협업 필터링 추천 시스템의
예측 정확도 향상에 관한 연구

연세대학교 대학원

경 영 학 과

김 병 호

협업 필터링 추천 시스템의
예측 정확도 향상에 관한 연구

지도 임 일 교수

이 논문을 석사 학위논문으로 제출함

2010년 7월 일

연세대학교 대학원

경 영 학 과

김 병 호

김병호의 석사 학위논문을 인준함

심사위원 임 일 인

심사위원 이 호 근 인

심사위원 손 재 열 인

연세대학교 대학원

2010년 7월 일

차 례

그 림 차 례

표 차 례

식 차 례

국문요약

협업 필터링 추천 시스템의
예측 정확도 향상에 관한 연구

추천 시스템이란 사용자의 요구에 부합하는 정보를 자동으로 검색하여 주는 시스템이다. 본 연구는 일대일 마케팅과 웹 개인화, 맞춤화 서비스를 가능하게 해 정보 제공자와 사용자 모두에게 부가가치를 제공하는 추천 시스템의 성능 향상에 목적을 두고 협업 필터링 추천 시스템에서 사용할 수 있는 새로운 방법을 제시하였다.

전통적으로, 협업 필터링 추천 시스템에서는 모든 사용자에게 동일한 기준을 적용하여 추천 리스트를 생성한다. 협업 필터링 추천 시스템의 성능에 가장 중요한 영향을 주는 유사 사용자 그룹의 선정에 있어서도 대부분 동일한 기준을 적용하여 같은 크기의 유사 사용자 그룹을 선정한다. 하지만, 사용자마다 다르게 나타나는 선호도 특성과 패턴으로 인해 유사한 사용자의 크기도 다를 수밖에 없다. 본 연구에서는 이러한 사용자들의 선호 이질성에 주목해, 최적의 유사 사용자 그룹의 크기 역시 사용자마다 차이가 있을 것이라고 생각하였다.

본 연구에서는 사전에 가지고 있는 사용자의 선호도 정보를 바탕으로 사용자마다 다른 최적의 유사 사용자 그룹을 예측하고, 실제 추천 리스트 생성 시 개인화된 유사 사용자 그룹의 크기를 적용하여 추천 리스트를 생성하는 새로운 방법의 협업 필터링 추천 시스템의 방법을 제시하였다. 시뮬레이션 기법을 사용해 각 사용자에게 적합한 유사 사용자 그룹의 크기를 예측한 후, 예측된 유사 사용자 그룹의 크기를 적용해 선호도를 예측하는

방법으로 새로 제안한 협업 필터링 추천 시스템의 추천 리스트 생성 방법을 검증하였다. 미국 최대의 비디오 대여 사이트인 Netflix에서 개최한 Netflix Prize에서 제공한 실제 사용자들의 비디오 선호도 정보를 사용하여 새로 제안한 방법을 검증하였고 검증 결과, 기존의 모든 사용자에게 동일한 크기의 유사 사용자 그룹의 크기를 적용한 방법에 비해 새로 제시한 방법이 약 22% 향상됨을 알 수 있었다.

본 연구를 통해 사용자들마다 다른 최적의 유사 사용자 그룹이 존재함을 확인할 수 있었고, 추천 리스트 생성 시 사용자마다 개인화된 유사 사용자 그룹을 적용하여 추천 시스템의 성능을 향상시킬 수 있음을 확인할 수 있었다.

핵심되는 말 : 추천 시스템, 협업 필터링 추천 시스템, 유사 사용자 그룹, 유사 사용자 그룹의 크기, Netflix, 시뮬레이션

제 1 장 서 론

1.1 연구배경

우리가 인터넷을 통해 얻을 수 있는 정보의 종류와 양은 셀 수 없을 정도로 많다. 작은 쇼핑 사이트부터 대형 쇼핑 사이트, G마켓(Gmarket)과 같은 오픈 마켓, 미투데이(me2day)나 싸이월드(Cyworld) 미니홈피와 같은 SNS 사이트, 다음(Daum)이나 네이버(Naver)와 같은 포털 사이트에 이르기까지 다양한 정보 제공자들이 존재하고, 이들을 통해 사용자들이 얻을 수 있는 옷, 음반, 서적, 뉴스 등의 아이템과 관련된 정보는 그 수를 일일이 헤아리기 어렵다.

인터넷의 등장과 함께 인터넷을 표현하기 위해 주로 사용되던 정보의 바다라는 표현은 이제 더 이상 적합하지 않다. 이제는 인터넷을 통한 정보의 홍수라는 표현이 오히려 더 자주 사용되고 있다. 정보의 홍수라는 표현 그대로 쏟아지는 정보들 속에서 사용자들은 자신이 원하는 정보를 얻기 위해 점점 더 많은 노력을 들여야 한다. 또한, 정보 제공자들은 수많은 정보들 안에서 자신들의 정보가 사용자의 눈에 더 잘 띄게 하기 위해 더 많은 노력을 들여야 한다. 자신들이 원하는 정보를 더 쉽고, 빠르게 얻으려고 하는 사용자들과 자신들이 제공하는 정보를 더 많은 사용자들이 소비하기 원하는 정보 제공자들이 목표를 달성하기가 점점 더 어려워지고 있는 것이다.

아이팟과 아이폰 등 애플사에서 만든 기기들에 설치된 애플 운영체제에서 사용할 수 있는 응용 프로그램을 판매하고 있는 애플 앱스토어(http://www.apple.com/iphone/apps-for-iphone)의 경우를 생각해 보자. 애플 앱스토어에는 많은 수의 정보 제공자들 즉, 개발자와 개발사들이 응용 프로그램을 개발하고, 판매하기 위해 자신이 만든 응용 프로그램에 대한 정보를 등록한다. 애플 아이튠즈라는 사이트 안에 존재하는 음악과 영화 등을 제외한, 애플 앱스토어에 등록된 응용 프로그램만 해도 2010년 현재 이미 200,000개 이상이 등록되어 있다(Distimo World Mobile Congress, 2010). 애플 앱스토어 관련 통계를 살펴보면, 애플 앱스토어를 통해 응용 프로그램을 구매하기 원하는 사용자들은 2010년 현재 총 58,000,000명이 넘어가고 있는데, 이들은 한 달에 평균 280,000,000개 이상의 응용 프로그램을 다운로드 한다고 한다(GigaOM, 2010). 즉, 한 사용자가 한 달에 평균 4.8개의 응용 프로그램을 다운로드 한다는 것이다. 하지만, 사용자들이 자신이 원하는, 자신에게 적합한 응용 프로그램을 정확히 찾아서 살펴보고 다운로드 한 것은 아닐 것이다. 현실적으로 모든 응용 프로그램에 관한 설명을 다 읽어보고 선택한다는 것은 불가능에 가깝기 때문이다. 그렇다면 애플 앱스토어를 사용하는 사용자들은 자신이 원하는, 자신에게 적합한 응용 프로그램을 찾기 위해 어떤 노력을 들일까? 어떻게 응용 프로그램을 찾고, 다운로드 하는 것일까? 애플에서는 사용자들이 더 쉽고, 더 적은 노력을 들여 응용 프로그램을 찾고, 더 많은 응용 프로그램을 다운로드 할 수 있도록 어떤 서비스를 제공해주고 있을까?

그림 1 GigaOM : The Apple App Store Economy

사용자들은 애플 앱스토어에 있는 응용 프로그램에 관한 정보들을, 다른 인터넷을 통한 정보들을 얻는 방법과 마찬가지로, 크게 두 가지 방법을 통해 얻을 수 있다. 애플 앱스토어 내부에서 제공해주는 직접적인 방법을 통해 정보를 얻는 것과, 애플 앱스토어 외부에서 정보를 얻는 방법이다. 애플 앱스토어 외부에 존재하는 아이팟, 아이폰 등과 관련된 커뮤니티들을 살펴보면, 대부분 추천 응용 프로그램이라는 좋은 응용 프로그램을 소개하고, 공유하기 위한 메뉴를 가지고 있다. 많은 사용자들이 공통적으로 좋은

응용 프로그램을 찾기 위해 많은 노력을 들이고 있기 때문에 그 노력을 줄이기 위한 방법으로 생겨난 메뉴일 것이다. 또한, 많은 블로그나 기사, 광고 등 애플 앱스토어가 아닌 다른 곳에서 사용자들은 응용 프로그램에 관한 정보를 얻을 수 있다.

애플 앱스토어 내부를 통해 사용자들이 더 쉽고, 빠르게 자신이 원하는, 자신에게 적합한 응용 프로그램을 찾는 데 도움을 받을 수 있는 것으로는 두 가지의 리스트가 있다. 하나는 새로운 응용 프로그램들 중에서 애플에서 선정한 좋은 응용 프로그램을 추천해주는 리스트이고, 다른 하나는 다운로드 횟수를 기준으로 제공해주는 리스트이다. 하지만, 기존의 응용 프로그램과 새로운 응용 프로그램들 중에서 이 리스트를 통해 확인할 수 있는 응용 프로그램의 수는 인터넷 사이트에 직접 적속해 여러번의 클릭을 해도, 최대 200여개에 불과하다. 아이팟이나 아이폰에서 직접 리스트를 통해 정보를 얻을 수 있는 응용 프로그램의 수는 이것보다 더 적다. 즉, 사용자들이 모든 응용 프로그램에 관한 정보를 살펴보고, 자신이 원하는, 자신에게 적합한 응용 프로그램을 찾아 선택하는 것은 불가능하다.

우리나라 최대 인터넷 오픈마켓인 G마켓(http://www.gmarket.co.kr)의 경우는 애플 앱스토어의 경우보다 사용자들이 자신이 원하는 아이템을 찾기가 더 힘들다고 할 수 있다. G마켓에는 셀 수 없을 정도로 수 없이 분류된 카테고리가 존재하고, 이 다양한 카테고리 안에 많은 아이템들이 존재한다. 여성 관련 카테고리만 하더라도 무려 100여 개에 달하며, 그 중 하나인 비치웨어의 경우 매일 70여 개 이상의 아이템들이 매일 새로 등록되고 있다(G마켓, 2010).

그림 2 G마켓 여성의류/패션 내 카테고리

G마켓 사이트에 접속한 후, 두 번 이상의 클릭을 통해 들어간 세부 카테고리 안에 매일 수십여 개 이상의 새로운 아이템들이 등록되고 있는 것이다. 하지만, 사용자들에게 제공되는 리스트의 수는 한 화면에 불과 십여 개에 불과하다. 게다가, 십여 개만 제공되는 리스트에도 해당 아이템들에 관한 정보를 다 보여주기 불가능하기 때문에, 간략한 이름, 설명, 작은 사진 정도만 보여주고 있다. 사용자가 관심을 갖는 아이템에 관해 더 자세한 정보를 얻기 위해 또 다른 클릭이 필요하다. G마켓을 통해 사용자들이 자신이 원하는 아이템을 찾기 위해선 어쩔 수 없이 많은 노력을 들여야 하고, 많은 노력을 들이더라도 자신이 원하는, 자신에게 적합한 아이템을 찾기란 매우 힘들다.

애플 앱스토어나 G마켓 이외에도 우리 생활에서 사용자들이 자신이 원하는 아이템을 찾는데 어려움을 겪는 경우는 수 없이 찾아볼 수 있다. 음반이나 서적의 경우에는 그나마 매일 셀 수 있을 정도의 새로운 아이템들이 생겨난다. 하지만, 뉴스와 같은 경우에는 셀 수조차 없을 정도로 많은 아이템들이 쏟아져 나온다. 이렇게 방대한 정보들 속에서 사용자들은 자신들이 원하는 정보를 더욱 빨리, 쉽게 찾기를 원하고, 정보 제공자들은 사용자들의 욕구를 만족시켜 주기 위해 많은 노력을 하고 있다.

정보 제공자들이 사용자들의 욕구를 만족시켜주기 위한 노력 중 가장 대표적인 것은 추천 시스템이다. 쏟아지는 새로운 정보들 안에서 일대일 마케팅(One-to-one marketing), 웹 개인화(Web personalization), 맞춤화된 고객 관계 관리(Customer Relationship Management: CRM) 등과 같은 것들을 가능하게 해주는 추천 시스템은 실무에서 뿐 아니라 연구에서도 그 중요성이 점차 높아지고 있다(Im and Hars, 2007; Mile and

Natter, 2002; Sarwar et al., 2000; Xiao and Benbasat, 2007).

본 연구는 인터넷을 통해 사용자들이 보다 쉽게 자신이 원하는 정보를 찾고, 자신에게 적합한 아이템을 찾을 수 있도록 도와주는 추천 시스템에 주목하였다. 구체적으로는 협업 필터링 방식을 사용하여 추천 리스트를 생성해주는 협업 필터링 추천 시스템의 성능 향상에 관심을 두고 있다.

1.2 연구목적 및 질문

사용자들의 잠재적인 정보 요구에 부합하는 자료들을 자동으로 검색하여 리스트 형식으로 제공해주는 시스템인 추천 시스템은 이미 많은 분야에서 사용되고 있다(Baumann and Hummel, 2005; Im and Hars, 2007). 또한, 추천 시스템의 성능 향상을 위한 연구들은 다양한 분야에서 계속 진행되고 있다(Chen and Aickelin, 2004; Hill et al., 1995; Miller et al., 2003; O'connor and Herlocker, 2001; Sarwar et al., 2000; Xiao and Benbasat, 2007).

우리나라의 대표적인 포털 사이트인 다음(daum)의 경우 추천 시스템을 사용해 제공해주는 리스트의 수가 첫 화면에만 무려 10여 개에 달한다. 추천 검색어와 이슈 검색어, 주요 뉴스, 유익한 정보 검색 등이 그것이다. 이외에도 대부분의 쇼핑 사이트나 오픈 마켓, 인터넷 서점 등 수 많은 사이트들이 추천 시스템을 통한 추천 리스트를 제공해 줌으로 인해 사용자들이 넘쳐나는 정보들 속에서 자신들이 원하는 정보를 손쉽게 찾을 수 있도록 도와주고 있고, 이를 통해 해당 사이트의 정보 소비량과 사용량이 늘어나길 바라고 있다.

기업과 학계에서는 다양한 방법의 투자와 연구를 통해 자신들이 사용하고 있는 추천 시스템의 성능 향상을 위해 노력하고 있다. 대표적인 미국의 비디오 대여 사이트인 Netflix의 경우, 총 상금 1,000,000 달러의 대회(Netflix Prize)를 통해 자신들의 추천 시스템 성능 향상을 위해 노력하였다(http://www.netflixprize.com).

추천 시스템은 기존의 연령이나 지역 등으로 사용자들을 구분해 마케팅을 진행하는 것이 아닌, 사용자 개개인에 맞춘 마케팅을 가능하게 해준다(Schafer et al. 2001). 또한, 사용자 개개인에 특화된 사이트를 제공하는 웹 개인화를 통해 개인별로 맞춤화된 서비스를 가능하게 해주며, 정보 과부하 문제를 해결할 수 있다(Changchien and Lu, 2001; Yuan and Chang, 2001). 즉, 추천 시스템은 여러 장점들을 통해 정보 제공자와 소비자 모두에게 부가가치를 제공해주는 것이다.

본 연구는 이러한 추천 시스템의 성능을 향상시킬 수 있는 새로운 방법을 제시하고, 검증해 보고자 한다. 본 연구의 목적인 추천 시스템의 성능 향상이 이루어진다면, 추천 시스템을 통해 지금보다 더 많은 부가가치를 정보 제공자와 소비자 모두에게 제공할 수 있을 것으로 판단된다.

1.3 연구방법 및 논문의 구성

본 연구에서는 추천 시스템 중 협업 필터링 방식을 사용하여 추천 리스트를 생성하는 협업 필터링 추천 시스템의 성능 향상에 관한 새로운 방법을 제시하는데 목적을 두고 있다. 따라서 본 연구는 협업 필터링 추천 시스템에서 사용 가능한 새로운 방법을 제시한 후, 시뮬레이션 기법을 통해 기존의 방법과 새로운 방법을 비교하여 개선 정도를 검증하였다. 같은 방법을 사용한 추천 시스템이라고 해도, 사용하는 자료에 따라 다른 결과 값을 나타낼 수 있기 때문에 기존의 방법으로 MAE, RMSE 값을 다시 계산하였고, 새로운 방법으로 같은 값을 계산하여 비교, 분석하였다.

본 논문은 서론, 문헌연구, 연구 가설, 연구 대상 및 연구 방법, 가설 검증 및 결과 분석, 결론의 총 6장으로 구성되어 있다. 제 1장인 서론에서는 본 연구의 배경과 목적 등을 제시하고 있다. 제 2장은 문헌연구로서 본 연구의 목적을 달성하기 위해 필요한 추천 시스템과 그에 관련된 여러 이전 연구들에 관한 내용을 담고 있다. 제 3장에서는 2장의 선행연구를 통하여 새롭게 제안하는 추천 시스템 방법에 관하여 소개하고 있다. 제 4장은 3장에서 제안하는 새로운 방법을 확인하기 위한 구체적인 방법을 소개하고 있다. 제 5장에서는 새로운 방법으로 도출된 결과를 이전 방법과 비교하여 추천 시스템의 성능 향상 여부를 검증하였다. 제 6장은 학문적, 실무적 시사점과 연구의 한계를 밝히고 향후 연구방향을 제안하고 있다.

제 2 장 문헌연구

2.1 추천 시스템

2.1.1 추천 시스템의 정의

추천 시스템은 정보 사용자가 관심을 가지는 아이템에 관한 정보나 정보 사용자의 인구통계학적 정보 또는 과거 구매행동의 분석을 토대로 정보 사용자가 원할 것이라고 예상되는 아이템을 미리 추천하여 주는 시스템이다(Sarwar et al., 2001). 즉, 정보 사용자의 잠재적인 정보 요구에 부합하는 자료를 자동으로 검색하여 제공하여 주는 시스템이라고 할 수 있다(Gediminas and Alexander, 2005). 이러한 추천 시스템을 고객들이 구매하고자 하는 상품을 쉽게 찾을 수 있도록 도와주는 정보 필터링 기술 혹은 고객들의 편의를 도모하고 교차판매 및 매출 증대에 초점을 맞춘 시스템으로 정의하기도 한다(Schafer et al., 1999; 박지선 외, 2000).

추천 시스템은 사용자가 자신이 필요로 하는 정보를 검색 키워드의 형태로 명시해야 하는 일반적인 검색 시스템에 비해, 보통 키워드의 입력 없이 미리 가지고 있는 정보를 토대로 사용자의 행동, 패턴 등에 따라 아이템에 대한 선호도를 예측해야 한다(Ricardo and Berthier, 1999). 그렇기 때문에 기본의 아이템들에 관한 정보 외에도 사용자들에 관한 정보, 구매이력들에 대한 정보들도 저장하고 있어야 하고, 그 정보들 안에서 보다 더 복잡한 알고리즘을 통해 소비자들에게 제공할 추천 리스트를 만들어 내야

하기 때문에, 보통 일반적인 검색 시스템들보다 더 많고 수준 높은 분석 작업이 요구된다(Gediminas and Alexander, 2005; 이경종 외, 2007).

2.1.2 추천 시스템의 장점

추천 시스템은 일대일 마케팅(One-to-one marketing)을 가능하게 해주고, 웹 개인화(Web personalization)와 맞춤화 서비스를 가능하게 해주며, 정보 과부하 문제를 해결해 정보 제공자와 소비자 모두에게 부가가치를 제공한다는 장점이 있다(Berson et al., 2000; Changchien and Lu, 2001; Yuan and Chang, 2001).

추천 시스템은 사용자들을 단순화하기 위해서 성별이나 나이, 지역 등의 인구통계학적인 정보를 토대로 나눌 필요성이 없다. 기존의 마케팅에서는 현실적인 어려움 때문에 사용자들을 몇몇 특성에 따라 그룹화하고, 그 그룹들 안에서 새로운 아이템을 제일 잘 소비할 것으로 예상되는 그룹을 선정해 그들을 목표로 하는 마케팅 기법을 사용했다(Stephen and Robert, 2004). 하지만, 추천 시스템은 사용자 그룹을 더 세분화할 수 있고, 개인별로 특성을 정의해 그에 맞는 마케팅을 가능하게 해준다. 추천 시스템을 통해 불가능에 가까웠던 일대일 마케팅이 가능해지는 것이다(Weng and Liu, 2004; Berson et al., 2000).

웹 개인화와 맞춤화 서비스를 가능하게 해주는 것도 추천 시스템이다. 모든 사용자 혹은 몇 그룹으로 나누어진 사용자들에게 일괄적으로 같은 화면 구성과 메뉴를 보여주는 것이 아닌, 웹 페이지의 내용과 화면 구성을

개인별로 사용자가 원하는 방식으로 보여주는 웹 개인화와 맞춤화 서비스는 추천 시스템의 사용으로 가능해 졌으며(Mobasher et al., 2002), 이를 통해 사용자들은 해당 웹 페이지에 더 오래 머물고 있다(Kravatz, 2000).

또한, 사용자들은 추천 리스트를 사용함으로 정보를 검색하는데 들어가는 노력이 줄어들기 때문에, 적합한 아이템을 추천받은 사용자들은 해당 서비스를 제공하는 사이트에 대한 충성도가 높아진다(Kim et al., 2002). 뿐만 아니라, 이런 과정을 통해 해당 사이트와 사용자간의 유대감이 높아진다. 즉, 추천 시스템이 고객 관리에 긍정적인 영향을 주는 것이다(Mile and Natter, 2002).

그리고 추천 시스템은 사용자가 자신이 원하는 정보를 직접적인 노력을 들여 찾아보기 이전에 자동으로 사용자가 원할 것이라고 예상되는 적합한 정보를 제공해주기 때문에, 많은 정보들 안에서 소비자가 원하는 정보를 더 빨리, 보다 적은 노력을 들여 찾을 수 있게 해해주기 때문에 정보 과부하 문제를 해결한다(Berson et al., 2000). 결국, 위의 장점들로 인해 추천 시스템은 정보 제공자와 사용자 모두에게 부가가치를 제공해 준다.

2.1.3 추천 시스템의 추천 리스트 생성 방법

추천 시스템에서 사용하는 추천 리스트 생성 방법에는 크게 세 가지가 있다. 아이템이나 사용자에 관련된 내용을 분석해 그것을 기반으로 추천 리스트를 생성해주는 내용 기반 추천 시스템과, 비슷한 성향을 보이는 다른 아이템과 다른 사용자들의 정보를 바탕으로 새로운 아이템을 추천해주

는 협업 필터링 추천 시스템, 그리고 여러 가지 다양한 방식을 혼합해 추천 리스트를 생성하는 하이브리드 방식의 추천 시스템이다.

2.2 내용 기반 추천 시스템

2.2.1 내용 기반 추천 시스템의 정의

내용 기반 추천 시스템은 정보 검색 기술에 바탕을 둔 추천 시스템으로 아이템이나 사용자에 관련된 내용을 분석하여 아이템과 아이템간의 유사성 혹은 아이템과 사용자간의 유사성을 바탕으로 새로운 아이템을 추천해주는 시스템이다(Resnick and Varian, 1997). 도서 검색에서 주로 사용되는 방식으로 문헌 정보학에서 가장 많이 사용하는 방식이며, 가장 활발한 연구가 이루어지고 있다. 내용 기반 추천 시스템은 텍스트가 내용을 자동으로 분석하기 가장 쉽기 때문에 주로 텍스트를 기반으로 한 아이템들을 추천해주는데 주로 사용되었지만(Krulwich and Burkey 1996; Lang, 1995) 최근에는 텍스트가 기반인 서적과 뉴스, URL 등에서 벗어나 영화, 음악에 이르기까지 다양한 아이템들을 추천하는데 사용되고 있다 (Baumann and Hummel, 2005).

내용 기반 추천 시스템은 [그림 3]과 같은 방법으로 추천 리스트를 생성하게 된다. 우선, 아이템이나 사용자가 가지고 있는 속성들을 분석하여 구조화된 정보로 저장한다. 그 후, 어떤 정보들을 어떻게 사용하여 아이템이나 사용자의 특성을 나타내야 사용자의 선호도를 예측하는데 유용할 것인지 정의하고, 그것에 따라 사용자의 선호도 특성을 프로파일한다. 마지막으로, 선정된 속성들을 다양한 방법과 공식들을 사용하여 아이템에 대한 선호도를 예측하고, 정해진 방법에 따라 추천 리스트를 생성한다.

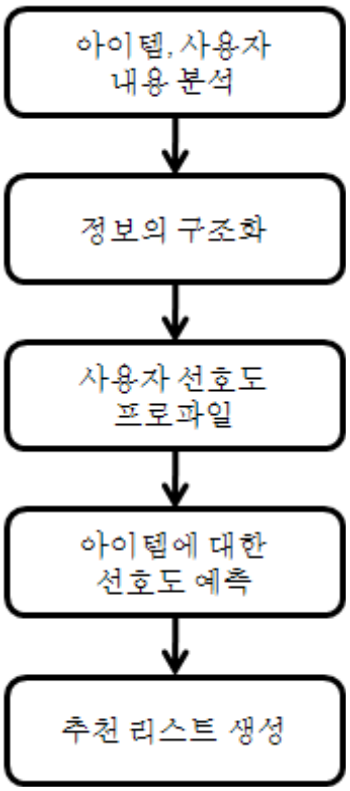

그림 3 내용 기반 추천 시스템의 추천 리스트 생성 과정

2.2.2 정보의 구조화

내용 기반 추천 시스템에서 아이템이나 사용자의 특성을 저장하고 사용하기 위해 가장 먼저 필요한 것은 수집된 정보의 구조화이다. 문서나 뉴스, URL 등의 경우 주로 텍스트로만 이루어져 있기 때문에 아이템 전체를 다 저장하고 사용할 것인지, 제목이나 목차, 키워드만 저장하고 사용할 것인지 등에 대한 비교적 간단한 방법을 통한 기준을 선정해 정보를 구조화 시킬 수 있다.

텍스트를 구조화하기 위해선 사전 작업이 필요하다. 특수 문자나 공백 등 불필요한 것들을 제거하고, 유의어를 통합하여 단어의 수를 줄이는 것 등이다. 텍스트를 수치로 변환하기도 하는데, 나타난 단어에는 1, 나타나지 않은 단어에는 0의 값을 준다던가, 단어의 출현 횟수를 센다던가 하는 것이다.

내용 기반 추천 시스템에서 가장 많이 사용하는 방법은 한 문서에서의 단어 출현 횟수와 전체 문서에서의 단어 출현 횟수에 가중치를 주는 TF-IDF(Term Frequency-Inverse Document Frequency)와 해당 아이템들 사이의 벡터 값을 계산하는 것이다.

[식 1]은 TF-IDF의 여러 공식 중 하나이다.

$$tf_{t,d}\log(\frac{N}{df_t})\times\sqrt{\Sigma_i(tf_{t_i,d})^2\log(\frac{N}{df_{t_i}})}$$

식 1 log-TF-IDF

$tf_{t,d}$는 t단어가 d문서에 나타나는 횟수를 의미하고, N은 전체 문서의 수, df_t는 t단어를 포함하고 있는 문서의 수를 의미한다. TF-IDF는 자주 출현하는 단어가 해당 문서를 대표하는데 많은 역할을 하지만, 다른 문서에서도 자주 출현하는 단어라면 그 역할이 중요하지 않다는 사실에 근거를 둔 공식이다. 예를 들어, '추천 시스템' 이라는 단어가 자주 등장하는 문서의 경우 '추천 시스템' 이라는 단어는 검색을 하고자 하는 문서들의 분류에 상관없이 해당 문서에서 출현 횟수가 많기 때문에 높은 TF값을 가지

게 된다. 하지만, 이 문서를 전체 문서들 사이에서 검색을 할 때는 높은 IDF값을 가지고 전체 문서들 사이에서 해당 문서의 특징을 대표하는 데 중요한 역할을 하는 키워드로 인식되지만, 웹 개인화나 일대일 마케팅에 관련된 문서들 사이에서는 '추천 시스템' 이라는 단어가 자주 등장하기 때문에 낮은 IDF값을 갖게 되고, 그 중요성이 낮다고 인식된다.

2.2.3 사용자 선호도 프로파일

내용 기반 추천 시스템에서 사용할 아이템과 사용자의 정보들을 적절한 구조로 변경하여 저장하고 난 후, 어떤 정보들을 사용하여 사용자의 아이템에 대한 선호도를 나타낼지 정하는 프로파일 과정을 거친다. 선호도 예측에 사용할 정보들은 직접적인 방법과 간접적인 방법을 통해 수집할 수 있다. 사용자가 구매했던 아이템 내역을 사용해 구매한 아이템의 경우 선호도에 대한 가중치를 높게 주고, 그렇지 않은 아이템의 경우 낮은 선호도 가중치를 줄 수 있다. 검색했던 단어를 포함하는 아이템의 선호도에 대한 가중치를 높게 주고, 그렇지 않은 아이템에 대한 선호도 가중치를 낮게 줄 수 있다. 또는, 직접적인 선호도 입력을 사용자에게 요구해 선호도 정보를 얻을 수 있다.

아이템에 대한 사용자의 선호도에 관한 정보들을 모은 후, 분석을 통해 사용자들이 좋아하는 아이템과 싫어하는 아이템을 알 수 있고, 이것들을 바탕으로 아이템과 아이템, 아이템과 사용자들과의 관계를 분석해 최종적으로 사용자의 선호도에 대한 프로파일을 한다.

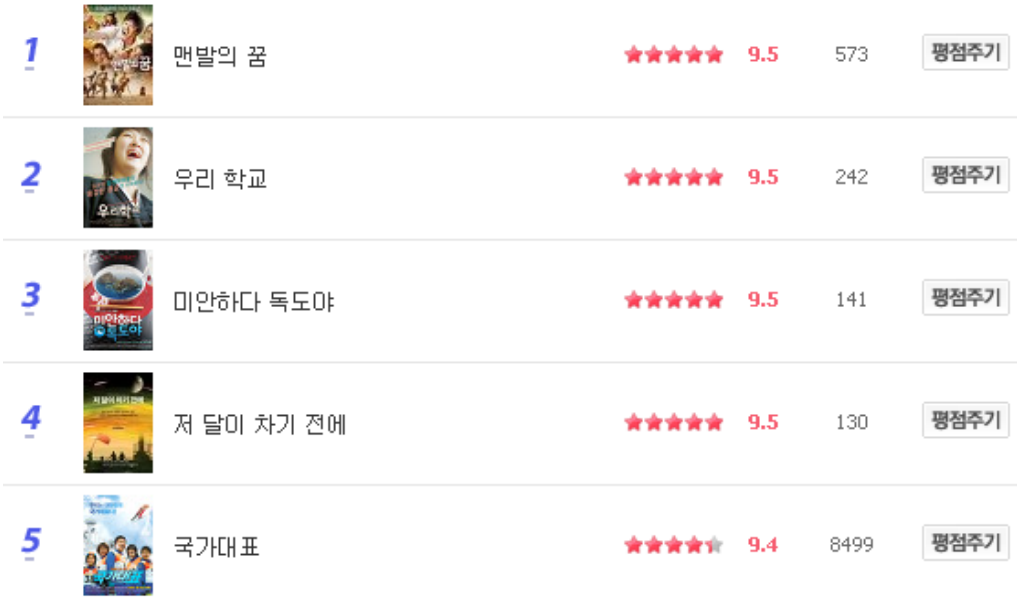

그림 4 Daum 영화 평점

2.2.4 선호도 예측

사용자의 선호도에 대한 예측에 관해 프로파일이 끝나게 되면, 이것을 기반으로 각 아이템에 대한 선호도를 예측하게 된다. 예측 대상이 되는 아이템은 추천 리스트에 들어갈 아이템으로, 사용자가 아직 선호도를 표시하지 않았거나, 구매하지 않은 아이템이다. 선호도 예측에 사용되는 방법은 Decision Tree, Rule Induction, Similarity Measure 등 여러 가지가 있다.

Decision Tree 방식의 예로 레스토랑에 대한 선호도를 예측한다고 생각해 보자. 그림으로 표현하면 다음과 같다.

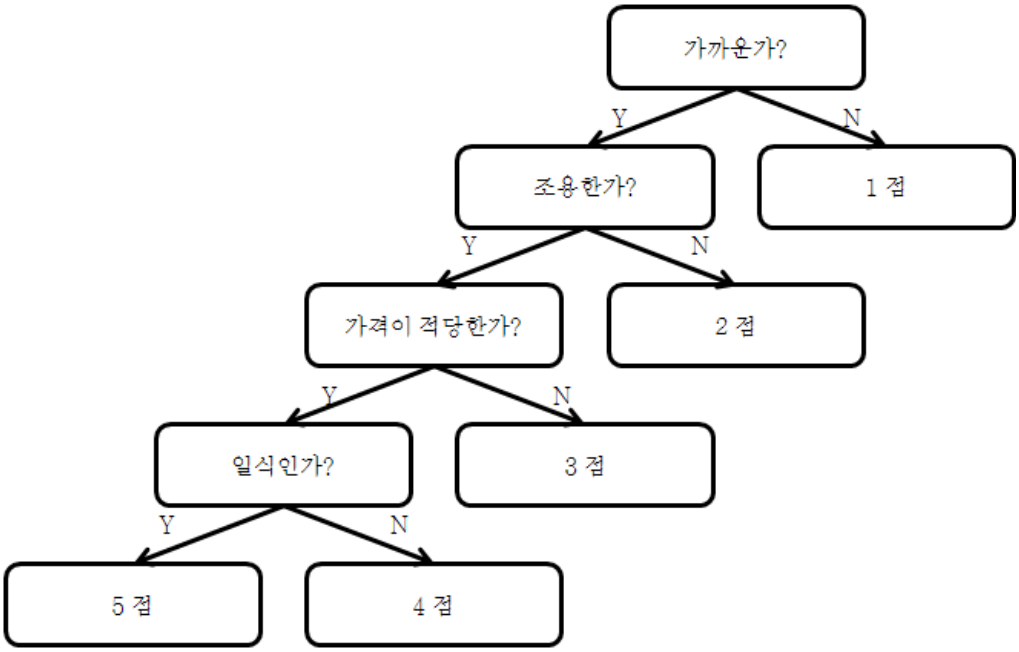

그림 5 Decision Tree

즉, 가깝고, 조용한, 적당한 가격의 일식 레스토랑의 경우 5점의 선호도를 예측하게 되는 것이고, 가깝고 조용하고, 적당한 가격이지만 일식이 아닌 레스토랑의 경우 4점의 선호도를 예측하게 되는 것이다. 이 방법에서 가장 중요한 것은 어떤 속성을 Tree 상단에 위치시킬 것인가에 관한 것이다(Alpaydin, 2004). 구조화가 용이한 아이템의 경우 적합한 방법이며, 텍스트와 같이 복잡한 경우에는 의사 결정 과정이 너무 많아져 효율적이지 못한 방법이다(Schafer et al, 2001; Quinlan, 1990).

Rule Induction은 사용자 선호도 프로파일과 기타 정보들을 바탕으로 사용자의 선호도에 관련된 규칙을 찾아내는 방식이다(Cohen, 1995;

Michael and Daniel, 2007). 학습 과정에서 어떤 규칙을 통해 사용자의 선호도가 가감되는지를 찾아내고, 이 규칙에 따라 새로운 아이템에 대한 선호도를 예측하는 것이다. 레스토랑을 추천하는 예를 Rule Induction으로 바꿔보면 다음과 같다.

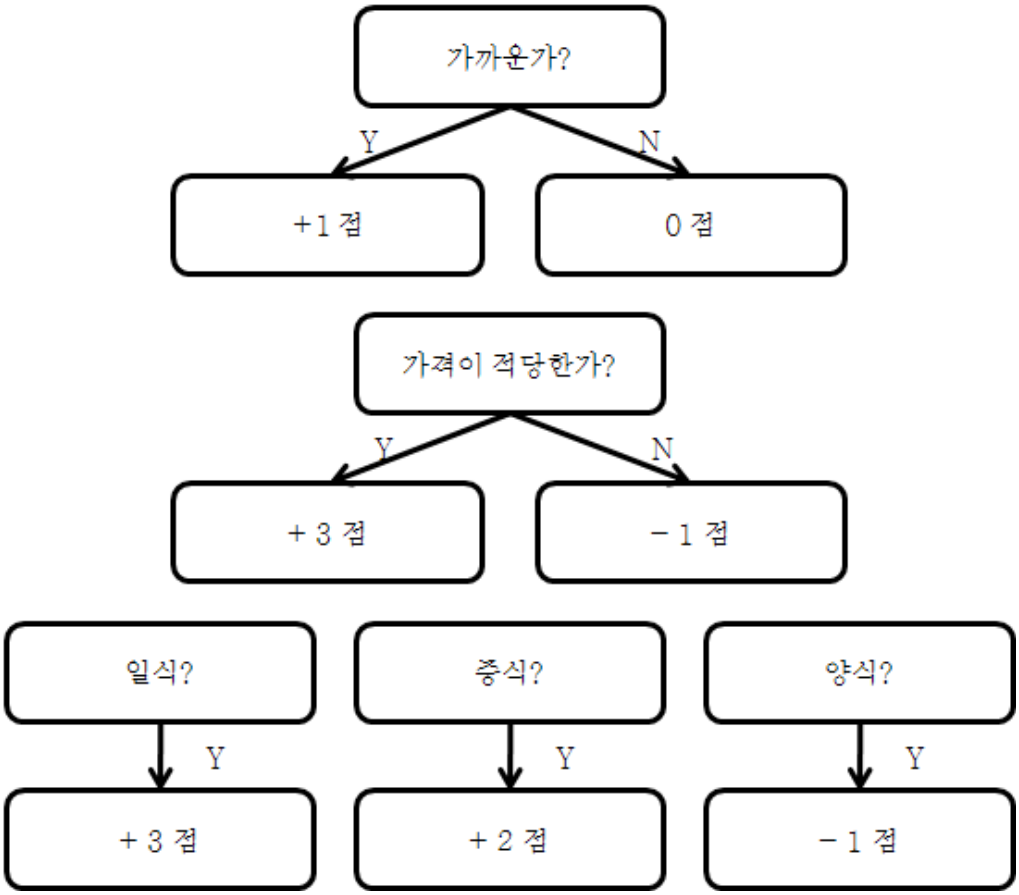

그림 6 Rule Induction

Similarity Measure는 두 아이템간의 유사성을 측정해 선호도를 예측하는 방법이다(Marko and Yoav, 1997). 예를 들어, A라는 아이템과 B

라는 아이템을 사용자 선호도 프로파일에 기초해 중간에 들어가는 가중치의 값들을 조절하며 유사성을 계산한다. A 아이템이 사용자 선호도가 높은 아이템일 경우, B 아이템 역시 사용자 선호도가 높을 것이고, 그렇지 않을 경우 사용자 선호도가 낮을 것이라고 예측하는 것이다. 주로 사용하는 유사도 계산 공식으로는 Cosine Similarity와 Euclidean Distance가 있는데, Cosine Similarity가 정규화 과정이 따로 필요하지 않기 때문에 더 자주 쓰인다(Mukund and George, 2004).

2.2.5 추천 리스트 생성

최종적인 추천 리스트는 예측된 선호도를 바탕으로 생성한다. 크게 두 가지 방법을 사용해 최종 리스트를 생성하는데, 순서대로 m 개의 아이템만을 보여주는 방식과 선호도 n 이상의 아이템만을 보여주는 방식이 있다. 상위 m개, 선호도 n은 추천 리스트의 용도, 화면 구성, 아이템의 특성 등에 따라서 임의로 결정한다.

2.2.6 내용 기반 추천 시스템의 단점

내용 기반 추천 시스템을 사용하기 위해서는 텍스트를 기반으로 하는 아이템이거나 텍스트가 아닌 경우 텍스트나 수치로 구조화하기 용이한 특성을 가지고 있는 아이템이어야 한다는 단점이 있다(Balabanovic and Shoham, 1997). 따라서 사용할 수 있는 아이템들이 적은 편이다. 문서나 뉴스와 같은 경우 원래 텍스트로 이루어진 아이템이기 때문에 큰 문제가

없지만, 음식의 맛과 향, 음악의 느낌, 영화의 재미, 옷의 아름다운 정도에 관한 것들을 텍스트와 같이 구조화하기는 힘들다.

또한, 내용 기반 추천 시스템은 기본적으로 사용자가 선호했던 아이템과 비슷한 아이템을 찾아 추천하는 방식이기 때문에, 기존에 선호했던 아이템과 비슷한 아이템만을 찾아 추천해주는 단점이 있다(Shoham and Citeseer, 1997). 즉, 아이템의 지나친 특성화가 발생하게 되는 것인데, 전혀 없었던 특성을 가진 새로운 특성의 아이템이나 그 특성을 정의하기 힘든 아이템의 경우에는 추천 리스트에 들어가기 어렵다. 이러한 단점을 극복하기 위한 방법으로 무작위로 추출된 아이템을 추천 리스트에 일정 부분 추가하거나, 유사도 계산 과정 중간에 랜덤하게 생성된 가중치를 적용하기도 한다(Sheth and Maes, 1993).

마지막으로, 대부분의 추천 시스템의 단점에 해당하는 것으로, 사용자 선호도 프로파일을 구축하기 위한 정보들이 꼭 필요하다는 것이다. 내용 기반 추천 시스템은 사용자 선호도 프로파일이 구축되어야만 추천이 가능한데, 이 사용자 선호도 프로파일을 구축하기 위해선 사용자의 이전 아이템에 대한 선호도를 예측할 수 있는 평점, 구매 내역 등이 필요하다. 즉, 자신의 정보를 공개하기 싫어하는 사용자나, 새로운 사용자에 대한 정확한 추천은 자료의 희박성 때문에 정확하게 하기 힘들다.

2.3 협업 필터링 추천 시스템

2.3.1 협업 필터링 추천 시스템의 정의

협업 필터링 방식의 추천 시스템은 내용 기반 방식의 추천 시스템과 함께 추천 시스템에서 자주 쓰이는 방식 중 하나로, 오프라인의 구전효과(Shardanand and Maes, 1995)를 자동화한 시스템이라고 할 수 있다(Konstan et al, 1997). 즉, 협업 필터링 추천 시스템은 목표 사용자가 아이템을 선호하는 패턴과 유사한 다른 사용자들을 찾아, 그들의 선호도를 바탕으로 목표 사용자의 선호도를 예측하는 방식이다. 이는 사용자의 아이템에 대한 선호도가 일반적인 유행을 따르거나, 일정한 패턴을 가지고 있다는 사실에 근거한 것이다. 실제로 많은 연구들이 이러한 사실에 근거해 연구들을 진행하였고, 긍정적인 연구 결과를 얻었다(Herlocker et al., 2004; Hofmann et al., 2004).

협업 필터링 추천 시스템과 관련된 연구는 정보 검색이나 인공 지능, 인지 과학 분야에서 처음 시작되었으나, 현재에는 그 사용처가 다양해짐에 따라 다양한 분야에서 성능 향상에 관한 연구가 이루어지고 있다.

협업 필터링 추천 시스템의 성능에 가장 큰 영향을 주는 부분은 목표 사용자와 유사한, 다른 사용자들을 찾아내는 방법이다. 유사한 사용자들을 찾는 방법에 따라 협업 필터링 추천 시스템을 사용자 기반 협업 필터링 추천 시스템과 아이템 기반 협업 필터링 추천 시스템으로 구분할 수 있다(Herlocker et al., 1999).

사용자 기반 협업 필터링 추천 시스템은 사용자와 사용자간의 선호도에 대한 유사성을 측정하여 유사성이 높은 사용자들의 선호도가 높은 아이템을 추천해주는 방식이다(Basu et al., 1998). 즉, 유사성이 높은 사용자들을 찾아내고 정의하는 것이 사용자 기반 협업 필터링 추천 시스템의 핵심이다. 하지만, 사용자들 간의 유사성을 찾아내기 위한 정보들이 부족할 경우 아이템 기반 추천 시스템의 성능이 떨어진다는 단점이 있다(Sarwar et al., 2001). 이러한 단점을 개선하기 위해 유사한 고객들을 찾은 후 그들이 평가한 아이템들 사이의 유사도를 사용해 성능을 향상시키는 방식이 고안되었고, 이러한 방식의 협업 필터링 추천 시스템을 아이템 기반 협업 필터링 추천 시스템이라 한다(Karypis, 2001; Wang et al, 2006).

협업 필터링 추천 시스템은 위에서 언급한 방식 외에도 다양한 방식을 사용하여 성능 향상을 위한 연구들이 진행되고 있다. 클러스터링 기법을 아이템 기반 협업 필터링 추천 시스템에 응용하기도 하였고(Kim et al., 2002), 선호도에 장바구니 정보를 추가하여 연구를 진행하기도 하였다(Weng and Liu, 2004). 또한, 방대한 정보를 사용하기 때문에 데이터마이닝 기법이나 SOM을 응용한 방식을 통한 협업 필터링 추천 시스템에 대한 연구도 있다(Fu et al., 2000; Roh et al., 2003).

협업 필터링 추천 시스템은 내용 기반 추천 시스템에 비해 다양한 장점을 가지고 있다. 내용 기반 추천 시스템은 그 내용이 수치화가 용이한 정보에 한해 추천에 사용할 수 있다는 단점이 있다. 하지만 협업 필터링 추천 시스템은 내용 기반 추천 시스템에서 다루기 힘든 정보들도 사용해 추천할 수 있다(Abhinandan et al., 2007). 음식의 맛에 관한 정보를 추천

시스템에서 사용하기 원할 경우 내용 기반 추천 시스템에서는 몇몇의 전문가에게 의뢰해 맛에 대한 정보를 수치화 하거나 사용 자체를 포기하기도 한다. 하지만 협업 필터링 추천 시스템에서는 많은 사용자들이 수치로 평가한 선호도를 사용할 수 있다. 음악의 느낌이나 영화의 재미와 같은 정보들도 협업 기반 필터링에서 추천 리스트 생성에 사용할 수 있는 것이다. 즉, 사용자들의 취향과 상품의 질을 내용 기반 추천 시스템에 비해 더 잘 반영할 수 있게 되는 것이다. 또한, 다른 사용자들의 경험을 기반으로 추천이 가능하기 때문에 내용 기반 추천시스템에 비해 결과에 대한 만족도가 높다(Konstan et al., 1997).

2.3.2 협업 필터링 추천 시스템의 알고리즘

협업 필터링 추천 시스템의 알고리즘은 [그림 7]과 같다.

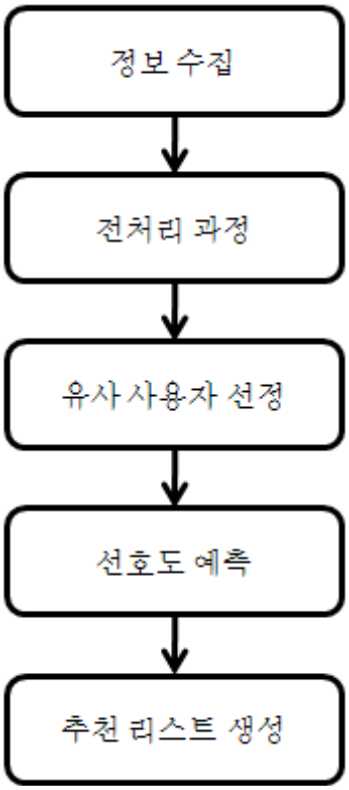

그림 7 협업 필터링 추천 시스템의 알고리즘

협업 필터링 추천 시스템에서 가장 먼저 필요한 것은 추천 과정에 필요한 정보를 수집하는 것이다. 내용 기반 추천 시스템에서 주로 사용하는 사용자나 아이템의 내용에 대한 정보는 꼭 필요하지는 않지만, 세부적인 방식에 따라 필요한 경우도 있다. 보통 아이템에 대한 사용자들의 선호도나

구매 내역 정보 등이 필수적으로 필요한 정보에 속한다. 이외에 상황에 따라 다른 정보들을 수집한다. 정보들을 수집 한 후 해당 협업 필터링 추천 시스템에서 사용하기 용이한 형태의 정보로 구조를 변경하는 전처리 과정을 거친다. 그 후 구조화 된 정보들을 사용하여 목표 사용자와 유사한 사용자 그룹을 찾아 정의하고, 유사 사용자 그룹의 선호도를 바탕으로 목표 사용자의 선호도를 예측한다. 마지막으로 예측한 선호도를 바탕으로 추천 리스트를 생성하여 미리 정의된 형식에 맞춰 보여주게 된다.

2.3.3 정보의 구조화

협업 필터링 추천 시스템에서 보통 필수적인 정보는 사용자의 아이템에 대한 선호도이다. 이 정보는 보통 7점 척도나 5점 척도로 입력 받는 것이 보통이며 이러한 정보의 경우 특별히 구조화를 위한 노력이 필요하지는 않다. 하지만 숫자로 입력 한 선호도가 아닌 텍스트의 경우 텍스트를 분석해 숫자로 변환하는 과정이 필요하기도 하다.

협업 필터링 추천 시스템에서는 그 방식에 따라 입력 정보가 희박할 경우 예측이 부정확해지는 단점을 보완해주기 위해 유사한 사용자나 아이템을 클러스터링 기법을 사용해 그룹화 하기도 한다. 이러한 경우 유사 사용자나 아이템에 대한 정의가 필요하고 그들을 동일한 사용자와 아이템으로 처리하기 위한 과정이 필요하다.

일반적인 협업 필터링 추천 시스템에서는 위와 같은 정보의 구조화 과정을 통해 사용자 아이디, 아이템 아이디 등 구분을 위한 정보와 구매 유

무, 선호도, 날짜 등과 같은 유사성을 측정하기 위한 정보들로 구분되어 구조화 한다.

2.3.4 유사도 계산 공식

협업 필터링 추천 시스템에서 핵심적인 부분이 유사한 사용자를 찾는 과정인 만큼 유사성을 계산하는 공식에 관한 다양한 시도들이 있어왔다. 이전 연구들에서 사용한 유사도 계산 공식은 Mean Squared Difference, Pearson Correlation Coefficient, Constrained Pearson Correlation Coefficient, Spearman Correlation Coefficient, Significance Weighting, Variance Weighting 등이 있다.

Mean Squared Difference는 사용자간에 공통으로 평가한 아이템 선호도의 차를 사용한 방식이다.

$$\sum_{i=1}^{m}(r_{a,i}-r_{b,i})^2$$

a : 목표 사용자　　b : 이웃 사용자

m : 사용자 a, b가 공통적으로 선호도를 평가한 아이템 수

$r_{a,i}$: 사용자 a의 아이템 i에 대한 선호도

$r_{b,i}$: 사용자 b의 아이템 i에 대한 선호도

식 1 Mean Squared Difference

[식 1]은 Mean Squared Difference를 나타낸 것이다. 유사도 계산 공

식 중 가장 간단한 공식으로 사용자 개개인의 선호도 분포는 고려하지 않고 단순히 선호도간의 차를 제곱하여 더한 값을 사용한다.

Pearson Correlation Coefficient는 상관관계 분석에서 가장 보편적으로 사용되는 공식이다(Konstan et al., 1997).

$$\sum_{i=1}^{m} \frac{Ave(r_{a,i} \times r_{b,i}) - (Ave(r_{a,i}) \times Ave(r_{b,i}))}{Std(r_a) \times Std(r_b)}$$

a : 목표 사용자　　b : 이웃 사용자

Ave : 평균　　Std : 표준편차

m : 사용자 a, b가 공통적으로 선호도를 평가한 아이템 수

$r_{a,i}$: 사용자 a의 아이템 i에 대한 선호도

$r_{b,i}$: 사용자 b의 아이템 i에 대한 선호도

r_a, r_b : 사용자 a, b의 전체 선호도

식 2 Pearson Correlation Coefficient

[식 2]는 Pearson Correlation Coefficient 공식이다. 이 공식의 결과로 나오는 상관계수는 −1부터 1까지 값을 갖는다. 즉, 계산된 아이템들 사이의 선호도가 유사할수록 1에 가까운 상관계수 값을 갖고, 반대의 경우 −1에 가까운 값을 갖는다. 상관관계가 성립하지 않을 경우 0의 값을 갖는다.

Constrained Pearson Correlation Coefficient는 선호도가 1점부터 양의 방향으로만 되어있는 것을 보완하기 위해 사용된 공식이다(Shardanand, 1995). 척도의 중간 값을 0점으로 변환한 후 선호도가 같

은 방향일 경우만 상관관계를 계산한다.

$$\sum_{i=1}^{m} \frac{Ave[(r_{a,i}-4)\times(r_{b,i}-4)]-(Ave(r_{a,i})\times Ave(r_{b,i}))}{Std(r_a)\times Std(r_b)}$$

a : 목표 사용자　　b : 이웃 사용자

Ave : 평균　　Std : 표준편차

m : 사용자 a, b가 공통적으로 선호도를 평가한 아이템 수

$r_{a,i}$: 사용자 a의 아이템 i에 대한 선호도

$r_{b,i}$: 사용자 b의 아이템 i에 대한 선호도

r_a, r_b : 사용자 a, b의 전체 선호도

식 4 Constrained Pearson Correlation Coefficient

[식 4]는 7점 척도로 선호도를 입력받아 중간값이 4점인 경우 사용하는 Constrained Pearson Correlation Coefficient 공식이다. 즉, 7점 척도에 5점과 6점의 선호도가 있을 경우 각각 1점과 2점으로 변환되며 같은 방향이기 때문에 상관관계를 계산하고 선호도 예측에 사용하지만, 2점과 5점이 있을 경우 −2점과 1점으로 변환되고 다른 방향이기 때문에 상관관계를 계산하지 않는다.

Spearman Correlation Coefficient는 평점을 순위로 환원시켜 상관관계를 계산한다.

$$\sum_{i=1}^{m} \frac{Ave[(rank_{a,i}) \times (rank_{b,i})] - (Ave(rank_{a,i}) \times Ave(rank_{b,i}))}{Std(r_a) \times Std(r_b)}$$

a : 목표 사용자　　b : 이웃 사용자

Ave : 평균　　Std : 표준편차

m : 사용자 a, b가 공통적으로 선호도를 평가한 아이템 수

$rank_{a,i}$: 사용자 a의 아이템 i에 대한 선호도의 순위

$rank_{b,i}$: 사용자 b의 아이템 i에 대한 선호도의 순위

r_a, r_b : 사용자 a, b의 전체 선호도

식 5 Spearman Correlation Coefficient

[식 5]는 Spearman Correlation Coefficient 공식이다. Pearson Correlation Coefficient는 아이템의 선호도 관계가 선형적이며 오차의 평균이 0이라는 가정을 하고 있는데 이러한 가정이 성립하지 않는 경우 예측 정확도가 떨어진다. 이러한 상황에서의 예측 정확도를 향상시키기 위한 목적으로 변형된 공식이지만, 이전 연구에 따르면 이 공식과 기존 Pearson Correlation Coefficient 공식을 사용한 경우 결과에 큰 차이는 없다(Herlocker et al., 1999).

Significance Weighting과 Variance Weighting은 기존의 공식에서 사용자에 따라 다른 가중치를 주도록 공식을 수정한 것이다. Significance Weighting은 사용자가 동일하게 평가한 아이템의 수에 따라 가중치를 다르게 적용한다. 아이템 선호도에 있어서 유사도가 같은 사용자가 있다고

하더라도 동일하게 평가한 아이템의 수가 더 많은 사용자에게 가중치를 높게 주는 방식이다. 예를 들어, 동일하게 평가한 아이템의 수가 50개 이하일 경우 n/50를 최종 유사도에 곱해 유사도의 가중치를 줄이는 식이다.

Variance Weighting은 분산에 따라 유사도에 다른 가중치를 주도록 공식을 수정한 것이다.

$$\frac{\sum_{i=1}^{m}(v_i \times z_{a,i} \times z_{b,i})}{\sum_{i=1}^{m} v_i}$$

$$v_i = \frac{var_i - var_{\min}}{var_{\max}},\ var_i = \frac{\sum_{b=1}^{n}(r_{b,i} - \overline{r_i})^2}{n-1}$$

a : 목표 사용자　　b : 이웃 사용자

m : 사용자 a, b가 공통적으로 선호도를 평가한 아이템 수

n : 전체 사용자 수

$z_{a,i}$: 사용자 a의 아이템 i에 대한 정규화 된 선호도

$z_{b,i}$: 사용자 b의 아이템 i에 대한 정규화 된 선호도

$r_{b,i}$: 사용자 b의 아이템 i에 대한 선호도

$\overline{r_i}$: 아이템 i에 대한 전체 사용자의 선호도의 평균

식 5 Variance Weighting

[식 5]는 Variance Weighting 공식으로 분산이 작은 아이템에는 1 미만의 수를 곱해 유사도의 가중치를 줄이고, 분산이 큰 아이템에는 1 이상의 수를 곱해 유사도의 가중치를 늘리는 방식이다. 여러 사용자들이 비슷

한 선호도를 보이는 경우보다 사용자마다 다른 선호도를 보이는 아이템에서 같은 선호도를 보이는 사용자들이 더 비슷할 것이기 때문이다.

2.3.5 유사 사용자 선정 방법

유사 사용자를 선정하는 방법에는 크게 세 가지가 있다. 유사성을 계산한 후 유사도가 큰 순서대로 n명의 사용자를 선정하는 방법과 일정 크기 이상의 유사도를 보인 사용자를 선정하는 방법, 벡터를 사용하여 선정하는 방법이다. 벡터를 사용하여 선정하는 방법은 목표 사용자와 유사도가 가장 큰 사용자를 선정한 후 목표 사용자와 유사도가 가장 큰 사용자의 평균 선호도에 대한 벡터 값을 계산하여 해당 벡터 값과 가장 가까운 사용자를 찾는 것을 일정 횟수 반복하는 방법이다(Sarwar, 2000).

이 외에도 유사 사용자 선정 과정에서 성능을 향상시키기 위한 다양한 시도들이 있어왔다. 동일하게 평가한 아이템의 수가 극히 적은 경우나 전반적으로 다른 사용자들과의 유사성이 낮은 사용자를 제외시키기도 한다.

위에 언급한 방법들은 모든 사용자들 사이의 유사도를 계산해야 하기 때문에, 사용자가 늘어날수록 연산 시간이 기하급수적으로 늘어난다는 단점이 있다. 따라서 이러한 단점을 보완하기 위한 연구들도 진행되었다. 대부분 클러스터링 기법을 사용하여 연산 시간을 줄이려고 노력하였으며 일반적인 K-means 클러스터링 기법부터 Bolzman 클러스터링이나 신경망 연구에서 주로 사용되는 SOM / Kohonen 클러스터링 기법을 사용하기도 하였다(Roh et al., 2003; Weng and Liu, 2004).

최근에는 악의적인 목적을 가진 사용자들이 추천 결과를 조작하려는 시도(Shilling attack)가 늘어나고 있어 이러한 사용자들을 찾아 유사 사용자에 포함시키지 않기 위한 연구도 진행되고 있다(김홍남 외, 2009).

2.3.6 선호도 예측

선호도 예측을 위한 방식으로는 주로 Weighted Average 방식이 쓰인다.

$$\frac{\sum_{b=1}^{m}(s_{a,b} \times r_{b,i})}{\sum_{b=1}^{m} s_{a,b}}$$

a : 목표 사용자　　b : 이웃 사용자　　m : 이웃 사용자의 수

$s_{a,b}$: 사용자 a와 사용자 b와의 유사도

$r_{b,i}$: 사용자 b의 아이템 i에 대한 선호도

식 6 Weighted Average

[식 6]은 아이템 i의 선호도를 예측하기 위한 Weighted Average공식이다. 이 방식은 비교적 계산이 간단하면서도 이전 연구에서 좋은 성능을 나타낸 방식으로, 각 사용자들의 선호도에 유사도를 가중치로 적용시켜 목표 사용자의 선호도를 예측하는 것이다.

사용자에 따라 비슷한 정도의 선호도를 보인다고 해도 전반적으로 선호도를 낮게 표시하는 사용자가 존재하고 그와 반대로 높게 표시하는 사용자

가 존재한다. 예를 들어 비슷한 영화를 보고 비슷한 분포로 선호도를 표시하지만 평균이 3점인 사용자와 5점인 사용자가 존재한다는 것이다. 같은 척도를 사용한다고 해도 사용자들이 받아들이는 수치에 관한 느낌이 다를 수 있기 때문이다. 이러한 점을 고려해 Weighted Average 공식을 수정한 것이 Bias-From-Mean Average 공식이다.

$$\frac{\sum_{b=1}^{m}(s_{a,b}\times r_{b,i})}{\sum_{b=1}^{m}s_{a,b}}+(Ave(r_a)-Ave(r_b))$$

a : 목표 사용자　　b : 이웃 사용자　　m : 이웃 사용자의 수

$s_{a,b}$: 사용자 a와 사용자 b와의 유사도

$r_{b,i}$: 사용자 b의 아이템 i에 대한 선호도

$Ave(r_a), Ave(r_b)$: 사용자 a와 사용자 b의 선호도의 평균

식 7 Bias-From-Mean Average

[식 7]은 Bias-From-Mean Average을 나타낸 것으로 Weighted Average 에 개별 사용자의 평점의 평균을 사용하여 평점을 정규화한 것이다. 이전 연구들에 따르면 Bias-From-Mean Average방식이 Weighted Average방식을 사용한 것보다 더 좋은 성능을 보였다(Miller, 2003).

이 외에도 평균과 분산을 고려하여 정규화한 Z-score Weighted Average방식과 예측 오차를 미리 예측해 선호도를 보정하는 방법 등 선호도 예측에 관한 다양한 연구가 진행되고 있다.

2.4 하이브리드 추천 시스템

내용 기반 추천 시스템은 선호도가 입력되지 않은 아이템에 대해서 추천이 가능하고, 협업 기반 추천 시스템은 새로운 성격의 아이템에 대해서 추천이 가능하다. 각 추천 시스템의 단점을 보완하고 장점을 살리기 위해 두 방식 모두를 사용한 추천 시스템에 대한 연구도 이루어지고 있다 (Balabanovic and Shoham 1997; Basu et al., 1998). 하나의 추천 리스트 안에 내용 기반 추천 시스템에서 얻은 추천 아이템들과 협업 필터링 추천 시스템에서 얻은 추천 아이템들을 적절히 섞어 보여주는 것이다.

2.5 기타 추천 시스템

각 아이템의 성분을 나누어 추천 리스트를 만들어내는 추천 시스템이 있다. 예를 들어 음악의 경우 장르, 연도, 음악가 등으로 아이템의 성분을 나누어 목표 사용자가 좋아하는 장르, 연도, 음악가의 음악을 추천해주는 것이다. 이러한 방식의 추천 시스템에서는 각 성분에 대한 가중치를 다르게 줌으로서 성능을 높이려는 시도가 있었다(Mukherjee et al., 2001).

클러스터링 기법과 협업 필터링 방식을 결합한 추천 시스템도 있다. 협업 필터링 추천 시스템에서 사용자가 늘어나면 기하급수적으로 연산이 늘어나 시스템이 느려지는 단점을 보완하기 위한 방법으로 클러스터링 기법을 통해 계산 할 사용자를 미리 줄여주는 방식이다(O'connor and Herlocker, 2001). 즉, 목표 사용자와 모든 사용자간의 유사도를 계산하는 것이 아닌 미리 구분된 그룹과의 유사도만을 계산하는 것이다.

적은 정보를 제공하는 사용자를 제거하는 방식과 반대로 많은 정보를 제공하거나 신뢰도가 높다고 판단되는 사용자에게 높은 가중치를 적용하는 방식의 추천 시스템도 있다(Golbeck and Hendler 2006; Golbeck 2006).

2.6 Coverage

Coverage는 해당 추천 시스템을 통해서 얼마나 많은 사용자에게 추천 리스트를 생성해 줄 수 있는가를 나타내는 정보이다. 내용 기반 추천 시스템과 협업 필터링 추천 시스템 등 추천 시스템의 성능이 아무리 좋다고 해도 적은 수의 사용자에 대한 추천 리스트만 생성해 줄 수 있는 방식이라면 좋은 추천 시스템이라고 말할 수 없기 때문이다. 예를 들어, 100명의 사용자 중 95명의 사용자에게 추천 리스트를 생성해 줄 수 있는 시스템의 경우 Coverage는 95%가 된다.

2.7 예측된 선호도에 대한 검증 방법

추천 시스템에 관한 연구들에서 사용하는 선호도 예측에 관한 검증 방법은 크게 MAE(Mean Absolute Error)와 RMSE(Root Mean Squared Error) 두 가지가 있다.

$$\frac{\sum_{i=1}^{n}|p_i - r_i|}{n}$$

n : 선호도가 예측된 아이템의 수

p_i : 아이템 i에 대해 예측된 선호도

r_i : 아이템 i의 실제 선호도

식 8 Mean Absolute Error

[식 8]은 MAE공식으로 실제 선호도와 예측 선호도 차이의 합의 평균 값을 구하는 것이다.

$$\sqrt{\frac{\sum_{i=1}^{n}(p_i - r_i)^2}{n}}$$

n : 선호도가 예측된 아이템의 수

p_i : 아이템 i에 대해 예측된 선호도

r_i : 아이템 i의 실제 선호도

식 9 Root Mean Squared Error

[식 9]는 RMSE 공식이다. MAE와 RMSE는 모두 예측한 선호도와 실제 선호도가 평균적으로 얼마나 떨어져 있는지를 나타내며, 그 값이 작을수록 결과가 좋다는 것을 의미한다. 많은 연구들에서 계산의 편리함 때문에 MAE를 많이 사용하고, 통계학에서는 RMSE를 많이 사용한다. 이전 연구들에 따르면 두 공식에 따른 큰 차이를 나타내진 않는다.

2.8 선호이질성

선호이질성은 아이템 특성에 따라 사용자들의 선호도가 다르게 나타날 수 있다는 것을 뜻한다. [그림 7]에서 보는 것과 같이 사용자들의 선호도가 다양하게 나타나는 아이템 군과 비슷하게 나타나는 아이템 군이 존재한다는 것이다(Miller et al., 1997).

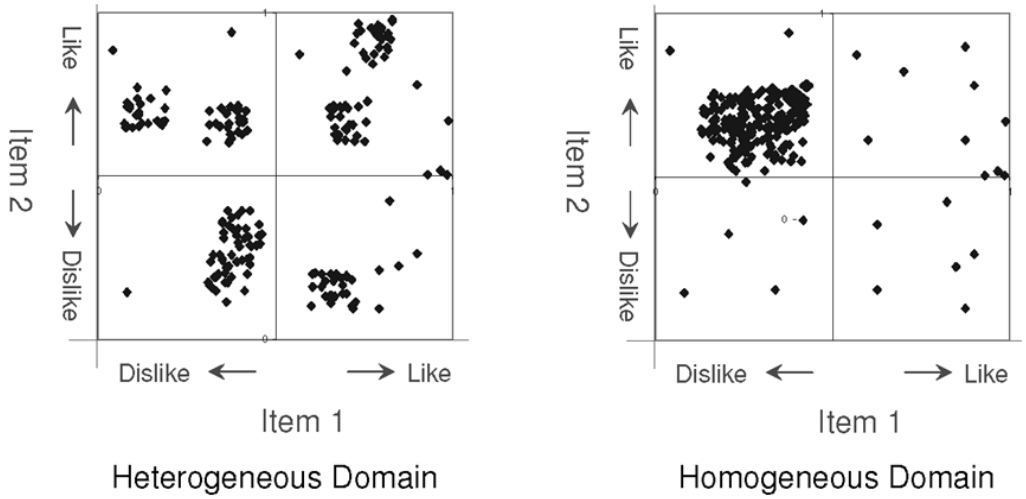

그림 7 선호이질성

이전 연구에 따르면, 사용자들의 선호이질성 정도에 따라 아이템에 대한 반응이 다르게 나타난다(Feick and Higie, 1992). 조금 단적인 예를 들어 영화와 성형수술을 생각해 보자. 이질적인 선호도를 보이는 군을 성형수술이라 하고 각각의 아이템을 코와 눈 이라고 생각하면 [그림 7]의 왼쪽 Heterogeneous Domain과 비슷한 결과를 보일 것이라고 생각할 수 있다. 동질적인 선호도를 보이는 군을 영화라 하고 각각의 아이템을 '해리포터'와 '슈렉'이라고 생각하면 [그림 7]의 오른쪽 Homogeneous Domain과 비슷한 결과를 보일 것이다.

제 3 장 연구 가설

3.1 이론적 가설 설정

추천 시스템에서 정확한 추천 리스트를 사용하기 위해 사용하는 정보의 종류는 매우 다양하다. 내용 기반 추천 시스템에서 주로 사용하는 정보로는 아이템의 이름, 종류, 특징과 사용자의 성별, 나이, 검색어 등이 있고, 협업 필터링 추천 시스템에서 주로 사용하는 정보로는 선호도와 구매에 관련된 정보들이 있다. 협업 필터링 추천 리스트에서 사용하는 선호도와 구매에 관련된 정보는 구체적으로 선호도 점수, 선호도를 표시한 날짜, 목표 사용자와 다른 사용자간에 선호도를 표시한 아이템 중 중복되는 아이템의 수와 선호도를 표시한 날짜나 요일의 차이, 구매한 아이템, 아이템 구매일, 목표 사용자가 구매한 아이템과 다른 사용자가 구매한 아이템 중 중복되는 아이템의 수와 구매일간의 날짜나 요일의 차이 등이 있다.

이론적으로 협업 필터링 추천 시스템에서는 위에서 언급한 정보들 이외에 대부분의 정보들을 사용해 추천 리스트를 생성할 수 있다. 중요한 점은 얻을 수 있는 정보들을 사용해 사용자마다 다른, 사용자에게 맞춘 추천 리스트를 생성한다는 것이다. 즉, 웹 개인화와 맞춤화가 추천 시스템의 가장 주된 사용 목적이라는 것이다. 하지만, 협업 필터링 추천 시스템의 추천 리스트 생성 과정에서 가장 중요한 유사 사용자 그룹 선정 과정에 대한 개인화는 아직 미흡하다.

이전 연구들은 협업 필터링 추천 시스템의 성능 향상에 있어서 유사 사용자 그룹의 최적화된 크기를 찾거나 유사도 계산 공식의 변형에 초점을 맞춰왔다. 즉, 유사도 계산 공식에 들어가는 변수나 계산 방법을 바꾸거나, 각 정보에 대한 가중치 적용 방식을 바꿔 협업 필터링 추천 시스템의 성능을 향상시키는데 연구의 목적을 두었다. 또한, 유사 사용자 그룹의 최적화된 크기를 찾는데 연구의 목적을 두었다. 하지만 이러한 대부분의 연구들은 모든 사용자에게 동일하게 적용 가능한 유사 사용자에 관한 기준이 존재한다고 가정하고 있고, 유사도 계산 공식과 그 가중치를 모든 사용자에게 동일하게 적용할 수 있다고 가정하고 있다. 예를 들어, 연구에서 사용할 최적의 유사 사용자 그룹의 크기가 25명이라고 정의하였다면, 모든 사용자들의 유사 사용자를 유사도가 높은 순서대로 25명을 뽑아 정의한다. 혹은, 중복되어 구매한 아이템을 10개까지 가중치를 주겠다고 정의한다고 하면, 모든 사용자의 유사도 계산 과정에서 10개까지 중복된 아이템에 관해 가중치를 준다. 이전 연구들에서 사용자마다 다른 선호도와 구매 패턴을 가지고 있고, 다른 성향을 가지고 있음이 밝혀졌지만, 실제 협업 필터링 추천 시스템에는 반영되지 못하고 있는 것이다.

본 연구에서는 이러한 미흡한 부분을 보완한다면 추천 시스템의 성능이 향상될 것이라고 가정했다. 즉, 사용자마다 다른 최적화된 유사 사용자 그룹의 크기와 가중치에 대한 기준을 찾아 적용한다면, 협업 필터링 추천 시스템의 성능이 향상될 것이다.

3.1.1 유사 사용자 그룹의 크기

유사 사용자 그룹의 크기는 협업 필터링 추천 시스템에서 사용하는 목표 사용자의 선호도 예측을 위한 유사 사용자 그룹에 들어가는 다른 사용자들의 수를 뜻한다. 즉, 목표 사용자의 선호도를 예측할 때, 몇 명의 유사한 사용자를 참고로 할 것인지를 정하는 값이다. 유사 사용자 그룹의 크기를 정하는 방법에는 크게 두 가지가 있는데, 유사도가 높은 순서대로 N명을 선정하는 best-N neighbors 방식과, 정해진 값 이상의 유사도를 보이는 사용자를 유사 사용자로 선정하는 thresholding 방식이다(Breese et al., 1998; Herlocker et al., 1999). 일반적으로 thresholding 방식보다 best-N neighbors방식을 사용한 협업 필터링 추천 시스템의 성능이 더 좋기 때문에, 본 연구에서는 best-N neighbors 방식을 사용하였다 (Herlocker et al., 1999).

이전 연구들에 따르면, best-N neighbors 방식을 사용하여 추천 리스트를 생성할 때, N 값에 따른 예측 정확도의 차이가 존재한다고 한다. 즉, 유사 사용자 그룹의 크기에 따라 예측된 선호도의 정확도 차이가 존재하고, 최적의 유사 사용자 그룹의 크기가 존재한다는 것이다. 실제로 Herlocker 등의 연구에 따르면 같은 아이템 안에서 최적의 유사 사용자 그룹의 크기가 존재한다는 것을 알 수 있다(그림 8 참조). 하지만, 최적의 유사 사용자 그룹의 크기는 아이템의 종류에 따라 다르기도 하다. 즉, 아이템의 성격에 따라 최적의 유사 사용자 크기는 다르다는 것을 알 수 있다 (Im and Hars, 2007).

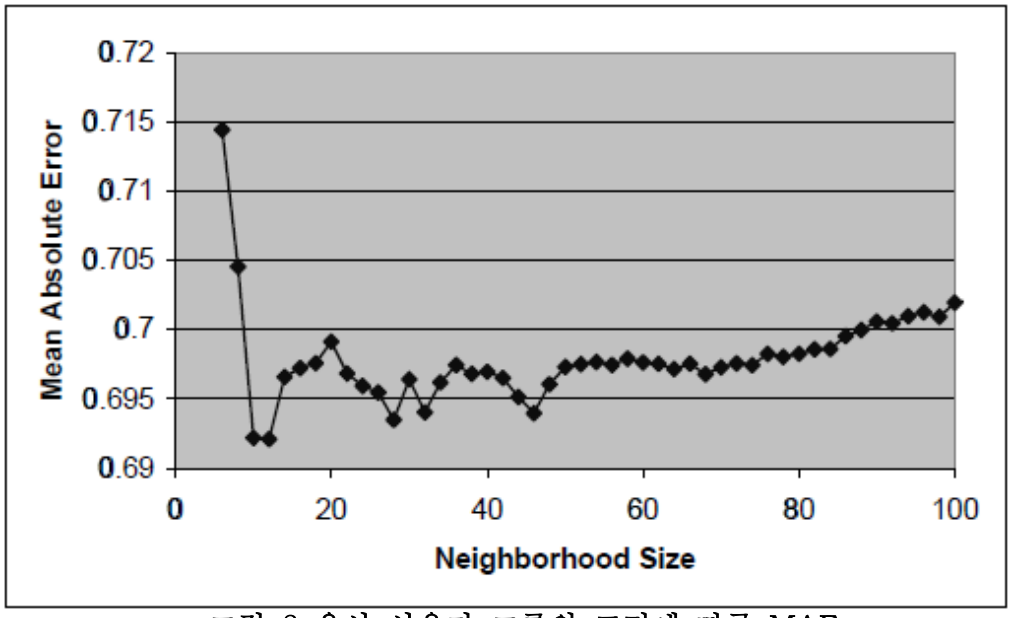

그림 8 유사 사용자 그룹의 크기에 따른 MAE

하지만, 대부분의 연구들을 살펴보면 유사 사용자 그룹의 크기에 관에 하나의 큰 가정을 가지고 있다. 모든 사용자에게 적용 가능한 유사 사용자 그룹의 크기가 존재한다는 것이다(Xue et al., 2005). 사용자들은 개개인에 따라 다른 특성을 보인다(Bell et al., 2007). 선호도에 있어서도 마찬가지일 것이다. 유사한 사용자가 많은 사람이 존재할 수 있고, 유사한 사용자가 적은 사용자가 존재할 수 있다. 즉, 대중적인 영화를 좋아하는 사용자가 있고, 단편 영화나 독립 영화 취향을 가진 사용자가 존재한다는 것이다.

본 연구에서는 사용자의 특성에 따라 최적의 유사 사용자 그룹의 크기가 다를 것이라 생각하고, 사용자마다 다른 최적 유사 사용자 그룹 크기를 적용한다면 협업 필터링 추천 시스템의 예측 정확도가 개선될 것이라고 생각한다.

제 4 장 연구 대상 및 연구 방법

4.1 연구 대상

본 연구에서는 미국의 최대 비디오 대여 사이트인 Netflix(http://www.netflix.com)의 사용자 선호도 정보를 사용하여 가설을 검증하였다. Netflix는 한 달에 30~40 달러를 내면 집까지 비디오와 DVD를 배달해주는 서비스를 제공하는데 현재 회원 수가 1,200만 명이 넘고 2009년 4분기에만 4억 4,450만 달러의 매출을 올렸다. Netflix에서는 자신들이 제공해주는 추천 시스템의 성능을 높이기 위해 많은 노력을 하고 있으며, 그 노력의 하나로 자신들이 사용하고 있는 추천 시스템의 성능을 10% 이상 개선시킨 팀에게 1,000,000달러의 상금을 주는 Netflix Prize(http://www.netflixprize.com)를 열었다. 본 연구에서는 Netflix Prize에서 제공한 Training set을 사용하여 연구를 제공하였다.

Netflix Prize의 Training set에는 480,189명의 고객과 17,770개의 영화에 관한 100,480,507건의 선호도와 선호도 입력 날짜에 관한 정보가 들어있다. 모든 정보를 사용하여 가설을 검증하기에는 너무 많은 시간이 소요될 것으로 예상되어 5,000명의 고객을 랜덤으로 선택하여 하나의 샘플링 그룹을 선정하였으며 총 10개의 샘플링 그룹을 대상으로 연구를 진행하였다.

4.2 연구 방법

본 연구에서는 시뮬레이션 기법을 사용하여 가설을 검증하였다. 랜덤으로 선정된 10개의 샘플링 그룹에서 각 100명의 목표 사용자를 선정하였으며, 목표 사용자당 1개의 영화에 대한 선호도를 최종적으로 예측하여 가설을 검증하였다. 비교를 위해 기존의 방법과 동일하게 5명부터 100명 까지 5명 단위로 유사 사용자 크기를 일괄적으로 적용한 후 예측 선호도의 MAE값과 RMSE값을 측정하였고, 새로운 방법을 적용한 후 예측 선호도의 MAE값과 RMSE값을 측정하여 비교하였다.

4.2.1 동일한 크기의 유사 사용자 그룹

새로운 방법과 기존의 방법을 비교하기 위해 Netflix Prize의 training set을 대상으로 동일한 크기의 유사 사용자 그룹을 사용한 협업 필터링 추천 시스템의 성능을 측정하였다. 해당 정보를 대상으로 알려진 적절한 크기의 유사 사용자 그룹이 없기 때문에, 5명 단위로 5명부터 100명까지 유사 사용자 그룹의 크기를 조절하여 선호도를 예측하였고 MAE값과 RMSE값을 모두 측정하였다.

4.2.2 개인별로 다른 크기의 유사 사용자 그룹 예측을 위한 전처리 과정

각 그룹별로 100명 씩 랜덤하기 선정된 총 1,000명의 목표 사용자들이 선호도를 표시한 영화들 중에서 한 사람당 한 개의 최종 목표 영화를 랜덤하게 선정하였다. 그 후, 해당 선호도 정보를 삭제하여 시뮬레이션 결과에 영향을 주지 않도록 하였다. 예를 들어, 그룹 1에서 A 사용자가 본 영화 a와 B 사용자가 본 영화 b를 최종적으로 선호도를 예측할 영화로 선정하였다면, 사용자 A와 다른 사용자들과의 유사도를 계산할 경우 [그림 9]와 같은 원래 선호도 정보에서 [그림 10]과 같이 영화 a에 대한 선호도 정보를 삭제한 후 사용자 A와 다른 사용자들과의 유사도를 계산하고, 사용자 B와 다른 사용자들과의 유사도를 계산할 경우 [그림 9]의 원래 선호도 정보에서 [그림 11]과 같이 영화 b에 대한 선호도 정보를 삭제한 후 사용자 B와 다른 사용자들과의 유사도를 계산하는 것이다.

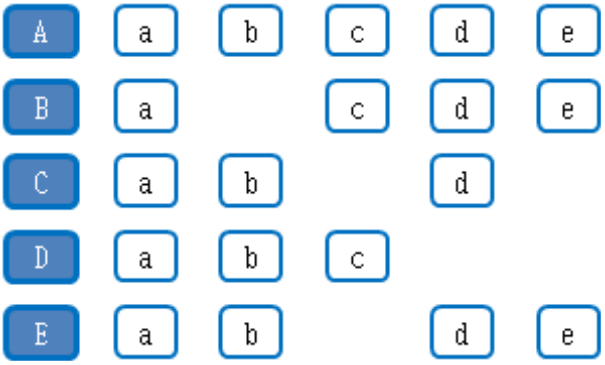

그림 9 원래 선호도 정보

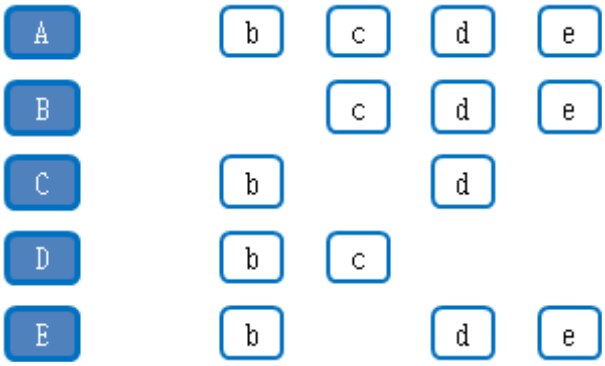

그림 10 영화 a에 대한 예측 시

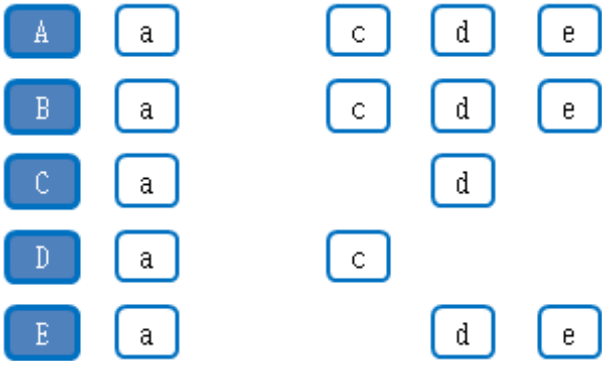

그림 11 영화 b에 대한 예측 시

4.2.3 개인별로 다른 크기의 유사 사용자 그룹 예측

개인별로 유사 사용자 그룹의 크기를 예측하기 위해 목표 사용자의 남아있는 선호도 예측 정보를 사용하였다. 사용자 A가 본 영화 a에 관해 선호도를 예측 할 경우 영화 a에 관한 선호도 정보는 전처리 과정에서 삭제된 상태이다. 이후, 다음과 같은 5가지 과정을 최대 100번까지 반복해 개인별 유사 사용자 그룹의 크기를 예측하였다.

과정 1 : 영화 b에 대한 선호도 정보 삭제
과정 2 : 유사 사용자 그룹의 크기를
4부터 25까지 1단위로 변경하여 선호도 예측
과정 3 : 예측된 선호도에 대한 RMSE값 계산
과정 4 : RMSE값을 바탕으로
최적의 유사 사용자 그룹 크기 선정
과정 5 : 영화 b에 대한 선호도 정보 복원

예를 들어, 사용자 A의 경우 전처리 과정에서 [그림 10]과 같은 영화 a에 관한 정보가 삭제된 상태의 training set에서부터 시뮬레이션 과정이 시작된다. 만약 사용자 A가 선호도를 입력했던 영화의 수가 100개 이상일 경우 랜덤하게 100개의 영화를 선정하고, 100개미만의 영화에 대한 선호도를 입력한 상태라면 모든 선호도에 대한 예측을 시도한다.

과정 1에서는 영화 b에 대한 선호도를 예측하기 위해 [그림 12]와 같이 영화 b에 관한 선호도 정보를 삭제한다. 즉, 최종 예측할 영화 a와 함

께 2개 영화에 관한 선호도 정보가 삭제된 상태이다.

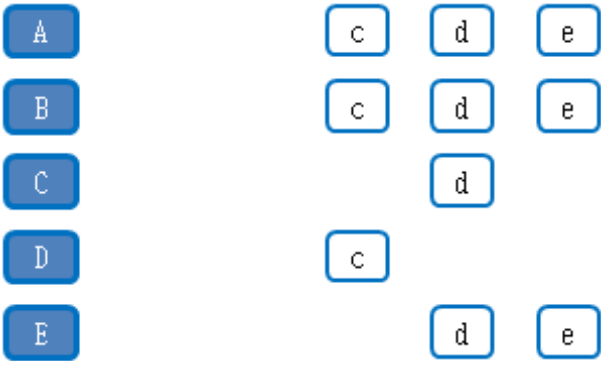

그림 12 유사 사용자 그룹의 크기 예측을 위한 과정 1

이러한 상태에서 다른 사용자들과 사용자 A와의 유사도를 계산하고, 유사도가 높은 순서대로 4명부터 25명까지 한 사용자씩 추가해가며 선호도를 예측한다. 그 후 각 크기의 유사도를 예측한 이후, 실제 영화 b에 대한 선호도와 비교해 RMSE값을 계산한다.

그리고 계산된 RMSE값 중 가장 적은 값을 갖는 유사 사용자 그룹의 크기를 계산한다. 이후, 영화 b에 대한 선호도 정보를 다시 복원시키고, 다음 선호도를 예측할 영화 c에 대한 선호도 정보를 삭제하여 [그림 13]과 같은 상태에서 위의 과정을 반복한다.

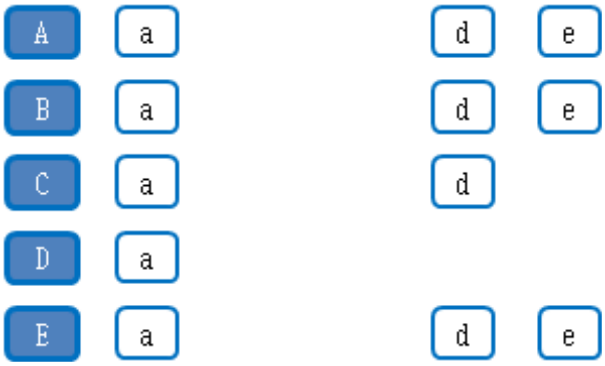

그림 13 유사 사용자 그룹의 크기 예측을 위한 과정 2

최대 100개까지 영화에 대한 시뮬레이션을 끝낸 후 저장되어진 100개의 최적 유사 사용자 그룹의 수를 평균 낸 것을 마지막 예측을 위한 유사 사용자 그룹의 크기로 정하는 것이다.

4.2.4 개인별로 다른 크기의 유사 사용자 그룹

각 그룹별로 100명 씩 랜덤하기 선정된 총 1,000명의 목표 사용자들이 선호도를 표시한 영화들 중에서 한 사람당 한 개의 최종 목표 영화를 랜덤하게 선정하였다. 그 후, 해당 선호도 정보를 삭제하여 시뮬레이션 결과에 영향을 주지 않도록 하였다. 예를 들어, 그룹 1에서 A 사용자가 본 영화 a와 B 사용자가 본 영화 b를 최종적으로 선호도를 예측할 영화로 선정하였다면, 사용자 A와 다른 사용자들과의 유사도를 계산할 경우 영화 a에 대한 선호도 정보를 삭제한 후 사용자 A와 다른 사용자들과의 유사도를 계산하고, 사용자 B와 다른 사용자들과의 유사도를 계산할 경우 영화 b에

대한 선호도 정보를 삭제한 후 사용자 B와 다른 사용자들과의 유사도를 계산하는 것이다.

4.2.5 최종 선호도 예측 및 결과 비교

4.2.4의 시뮬레이션 과정을 통해 얻은 개인별 유사 사용자 크기를 바탕으로 최종 선호도를 예측하고, 기존 방식과 비교하여 성능 향상 여부를 검증한다. 즉, 4.2.4의 과정을 통해 사용자 A에게 적합한 유사 사용자 그룹의 크기가 10명이라고 예측되었고, 사용자 B에게 적합한 유사 사용자 그룹의 크기가 16명이라고 예측되었다면, 기존의 방법에서 사용자 A와 사용자 B에게 동일한 유사 사용자 크기인 13명을 적용하여 최종 선호도를 예측하는 것이 아닌, 각각에게 적합하다고 예측된 10명과 16명을 다르게 적용하여 선호도를 예측하는 것이다.

제 5 장 가설 검증 및 결과 분석

5.1 연구 대상 분석

전체 표본의 사용자와 영화, 선호도에 대한 수와 랜덤 샘플링을 통해 선정된 10개 그룹의 영화, 선호도에 대한 수는 다음 [표 1]과 같다. Original은 Netflix에서 제공한 전체 샘플에 관한 정보이고 Group 1 ~ Group 10 은 각 샘플링 그룹에 관한 정보이다. Total Movie는 각 그룹의 전체 영화 수, Total Ratings는 전체 선호도 수, Total Customer는 전체 사용자의 수이다. AVG (Ratings)는 선호도의 평균이며 STDEV (Ratings)는 선호도의 표준편차, VAR(Ratings)는 선호도의 분산이다. T-test 결과 각 그룹 간에 차이는 없는 것으로 나타났다.

	Original	Group 1	Group 2	Group 3	Group 4	Group 5
Total Movie	17770	16164	15926	16443	16460	17700
Total Ratings	100480507	1082624	1050965	1032734	1074120	1056430
Total Customer	480189	5000	5000	5000	5000	5000
Ratings / Customer	209.252	216.525	210.193	206.547	214.824	211.286
AVG (Ratings)	3.604	3.621	3.613	3.609	3.608	3.588
AVG(Ratings) / Customer	3.674	3.672	3.674	3.669	3.681	3.683
AVG(Ratings) / Movie	3.228	3.318	3.307	3.274	3.291	2.901
STDEV (Ratings)	1.085	1.077	1.082	1.076	1.083	1.103
STDEV(Ratings) / Customer	0.998	0.994	0.998	0.996	0.999	0.996
STDEV(Ratings) / Movie	1.101	1.007	0.992	1.005	1.026	1.118
VAR(Ratings)	1.178	1.16	1.171	1.158	1.172	1.217
VAR(Ratings) / Customer	1.059	1.052	1.056	1.056	1.063	1.054
VAR(Ratings) / Movie	1.229	1.166	1.128	1.17	1.218	1.428

표 1 연구 대상 분석 (1)

	Original	Group 6	Group 7	Group 8	Group 9	Group 10
Total Movie	17770	16244	16743	15980	16026	16623
Total Ratings	100480507	1034707	1073998	1063262	1026716	1073348
Total Customer	480189	5000	5000	5000	5000	5000
Ratings / Customer	209.252	206.941	214.8	212.652	205.343	214.67
AVG (Ratings)	3.604	3.612	3.589	3.625	3.615	3.58334
AVG(Ratings) / Customer	3.674	3.682	3.675	3.678	3.673	3.67055
AVG(Ratings) / Movie	3.228	3.325	3.204	3.369	3.303	3.10956
STDEV (Ratings)	1.085	1.084	1.086	1.076	1.083	1.08671
STDEV(Ratings) / Customer	0.998	0.994	0.998	0.996	1.001	0.99418
STDEV(Ratings) / Movie	1.101	0.986	1.027	0.984	1.004	1.05479
VAR(Ratings)	1.178	1.174	1.18	1.157	1.173	1.18094
VAR(Ratings) / Customer	1.059	1.05	1.057	1.053	1.064	1.05143
VAR(Ratings) / Movie	1.229	1.117	1.202	1.111	1.158	1.27313

표 1 연구 대상 분석 (2)

5.2 동일한 크기의 유사 사용자 그룹

새로운 방법과 기존의 방법과의 성능을 예측하기 위해 동일한 유사 사용자 크기를 적용한 협업 필터링 추천 시스템의 성능을 측정하였다. 유사 사용자 그룹의 크기는 5명부터 100명 까지 5명 단위로 증가시켜가며 선호도를 예측하였다.

[표 2]와 [표 3]은 예측한 선호도와 실제 선호도를 MAE와 RMSE로 검증한 것으로, 두 가지 모두 유사 사용자 그룹의 크기가 20명일 때 가장 좋은 결과를 나타냈다.

Accuracy (MAE)	Reference group size									
	5	10	15	20	25	30	35	40	45	50
Sampling 1	0.796	0.745	0.741	0.744	0.759	0.772	0.781	0.763	0.771	0.772
Sampling 2	0.643	0.607	0.591	0.593	0.603	0.619	0.620	0.600	0.601	0.600
Sampling 3	0.720	0.729	0.715	0.686	0.675	0.684	0.689	0.695	0.683	0.684
Sampling 4	0.703	0.643	0.656	0.611	0.621	0.630	0.624	0.608	0.610	0.607
Sampling 5	0.626	0.612	0.627	0.622	0.616	0.613	0.606	0.616	0.614	0.616
Sampling 6	0.631	0.656	0.602	0.582	0.573	0.574	0.582	0.569	0.572	0.580
Sampling 7	0.620	0.609	0.592	0.594	0.601	0.599	0.581	0.576	0.574	0.587
Sampling 8	0.584	0.624	0.618	0.619	0.624	0.637	0.640	0.635	0.636	0.638
Sampling 9	0.760	0.723	0.718	0.726	0.731	0.738	0.706	0.705	0.708	0.709
Sampling 10	0.676	0.674	0.676	0.668	0.665	0.674	0.753	0.679	0.701	0.674
AVG	0.676	0.662	0.653	0.645	0.647	0.654	0.658	0.645	0.647	0.647

표 2 동일한 크기의 유사 사용자 그룹을 적용한 선호도 예측 결과 – MAE (1)

Accuracy (MAE)	Reference group size									
	55	60	65	70	75	80	85	90	95	100
Sampling 1	0.775	0.777	0.769	0.760	0.764	0.763	0.763	0.763	0.765	0.767
Sampling 2	0.600	0.603	0.628	0.610	0.605	0.605	0.611	0.608	0.611	0.609
Sampling 3	0.682	0.677	0.667	0.668	0.670	0.670	0.672	0.677	0.677	0.679
Sampling 4	0.610	0.611	0.617	0.623	0.622	0.632	0.742	0.645	0.635	0.637
Sampling 5	0.620	0.625	0.613	0.614	0.625	0.782	0.620	0.619	0.625	0.625
Sampling 6	0.568	0.565	0.565	0.563	0.564	0.564	0.567	0.577	0.580	0.575
Sampling 7	0.747	0.585	0.584	0.588	0.588	0.600	0.597	0.606	0.609	0.607
Sampling 8	0.640	0.642	0.647	0.649	0.655	0.652	0.657	0.657	0.655	0.654
Sampling 9	0.708	0.713	0.717	0.726	0.749	0.991	0.718	0.714	0.716	0.721
Sampling 10	0.671	0.677	0.678	0.680	0.689	0.688	0.686	0.689	0.677	0.684
AVG	0.662	0.647	0.648	0.648	0.653	0.695	0.663	0.655	0.655	0.656

표 2 동일한 크기의 유사 사용자 그룹을 적용한 선호도 예측 결과 – MAE (2)

Accuracy (RMSE)	Reference group size									
	5	10	15	20	25	30	35	40	45	50
Sampling 1	1.010	0.907	0.897	0.901	0.919	0.931	0.946	0.915	0.923	0.927
Sampling 2	0.808	0.789	0.763	0.768	0.775	0.809	0.808	0.768	0.773	0.778
Sampling 3	0.940	0.943	0.919	0.870	0.861	0.870	0.870	0.872	0.846	0.848
Sampling 4	0.896	0.840	0.931	0.820	0.829	0.838	0.820	0.805	0.805	0.802
Sampling 5	0.855	0.739	0.763	0.747	0.736	0.735	0.730	0.742	0.745	0.747
Sampling 6	0.797	0.989	0.768	0.744	0.726	0.735	0.745	0.737	0.742	0.747
Sampling 7	0.802	0.766	0.764	0.763	0.782	0.792	0.768	0.765	0.767	0.780
Sampling 8	0.735	0.772	0.774	0.776	0.789	0.797	0.806	0.799	0.800	0.808
Sampling 9	0.948	0.915	0.911	0.915	0.922	0.926	0.885	0.887	0.892	0.892
Sampling 10	0.854	0.860	0.852	0.847	0.853	0.864	1.199	0.876	0.917	0.873
AVG	0.865	0.852	0.834	0.815	0.819	0.830	0.858	0.817	0.821	0.820

표 3 동일한 크기의 유사 사용자 그룹을 적용한 선호도 예측 결과 – RMSE (1)

Accuracy (RMSE)	Reference group size									
	55	60	65	70	75	80	85	90	95	100
Sampling 1	0.930	0.933	0.925	0.917	0.922	0.919	0.921	0.924	0.926	0.926
Sampling 2	0.783	0.791	0.835	0.787	0.783	0.784	0.790	0.788	0.789	0.787
Sampling 3	0.843	0.840	0.830	0.833	0.835	0.837	0.841	0.845	0.846	0.848
Sampling 4	0.811	0.809	0.814	0.815	0.816	0.823	1.398	0.848	0.832	0.837
Sampling 5	0.755	0.764	0.751	0.751	0.762	1.728	0.758	0.761	0.763	0.767
Sampling 6	0.734	0.735	0.732	0.729	0.732	0.731	0.732	0.743	0.748	0.748
Sampling 7	1.696	0.778	0.777	0.781	0.784	0.792	0.791	0.797	0.799	0.800
Sampling 8	0.811	0.815	0.818	0.819	0.826	0.823	0.828	0.828	0.828	0.828
Sampling 9	0.891	0.895	0.896	0.907	0.957	2.802	0.902	0.902	0.904	0.909
Sampling 10	0.875	0.880	0.883	0.885	0.894	0.893	0.892	0.894	0.886	0.893
AVG	0.913	0.824	0.826	0.822	0.831	1.113	0.885	0.833	0.832	0.834

표 3 동일한 크기의 유사 사용자 그룹을 적용한 선호도 예측 결과 – RMSE (2)

5.3 사용자별 최적 유사 사용자 그룹의 크기 예측

시뮬레이션 기법을 통해 찾은 사용자별 최적 유사 사용자 그룹의 크기는 [표 4]와 같다. 유사 사용자 그룹 크기의 평균은 이전 연구들과 비슷한 22.5명으로 나타났다. 샘플 9의 경우 18명으로 가장 작았고, 샘플 6의 경우 26명으로 제일 큰 그룹의 크기를 예측했다.

	N
Sampling 1	21.183
Sampling 2	23.868
Sampling 3	19.472
Sampling 4	26.111
Sampling 5	23.083
Sampling 6	26.472
Sampling 7	20.615
Sampling 8	22.935
Sampling 9	18.121
Sampling 10	22.972
AVG	22.483

표 4 예측된 최적 유사 사용자 그룹의 크기

기존 연구와 평균적으로는 유사한 크기의 유사 사용자 그룹을 예측했지만, 자료를 자세히 살펴보면 큰 차이를 알 수 있다. 다음 [그림 14]부터 [그림 17]까지는 그룹 1부터 그룹 4까지에서 예측된 유사 사용자 그룹의 크기를 그래프로 정리한 것이다.

그림 14 Group 1의 예측된 유사 사용자 그룹의 크기 빈도 그래프

그림 15 Group 2의 예측된 유사 사용자 그룹의 크기 빈도 그래프

그림 16 Group 3의 예측된 유사 사용자 그룹의 크기 빈도 그래프

그림 17 Group 4의 예측된 유사 사용자 그룹의 크기 빈도 그래프

평균은 19명부터 24명으로 나타났지만 5명 내외의 최적 유사 사용자 그룹의 크기를 갖는 사용자들과 매우 많은 크기의 유사 사용자 그룹의 크기를 갖는 사용자들이 존재함을 알 수 있다. 작은 크기의 최적 유사 사용자 그룹의 크기를 보이는 사용자들의 예측 정확도 변화와, 큰 크기의 최적 유사 사용자 그룹의 크기를 보이는 사용자들의 예측 정확도 변화를 정리하면 [그림 18]과 [그림 19]와 같다.

[그림 18]과 [그림 19]에서 알 수 있듯이, 작은 크기의 최적 유사 사용자 그룹의 크기를 보인 사용자들의 경우, 유사 사용자 그룹의 크기를 늘려줄수록 예측 정확도가 나빠짐을 알 수 있다. 또한, 큰 크기의 최적 유사 사용자 그룹의 크기를 보인 사용자들의 경우, 유사 사용자 그룹의 크기를 늘려줄수록 예측 정확도가 좋아짐을 알 수 있다.

그림 18 작은 최적 유사 사용자 그룹의 크기를 갖는 사용자

그림 19 큰 최적 유사 사용자 그룹의 크기를 갖는 사용자

5.4 사용자별 최적 유사 사용자 그룹의 크기 적용

시뮬레이션 과정을 거쳐 개인별로 예측한 유사 사용자 그룹의 크기를 적용해 추천 리스트를 생성해 줄 경우 평균 22%의 성능 향상을 나타냄을 알 수 있었다. MAE 방식으로 성능을 평가한 결과 10 그룹 평균 0.630에서 0.404 성능이 향상되었고, RMSE 방식으로 성능을 평가한 결과 평균 0.798에서 0.622로 성능이 향상됨을 알 수 있었다.(표 5 참조)

	Accuracy			
	MAE		RMSE	
	Old	New	Old	New
Sampling 1	0.741	0.485	0.897	0.698
Sampling 2	0.591	0.338	0.763	0.528
Sampling 3	0.667	0.460	0.830	0.698
Sampling 4	0.607	0.399	0.802	0.612
Sampling 5	0.606	0.388	0.730	0.649
Sampling 6	0.563	0.357	0.726	0.544
Sampling 7	0.574	0.376	0.763	0.602
Sampling 8	0.584	0.347	0.735	0.530
Sampling 9	0.705	0.485	0.885	0.726
Sampling 10	0.665	0.407	0.847	0.637
AVG	0.630	0.404	0.798	0.622

표 5 사용자별 최적 유사 사용자 그룹을 적용한 결과

5.5 Coverage

협업 필터링 추천 시스템에서 예측 정확도와 함께 성능을 나타내는 중요한 척도는 Coverage이다. 본 연구에서 제안한 새로운 방법은 기존의 방법과 큰 차이가 없는 것으로 나타났다.

N	Old	New
Sampling 1	99	98
Sampling 2	98	95
Sampling 3	99	97
Sampling 4	100	98
Sampling 5	100	98
Sampling 6	98	95
Sampling 7	100	99
Sampling 8	99	97
Sampling 9	97	96
Sampling 10	99	95
AVG	98.9	96.8

표 6 Coverage 비교

제 6 장 결론

6.1 연구 결과 및 토의

협업 필터링 추천 시스템의 성능 향상은 경영학 뿐 아니라 문헌정보학, 컴퓨터 공학 등의 여러 분야에서 연구가 진행되고 있는 중요한 연구주제 중 하나이다. 본 연구는 협업 필터링 추천 시스템의 성능 향상을 위해 유사도 계산 과정에 필요한 변수를 개인화 시키는 새로운 방법을 제안하였고, 검증 결과 기존의 방식에 비해 더 좋은 성능을 보이는 것으로 확인되었다.

6.2 연구의 시사점

본 연구는 협업 필터링 추천 시스템의 선호도 예측에 있어서 사용자들의 선호도에 대한 성향이 다름을 확인하였다. 또한, 유사 사용자 그룹의 크기에 따라 그 성향이 다름을 확인하였다. 이러한 사실에 근거해 협업 필터링 추천 시스템의 성능 향상에 관한 새로운 방법을 제시하였다.

6.3 연구의 한계 및 향후 연구 방향

협업 필터링 추천 시스템에 있어서 자료의 희박성 문제는 항상 존재하였고, 해결하기 위한 다양한 연구들이 진행되고 있다. 본 연구에서 새로 제시한 방법이 예측 정확도를 향상시킬 수는 있지만, 희박성 문제를 해결하기는 어렵다는 연구의 한계가 있다.

또한, 모든 사용자에 있어서 본 연구에서 제시한 방법으로 유사 사용자 그룹의 크기를 예측 할 경우, 시간이 오래 걸린다는 단점이 있다. 모든 사용자의 정보를 활용해 시뮬레이션 기법을 통해 유사 사용자 그룹의 크기를 예측해야 하는데 현업에서 많은 사용자를 대상으로 실시간으로는 구하기 힘들다는 단점이 있다.

참고문헌

김홍남, 하인애, 조근식, "조작된 선호도에 강건한 협업적 여과 방법," 인터넷정보학회논문지 10(6), 2009, 81-98.

박지선, 김택헌, 류영석, 양성봉, "추천시스템을 위한 2-way 협동적 필터링 방법을 이용한 예측알고리즘," 정보과학회지 29(9,10), 2002, 669-675.

이경종, 공기현, 이상구, "사용자 선호도와 태그 간 상관도 분석을 통한 태그 기반 협력적 필터링 기법," 정보과학회지 34(2), 2007, 72-77.

Abhinandan S. Das, Mayur Datar, Ashutosh Garg and Shyam Rajaram, "Google news personalization: scalable online collaborative filtering," International World Wide Web Conference, 2007, 271-280.

Alpaydin, E., "Introduction to Machine Learning," MIT Press, 2004.

Balabanovic, M., and Shoham Y., "Fab: Content-Based, Collaborative Recommendation," Communications of the Association for Computing Machinery 40(3), 1997, 66-72.

Basu, C., Hirsh, H., and Cohen, W., "Recommendation as Classification: Using Social and Content-Based Information in Recommendation," Proceedings of the fifteenth national/tenth conference on Artificial intelligence/Innovative applications of artificial intelligence, 1998, 714-720.

Baumann, S., and Hummel, O., "Enhancing Music Recommendation Algorithms Using Cultural Metadata," Journal of New Music Research 34(2), 2005, 161-172.

Bell, R., Koren, Y. and Volinsky, C., "Chasing $1,000,000: How we won the Neflix progressive prize," Statistical Computing and Graphics 18(2), 2007, 4-12.

Berson, A., K. Smith, and K. Thearing, "Building Data Mining Applications for CRM," McGraw-Hill, New York, 2000.

Breese, J. S., Heckerman, D. and Kadie, C., "Empirical analysis of predictive algorithms for collaborative filtering," In Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence (Madison, WI, July 1998).

Changchien, S.W., and T.-C. Lu, "Mining Association Rules Procedure to Support Online Recommendation by Customers and Products Fragmentation," Expert Systems with Applications 20(4), 2001, 325-335.

Chen, Q., and Aickelin, U., "Movie Recommendation Systems using an Artificial Immune System," Poster Proceedings of ACDM, 2004.

Cohen, W.W., "Fast Effective Rule Induction," Proceedings of the Twelfth International Conference on Machine Learning, 1995.

Feick, L. and Higie, R.A., "The effect of preference heterogeneity and

source characteristics on ad processing and judgments about endorsers," J. Advert. 21(2), 1992, 9-24.

Gediminas Adomavicius, Alexander Tuzhilin, "Using Data Mining Methods to Build Customer Profiles," Communications of the ACM 48(10), 2005, 80-90.

Golbeck, J., "Generating Predictive Movie Recommendations from Trust in Social Networks," iTrust 2006, lecture Notes in Computer Science, 2006, 93-104.

Golbeck, J., and Hendler, J., "FilmTrust: Movie Recommendations using Trust in Web-based Social Networks," IEEE CCNC 2006 proceedings., 2006.

Herlocker, J., J.A. Konstan, A. Borchers, and J. Riedl, "An Algorithmic Framework for Performing Collaborative Filtering," Proceedings of the 1999 Conference on Research and Development in Information Retrieval, 1999.

Hill, W., Stead, L., Rosenstein, M., and Furnas, G., "Recommending and Evaluating Choices in a Virtual Community of Use," Proceedings of the SIGCHI conference on Human factors in computing systems, 1995, 194-201.

Im, I. and Hars, A. "Does a one-size recommendation system fit all?: The effectiveness of collaborative filtering based recommendation systems across different domains and search modes," ACM

Transactions on Information Systems 26, 1 (November 2007).

Sarwar, B.M., G. Karypis, J.A. Konstan, and J. Riedl, "Analysis of Recommendation Algorithms for e-Commerce," Proceedings of the ACM E-Commerce 2000 Conference (2000), 158-167.

Kim, J.K., Y.H. Cho, W.J. Kim, J.R. Kim, and J.H. Suh, "A Personalized Recommendation Procedure for Internet Shopping Support," Electronic Commerce Research and Applications, Vol. 1, No. 3/4(2002), 301-313.

Konstan, J.A., B. Miller, D. Maltz, J.L. Herlocker, L.R. Gordon, and J. Riedl, "GroupLes: Applying Collaborative Filtering to Usenet News," Communications of the ACM, Vol. 40, No. 3(1997), 77-87.

Krulwich, B., and Burkey, C. 1996. Learning user information interests through extraction of semantically significant phrases. Proceedings of the AAAI Spring Symposium on Machine Learning in Information Access.

Lang, K. 1995. Newsweeder: Learning to filter netnews. Proceedings of the 12th International Conference on Machine Learning.

Mitchell, T.M. 1997. Machine Learning. McGraw-Hill Education.

Mild, A., and M. Natter, "Collaborative Filtering or Regression Models for Internet Recommendation Systems?" Journal of Targeting, Measurement and Analysis of Marketing, Vol.10, No. 4(2002),

304-313.

Miller, B.N., Albert, I., Lam, S.K., Konstan, J.A., and Riedl, J. 2003. MovieLens Unplugged: Experiences with an Occasionally Connected Recommender System. Proceedings of the 8th international conference on Intelligent user interfaces, 263-266.

Mobasher, B., H. Dai, T. Luo, and M. Nakagawa, "Discovery and Evaluation of Aggregate Usage Profiles for Web Personalization," Data Mining and Knowledge Discovery, Vol. 6, No.1(2002), 61-82.

Mukherjee, R., Dutta, P. S., Jonsdottir, G., and Sen, S. 2001. MOVIES2GO - An Online Voting Based Movie Recommender System. Proceedings of the fifth international conference on Autonomous agents, 114-115.

O'Connor, M., and Herlocker, J. 2001. Clustering Items for Collaborative Filtering. Proceedings of SIGIR-2001 Workshop on Recommender Systems.

Resnick, P., and Varian, H. R. 1997. Recommender Systems. Communications of the Association for Computing Machinery 40(3), 56-58.

Roh, T.H., K.J. Oh, and I. Han, "The Collaborative Filtering Recommendation Based on SOM Cluster-Indexing CBR," Expert Systems with Applications, Vol. 25, No. 3(2003), 413-423.

Sarwar, B.M., G. Karypis, J.A. Konstan, and J. Riedl, "Analysis of Recommendation Algorithms for e-Commerce," Proceedings of the ACM E-Commerce 2000 Conference (2000), 158-167.

Sarwar, B.M., G. Karypis, J.A. Konstan, and J. Riedl, "Item-Based Collaborative Filtering Recommendation Algorithms," Proceedings of the 10th International World Wide Web Confernece(2001), 285-295.

Schafer, J.B., J.A. Konstan, and J. Riedl, "E-Commerce Recommendation Applications," Journal of Data Mining and Knowledge Discovery, Vol. 5, No. 1(2001), 115-152.

Schafer, J.B., J.A. Konstan, and J. Riedl, "Recommender Systems in e-Commerce," Proceedings of the ACM Confernece on Electronic Commerce(1999).

Shardanand, U., and Maes, P. 1995. Social Information Filtering: Algorithms for Automating "Word of Mouth". Proceedings of the SIGCHI conference on Human factors in computing systems, 210-217.

Sheth, B., and Maes, P. 1993. Evolving agents for personalized information filtering. Proceedings of the 9th IEEE Conference on Artificial Intelligence for Applications.

Weng, S.-S., and M.-J. Liu, "Feature-Based Recommendations for One-to-One Marketing," Expert Systems with Applications, Vol. 26,

No. 4(2004), 493-508.

Xiao, B. and Benbasat, I. 2007 E-commerce product recommendation agents: Use, characteristics, and impact. MIS Quarterly 31, 1 (March 2007), 137-209.

Yuan, S., and W. Chang, "Mixed Initiative Synthesized Learning Approach for Web Based CRM," Expert Systems with Applications, Vol. 20, No. 2(2001), 187-200.

Abstract

A Study on the Improving Prediction Accuracy of the CF based Recommendation Systems

Kim, Byung ho
The School of Business
The Graduate School
Yonsei University

Recommendation systems are useful method for one-to-one marketing and web-personalization. Recently, many researchers on recommendation systems and CF-based recommendation systems have been proceeding in both research and practice. However, previous studies did not reflect the preference heterogeneity and traditional CF-based recommendation systems have applied constant setting such as a reference group size. To overcome this limitation, this paper proposed new methodology for CF-based recommendation systems. The proposed methodology uses optimal personalized settings for each user and applies them to generating recommendations for individual users. The new methods, personalized settings, are compared with old methods, constant settings, using Netflix data. The results of the simulation show that the proposed method outperforms the traditional, ordinary CF-based

recommendation systems. Limitations, implications and future research directions are also dicussed.

Key words : Recommendation systems, CF-based recommendation systems, personalized, netflix, simulation, reference group size