개인화 추천시스템의 성능 향상 적용 알고리즘 분석

An Analysis of Performance Improvement Algorithm for Personalized Recommender System

윤수진, 윤희병
국방대학교 전산정보학과

Sujin Yun, Heebyung Yoon
Dept. of Computer and Information Science, Korea National Defense University
E-mail : imix96,hbyoon@kndu.ac.kr

요 약

무수히 많은 정보 중에서 특정 사용자에게 가장 유용할 것으로 판단되는 정보를 추천하여 제공함으로써 특정 사용자의 편의를 돕는 시스템이 추천시스템이다. 이러한 추천시스템에 성공적으로 적용된 알고리즘이 협력적 필터링이며 이것은 다른 사용자로부터 먼저 평가된 웹문서를 제공받아 이를 축적하고 다시 사용자에게 환원하는 알고리즘이다. 하지만 이 알고리즘은 초기 평가, 희소성, 확장성 등의 문제점을 내포하고 있다.

따라서 본 논문은 이러한 문제점을 해결하고 성능 향상을 하기 위해 적용된 개인화 추천시스템 관련 최신 알고리즘들을 비교하고 분석한 결과를 제시한다. 이를 위해 먼저 최근에 발표된 협력적 필터링과 최근접 이웃 알고리즘, 인공 지능기술을 이용한 알고리즘, 군집화 알고리즘 등 각각에 대한 기술적 분석 결과를 수행한다. 그런 후 이들 다양한 알고리즘들의 조합을 통한 성능 향상 결과에 대한 비교분석과 각각의 조합에 대한 장단점 분석 결과도 또한 제시한다.

1. 서론

무수히 많은 인터넷의 정보 중에서 특정 사용자에게 가장 유용할 것으로 판단되는 정보를 추천하여 제공함으로써 사용자의 편의를 돕는 시스템이 추천시스템이다.

추천시스템은 정보 필터링(Information Filtering) 기술을 적용하여 사용자의 요구사항을 기반으로 추천을 하고 있으며, 추천방법은 크게 내용-기반 필터링(CBF: Contents-Based Filtering)과 사회 필터링(SF: Social Filtering)의 2 가지로 나눌 수 있다.

내용-기반 필터링이 사용자 프로파일을 바탕으로 사용자가 표시한 선호도를 참고하여 항목을 추천하는 반면, 사회 필터링은 유사한 사용자들은 유사한 항목을 선호한다는 원칙을 적용하여 특정 항목을 추천하는 협력적 필터링 (Collaborative Filtering) 방법이다.

협력적 필터링은 94년 Tapestry에 최초로 적용되었고, 뒤이어 음악앨범 추천용 링고시스템, 영화 추천용 무비렌즈시스템, 인터넷 서점, 인터넷 CD 판매사이트, 인터넷 영화 추천사이트 등으로 확대 적용되어 현재 추천시스템에 가장 성공적으로 적용된 알고리즘으로 평가 받고 있으며, 그림 1에 협력적 필터링의 연도별 발전과정이 도시되어 있다. 그러나 협력적 필터링은 초기 평가 문제, 희소성 문제, 확장성 문제를 가지고 있으며, 이를 해결 하기위해 다양한 알고리즘과의 결합을 통해서 추천시스템의 성능 향상 및 문제점 해결을 꾀하고 있다[1, 2].

이러한 추천시스템의 개선방법을 비교하고, 각 방법의 장단점을 분석하여 제시한다.

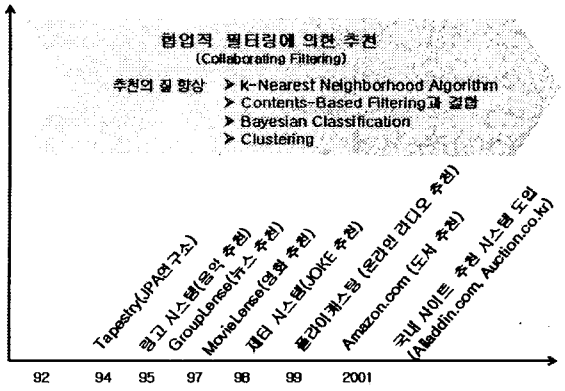

그림 1 협력적 필터링의 발전

2. 관련 연구

2.1 개인화 추천시스템

인터넷에서 개인화의 의미는 모든 사용자가 동일한 사용자로 취급되고 동일한 형태의 정보를 제공받는 것에서 발전하여, 각 개인 사용자와 연관되어 개인의 특성에 따라 기능, 인터페이스, 정보 내용, 또는 시스템적 변화가 적응적으로 나타나는 과정이다[3].

개인화 추천시스템은 수많은 정보 중에서 특정 개인에게 가장 적합할 것으로 판단되는 정보를 추천하여 제공함으로서 사용자의 편의를 돕는 시스템이다. 이러한 개인화 추천시스템은 고객이 자신의 선호, 관심, 구매 경력과 같은 정보를 웹사이트에 제공하면 웹사이트는 고객이 제공한 정보를 기초로 고객에게 가장 알맞은 정보를 제공하게 되고, 이를 통해 웹사이트 운영자는 고객의 지속적인 이용을 얻을 수 있고 사용자는 자신에게 가장 알맞은 정보를 보다 편리한 방법으로 얻을 수 있게 된다.

2.2 내용-기반 필터링

내용-기반 필터링은 사용자들이 입력한 선호하는 프로파일에 근거하여 사용자들의 특성을 파악하고, 사용자가 표현한 선호 정보에 가장 적합한 항목 콘텐츠의 내용분석을 통해 추천하는 방식이다. 사용자들이 입력한 개인 프로파일은 추천의 가장 중요한 근거가 되나 사용자가 직접 입력한 정보의 정확성에는 한계가 있고, 사용자의 선호가 변할지라도 프로파일 정보는 변화되지 않을 수 있으므로 정확성이 떨어진다.

2.3 협력적 필터링

협력적 필터링은 유사한 기호를 가진 다른 사용자들이 웹문서에 대해 산술적인 크기로 매긴 명시적 의견이나, 로그분석 등을 통해 암묵적으로 측정된 선호도로 나타나는 의견을 바탕으로 추천을 제공한다.

그림 2는 협력적 필터링 프로세스의 개략적인 도해로서[1], 선호도 매트릭스인 $m \times n$ 사용자-항목 데이터로 나타난 선호도 테이블에 적용되어 n개의 추천 리스트를 생성하거나 특정 항목에 대한 선호도를 예측하게 된다.

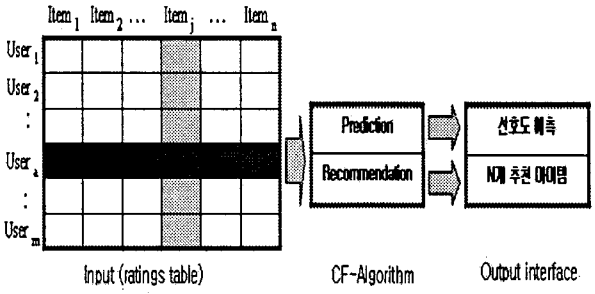

그림 2 협력적 필터링 프로세스

협력적 필터링의 사용자는 아래의 3단계를 통해 추천을 받게 되고[2],각 단계는 다음과 같다.

(1) 사용자의 선호 프로파일 정보 입력

선호 정보는 사용자의 행동에서 나타난 페이지-뷰 횟수, 시간 등에 의해 간접적으로 측정되거나, 직접적으로 선호도를 표현한 항목에 대하여 산술적 척도 등의 선호 프로파일 정보를 수집한다.

(2) 유사한 프로파일을 이용한 이웃 선정

유사한 프로파일을 가진 사용자를 결정하는 단계로 협력적 필터링의 성공을 위한 가장 중요한 단계이다. 만약 가장 유사한 프로파일을 가진 사용자를 선정할 수 있다면 추천의 결과는 가장 최선의 것이 될 수 있다. 이 단계에서 사용자간 유사도를 결정하기 위하여 유사도 계산 기법을 활용한다.

(3) 이웃들의 선호를 추천형태로 결합

선호도를 가중치합의 방법으로 구하거나 회귀분석(선형 회귀, 다중 회귀 등)을 통해 특정 항목에 대한 사용자의 선호도를 예측하게 된다.

3. 협력적 필터링의 문제점

3.1 초기평가 문제(Early rate problem)

순수한 협력적 필터링은 다른 사용자의 평가를 기반으로 예측을 수행하기 때문에, 기존에 평가가 없는 경우 예측을 제공할 수 없다. 또한 이와 유사하게 새로운 사용자가 가입하는 경우

충분한 예측을 할 수 없다.

3.2 희소성 문제(Sparsity problem)

실제로, 많은 상업적 추천시스템은 대단히 많은 수의 평가항목 집합을 가지고 있어, 모든 사용자들은 모든 항목에 대해 평가를 할 수 없으며, 항목에 대한 평가는 매우 희박하다. 이에 따라 추천시스템은 유사한 사용자를 찾을 수 없는 특정 사용자에게 추천을 하는 것이 불가능 하게 되고 추천의 질은 낮아지게 된다.

3.3 확장성 문제(Scalability problem)

비교하고자 하는 사용자의 항목 수가 많아질수록 협력적 필터링을 수행하기 위한 시간은 사용자 수와 항목 수에 비례하여 더욱 많이 걸리게 된다. 온라인 환경에서 많은 사용자에게 빠르고 정확한 결과를 제공하고자 하는 추천시스템에서는 단점이 된다.

4. 협력적 필터링의 개선

협력적 필터링이 가진 문제점을 해결하기 위하여 다양한 알고리즘과의 결합이 시도되고 있다.

초기값 문제를 해결하기 위해 내용-기반 필터링 또는 최근접 이웃 알고리즘, 군집화 등이 적용되기도 하고, 협력적 필터링에서 추천의 질을 결정짓는 유사한 사용자를 선정하는 과정에서 각 알고리즘을 적용하여 우수한 이웃을 선정하여 추천의 질을 개선한다. 또한 군집화 알고리즘과 결합하여 희소성 문제를 완화하기도 한다. 그림 3은 추천시스템의 문제점과 그에 따른 개선을 개략적인 그림으로 나타내었다.

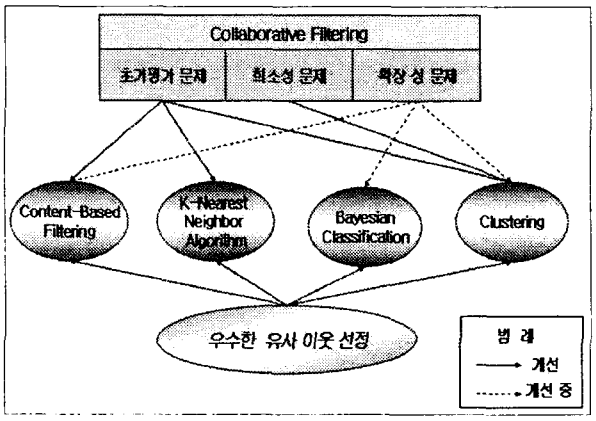

그림 3 협력적 필터링의 개선

4.1 내용-기반 필터링과의 결합

기존 협력적 필터링과 내용-기반 필터링의 기술을 결합한 혼합형 추천시스템은 내용-기반 필터링의 사용자 프로파일 정보를 협력적 필터링과 결합하고 있는데[4,5] 이는 사용자 프로파일 정보에서 초기값을 지정하여 초기값 문제를 해결하고, 이웃 선정시 사용자 프로파일 정보와 항목평가 정보를 동시에 이용하여, 보다 밀접한 이웃 선정에 영향을 미친다.

또한 사용자 프로파일이 사용자의 성향 변화를 정확히 반영하지 못하여 추천의 질이 낮아짐으로 웹 사용 마이닝(Web Usage Mining)을 이용하여 프로파일을 동적으로 변환시키는 방법도 사용되고 있다[6].

4.2 최근접 이웃 알고리즘

n-차원의 점으로 표현된 미지의 샘플은 패턴 공간에서 가장 근접한 이웃 k개 중에 가장 공통 특성을 가진 클래스로 할당하기 위해 유클리드 거리 함수를 사용하여 근접성을 판단한다.

최근접 이웃 알고리즘은 미지의 샘플이 주어졌을 경우 이러한 알고리즘에 따라 사용자와 가장 유사한 이웃을 결정할 수 있고, 실제 값을 예측하기 위해서도 사용되는데 이 경우 분류기는 미지의 샘플의 k-인접 이웃에 연관된 실제 레이블의 평균을 출력함으로서, 초기평가 문제를 해결할 수 있다[1,7].

그러나 근접 이웃 분류기는 새로운 샘플이 분류되기를 요청할 때까지 분류기를 생성하지 않는다는 점에서 인스턴스 기반(Instance Based) 또는 게으른 학습기(Lazy Learner)로서 모든 계산이 분류하는 시간까지 지연되기 때문에 분류 속도는 더욱 느리다. 따라서 최근접 이웃 알고리즘은 분류하고자 하는 샘플들을 비교하는 잠재적 이웃들(저장된 훈련 샘플들)의 수가 많은 경우 매우 큰 비용을 수반하며 이러한 특성 때문에 확장성에서 문제가 될 수 있다.

4.3 베이지안 분류기(Bayesian Classification)

베이지안 분류기는 유사한 사용자가 정해지지 않은 특정 사용자를 확률 이론에 따라 해당 클래스로 분류하게 됨으로 가장 유사하고 우수한 이웃을 결정할 수 있게 되고 질 높은 추천을 수행하는데 기여한다[8].

베이지안 분류기는 이론적으로 다른 의사결정 트리나 신경망 분류기와 비교했을 때 비교적 최소 오류율을 가지고 이론적 근거를 제공하는 점에서 유용하다. 베이지안 분류기는 통계적 분류기로서 주어진 샘플이 특정 클래스에 속할 확률과 같이 어떤 특정 항목이 특정 클래스에 속할

확률을 예측한다.

4.4 군집화(Clustering)

군집화는 특정 항목에 대한 평가 데이터는 적다고 할지라도 이 항목이 속한 군집에 대한 평가 데이터는 상대적으로 많다는 것에 착안하여 이 평가 데이터를 활용하여 특정 항목의 평가를 예측할 수 있으므로 희소성 문제를 해결하여 추천의 질을 향상한다[9,10]. 또한 군집 형성 후 군집에 의한 기본값을 새로운 사용자에 대한 기본값으로 부여함으로서 초기평가 문제를 해결하는데 활용된다.

주로 사용되는 군집화는 연관관계를 이용하거나 자카드 계수(Jaccard coefficient)를 활용하며 군집간 객체끼리는 높은 유사성을 갖게 하고 다른 군집의 객체와는 큰 상이성을 갖도록 유사한 객체끼리 군집화를 한다.

표 1은 각 개선 알고리즘별 특징을 나타내었다.

구 분	특 징
내용기반 필터링	• 사용자 프로파일에서 초기값 지정으로 초기값 문제 해결 • 이웃 선정시 사용자프로파일 및 아이템 정보 동시적용으로 우수한 이웃 선정
최근접 이웃 알고리즘	• 실제 값 예측으로 초기평가 문제 완화 • 샘플 수 증가시 급격한 계산양 증가
베이지안 분류	• 최소오류율의 분류기(정확성)로서 우수한 이웃 선정 가능 • 이론적 근거-베이즈 이론
군집화	• 상대적으로 많은 군집의 데이터를 활용하여 희소성 문제 해결 • 군집의 기본값 부여 초기평가 문제 해결

표 1 개선 알고리즘 특징

5. 결론 및 향후 연구과제

본 논문에서는 현재 추천시스템에 가장 성공적으로 적용된 협력적 필터링의 성능 개선을 위해 사용된 알고리즘들을 제시하였다. 협력적 필터링은 성능 개선을 위하여 내용-기반 필터링, 최근접 이웃 알고리즘, 베이지안 분류기, 군집화 등 여러 알고리즘과의 조합이 사용되고 있고, 이 알고리즘들은 다양한 형태의 데이터베이스에 적용되어 추천의 성능을 개선하고 있다. 또한 최근의 추세는 협력적 필터링의 프로세스에서 동시에 여러 알고리즘을 적용하여 협력적 필터링의 문제를 해결하고 있다.

이러한 개선 알고리즘의 분석을 토대로 초기평가 문제, 희소성 문제, 확장성 문제를 해결할 수 있는 알고리즘의 조합이나 새로운 알고리즘을 제시하는 것이 향후 연구해야 할 과제가 될 것이다.

6. 참고문헌

[1] B.Sarwar, G.Karypis, J.Konstan, J.Riedl, "Item Based Collaborative Filtering Recommendation Algorithms," World Wide Web 10th(WWW10), 2001.

[2] B.Sarwar, G.Karypis, J.Konstan, J.Riedl, "Explaining Collaborative Filtering Recommendation," CSCW'00, 2000.

[3] Jan Blom, "A Theory of Personalized Recommendation," CHI2002, ACM, 2002.

[4] M.balbanovic, Y.Shoham, "Content-based collaborative recommendation," CACM, Vol. 40, 1997.

[5] 김병만, 이경, 김시관, "추천시스템을 위한 내용기반 필터링과 협력필터링의 새로운 결합기법," 한국정보과학회, 2004.

[6] 안계순, 고세진, 정준,"웹 사용 마이닝 기반의 동적 사용자 프로파일 생성," 한국정보처리학회, 2002.8.

[7] J.Herllocker, J.Riedl, "An Algorithmnic Framework for Performing Collaborative Filtering," Conference on Research and Development in information Retrieval, 1999.

[8] J.S.Breese, D.Heckerman, "Empirical analysis of predictive Algorithm for collaborative filtering," 14th UAI, 1998.

[9] 김진수, 김태용, "사용자 로그분석과 클러스터 내의 문서 유사도를 이용한 동적 추천시스템," 한국정보과학회, 2004.

[10] 정경용, 김진현, "개인화 추천시스템에서 연관 관계 군집에 의한 아이템 기반의 협력적 필터링 기술," 한국정보과학회, 2004.