쉽지 않은데, 이를 위해 사용자에게 아는지 모르는지 일일이 피드백을 요구할 수 없다. 이 때문에 본 논문에서는 사용자 인지도를 추천 시스템에서 알기 위해 상품의 추천 내역을 이용한다. 사용자에게 추천된 적이 있는 상품은 사용자들이 인식하고 있을 가능성이 그렇지 않은 것보다 높으며, 상품의 추천 순위가 높은 것이 낮은 것보다 상품을 인지할 확률이 더 높다. 이러한 정보들을 인지도 분석 요소로서 이용한다.

추천 시스템을 통해서 한 사용자의 유사 사용자 집합 M이 총 *x*번의 추천이 수행되었고 이 중에 상품 *i*를 포함한 추천의 수가 *y*번, 이들 추천 중 *i*를 선택한 횟수가 *z*번이라고 했을 때 사용자가 *i*를 알고 있을 확률인 사용자 인지도는 (식4), (식5)와 같다.

$$A(i) = \frac{y_i}{x_i} \times \frac{z_i}{y_i} \times r_i \tag{4}$$

$$A(i) = \frac{z_i}{x_i} \times r_i \tag{5}$$

위의 (식4)에서 y_i/x_i는 전체 추천 중 *i*가 포함될 확률, z_i/y_i는 포함된 추천 중 *i*가 선택될 확률, *r*은 *i*가 추천되었을 때 상대적인 순위를 의미한다. 인지도 계산을 위한 나머지 요소 중 *r*에 대해 예를 들면, 100개의 상품이 추천이 되었을 때 *i*의 순위가 90번째라면 이 상품의 *r*값은 0.1이다. *r*을 구하는 식은 (식6)과 같다.

$$r_i = 1 - \frac{\sum_{t=1}^{v} l_i}{y_i} \tag{6}$$

여기서 l은 *i*가 추천되었을 때의 평균 순위값이다.

(식4)는 (식5)를 정리한 것이다. 이로서 *A(i)*를 총 추천 프로세스 중 *i*가 추천에 포함될 확률, 추천되었을 때의 선택 확률, *i*의 상대적인 순위를 통해 계산할 수 있음을 보인 것이다. 이를 통해 추천 빈도와 추천 순위에 따라 상품에 대한 사용자의 인지도를 알 수 있게 된다.

3.2.3 상품 품질 측정

본 절에서는 이 상품 품질을 인기도와 사용자 인지도를 통해 계산하는 방법을 설명한다.

이전 장에서 인기도, 인지도, 상품 품질에서 (식1)과 (식2)의 관계를 설명하였다. 이 관계를 통해 상품 품질 계산식을 (식6)과 같이 유도할 수 있다.

$$Q(i) = \frac{z_i}{y_i} \times \frac{z_i \times r_i}{x_i} = \frac{z_i^2 \times r_i}{x_i \times y_i} \tag{7}$$

(식7)은 상품 품질을 구하기 위해서 (식1)에 (식3)과 (식4)를 적용하여 유도한 것이다. 그러나 이 공식은 어디까지나 확률을 나타낼 뿐이기 때문에 얼마나 가중되어야 하는지 그 정량적인 수치를 제공하지 못한다. 이에 대한 것이 <표 3>에 제시되어 있다.

〈표 3〉 상품 품질 수치의 예

x	y	z	r	IQ
420,443	170,504	1,405	0.5	0.00001377
	102,724	256	0.5	0.00000076
	147,980	857	0.5	0.00000590
	154,802	2,013	0.5	0.00003113

위의 <표 3>에서 제시한 상품 품질에 대한 수치를 보고 저것이 무엇을 의미하는지 알 수 없는데, 이는 <표 3>의 각 상품 품질에 비해 얼마만큼의 품질인지 비교 판단할 수 있는 기준치가 제시되지 않았기 때문이다. 이를 위해서 각 상품 품질 평가 시 그 상품이 가질 수 있는 최대 상품 품질을 해당 상품의 기준으로 정의한다. 최대 상품 품질은 i가 유사 사용자 집합 M에 있는 모든 사용자들이 i를 선택한 경우이다. 이를 통해 최대 상품 품질에 비해 상품 품질이 얼마나 되는지를 비교하면 정량적인 수치를 얻을 수 있게 된다.

3.3 신규 사용자 추천 문제 해결 기법

추천 시스템은 이러한 신규 사용자 추천 문제는 반드시 해결되어야만 한다. 이를 위해 본 논문에서는 사용자 프로파일에 기록된 사용자 정보를 토대로 신규 사용자 추천 문제에 접근하는 신규 사용자 추천 문제 해결 기법을 제안한다.

3.3.1 사용자 프로파일 기반 사용자 그룹을 통한 추천

본 논문에서는 신규 사용자 추천 문제를 해결하기 위해서 사용자 프로파일을 이용한다. 일반적으로 사용자 프로파일은 사용자의 생년월일, 나이, 주소, 성별과 같은 기본적인 사용자 정보를 포함하며 사용자를 나타내는 정적 정보이다. 사용자 측면에서 신규 사용자 추천 문제가 발생하는 시점은 추천 시스템에 신규 사용자가 등록되어 서비스를 이용하려 할 때이기 때문에 추천 시스템은 사용자의 정적 정보 이외의 정보는 알 수 없으며, 새로운 사실을 추론할 수 없다. 이러한 문제로 인해서 추천 시스템의 추천 프로세스는 사용자 프로파일로 제한한다.

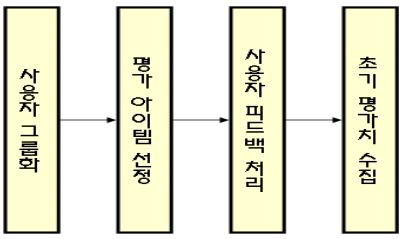
본 논문에서는 이러한 한계를 해결하기 위해서 사용자 프로파일에 기반하여 사용자를 군집하고 이를 통해 초기 평점 획득을 위한 상품을 선정하도록 한다. 이를 위해서 본 논문은 협력적 필터링 기법의 아이디어를 통해 신규 사용자 문제에 접근한다. 협력적 필터링 기법의 기본 아이디어는 특정 사용자 u와 비슷한 취향을 가진 사용자들 U의 평가는 유사한 점으로 협력적 필터링 기법의 아이디어와 동일하며 이를 신규 사용자 문제를 해결하기 위해 적용한다. 즉, 사용자 프로파일을 통해 비슷한 취향을 가진 사용자들을 분류하고 이들의 평가에 기반하여 추천을 수행한다. 이와 같이 비슷한 취향을 가진 사용자의 분류는 프로파일에 기술된 나이, 성별, 직업 등의 정보를 이용하며, (그림 1)과 같은 데이터를 포함하고 있다. (그림 1)은 MovieLens Project의 사용자 프로파일이다.

UserID	Gender	Age	Occupation	ZipCode
5858	M	25	4	01002
5859	M	25	7	89117
5860	F	25	4	91428
5861	F	50	1	98499
5862	F	25	9	76120
5863	F	25	14	89511
5864	F	50	1	28043
5865	M	25	17	95926
5866	F	25	6	06114
5867	F	45	3	91306
5868	M	35	14	85331
5869	F	25	7	19103
5870	M	25	14	98109
5871	F	25	4	01002
5872	M	25	17	61265
5873	M	18	12	10021
5874	M	25	4	01002
5875	M	25	4	19103
5876	M	50	7	30066
5877	M	18	12	02138
5878	F	25	0	60640

(그림 1) MovieLens 사용자 프로파일

3.3.2 신규 사용자 피드백 처리

신규 사용자 문제를 해결하기 위해서 본 논문은 사용자 그룹화, 평가 상품 선정, 사용자 피드백 처리, 초기 평점 수집 단계를 통해 문제에 접근한다. 이를 위한 처리 흐름도는 (그림 2)와 같다.



(그림 2) 신규 사용자 문제 처리 흐름도

신규 사용자 추천 문제 중 신규 사용자에 대한 문제를 처리하기 위해 (그림 2)에 제시된 것과 같이, 4단계에 대한 순차적 실행을 통해 신규 사용자에게 추천을 하기 위한 초기 평점을 생성한다. 이하의 절에서 각 단계에 대해 상세히 설명한다.

3.3.3 사용자 그룹화

본 논문에서는 프로파일을 통한 사용자 그룹화를 위해서 나이, 성별, 직업 정보를 활용하여, 비슷한 나이의 사용자들끼리 그룹화하고, 같은 성별로 그룹화한 뒤에 같은 직업을 가진 사용자들로 그룹화 한다. 이러한 방식의 그룹화는 협력적 필터링의 관점에서 같은 성별 및 직업을 갖는 비슷한 나이의 사용자들은 상품에 대해 비슷한 평가를 할 것이란 전제에서 비롯된 것이다.

위의 두 가지 전제에 기반하여 본 논문은 사용자들을 그룹별로 나누는데, 이러한 그룹화는 3가지의 각기 다른 그룹으로 분류한 상태에서 그룹을 병합하는 방식이 아니라 한 그룹에 대해 분류하고 그 분류 속에서 또 다시 다른 정보를 통해 분류하는 방식으로 처리한다.

3.3.4 평가 항목 선정

사용자 프로파일의 성별, 나이, 직업에 의해 사용자들을 그룹별로 분류한 뒤에 신규 사용자에게 초기 평점 구성을 위한 평가 항목을 선정하는데 이는 사용자에게 추천이 이루어지기 위해서 반드시 해당 사용자의 평가 정보가 있어야 하기 때문이다. 따라서 추천의 계산 영역은 같은 그룹으로 분류된 사용자들이 평가한 항목 평점이다. 예를 들면, 사용자 *A, B, C*와 신규 사용자 *D*가 있을 때, *A, B, C*가 평가를 수행한 항목들의 평점에 대해 계산된다.

신규 사용자를 위한 유사 사용자 그룹과 평점이 수집되면, 신규 사용자에게 평가 항목으로 제공할 항목들을 선정하기 위해서 유사 사용자들의 평점에 기반하여 항목 선정을 위한 항목 선별을 수행한다. 항목의 선별은 유사 사용자 *A, B, C*가 항목에 대해 부여한 평점의 평균을 통해 이루어지게 된다.

3.3.5 사용자 피드백을 이용한 초기 평점 획득

평가 항목이 사용자에게 제공되면, 사용자는 평가 항목을 평가하여 실제적으로 사용자의 선호가 얼마나 되는지 평가하고, 평가된 사용자의 선호 정보를 해당 사용자를 위한 초기 평점으로 활용한다. 그러나 평가 항목이 너무 많은 경우 사용자는 성실하게 초기 평가에 응하지 않는다. 따라서 적당한 수의 평가 항목이 제시되어야 한다. 논문에서는 적정한 평가 항목의 수를 20개로 정하였고, 별점 부여 방식을 적용하여 사용자 피드백을 추천 시스템에 제공토록 하였다.

4. 성능 평가

4.1 협력적 필터링 기법별 추천 정확도 비교

본 논문에서 제시한 성능 평가는 추천 정확도의 개선을 입증하기 위해 6,040명의 사용자에 대한 사용자간 유사도를 피어슨 상관계수(Pearson Coefficient) 기법을 통해 구하고, 이 사용자 유사도를 순수 협력적 필터링 알고리즘인 NBCFA, CMA(Correspondence Mean Algorithm) Type1과 제안한 기법을 적용했을 경우를 비교한다.

4.1.1 추천 정확도 비교

세 그룹의 실험 데이터를 통해 순수 협력적 필터링 기법의 대표적인 알고리즘인 NBCFA와 제안한 상품 품질을 적용한 NBCFA에 대한 추천을 수행 정확도 비교를 수행하였다.

피어슨 상관계수를 유사도 가중치로 이용한 NBCFA에서 MAE는 <표 4>과 같다.

<표 4>에 제시된 바와 같이 3개의 실험데이터에 대한 MAE를 비교하였다. 그 결과 상품 품질을 가중치로 적용한 NBCFA'과 CMA Type1'의 추천 정확도가 우수하게 나타났다. [34]에 제시된 것처럼 기존 NBCFA의 기법은 0.73의 MAE에서 향상되지 않음을 볼 수 있는데, 이에 비해 NBCFA'에서는 0.73보다 낮은 MAE를 보임에 따라 추천 정확도가 기존보다 개선되었음을 확인할 수 있다.

〈표 4〉 NBCFA 기반 예측 성능(MAE) 비교

구분	NBCFA	NBCFA'	CMA Type1	CMA Type1'
데이터1	0.723	0.69	0.71	0.68
데이터2	0.722	0.7	0.7	0.68
데이터3	0.724	0.68	0.71	0.67

〈표 5〉 기법별 평점 예측

영화 ID	실제 평점	NBCFA	NBCFA'	CMA Type1	CMA Type1'
595	5	4.753	5.012	3.753	4.057
720	3	5.484	5.131	4.484	4.184
938	4	5.571	5.203	4.571	4.262
1566	4	3.774	4.054	2.774	3.083
1962	4	4.879	4.566	3.879	4.158
2355	5	5.800	5.561	4.800	5.101
3105	5	4.446	4.727	3.446	3.723

이렇게 개선된 MAE는 MovieLens 데이터베이스 중 일부에서 <표 5>와 같은 예측 평점을 제공했다.

<표 5>는 데이터1의 테스트 집합에서 사용자 ID가 21번인 사용자가 아직 평가하지 않은 영화에 NBCFA, CMA Type1과 상품 품질을 가중치로 적용했을 때의 예측 평점이다. <표 5>에 제시된 것처럼 모든 기법에서 상품 품질이 가중된 경우가 정확함을 확인할 수 있다.

5. 결 론

본 논문에서는 협력적 필터링 기법에서 가중치로 적용할 수 있는 상품 품질을 제안하였다. 이 개념은 상품의 상대적인 인기도 변화와 사용자들의 인식률에 따라 변화하는 특성을 그대로 적용할 수 있기 때문에 유행에 따른 추천이 가능하다. 또한 상품의 상대적인 품질이 추천에 적용됨에 따라서 사용자가 상품을 선택했을 때 만족할 확률을 높일 수 있다는 장점을 가지고 있다.

또한 본 논문에서는 신규 사용자 추천 문제로 명명된 신규 사용자 추천 문제에 대한 해결 기법을 제안함으로서 신규 사용자에 대한 추천이 가능하도록 하였다. 이는 추천 시스템이 보유한 사용자 프로파일의 인구통계학적 정보를 기반으로 평가 항목을 선정하고 이를 사용자에게 제시하여 피드백 받음으로서 유사 사용자를 구별하기 위한 초기 평가 집합을 구성함에 따라 가능하게된 것이다. 이로서 순수 협력적 필터링 알고리즘이 가지는 신규 사용자 추천 문제를 해결하였다.

참 고 문 헌

[1] Sarwar, B., Karypis, G., Konstan, J. and Riedl, J.,

"Application of Dimensionality Reduction in Recommendation System-A Case Study", ACM WebKDD 2000 Web Mining for E-Commerce Workshop, http://robotics.stanford.edu/~ronnyk/WEBKDD2000/papers/, 2000.

[2] J.A. Konstan, B.N. Miller, D. Maltz, J.L. Herlocker, L.R. Gordon, and J. Riedl, "GroupLens: Applying Collaborative Filtering to Usenet News," Comm. ACM, Vol.40, No.3, pp. 77-87, 1997.

[3] D. Goldberg, D. Nichols, B.M. Oki, and D. Terry, "Using Collaborative Filtering to Weave an Information Tapestry," Comm. ACM, Vol.35, No.12, pp.61-70, 1992.

[4] U. Shardanand and P. Maes, "Social Information Filtering: Algorithms for Automating 'Word of Mouth'," Proc. Conf. Human Factors in Computing Systems, 1995.

[5] L. Getoor and M. Sahami, "Using Probabilistic Relational Models for Collaborative Filtering", Proc. Workshop Web Usage Analysis and User Profiling (WEBKDD '99), August. 1999.

[6] K. Goldberg, T. Roeder, D. Gupta, and C. Perkins, "Eigentaste: A Constant Time Collaborative Filtering Algorithm," Information Retrieval J., Vol.4, No.2, pp.133-151, July, 2001.

[7] Breese, J. S., Heckerman, D., Kadie. C. 1998. "Empirical Analysis of Predictive Algorithms for Collaborative Filtering", In Proceedings of the Fourteenth Annual Conference on Uncertainty in Artificial Intelligence, pp.43-52, Madison, Wisconsin, July, 1998.

[8] Resnick P. and Varian H.R., "Recommender systems", Communications of the ACM Vol.40, No.3, pp.56-58, 1997.

[9] M. Pazzani, "A Framework for Collaborative, Content-Based, and Demographic Filtering," Artificial Intelligence Rev., pp.393-408, Dec., 1999.

[10] Resnick P. and Varian H.R., "Recommender systems", Communications of the ACM Vol.40, No.3, 56-58, 1997.

[11] G. Salton, Automatic Text Processing. Addison-Wesley, 1989.

[12] R.J. Mooney, P.N. Bennett, and L. Roy, "Book Recommending Using Text Categorization with Extracted Information," Proc. Recommender Systems Papers from 1998 Workshop, Technical Report WS-98-08, 1998.

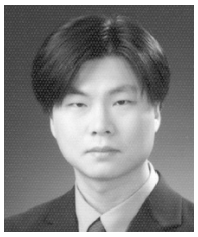
[13] M. Pazzani and D. Billsus, "Learning and Revising User Profiles:The Identification of Interesting Web Sites," Machine Learning, Vol.27, pp.313-331, 1997.

[14] D. Goldberg, D. Nichols, B.M. Oki, and D. Terry, "Using Collaborative Filtering to Weave an Information Tapestry," Comm. ACM, Vol.35, No.12, pp.61-70, 1992.

[15] M. Claypool, A. Gokhale, T. Miranda, P. Murnikov, D. Netes, and M. Sartin, "Combining Content-Based and Collaborative Filters in an Online Newspaper," Proc. ACM SIGIR'99 Workshop Recommender Systems: Algorithms and Evaluation, Aug., 1999.

[16] Gediminas Adomavicius, Alexander Tuzhilin, "Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions", Knowledge and Data Engineering, IEEE Transaction, Vol.17, pp.734-749, 2005.

[17] U. Shardanand and P. Maes, "Social Information Filtering: Algorithms for Automating 'Word of Mouth'," Proc. Conf. Human Factors in Computing Systems, 1995.



조 성 훈

e-mail : shcho@kaeri.re.kr

2001년 2월 한남대학교 컴퓨터공학과 (공학사)

2002년 11월~2003년 12월 씨드텔레콤 소프트웨어 개발팀

2003년 2월 한남대학교 컴퓨터공학과 (공학석사)

2008년 8월 한남대학교 컴퓨터공학과(공학박사)

2009년 3월~현 재 한국원자력연구소 하나로운영부 연구원

관심분야 : Ubiquitous, Data Mining, Recommender System, Ontology Modeling, Pattern Recognition

이 무 훈

e-mail : mhlee@dblab.hannam.ac.kr

2002년 2월 한남대학교 컴퓨터공학과 (공학사)

2004년 2월 한남대학교 컴퓨터공학과 (공학석사)

2004년 3월~현 재 한남대학교 컴퓨터 공학과 박사과정

2009년 3월~현 재 한국전자통신연구원 네트워크로봇팀 연구원

관심분야 : Web Search Engine, Data Mining, Distributed Computing, Context-Awareness, Machine learning

김 정 석

e-mail : jskim@dblab.hannam.ac.kr

2007년 2월 한남대학교 컴퓨터공학과 (공학사)

2009년 2월 한남대학교 컴퓨터공학과 (공학석사)

2009년 3월~현 재 한남대학교 컴퓨터 공학과 박사과정

관심분야 : Database, Semantic Web, Web Search, Ubiquitous Computing, Context-Awareness, U-Learning

김 봉 회

e-mail : kbh@ubnc.net

1984년 2월 중앙대학교 전자계산학과 (이학석사)

2003년 2월 대전대학교 컴퓨터공학과 (공학박사)

2005년~현 재 (주)유비엔씨 대표이사

관심분야 : Ubiquitous, Ubiquitous Security, USN, RFID

최 의 인

e-mail : eichoi@hnu.kr

1995년 홍익대학교 전자계산학과(이학박사)

1992년~1996년 명지전문대학교 전자계산과 조교수

1996년~현 재 한남대학교 컴퓨터공학과 교수

2003년 UCLA visiting Scholar

관심분야 : Ubiquitous Computing, Web search engine, Semantic Web, Context Modeling, Grid Computing