



저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:



저작자표시. 귀하는 원저작자를 표시하여야 합니다.



비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.



변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 [이용허락규약\(Legal Code\)](#)을 이해하기 쉽게 요약한 것입니다.

[Disclaimer](#)

공 학 박 사 학 위 논 문

아이템 특성을 기반으로 한 상품의
추천 기법



2011년 8월

부 경 대 학 교 대 학 원

전자상거래협동과정

윤 소 영

공 학 박 사 학 위 논 문

아이템 특성을 기반으로 한 상품의 추천 기법

지도교수 윤 성 대

이 논문을 공학박사 학위논문으로 제출함.



2011년 8월

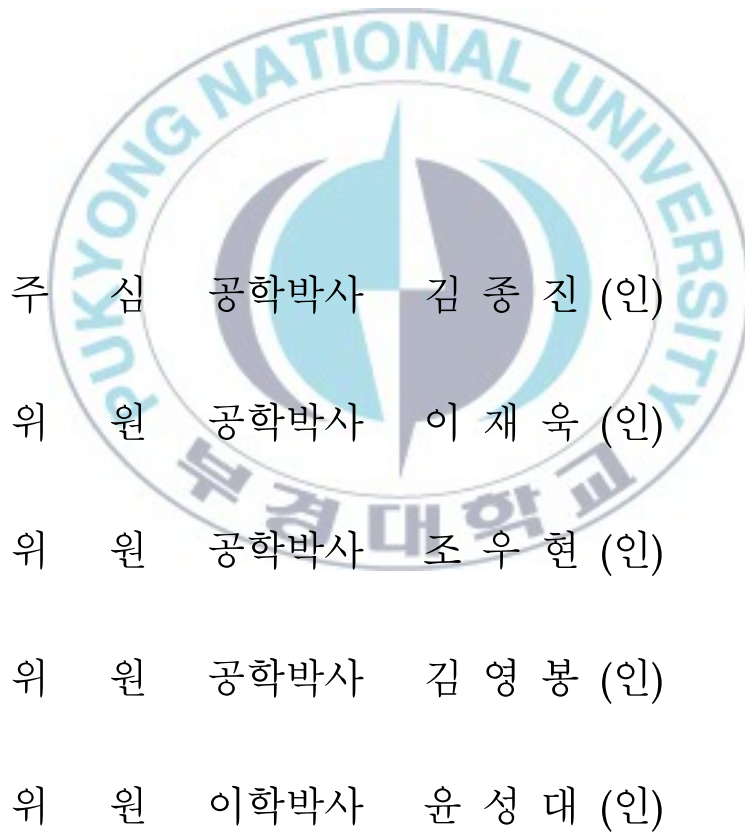
부 경 대 학 교 대 학 원

전자상거래협동과정

윤 소 영

윤소영의 공학박사 학위논문을 인준함

2011년 8월 26일



목 차

표 차례	iii
그림 차례	v
Abstract	vi
1. 서 론	1
2. 관련 연구	5
2.1 추천 시스템	5
2.1.1 내용 기반 추천 기법	6
2.1.2 규칙 기반 추천 기법	9
2.1.3 사례 기반 추천 기법	11
2.1.4 협업 필터링	14
2.2 협업 필터링	15
2.2.1 협업 필터링 분류	15
2.2.2 협업 필터링 알고리즘	19
2.2.2.1 이웃 기반 알고리즘	19
2.2.2.2 장르 기반 알고리즘	25
2.2.3 추천 성능 평가	27
2.2.3.1 예측 평가 방법	27
2.2.3.2 추천 평가 방법	28
2.2.4 협업 필터링의 문제점	30
3. 아이템 특성을 가중치로 이용한 추천 기법 제안	32
3.1 아이템 특성 추출 단계	32

3.1.1 장르 특성 추출	33
3.1.2 아이템에 대한 사용자 정보 특성 추출	37
3.2 아이템 추천 단계	47
3.2.1 아이템 간 유사도 측정	47
3.2.2 아이템 예측값 생성	50
3.2.3 아이템 추천	52
4. 실험 및 평가	57
4.1 실험 환경	57
4.2 실험 방법	58
4.3 성능 평가	60
4.3.1 예측 성능 평가	61
4.3.2 추천 성능 평가	63
4.3.3 추천 효율성 평가	68
5. 결 론	74
[참고문헌]	77

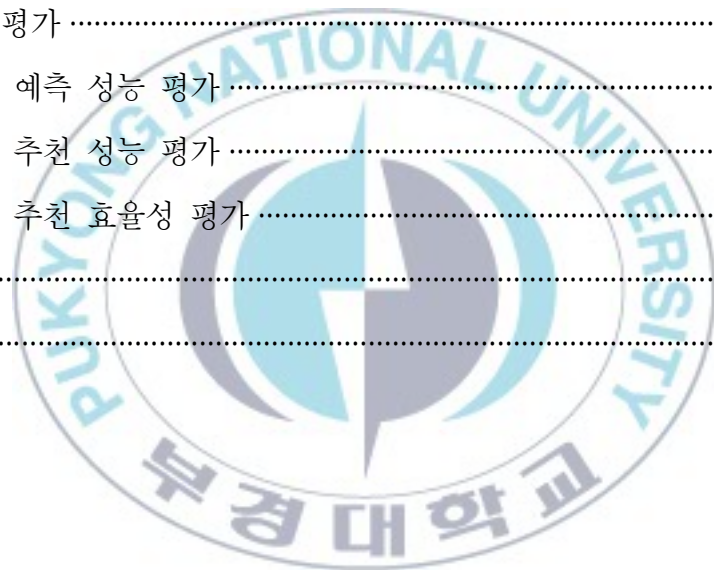


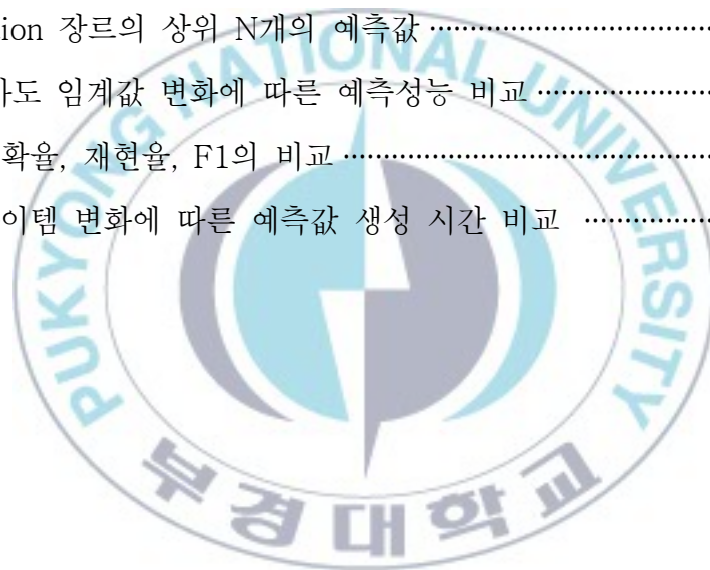
표 차례

<표 1> 사용자 × 아이템 매트릭스	19
<표 2> 사용자 × 아이템 매트릭스 예	20
<표 3> 목표 사용자 C1과 다른 사용자들 간의 유사도	23
<표 4> 목표 사용자 C1의 아이템 평가 예측	24
<표 5> 장르별 아이템 수	34
<표 6> 장르 분류의 적용	35
<표 7> Action 장르의 아이템들	36
<표 8> 연령과 직업의 분류	37
<표 9> 장르별 직업별 인원수	38
<표 10> 장르별 성별 인원수	39
<표 11> 장르별 연령별 인원수	39
<표 12> Action 장르 중에서 아이템 2	41
<표 13> 직업별 인원	42
<표 14> 성별 인원	42
<표 15> 연령별 인원	42
<표 16> 장르별 비율	43
<표 17> 성별 비율	43
<표 18> 연령별 비율	43
<표 19> Action 장르의 아이템들의 특성	44
<표 20> 아이템 2와 비교 아이템간의 유사도	49
<표 21> 제안하는 기법의 추천 아이템	54

<표 22> 아이템 기반 기법의 추천 아이템	55
<표 23> 장르 기반 기법의 추천 아이템	55
<표 24> 사용자 아이디 7이 선택한 아이템들	55
<표 25> 유사도 임계값 변화에 따른 예측 성능 비교	62
<표 26> 아이템 기반 기법의 정확율과 재현율	63
<표 27> 장르 기반 기법의 정확율과 재현율	64
<표 28> 제안하는 기법의 정확율과 재현율	64
<표 29> 아이템 기반 기법의 F1	66
<표 30> 장르 기반 기법의 F1	66
<표 31> 제안하는 기법의 F1	67
<표 32> 아이템 특성 추출 시간	69
<표 33> 트랜잭션 변화에 따른 매트릭스 생성 시간 비교	70
<표 34> 비교 기법들의 트랜잭션 변화에 따른 매트릭스 생성 시간 비교	70
<표 35> 장르별 예측값 생성 시간	71
<표 36> 아이템 변화에 따른 예측값 생성 시간 비교	72

그림 차례

(그림 1) 내용 기반 추천 시스템의 일반적인 구조	8
(그림 2) 규칙 기반 추론 구성	10
(그림 3) CBR Cycle	12
(그림 4) 아이템의 특성 추출 알고리즘	45
(그림 5) 새로운 아이템 × 사용자 매트릭스	46
(그림 6) 아이템 유사도 계산과 예측값 계산 알고리즘	51
(그림 7) 아이템 예측값 채우기 알고리즘	52
(그림 8) Action 장르의 상위 N개의 예측값	53
(그림 9) 유사도 임계값 변화에 따른 예측성능 비교	62
(그림 10) 정확율, 재현율, F1의 비교	67
(그림 11) 아이템 변화에 따른 예측값 생성 시간 비교	72



A Recommendation Technique of Product based on the Item Characteristic

So-Young Yun

Dept. of Interdisciplinary Program of Electronic Commerce of
Pukyong National University

Abstract

Now users have to pour more endeavor to find information they want for the drastic increase of information and the expansion of various information communication devices. Companies are using a recommendation system which recommends an item to the user in order to settle the problem of such information overload and secure more users.

Collaborative filtering is the most widely used method among the recommendation systems and is based on user or item evaluation. However, collaborative filtering exhibits problems such as sparsity, scalability, and cold start which reduce the accuracy of recommendation.

To get rid of the problems of collaborative filtering, this dissertation suggests a method to use item characteristics as the weighted value. The proposed method consists of steps for item

characteristic extraction and item recommendation. The item characteristic extraction step classifies items by genres and uses only the data of which rating is 4 or higher to analyze the user information over the item.

The item recommendation step computes similarity by making use of item characteristic. After computing similarity, only the data which satisfies the condition is designated as nearest neighbor. It computes the predicted value of the targeted item using the rating of k-nearest neighbors and designates the rating of the unrated cell using only the predicted value of Top-N. When it recommends the item to the user, only the items of which the predicted value is 3 or higher are recommended.

The technique suggested through this method not only reduces sparsity problem but enhances accuracy as well. And when a new item and user are registered, it is possible to conduct fast classification and recommendation based on analyzed information and also reduce the cold start problem.

The experiment result using MovieLens data set showed that the suggested technique has been more enhanced the accuracy, appropriacy, and efficiency than item-based and genre-based method.

1. 서 론

인터넷의 보급과 e-commerce의 도입으로 많은 사용자들은 자신이 원하는 정보를 보다 빠르고 쉽게 얻을 수 있게 되었다. 그러나 다양한 정보 통신 기기의 확산과 정보의 급격한 증가로 인해 사용자들은 자신이 원하는 적합한 정보를 찾기 위해 이전 보다 많은 노력을 기울여야만 하게 되었다. 기업은 이러한 정보 과잉으로 인해 발생하는 문제를 해결하고 더 많은 사용자를 확보하기 위해 추천 시스템을 사용하고 있다. 추천 시스템은 사용자들에게 그들이 관심 있고 좋아할 만한 아이템을 추천해 주어 원하는 아이템을 쉽고 빠르게 찾을 수 있도록 돕는 역할을 한다. 추천 시스템은 온라인 뉴스, 영화, 다양한 형태의 web resource들을 추천하며 Amazon.com, CDNow, DangDang, Sinforyou와 같은 많은 e-commerce 사이트에서 사용되고 있다[1].

이러한 추천 시스템에는 내용 기반 추천 기법, 규칙 기반 추론 기법, 사례 기반 추론 기법, 협업 필터링 등 다양한 기법들이 사용되고 있다. 내용 기반 추천 기법은 아이템이나 사용자 정보의 내용에 기반하여 아이템의 특성 혹은 사용자의 프로파일을 구성하고 학습하여 추천에 이용한다. 이 추천 기법은 신문 기사나 논문과 같이 내용 정보가 풍부한 아이템에 적용했을 경우, 비교적 정확한 추천을 할 수 있지만 내용 정보가 빈약한 영화나 음반 같은 아이템은 특징을 적절히 표현할 수 없어 올바른 추천이 어렵다는 문제점이 있다[2].

규칙 기반 추론 기법은 전문가 시스템에서 전문가가 가지는 지식을 규

칙 형식으로 규칙 기반에 저장하여 두고 저장된 규칙에 따라서 주어진 문제를 해결하는 기법이다. 이 기법에서는 문제영역의 규칙을 인간 전문가로부터 모두 추출한 다음 관리자가 규칙을 정리하여 규칙 기반으로 구현하고 이 규칙에 의해 추론하여 해를 얻는다. 그러나 실제로 문제를 해결할 때 미리 모든 규칙을 구축할 수 없는 경우가 많으며, 문제와 규칙이 일치하지 않을 경우에는 문제를 해결하기 어렵다. 또 이 기법은 문제가 주어질 때마다 이를 해결하기 위해 관련된 규칙을 순서대로 처리해야 하므로 규칙의 수가 증가할수록 성능이 저하되고[66] 이 과정에서 시간이 상당히 소모되어 지식획득의 병목 현상이 발생하는 문제점을 가진다[67].

사례 기반 추론 기법은 과거의 어떤 문제를 해결하기 위해 사용했던 경험을 바탕으로 새로운 문제를 해결하는 기법이다. 이 기법은 기존의 사례가 풍부할 때는 유용하지만 사례의 연관성이 검색되지 않을 경우 추천이 불가능하다는 문제점을 가진다[71].

협업 필터링은 유사한 사용자들은 유사한 성향을 가진다는 것에 기반하며 사용자들의 평가 정보를 포함한 데이터베이스를 구축하고 목표 사용자와 유사한 선호도를 가진 사용자들을 데이터베이스로부터 찾아내어 이들의 선호에 기반해 새로운 평가를 예측하고 이를 목표 사용자에게 추천하는 방식이다. 추천 시스템들 중에서 가장 널리 사용되고 있다.

협업 필터링은 이 기법을 적용한 초기 연구인 GroupLens 시스템, Tapestry 시스템을 포함해 다양한 기법들이 꾸준히 연구되고 있으며 e-commerce에서 성공적으로 널리 사용되고 있지만 희소성(sparsity), 확장성(scalability), 초기 평가(cold start)등의 문제점을 가진다.

희소성은 사용자들이 데이터베이스에서 이용 가능한 전체 아이템들 중

소수의 아이템만을 평가하기 때문에 발생하며[3], 이로 인해 사용자와 아이템 사이에 선호도 정보가 매우 적게 나타난다. 데이터가 희소성을 나타낼 경우 사용자의 성향 분석을 위한 정보가 희소하여 사용자의 성향 분석을 이용해 최근접 이웃을 선정하는 것을 어렵게 만든다. 그러므로 희소성은 추천의 정확성을 떨어뜨리는 가장 큰 요인이다.

확장성은 아이템의 수와 사용자 수가 증가함에 따라 목표 사용자의 최근접 이웃을 찾기 위한 연산이 급격히 증가하는데 이에 따른 데이터 처리는 신속히 이루어지지 못함으로써 발생하는 문제이다[4].

초기 평가 문제는 새로운 사용자가 가입했을 경우 사용자의 아이템에 대한 선호도 정보가 없어 유사 사용자를 찾을 수 없고 성향 분석도 할 수 없어 아이템 추천이 불가능하여 발생하게 되는 문제이다.

협업 필터링의 문제점인 희소성과 확장성, 초기 평가 문제를 해결하기 위해 인구 통계 정보를 이용한 연구[5], 사용자 평가 매트릭스를 이용한 연구[6], 협업 필터링과 내용기반 필터링 기법의 장점을 결합한 하이브리드 연구[5, 7, 8, 9], 장르 기반 연구[10, 63]등 다양한 연구들이 진행되고 있다.

본 논문에서는 추천의 정확성에 심각한 문제를 발생시키는 데이터의 희소성과 확장성을 줄이고 초기 평가 문제를 완화시켜 정확성을 높이기 위해 아이템의 분류 정보와 사용자 정보(user profile)를 활용한 기법을 제안한다.

제안하는 기법은 아이템의 분류 속성으로 영화의 장르 속성을 사용하고 사용자들의 정보 속성으로 성별, 나이, 직업 속성을 사용하여 아이템에 대한 사용자 특성을 분석하고 이 특성들을 가중치로 사용하여 아이템의 유

사도를 계산한다. 유사도 계산 후 조건을 지정하여 이를 만족하는 데이터들만을 사용하여 예측값(아이템에 대한 선호도 예측값)을 계산하고 그 결과 중에서 Top-N 개의 값만을 사용하여 평가되지 않은 아이템에 예측값을 부여한다. 평가되지 않은 아이템에 예측값을 지정함으로써 데이터의 희소성을 줄이고 추천의 정확성을 높일 수 있다. 또한 아이템들에 대해 분석한 특성을 이용하여 새로운 아이템이 입력되었을 때 유사 아이템을 선호하는 사용자에게 빠르게 추천이 이루어질 수 있으며 새로운 사용자가 가입했을 때도 사용자들에 대한 분류를 기준으로 빠르게 추천이 이루어질 수 있다.

본 논문의 구성은 다음과 같다. 2장에서는 관련 연구인 추천 시스템과 협업 필터링에 대해 기술하고, 3장에서는 사용자 정보를 활용하여 아이템 특성을 추출하는 제안하는 기법에 대해 살펴본다. 4장에서는 실제 데이터를 대상으로 제안하는 기법의 성능을 평가하여 실험 결과를 기술하고 마지막 5장에서는 결론 및 향후 연구를 제시한다.

2. 관련 연구

2.1 추천 시스템

정보 통신 기술의 발달과 인터넷 확산으로 기업들은 넘쳐나는 정보로부터 고객이 원하는 정보를 빠르게 찾아 사용자에게 제공하기 위해 추천 시스템을 도입했다. 추천 시스템은 새로이 등장한 개인화 시스템으로 사용자의 취향에 적합한 상품이나 아이템을 찾아 사용자에게 추천하는 시스템이다[11].

추천 시스템의 개념은 1992년 Goldberg[12]에 의해 최초로 제안되었으며 초기에는 많은 양의 문서에서 자신의 선호도에 적합한 문서만을 필터링하기 위한 도구로 활용되었다.

초기 추천 시스템은 usenet news의 기사나 웹 페이지와 같은 대량의 문서에서 사용자들이 선호하는 내용을 자동적으로 선별하도록 고안되었으며 정보검색과 인공지능 분야에 활용되었다[11, 13]. 이후 추천 시스템은 아이템, 서비스를 사용자에게 추천하기 위해 다양한 응용분야에 이용되고 있다.

특히 전자상거래에서 추천 시스템은 인터넷과 접목되면서 사용자와 제품 간의 관계에서 얻어지는 거래 데이터와 인터넷 상에서 실시간으로 얻어지는 사용자 행동에 대한 데이터를 분석하여 아이템을 추천하기 위한 핵심적인 정보를 수집하게 되었다. 사용자와 아이템 간의 관계 데이터는 사용자의 인구 통계학적 자료와 기업이 보유하고 있는 사용자 정보와 아이템 정보로부터 얻게 된다. 이러한 데이터들은 사용자의 선호도 예측 모

형과 알고리즘을 이용하여 사용자가 원하는 아이템을 추천하기 위한 추천 시스템의 추천 엔진에 입력되고 분석되어 사용자의 구매에 영향을 주게 된다[64].

추천 시스템은 사용자와 기업 모두에게 이익을 발생시킨다. 사용자는 본인이 원하는 아이템이나 서비스를 보다 빠르게 이용할 수 있어 원하는 것을 찾기 위한 시간과 노력을 줄일 수 있는 이점이 있다. 그리고 기업은 이전까지의 단순 방문자를 구매자로의 전환을 유도할 수 있으며 교차 판매의 효과와 사용자 충성도를 증진시킬 수 있는 이점을 얻을 수 있다[14].

추천 시스템은 이미 Amazon.com, CDnow 등과 같은 전자상거래 기업들에 성공적으로 적용되어 사용자의 구매를 촉진시켰으며, 많은 전자상거래 사이트에서의 직접적인 판매 향상을 확인할 수 있다[15, 16].

추천 시스템의 도입으로 웹상에서의 전자상거래는 더욱 활성화되었으며, 이제는 유비쿼터스 컴퓨팅 환경과 같은 새로운 분야에까지 확장하여 적용하기 위한 연구가 활발히 진행되고 있다.

이러한 추천 시스템이 데이터를 처리하여 추천목록을 생성하는 기법으로는 내용 기반 추천 기법, 규칙 기반 추론 기법, 사례 기반 추론 기법, 협업 필터링 등 다양한 방법이 있지만 그 중에서 협업 필터링이 가장 성공적으로 적용된다[5, 17].

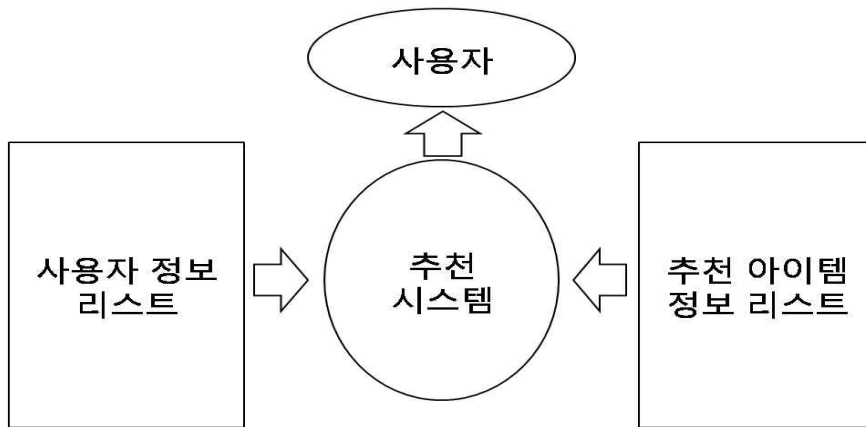
2.1.1 내용 기반 추천 기법

내용 기반(content-based) 추천 기법은 정보검색 분야에서 기본적인

아이디어를 가져왔다고 할 수 있다. 검색 시스템에서 사용자는 시스템에 질의를 던지고 시스템은 사용자에게 질의의 결과로 문서를 순위화하여 보여준다. 순위를 매기는 기준은 사용자의 질의와 질의의 대상이 되는 문서가 얼마나 유사한가이다. 이때 유사도를 측정하는 방법은 여러 가지가 있지만 기본적으로 문서 안에 있는 단어들과 질의가 가진 단어들을 비교하는 것으로 이루어진다. 내용 기반 추천에서는 사용자가 질의를 하기 전에 시스템이 추천을 해야 하므로 사용자의 질의는 사용자 정보로 대체된다. 사용자 정보는 개념적으로 장기간 동안 변하지 않는 질의라고 생각할 수 있다[65].

내용 기반 추천 기법은 아이템이나 사용자 정보의 내용에 기반하여 아이템의 특성 혹은 사용자의 프로파일을 구성하고 학습하여 추천에 이용한다. 내용 기반 추천은 아이템 간의 특성에 대한 유사성을 이용하는 아이템 상관관계 추천 방식과 사용자 프로파일과 아이템 특성의 유사성을 이용하는 속성 기반 추천 방식이 존재한다[18]. 일반적으로 내용 기반 추천 기법은 속성 기반 추천 방식을 의미하는 경우가 많다. 속성기반 추천 방식은 아이템의 속성을 문자화시켜 고객의 프로파일과 일치도가 높은 아이템을 찾는 정보검색 방법을 적용한 것으로 정보검색을 위해 이웃의 기능과 분류화기법 등이 적용되어 아이템의 특성을 문자화시키기 위해 사용된다[19].

그림 1은 내용 기반 추천 시스템의 일반적인 구조이다. 내용 기반 추천에서 가장 중요한 두 가지 요소는 사용자 정보와 추천의 대상이 되는 아이템에 대한 정보로서 사용자 정보는 사용자가 자신의 선호도를 표시한 것들로 이루어진다[65].



(그림 1) 내용 기반 추천 시스템의 일반적인 구조
 (Figure 1) General structure of CBR systems

내용 기반 추천 기법은 베이저안 분류기(Bayesian classifiers)를 비롯한 다양한 기계학습(machine-learning) 기술들도 사용된다[8].

내용 기반 추천 기법은 신문기사나 논문과 같이 내용 정보가 풍부한 아이템에 적용했을 경우, 비교적 정확한 추천을 할 수 있다는 장점이 있다.

그러나 내용 기반 추천은 몇 가지 문제를 가진다. 첫째는 영화나 음반과 같은 멀티미디어 정보처럼 내용 정보가 빈약한 아이템의 경우 아이템의 특징을 적절히 표현할 수 없기 때문에 유사도 측정 및 올바른 추천이 어렵다는 문제이다[2]. 예를 들어 영화는 해당 영화가 어느 나라에서 만든 영화인지, 장르가 무엇인지, 감독이 누구인지, 주연 배우가 누구인지 등을 알아야 각각의 정보에 대한 선호도에 대응하여 추천이 가능하다.

둘째는 이 기법은 두 개의 서로 다른 상품들이 같은 특징 집합으로 표현되었다면, 구분되지 않는다는 문제이다[2].

셋째는 고객의 선호도나 상품에 대한 흥미가 다른 고객들의 취향이나

선호도에 영향을 받음에도 불구하고 상품 추천을 위해 개별 고객 자신의 우선적인 경험만을 이용한다는 문제이다[9]. 예를 들어 어떤 사용자가 액션 영화를 좋아한다고 선호도를 밝혔다면 추천이 되는 영화 중 상위에 랭크되는 영화는 대부분 액션 영화일 가능성이 많아 사용자가 다른 장르를 추천 받을 가능성이 낮아진다.

마지막으로 사용자가 의사 표현을 얼마나 많이 하느냐에 따라 추천의 질이 달라진다는 문제이다. 사용자가 추천한 내용에 대해 평가를 내리면 시스템은 이를 바탕으로 사용자 정보를 수정할 수 있다. 사용자의 의사 표현은 많으면 많을수록 시스템이 정확한 추천을 하는데 도움이 되지만 사용자에게는 부담스러운 일일 수 있다[65].

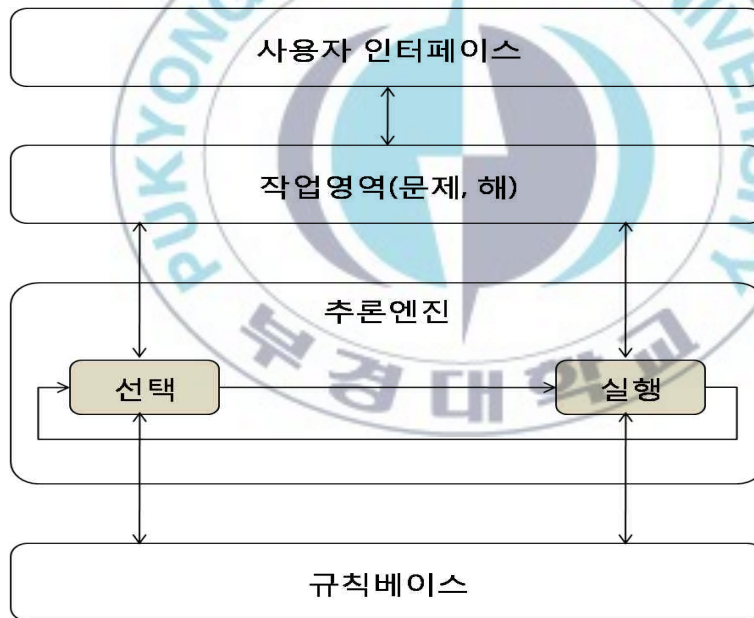
2.1.2 규칙 기반 추론 기법

규칙 기반 (ruled-based) 추론 기법은 인터넷 사이트 사용자에게 연속적인 질문을 던지고 이에 대한 사용자의 반응과 기존에 존재하는 사용자의 구매나 인구통계정보 등을 활용해서 적절한 아이템(상품, 웹 페이지, 광고 등)을 추천하는 기법으로 사용자에 의해서 입력되고 사용자에 의해서 생성된 데이터를 활용해서 세밀한 분석과 추론과정을 통해 규칙을 생성하게 된다. 예를 들어 컴퓨터 판매 사이트에서 사용자가 접속을 하면 사용자에게 적합한 컴퓨터를 추천해 주기 위해서 사용용도, 가격대, 컴퓨터 선호도에 대한 질문 등을 연속적으로 던지고 사용자의 이전 구매나 성향정보를 통합해서 최종적으로 그 사용자에게 적절하다고 판단되는 컴퓨터를

추천해 주는 것이다.

사용자에게 적합한 개인화된 추천을 위해서 사용되는 비즈니스 규칙은 웹 사이트의 마케팅 관리자에 의해서 새로운 정보가 축적될 때마다 주기적으로 갱신되고 수정되어 진다. 규칙 기반 추론 기법은 데이터마이닝 기법들을 많이 사용하는데 고객들을 분류해 내기 위해서 의사결정나무(decision tree)가 사용되고 여러 아이টে에 대해서 고객들의 아이টে 구매 가능성을 찾아내기 위해서 신경망(neural network)을 사용하게 되며 고객의 세분화를 위해서 군집화(clustering)가 사용된다.

규칙 기반 추론 기법은 사이트가 능동적으로 사용자에게 행동을 요구하게 되고 최종적으로 정보도 에이전트가 사용자에게 전달하게 된다[73].



(그림 2) 규칙 기반 추론 구성
(Figure 2) Constitution of rule-base reasoning

그림 2와 같이 규칙 기반 추론은 규칙베이스, 작업영역, 추론엔진 등으로 구성된다[20].

규칙 기반 추론 기법은 모듈성, 균일성, 자연성의 장점을 가지고 있다. 그러나 이 기법은 지식획득 병목현상이 발생하며, 문제의 이해가 쉽지 않고 계속적으로 변화하는 문제 영역에 부적절하다는 문제점이 있다[67].

2.1.3 사례 기반 추론 기법

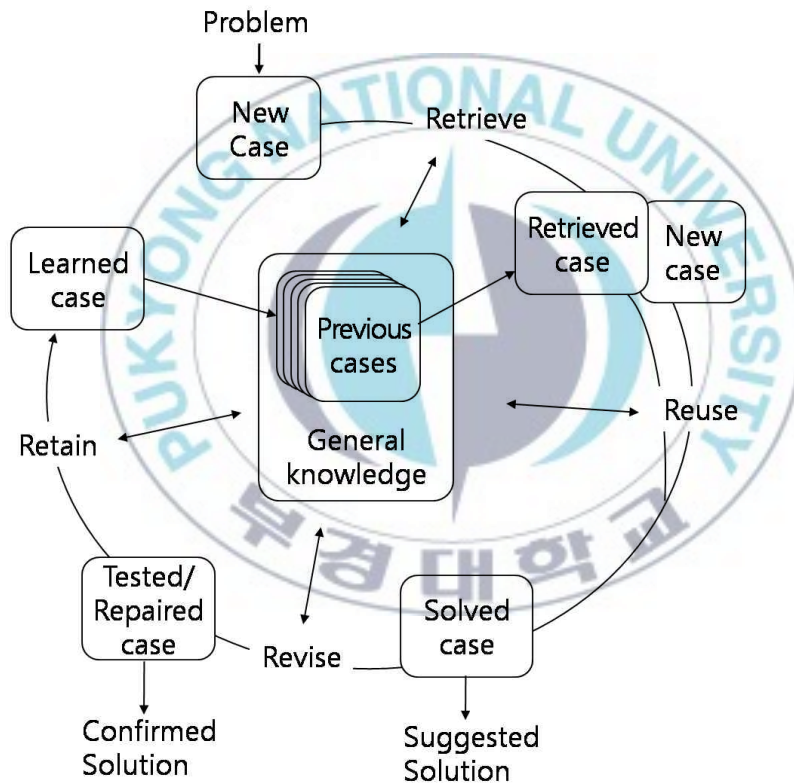
사례 기반 추론(Case-Base Reasoning) 기법은 과거의 어떤 문제를 해결하기 위해 사용했던 경험을 바탕으로 새로운 문제를 해결하는 기법으로 규칙 기반 필터링 기법의 문제점을 해결할 수 있다.

사례 기반 추론 기법은 새로운 문제에 직면했을 때, 이를 해결하는 방법으로 과거의 비슷한 상황을 기억하고 그 상황에서 사용했던 정보와 지식을 재사용함으로써 새로운 문제를 해결한다[20].

사례 기반 추론 기법은 두 개의 기본 사상에 기반을 하는데 하나는 유사한 문제는 유사한 해법을 가진다는 것이고, 다른 하나는 한번 발생한 문제는 자주 발생할 수 있다는 것이다. 따라서 과거에 현재의 문제와 유사한 문제가 존재하였고 그것이 어떻게 해결됐는지를 안다면 과거의 경험을 바탕으로 현재 문제의 해결책을 추론할 수 있다는 것이다. 사례 기반 추론 기법의 문제 해결 방식은 인간의 문제 해결 방식과 유사하기 때문에 그 결과를 이해하기 쉽고, 새로운 사례를 단순히 저장하는 것만으로도 추가적인 작업 없이 학습이 진행된다는 장점을 가진다[21].

사례 기반 추론 기법은 법률, 비즈니스, 의학, 경영, 상식에 의한 판단 등의 분야에서 이용되고 있으며[22], 고장진단[23, 24, 25], 헬프데스크 [26], 전략수립[27], 유비쿼터스 컴퓨팅 시스템의 상황인식 기능 및 개인화 서비스 구현[28] 등에도 이용되고 있다.

Aamodt & Plaza[29]는 사례 기반 추론 기법의 과정을 아래 그림 3과 같이 검색(Retrieve), 재사용(Reuse), 수정(Revise), 유지(Retain)의 4단계로 설명하였다.



(그림 3) CBR Cycle
(Figure 3) CBR Cycle

- 1) 검색(Retrieve) : 현재 문제와 가장 유사한 과거 사례들을 사례베이스로부터 찾아내는 단계이다.
- 2) 재사용(Reuse) : 검색으로 찾은 과거 유사 사례들의 해법을 현재 문제 해결을 위해 사용하는 단계이다.
- 3) 수정(Revise) : 현재 문제의 해결을 위해 검색된 유사 사례들의 해법을 현재 문제에 적합한 형태로 조정하는 단계이다.
- 4) 유지(Retain) : 새롭게 해결된 문제와 해법을 향후 새로운 문제 해결을 위한 목적으로 사례베이스에 저장하는 단계이다.

사례베이스로부터 유사사례를 찾기 위해 가장 많이 사용되는 유사도 측정기법은 k-NN(k-Nearest Neighbors)이다. k-NN은 현재 사례와 가장 유사한 k개의 사례를 검색하는 기법이다. 일반적으로 사용되는 유사도 측정식은 다음의 식 (1)과 같다.

$$Similarity(N, C) = \frac{\sum_{i=1}^n f(N_i, C_i) \times W_i}{\sum_{i=1}^n W_i} \quad (1)$$

N은 새로운 사례, C는 사례베이스에 저장된 과거 사례, n은 사례가 가지는 속성의 개수, N_i 는 새로운 사례의 i 번째 속성값, C_i 는 과거 사례의 i 번째 속성값, $f(N_i, C_i)$ 는 N_i 와 C_i 사이의 거리 측정 함수, W_i 는 i 번째 속성에 대한 가중치이다.

사례간의 유사도는 일반적으로 0에서 1사이의 정규화된 실수값으로 표

현되는데, 0에 가까울수록 두 사례의 유사성이 낮다는 것을 의미하고, 1에 가까울수록 유사성이 높다는 것을 의미한다[70].

그러나 이 기법은 기존의 사례를 성공적으로 평가하여 유사한 사례를 제시해야 하는데 이는 기존의 사례가 충분할 때만 가능하다. 만약 사례의 연관성이 검색되지 않을 경우 어떠한 추천도 불가능하며, 단순한 구매패턴에 의존한 방식이므로 고객의 특성을 고려하지 못하는 단점이 있다[71].

2.1.4 협업 필터링

협업 필터링(collaborative filtering)기법은 사용자가 선호하는 패턴과 유사한 다른 사용자들의 선호도를 이용하여 사용자에게 관련된 아이템이나 서비스를 추천하는 기법이다. 이 기법은 추천 시스템 분야에서 가장 성공적인 추천 기법으로 전자상거래 기업이 가장 널리 이용하고 있다[30].

그러나 이 기법은 비슷한 흥미를 갖는 사용자의 수가 적을 경우 정확한 사용자의 흥미를 예측하기 어렵고 새로운 사용자가 자신의 흥미에 대한 정보 제공을 하지 않았을 경우 추천을 할 수 없다는 초기 평가의 문제들이 있다[31]. 협업 필터링은 다음의 2.2절에서 자세히 기술한다.

2.2 협업 필터링

협업 필터링은 사용자와 아이템이 지닌 특성을 무시하고 사용자와 아이템간의 상호관계에 대한 데이터만을 이용하는 접근법을 취한다[11]. 이 상호관계 데이터는 일반적으로 명시적 성격을 지닌 수치 데이터를 활용하지만 암묵적 성격의 데이터 또한 명시적 데이터로 전환하여 사용할 수 있다. 사용자와 아이템의 상호 관계 데이터는 매트릭스의 형태로 표현되어 분석된다[13].

협업 필터링을 이용한 추천 시스템은 학계와 산업계에서 많이 개발되었다. Tapestry 시스템[12]은 사용자가 문서에 주석을 달 수 있으며 문서를 검색할 때 키워드 매칭과 주석을 통한 검색이 가능했다. GroupLens, Video Recommender, Ringo는 예측을 자동화하기 위해 협업 필터링 알고리즘을 사용한 첫 번째 시스템들이다[2].

2.2.1 협업 필터링 분류

협업 필터링은 데이터의 처리 방식에 따라 메모리 기반(memory-based) 알고리즘과 모델 기반(model-based) 알고리즘으로 분류되고[13] 사용자와 아이템의 관계에서 무엇을 중심으로 선호도를 예측하느냐에 따라 사용자 기반(user-based) 알고리즘[8]과 아이템 기반(item-based) 알고리즘으로 분류되며[32], 다른 기법과의 결합 또는 가중치 적용에 따라 하이브리드 기법과 가중치를 적용한 기법으로 분류된다.

메모리 기반 알고리즘은 사용자들이 아이템에 대해 평가한 전체 선호도 평가값을 기반으로 특정 아이템에 대한 목표 사용자의 선호도를 예측하는 방법으로 예측을 위해 사용자 데이터베이스의 모든 평가값을 이용하게 된다. 따라서 선호도 예측을 위해 예측 대상 아이템에 대해 선호도를 평가한 모든 사용자들의 정보를 이용하기보다는 선호도 평가값의 패턴에 따라 가장 유사한 일정수의 사용자를 선정하여 선호도를 예측하는 k-NN 방법을 이용하기도 한다[33].

그러나 이 알고리즘은 사용자가 아이템에 부여한 평가값의 수가 부족하여 추천의 성과가 떨어지는 희소성 문제점과 사용자의 수와 거래 데이터가 증가함에 따라 목표 사용자의 최근접 이웃을 찾기 위한 연산이 기하급수적으로 늘어난다는 확장성의 문제가 있다[12].

메모리 기반 알고리즘의 문제점들을 해결하기 위해 평가값이 비어 있는 셀에 기본 평가값을 부여하는 알고리즘[13], 개별 사용자들 간의 유사도를 계산하는 방법 대신 아이템 간의 유사도를 계산하는 알고리즘[32] 등 여러 방법들이 제안되었다.

모델 기반 알고리즘은 선호도 예측을 위해 사용자들의 선호도 평가값의 패턴을 기반으로 소규모 계층으로 사용자들을 구분한다. 즉, 목표 사용자의 선호도 평가값을 예측하기 위해 목표 사용자를 하나 또는 다수의 소규모 계층에 분류하여 추천 아이템에 대해 각 계층에 의해 예측된 값을 목표 사용자의 선호도 예측값으로 사용한다[64]. 모델 기반 알고리즘에서는 Aspect 기법[34], Bayesian network 기법[13], Clustering 기법[35], Graph model 기법[36] 등의 다양한 알고리즘이 사용된다.

그러나 모델 기반 알고리즘은 클래스의 모델 설정에 소요되는 시간이

매우 크며 또한 추가 데이터에 따라 모델의 재설정에 소요되는 시간이 매우 크므로 사용자 선호 모델들이 빈번하고 급속하게 업데이트 되어야만 하는 환경에는 적절하지 않은 문제점이 있다[37].

사용자 기반 알고리즘은 사용자간의 유사성을 측정하여 선호도가 비슷한 다른 사용자들이 평가한 아이템을 기반으로 목표 사용자가 선호할 만한 아이템을 추천하는 방식이다. 예측을 위해서는 전체 사용자들이 목표 사용자와 얼마나 유사한가를 알아야 하며, 이러한 유사 정도를 유사도(similarity)라 한다.

사용자 기반 알고리즘은 선호도가 비슷한 유사 사용자들이 동일하게 평가한 아이템에 대하여 상대적으로 높은 예측력을 보이고 있고[30], 새로운 아이템에 대한 예측도 가능하며, 데이터가 많은 경우에는 다른 기법에 비해 상대적으로 정확한 예측을 한다는 장점을 가진다. 그러나 둘 또는 그 이상의 사용자가 모두 평가를 내린 동일 아이템이 있어야 하는데, 사용자들이 서로 이질적인 평가를 한 아이템에 대해서는 단지 두 사용자 사이에서만 상관관계를 구하므로 예측의 정확성이 떨어질 가능성이 있다. 또, 사용자의 집단이 커지면 확장성의 문제가 발생하게 되어 연산처리를 많이 해야 하는 문제점이 있다. 사용자 기반 알고리즘에서 사용자 간의 연관성을 토대로 목표 사용자와 선호도가 비슷한 이웃들을 선정하는 기법으로 클러스터링, 최근접 이웃 추출법, 베이지안 네트워크 등이 있다[72].

아이템 기반 알고리즘은 사용자 기반 알고리즘의 문제점인 확장성 문제를 해결하기 위해 제안되었다[32]. 이 알고리즘은 대부분의 사람들이 과거에 자신이 좋아했던 아이템과 유사한 아이템을 선호하는 경향이 있고, 반대로 싫어하거나 선호하지 않았던 아이템과 유사한 아이템은 선호하지

않는 경향이 있다는 점을 근간으로 하고 있다. 아이템 기반 알고리즘은 아이템 간의 유사성을 측정하여 어떤 특정 사용자가 어떤 아이템을 선호하는지 예측하는 방식이다. 즉, 예측하고자 하는 아이템과 유사한 아이템들에 대하여 사용자가 높은 점수로 평가하였다면 그 아이템도 높게 평가할 것이며, 반면에 낮은 점수로 평가하였다면 그 아이템도 역시 낮은 점수로 평가할 것이라고 예측하는 방식이다[72]. 사용자 기반에서는 사용자의 증가속도를 시스템 설계자가 통제할 수 없지만 아이템 기반에서는 시스템 설계자가 통제가 가능하여 증가속도에 있어서도 거래 아이템의 증가속도가 사용자의 증가속도에 비해 느리다.

그러나 사용자들 간의 유사도가 전혀 고려되지 않기 때문에, 어떤 특정 사용자와 전혀 선호도가 다른 사용자들의 평가를 기반으로 한다면 아이템들 간의 상관관계 정확도가 떨어지게 되어 추천의 예측력과 추천 능력이 떨어질 수 있다[72].

사용자 기반 알고리즘과 아이템 기반 알고리즘은 기본적으로 아이템을 추출하기 위해 이웃 기반 알고리즘을 사용하는데 이웃을 구하기 위한 기준이 사용자인가 아이템인가의 차이만 존재한다.

하이브리드 기법은 협업 필터링에 다른 추천 기법을 결합하여 두 기법의 장점을 취하거나 사용자 기반 알고리즘과 아이템 기반 알고리즘을 결합하여 협업 필터링의 단점을 보완하기 위해 제안된 기법이다.

가중치 적용 기법은 협업 필터링의 성능을 향상시키기 위해 사용자 기반 알고리즘이나 아이템 기반 알고리즘에서 특정 속성을 가중치로 적용하는 기법이다. 사용자의 인구통계학적 정보나 아이템의 장르 정보를 가중치로 많이 사용한다.

2.2.2 협업 필터링 알고리즘

2.2.2.1 이웃 기반 알고리즘

협업 필터링기법으로 아이템을 추천하는 과정은 다음의 세 단계로 이루어진다.

1단계 : 평가를 위한 매트릭스 생성

1단계에서는 n명의 사용자들이 m개의 아이템에 부여한 평가값을 정리하여 표 1과 같은 $n \times m$ 의 사용자 \times 아이템 매트릭스를 생성한다.

<표 1> 사용자 \times 아이템 매트릭스

<Table 1> user \times item matrix

아이템 고객	Item1	Item2	Item3	Item4
User1	$R_{1,1}$		$R_{1,3}$	$R_{1,4}$
User2	$R_{2,1}$	$R_{2,2}$	$R_{2,3}$	
User3		$R_{3,2}$	$R_{3,3}$	$R_{3,4}$
User4	$R_{4,1}$		$R_{4,3}$	$R_{4,4}$
User5	$R_{5,1}$	$R_{5,2}$		$R_{5,4}$

$R_{1,1}$ 는 사용자 1이 아이템 1에 부여한 선호도 평가값을 의미하고 비어 있는 셀은 사용자가 아이템에 대한 선호도 평가를 하지 않았음을 의미한다.

<표 2> 사용자 × 아이템 매트릭스 예
 <Table 2> An example of user × item matrix

	I1	I2	I3	I4	I5	I6	I7		I13	I14	I15	I16	I17	I18	I19	I20
C1			4	3	3	5	4		5	5	5	5	3	4	5	4
C2	4	3			1					3			4			
C3	4		1							5	3		1		4	
C4				5			5								1	
C5	5			2					5	5	4				3	
C6			1			5	4				1					
C7	4			4			4		3			4				
C8											5					
C9							5		4	3	4			3	5	
C10		3				2				4		3				
C11							5		4	3	4			3	5	
C12	1	1			2		1							1		3
C13			1				5				5					1
C14	4						4		3							
C15	5				1				5	5	4				3	

표 2는 1단계의 예로써 목표 사용자 C1에게 아이템을 추천하기 위해 C1을 포함한 15명의 사용자가 20개의 아이템에 대해 평가한 트랜잭션 데이터베이스로부터 사용자 × 아이템의 매트릭스를 생성한 것이다.

2단계 : 이웃 선정

이 단계에서는 다른 사용자들과 목표 사용자의 유사도를 계산한 후 이웃을 선정한다. 일반적으로 유사도는 피어슨 상관계수(Pearson Correlation Coefficient)와 벡터 유사도(Vector similarity, Cosine)를 사용하여 계산한다.

두 사용자 a와 b의 선호도 유사 정도를 나타내는 피어슨 상관계수는 다음의 식 (2)와 같다[11].

$$Sim_{a,b} = \frac{\sum_{i=1}^m (R_{a,i} - \bar{R}_a)(R_{b,i} - \bar{R}_b)}{\sqrt{\sum_{i=1}^m (R_{a,i} - \bar{R}_a)^2} \sqrt{\sum_{i=1}^m (R_{b,i} - \bar{R}_b)^2}} \quad (2)$$

$R_{a,i}$ 와 $R_{b,i}$ 는 목표 사용자 a와 이웃 사용자 b가 공통으로 평가한 아이템 i 의 평가값을 의미하고 \bar{R}_a , \bar{R}_b 는 사용자 a와 b가 각각 선호도를 평가한 아이템들의 선호도 평가값의 평균이다.

벡터 유사도는 두 문서 간 유사성을 계산하기 위해 각 문서에서 단어의 출현 빈도를 벡터로 처리하여 계산한다. 이때의 코사인 벡터를 협업 필터링에 적용하여 이웃 사용자 a와 b가 공통으로 평가한 아이템들의 평가값을 q차원 공간에 벡터화 한 후 두 벡터 사이 각의 Cosine 값을 구할 수 있는데 그 식은 다음의 식 (3)과 같다.

$$Sim_{a,b} = \frac{\sum_{i=1}^m R_{a,i} R_{b,i}}{\sqrt{\sum_{i=1}^m R_{a,i}^2} \sqrt{\sum_{i=1}^m R_{b,i}^2}} \quad (3)$$

피어슨 상관계수와 벡터 유사도는 모두 사용자 간의 유사도를 나타내며 피어슨 상관계수는 두 사용자의 유사도를 1에서 -1까지의 양과 음의 관계로 표현하는 반면 벡터 유사도는 음의 값이 존재하지 않으며 최대 1의 유사도 가중치 값으로 정의된다.

Herlocker[38]는 사용자 간의 유사도를 계산할 때, 피어슨 상관계수를

사용하는 것이 Cosine 유사도를 사용하는 것보다 높은 추천 성과를 보인다고 하였다.

목표 사용자에게 대한 다른 사용자들의 유사도를 계산한 후 목표 사용자의 최근접 이웃을 선정한다. 최근접 이웃은 일반적으로 임계값(Threshold) 이나 k-NN 기법을 이용하여 선정한다. 임계값은 목표 사용자와의 유사도가 미리 지정한 임계값 이상인 사용자들만을 최근접 이웃으로 지정하는 방법이다. k-NN은 목표 사용자와의 유사도가 높은 상위 k명의 사용자들만을 목표 사용자의 최근접 이웃으로 지정하는 방법이다[11].

다음은 2단계의 예로써 표 2의 목표 사용자 C1과 다른 사용자 C2의 유사도를 피어슨 상관계수를 사용하여 계산한 결과이다.

$$\begin{aligned}
 Sim_{c1,c2} &= \frac{\sum_{i=1}^5 (R_{c1,i} - \overline{R_{c1}})(R_{c2,i} - \overline{R_{c2}})}{\sqrt{\sum_{i=1}^5 (R_{c1,i} - \overline{R_{c1}})^2} \sqrt{\sum_{i=1}^5 (R_{c2,i} - \overline{R_{c2}})^2}} \\
 &= \frac{(3-4)(1-3) + (4-4)(3-3) + (5-4)(4-3) + (5-4)(3-3) + (3-4)(4-3)}{\sqrt{(3-4)^2 + (4-4)^2 + (5-4)^2 + (5-4)^2 + (3-4)^2} \sqrt{(1-3)^2 + (3-3)^2 + (4-3)^2 + (3-3)^2 + (4-3)^2}} \\
 &= \frac{2+0+1+0+(-1)}{\sqrt{4}\sqrt{6}} \\
 &= 0.41
 \end{aligned}$$

목표 사용자 C1과 나머지 사용자들 간의 유사도 또한 같은 식을 사용하여 계산하면 다음 표 3과 같다. 구해진 유사도를 기준으로 k-NN기법을 사용하여 상위 5명의 사용자 C3, C5, C10, C13, C15만을 최근접 이웃으

로 지정한다.

<표 3> 목표 사용자 C1과 다른 사용자들 간의 유사도

<Table 3> Similarity between different user and target user C1

	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15
C1	0.41	0.85	-0.14	0.77	0.14	0.00	0.50	0.00	0.77	0.00	-0.17	0.58	-1.00	0.63

3단계 : 평가값 예측과 추천

3단계에서는 목표 사용자와 최근접 이웃들이 지정되면 목표 사용자가 평가하지 않은 아이템에 대해 이웃 사용자의 평가 정보를 이용하여 예측값을 계산한다. 예측값 $P_{a,i}$ 는 목표 사용자 a의 아이템 i에 대한 예측값으로 고객 a의 최근접 이웃의 평가값들을 가중평균으로 식 (4)와 같이 계산한다.

$$P_{a,i} = \bar{u}_a + \frac{\sum_{b=1}^n (u_{b,i} - \bar{u}_b) \times Sim_{a,b}}{\sum_{b=1}^n Sim_{a,b}} \quad (4)$$

\bar{u}_a 는 목표 사용자 a가 평가한 아이템들의 평가값의 평균이고, \bar{u}_b 는 사용자 a의 이웃 사용자 b가 평가한 아이템들의 평가값의 평균이다. $u_{b,i}$ 는 이웃 사용자 b의 아이템 i에 대한 평가값이다. $Sim_{a,b}$ 는 목표 사용자 a와 이웃 사용자 b의 유사도이고 n은 이웃들의 수이다.

목표 사용자가 평가하지 않은 아이템들의 평가값을 예측한 후 Top-N

기법으로 추천목록을 생성하여 수치가 높은 상위 N개의 아이템만을 목표 사용자의 예측값으로 사용한다.

다음은 2단계 예에서 추출한 최근접 이웃을 사용하여 목표 사용자 C1이 평가하지 않은 아이템 I1, I2, I8, I9를 예측식으로 계산하는 과정이다. 예측식을 이용하여 목표 사용자 C1의 아이템 I1에 대한 평가값을 예측하면 다음과 같다. I2, I8, I9를 같은 방법으로 예측하면 그 결과는 표 4와 같다.

$$P_{c1,I1} = \overline{R_{c1}} + \frac{\sum_{j=1}^3 (R_{j,I1} - \overline{R_b}) \times Sim_{c1,j}}{\sum_{j=1}^3 Sim_{c1,j}}$$

$$= 4.06 + \frac{(4-3.11) \times 0.85 + (5-4.14) \times 0.77 + (0-2.86) \times 0.77 + (0-3.67) \times 0.58 + (5-4) \times 0.63}{0.85 + 0.77 + 0.77 + 0.58 + 0.63}$$

$$= 4.06 + \frac{0.76 + 0.66 + (-2.21) + (-2.12) + 0.63}{3.61}$$

$$= 4.06 + (-0.63) = 3.43$$

<표 4> 목표 사용자 C1의 아이템 평가 예측
 <Table 4> Prediction of C1's item rating

	I1	I2	I8	I9
$P_{c1,x}$	3.43	1.18	4.21	2.93

최종적으로 예측된 값들 중에서 평가 값이 높은 상위 N개의 아이템만을 추천한다. N을 2로 지정하였다면 목표 사용자 C1에게 아이템 I1과 I8

을 추천한다.

2.2.2.2 장르 기반 알고리즘

장르 기반 알고리즘은 협업 필터링의 희소성 문제를 해결하기 위해 아이템의 장르를 활용한 알고리즘이다[10]. 장르 기반 알고리즘은 같은 장르에 속한 아이템이 다른 장르에 속한 아이템들보다 더 유사하다는 점에 초점을 둔 연구로서 아이템의 장르에 기반해 이웃 아이템들의 후보를 선정 후 평가 매트릭스를 통해 목표 아이템과 이웃 아이템들의 후보 간에 유사도를 계산하고 근접 이웃들의 집합을 추출한다. 장르 기반 알고리즘은 후보 아이템 생성 시 목표 아이템을 높게 평가한 사용자들의 그룹은 목표 아이템과 장르가 같은 다른 아이템들도 높게 평가할 것이라는 점을 이용하여 아이템의 평가값이 4이상인 데이터들만을 사용하였다.

장르 기반 알고리즘은 다음의 7단계로 이루어진다.

첫째, 랜덤하게 사용자들을 입력하고 사용자가 평가하지 않은 아이템 집합 I_{unrate} 를 가져온 후 I_{unrate} 의 속성이 될 목표 아이템 I_{aim} 을 선택한다.

둘째, 데이터베이스에서 목표 아이템을 높게 평가한 사용자들의 그룹을 선택한다. 목표 사용자가 평가한 아이템은 I_{other} 이다.

셋째, 모든 I_{other} 의 장르 수를 계산하고 목표 아이템과 I_{other} 의 장르 사이에 유사도 $simattri(i,j)$ 를 계산한다.

넷째, 유사도를 계산하는 3가지 방법(correlation, cosine, adjusted cosine)으로 목표 아이템과 I_{other} 간의 유사도 $simrat(i,j)$ 를 계산한다.

다섯째, 복합적 유사도 $sim_{inte}(i,j)$ 를 계산하고 목표 아이템을 위한 근접 이웃 아이템들의 집합 NI에서 상위 N개의 이웃 아이템들을 선택한다. 다음 식 (5)는 복합 유사도를 구하는 식이다.

$$sim_{inte}(i,j) = (1 - \alpha)sim_{attri}(i,j) + \alpha sim_{rat}(i,j) \quad (5)$$

α 는 0에서 1사이의 가중계수이다.

여섯째, $sim_{inte}(i,j)$ 와 목표 아이템과 가장 근접한 이웃 아이템들의 집합 NI에 속한 아이템들을 평가한 사용자들에 의한 예측값 $P(user, I_{aim})$ 을 다음의 식 (6)으로 계산한다.

$$P(user, I_{aim}) = \frac{\sum_{j \in NI} sim_{inte}(I_{aim}, J) \times R_{user,j}}{\sum_{j \in NI} |sim_{inte}(I_{aim}, J)|} \quad (6)$$

$R_{user,j}$ 는 아이템 j에 대한 목표 사용자의 평가값이고, NI는 가장 유사한 아이템들의 집합이다. $sim_{inte}(I_{aim}, J)$ 는 I_{aim} 번째 아이템과 J번째 아이템 사이의 유사도이다.

마지막으로 위의 두 번째에서 다섯 번째까지 과정을 반복하여 사용자가 아이템을 평가하지 않은 모든 예측값을 계산한 후 예측값 $P(user, I_{aim})$ 를 정렬하고 사용자에게 N개의 예측값을 추천한다.

2.2.3 추천 성능 평가

추천 시스템의 정확성을 평가하기 위한 방법으로 통계적 정확성 측정기법(statistical accuracy metrics)과 의사결정지원 측정기법(decision-support metrics)이 사용된다. 통계적 정확성 측정기법은 사용자가 평가한 값들과 예측된 값들을 비교하여 예측의 정확성을 측정하는 예측 평가 기법이고 의사결정지원 측정기법은 사용자가 양질의 아이템들을 선택하는 것을 얼마나 잘 지원하는가를 측정하는 추천 평가 기법이다[39].

2.2.3.1 예측 평가 방법

고객이 실제로 부여한 평가값과 추천 알고리즘에 의해 예측된 값의 차이로 추천 시스템의 성능을 측정할 수 있다. MAE(Mean Absolute Error)는 이러한 예측 평가를 위해 사용되는 기법으로 아이템에 대한 사용자의 실제 평가값과 추천 시스템의 예측값의 차이에 대한 절대 평균으로 추천의 성능을 평가하며 식 (7)과 같이 계산된다.

$$\text{MAE} = \frac{\sum_{i=1}^n |p_i - q_i|}{n} \quad (7)$$

n 은 예측한 아이템의 수이고, p_i 는 예측값이며, q_i 는 사용자의 평가값이다.

MAE의 값이 작을수록 예측의 정확성이 더 높아 우수한 성능의 추천 시스템으로 평가한다. 하지만 MAE가 낮다는 것이 추천의 적중률이 높다는 것을 보장 하는 것은 아니므로 MAE를 이용한 추천 시스템의 성능 평가에는 한계가 있다.

추천 시스템의 목적은 고객들에게 아이템을 선정하여 추천하는 것이다. 즉, 여러 개의 아이템들 중에서 예측된 값이 높은 상위 1개, 5개 또는 10개를 선정하여 추천하는 것이다. 추천된 아이템들 중에서 실제로 고객이 구매한 아이템이 속해 있으면 추천은 적중한 것이다.

하지만 MAE를 이용한 평가에서는 추천 시스템을 이와 같은 방법으로 평가하지 않는다. 그러므로 MAE가 아무리 낮게 나오는 추천 시스템이라고 해도 상위로 예측된 값의 아이템들이 고객이 실제로 높게 평가한 아이템들과 일치한다는 보장이 없다. 즉, MAE는 단지 예측된 값들과 실제 고객의 평가값들이 평균적으로 흡사하다는 것을 나타내는 것이지, 각 아이템 별로 평가값을 비교하는 것은 아니기 때문이다[63].

2.2.3.2 추천 평가 방법

Top-N 추천의 정확성을 평가하기 위해서 정확율(Precision), 재현율(Recall), F척도(F-measure)가 사용된다. 정확율과 재현율은 정보 검색 시스템의 성능을 평가하는 매우 유용한 평가 척도 중의 하나이며 추천 시스템의 성능 평가를 위해 많이 사용되고 있다[74].

아이템을 테스트 집합과 Top-N 집합으로 분류하고 이 두 집합에 모두 나타나는 집합을 hit 집합이라고 정의하면, 정확율은 추천된 아이템이 적합할 확률[74]로서 추천된 Top-N개의 집합에 대한 히트 집합의 비율이다. 재현율은 적합한 아이템이 선택될 확률[74]로서 테스트 집합에 대한 히트 집합의 비율이다.

다음 식 (8)과 식 (9)는 정확율과 재현율을 나타낸다[4].

$$\text{Precision} = \frac{\text{size of hit set}}{\text{size of top-N set}} = \frac{|test \cap top-N|}{N} \quad (8)$$

$$\text{Recall} = \frac{\text{size of hit set}}{\text{size of test set}} = \frac{|test \cap top-N|}{|test|} \quad (9)$$

그러나 정확율과 재현율은 상반되는 경향이 있어 N의 수가 증가하면 재현율은 높아지나, 정확율은 감소하는 경향이 있다. 따라서 정확율과 재현율을 병합한 식 (10)과 같은 F1 척도[4]를 사용한다. F1 척도는 정확율과 재현율에 같은 가중치를 부여하여 성능을 평가하는 기준이다[74].

$$F1 = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}} \quad (10)$$

2.2.4 협업 필터링의 문제점

추천 시스템 중에서 가장 널리 사용되는 협업 필터링은 초기 평가 문제를 포함한 다음의 몇 가지의 문제점을 가지고 있다.

첫째는 희소성의 문제이다[4, 38]. 희소성은 사용자가 아이템에 부여한 평가의 개수가 부족한 경우에 발생하는 문제이다. 이는 구매활동이 매우 적극적인 사용자들조차도 아이템에 대한 평가는 거의 하지 않기 때문에 발생한다. 사용자의 아이템 선호도가 나타난 데이터 집합에 희소성이 높을 경우 추천의 정확성이 떨어진다.

둘째는 확장성의 문제이다[4, 38]. 확장성은 사용자의 수와 거래 데이터가 증가함에 따라 목표 사용자의 최근접 이웃을 찾기 위한 연산이 기하급수적으로 늘어나는 문제로서 추천 시스템의 데이터 크기가 광범위할 경우 사용자의 유사성 비교에 많은 연산 시간이 필요하게 되고 새로운 사용자가 입력되었을 때 새로운 사용자의 선호도 정보 추가를 위해 전체 데이터를 갱신해야하는 문제로 인해 발생한다. 확장성 문제는 추천 시스템의 효율성을 떨어지게 한다.

셋째는 초기 평가의 문제이다[5]. 초기 평가 문제는 추천 시스템에 처음으로 접속한 초기 사용자는 아이템에 대하여 평가한 정보가 없기 때문에 유사한 사용자를 찾을 수 없으므로 적절한 추천을 제공받을 수 없다는 것이다. 초기 평가 문제는 새로운 사용자의 구매활동을 적극적으로 유도할 수 없게 한다.

협업 필터링 추천 시스템에서 추천의 정확성, 효율성, 신뢰성을 떨어뜨리는 희소성, 확장성, 초기 평가 문제를 해결하기 위해 인구 통계 정보를

이용한 연구[5], 사용자 평가 매트릭스를 이용한 연구[6], 협업 필터링과 내용기반 필터링 기법의 장점을 결합한 하이브리드 연구[5, 7, 8, 9], 장르 기반 연구[10, 63]등 다양한 연구들이 이루어지고 있다.



3. 아이템 특성을 가중치로 이용한 추천 기법 제안

본 장에서는 협업 필터링의 희소성, 확장성, 초기 평가 문제들을 감소시키기 위해 제안하는 아이템 특성을 이용한 추천 기법에 대해 기술한다.

협업 필터링의 문제점들을 해결하기 위해 기존의 사용자 기반이나 아이템 기반 알고리즘에 가중치를 적용한 연구, 하이브리드 연구들이 지속적으로 이루어지고 있다. 이러한 연구들은 전통적인 협업 필터링 기법들보다 희소성, 확장성 문제를 개선하여 추천 성능이 향상되었지만 여전히 초기 평가 문제를 가지고 있다.

본 논문에서는 아이템의 분류 정보와 사용자 정보를 사용하여 아이템의 특성을 추출하고 이 특성을 가중치로 적용하여 희소성과 확장성, 초기 평가 문제를 더 효과적으로 감소시키고자 한다.

3.1 아이템 특성 추출 단계

제안하는 추천 기법은 아이템 특성을 추출하기 위한 데이터 전처리 단계와 아이템 특성 추출 후 각종 계산을 통한 아이템 추천 단계로 이루어진다. 본 절에서는 데이터베이스를 이용하여 아이템의 특성을 추출하는 단계에 대해 설명한다.

3.1.1 장르 특성 추출

아이템 가중치를 사용한 알고리즘들은 대부분 아이템의 분류 정보를 가중치로 사용한다. 분류를 가중치로 사용한 연구들은 협업 필터링의 희소성과 확장성 문제를 개선하여 전통적인 협업 필터링 기법들보다 추천 성능이 향상되었다.

본 논문에서는 아이템의 분류 정보에 사용자 정보를 결합한 아이템 특성을 이용하여 희소성, 확장성뿐 아니라 초기 평가 문제를 개선하는 기법을 제안한다. 이 기법은 장르 속성과 사용자 정보 속성을 같이 사용하여 아이템의 특성으로 지정하기 때문에 새로운 아이템이 입력되었을 때, 그룹화된 장르와 아이템 특성을 사용하면 빠르게 추천이 이루어질 수 있다. 새로운 사용자가 등록되었을 때도 회원가입 시 선택한 선호 장르와 사용자 정보를 사용하여 빠르게 추천이 이루어질 수 있다.

아이템의 특성을 추출하기 위해 아이템의 분류로서 영화의 장르를 사용하고 사용자 정보로서 성별, 직업, 연령 정보를 사용한다. 본 논문에서는 이러한 특성을 추출하기 위해 MovieLens 데이터 집합을 사용한다.

아이템 특성을 추출하기 위해 장르와 사용자 정보를 사용한 2단계의 처리 과정이 필요하다. 첫 단계는 고객 평가 데이터 집합을 장르별로 그룹화하는 것이다. 데이터 집합의 각 아이템 장르는 미국의 최대 영화데이터베이스인 IMDB의 기준에 따라 장르를 알 수 없는 2개의 아이템을 제외하고 Action, Adventure, Animation, Children's, Comedy, Crime, Documentary, Drama, Fantasy, Film-Noir, Horror, Musical, Mystery, Romance, Sci-Fi, Thriller, War, Western, 18개로 분류되었다.

<표 5> 장르별 아이템 수

<Table 5> The number of items by genre

장르	편수	장르	편수	장르	편수
Action	251	Documentary	50	Mystery	61
Adventure	135	Drama	725	Romance	247
Animation	42	Fantasy	22	Sci-Fi	101
Children's	122	Film-Noir	24	Thriller	251
Comedy	505	Horror	92	War	71
Crime	109	Musical	56	Western	27

전체 아이템은 1,860개이고 위의 표 5는 각 장르별 영화의 수를 나타낸 것이다. 전체 아이템 수와 표 5를 보면 알 수 있듯이 각 아이템은 여러 개 (1개~5개)의 장르에 중복적으로 속해있다.

본 논문에서는 장르를 기준으로 데이터를 그룹화하고 장르 정보를 가중치로 사용하기 위해 장르를 대, 중, 소 3개로 분류한다. 분류의 세분화는 불필요한 계산시간을 증가시키고 근접 이웃의 수를 축소시켜 오히려 예측 결과의 정확성을 떨어뜨리는 경우가 많다. 따라서 본 논문에서는 대분류를 기존의 18개의 장르에서 전체 아이템 중 5% 이하의 아이템만을 가지며 독립적 장르로 사용되는 비중이 적은 8개 장르(Animation, Documentary, Fantasy, Film-noir, Musical, Mystery, War, Western)를 Etc로 묶어 11개 장르로 축소한다. 중분류와 소분류는 대분류 속에 포함되는 하위분류로서 중분류는 중복 장르를 2개씩 가진 아이템들을 분류한 것이고 소분류는 중복 장르가 3개 이상인 아이템들을 분류한 것이다.

기존 데이터베이스의 장르를 분류에 맞게 변경한다. 표 6은 데이터베이스에 변경된 장르 분류를 적용한 것이다. 데이터베이스를 수정한 후 아이

템을 대분류를 기준으로 11개의 장르별로 분류한다. 표 7은 표 6의 데이터베이스를 11개의 장르별로 분류한 데이터베이스 중 Action 장르의 데이터베이스를 나타낸 것이다.

<표 6> 장르 분류의 적용

<Table 6> Application of genre classification

item	user	genre	rating	gender	age	job
1	5	CH-CO	4	F	A4	J14
1	6	CH-CO	4	M	A5	J7
1	10	CH-CO	4	M	A6	J10
1	15	CH-CO	1	F	A5	J4
1	17	CH-CO	4	M	A4	J15
1	18	CH-CO	5	F	A4	J14
1	20	CH-CO	3	F	A5	J9
1	21	CH-CO	5	M	A3	J21
1	23	CH-CO	5	F	A4	J2
⋮						
1670	782	CO-TH	3	F	A3	J2
1671	787	DR	1	F	A2	J19
1672	828	DR	2	M	A3	J11
1672	896	DR	2	M	A3	J21
1673	835	AC-TH	3	F	A5	J7
1674	840	DR	4	M	A4	J2
1675	851	DR	3	M	A2	J14
1676	851	DR	2	M	A2	J14
1677	854	DR	3	F	A3	J19
1678	863	DR	1	M	A2	J19
1679	863	RO-TH	3	M	A2	J19
1680	863	DR-RO	2	M	A2	J19
1681	896	CO	3	M	A3	J21
1682	916	DR	3	M	A3	J5

<표 7> Action 장르의 아이템들
 <Table 7> Items of Action genre

item	user	genre	rating	gender	age	job
2	5	AC-AD-TH	3	F	A4	J14
2	13	AC-AD-TH	3	M	A5	J4
2	22	AC-AD-TH	2	M	A3	J21
2	42	AC-AD-TH	5	M	A4	J1
2	49	AC-AD-TH	1	F	A3	J19
2	64	AC-AD-TH	3	M	A4	J4
2	87	AC-AD-TH	4	M	A5	J1
2	92	AC-AD-TH	3	M	A4	J6
2	95	AC-AD-TH	2	M	A4	J1

1610	782	AC-CR	1	F	A3	J2
1613	489	AC-DR	4	M	A6	J14
1615	497	AC-AD-CH	3	M	A3	J19
1615	660	AC-AD-CH	2	M	A3	J19
1615	699	AC-AD-CH	3	M	A5	J14
1615	727	AC-AD-CH	1	M	A3	J19
1615	749	AC-AD-CH	4	M	A4	J14
1615	782	AC-AD-CH	3	F	A3	J2
1615	804	AC-AD-CH	4	M	A4	J4
1615	807	AC-AD-CH	4	F	A5	J8
1615	852	AC-AD-CH	2	M	A5	J1
1618	528	AC-CR	1	M	A2	J19
1646	655	AC-DR	3	F	A6	J8
1646	828	AC-DR	4	M	A3	J11
1657	727	AC-DR	3	M	A3	J19
1673	835	AC-TH	3	F	A5	J7

3.1.2 아이টে에 대한 사용자 정보 특성 추출

아이টে에 대한 사용자 정보를 분석하기 위해 각 장르에서 사용자들이 아이টে에 대해 평가한 값이 4(만족)이상인 데이터들만을 선택한다. 이는 특정 아이টে에 대해 4이상으로 평가한 사용자는 같은 특성을 가진 아이টে을 다시 선택할 확률이 높기 때문이다. 사용자 정보는 회원 가입 시 입력한 성별, 나이, 직업 정보를 사용한다. 나이는 실제 나이를 사용하여 비교하지 않고 비슷한 나이의 사용자들은 비슷한 성향을 가지므로 나이를 7개의 연령대로 분류하여 A1~A7로 나타낸다. 직업은 21개의 직업을 J1~J21로 나타낸다. 표 8은 연령과 직업의 분류를 나타낸다.

<표 8> 연령과 직업의 분류
 <Table 8> The Classification of age and job

연령		직업			
A1	13세 이하	J1	administrator	J12	marketing
A2	14세~19세	J2	artist	J13	none
A3	20세~29세	J3	doctor	J14	other
A4	30세~39세	J4	educator	J15	programmer
A5	40세~49세	J5	engineer	J16	retired
A6	50세~59세	J6	entertainment	J17	salesman
A7	60세 이상	J7	executive	J18	scientist
		J8	healthcare	J19	student
		J9	homemaker	J20	technician
		J10	lawyer	J21	writer
		J11	librarian		

사용자들 중에 특정 성별, 연령, 직업을 가진 사용자가 많을 경우 단순한 수치로 비교해서는 소수의 성별, 연령, 직업 정보는 아이템의 특성으로 선택될 수 없으므로 특정 아이템을 선호하는 성별, 연령별, 직업별 속성 정보를 추출할 때 단순한 수치가 아닌 비율을 사용한다.

<표 9> 장르별 직업별 인원수
 <Table 9> The number of people by job in genre

직업	장 르										
	Ac	Ad	Ch	Co	Cr	Dr	Ho	Ro	Sc	Th	Et
J1	71	64	54	76	66	79	41	78	64	75	32
J2	27	26	17	27	23	28	15	28	23	27	15
J3	7	6	5	7	6	7	4	7	6	6	3
J4	88	73	60	90	81	93	42	88	68	87	50
J5	65	60	50	66	56	67	46	66	59	65	32
J6	17	14	9	17	17	18	11	17	18	18	11
J7	32	32	22	32	27	32	17	31	29	31	8
J8	15	13	11	15	14	16	10	16	13	16	8
J9	7	5	4	6	5	7	3	7	6	7	1
J10	10	9	9	12	8	12	5	12	7	12	7
J11	49	37	29	47	41	49	24	49	32	48	28
J12	23	17	15	24	20	26	11	24	17	22	10
J13	8	8	7	8	7	8	7	8	7	9	3
J14	95	86	75	99	90	105	53	101	75	97	48
J15	65	60	48	65	60	66	40	64	61	64	34
J16	12	10	7	14	14	14	5	14	11	14	7
J17	12	8	4	9	12	12	5	11	6	12	4
J18	28	24	17	30	27	31	10	30	26	30	12
J19	191	173	125	185	167	196	124	191	165	189	89
J20	27	24	19	27	24	27	22	27	25	25	16
J21	42	38	31	43	42	45	28	44	37	42	26
합계	891	787	618	899	807	938	523	913	755	896	444

<표 10> 장르별 성별 인원수

<Table 10> The number of people by gender in genre

장르	성 별		합계
	남	여	
Action	638	253	891
Adventure	578	209	787
Children's	435	183	618
Comedy	642	257	899
Crime	584	223	807
Drama	669	269	938
Horror	396	127	523
Romance	646	267	913
Sci-fi	571	184	755
Thriller	641	255	896
Etc.	328	116	444

<표 11> 장르별 연령별 인원수

<Table 11> The number of people by age in genre

장르	연 령 대							합계
	A1	A2	A3	A4	A5	A6	A7	
Action	8	68	320	227	157	84	57	891
Adventure	7	59	292	204	135	70	20	787
Children's	6	44	222	165	108	57	16	618
Comedy	7	61	322	228	161	91	29	899
Crime	4	56	296	202	141	79	29	807
Drama	8	69	331	240	166	93	31	938
Horror	3	50	215	136	70	38	11	523
Romance	7	66	326	232	160	91	31	913
Sci-fi	7	55	283	196	136	59	19	755
Thriller	7	68	323	228	153	87	30	896
Etc.	1	27	165	121	74	46	10	444

표 9, 표 10, 표 11은 각 장르에 대한 직업별, 성별, 연령별 인원수를 나타낸 것이다.

각 속성의 비율을 계산할 때도 단순한 평균을 사용하면 속성의 요소들 중에서 항상 특정 요소만 선택이 되므로 해당 아이템을 선택한 사용자들의 수의 합과 전체 사용자 수를 사용하여 조화평균으로 구한다. 식 (11)은 사용자 정보 속성의 비율을 구하는 식이다.

$$R_{i,u_{cx}} = \left(2 \times \frac{u_{cx}}{\sum_{j=1}^n u_c} \times \frac{u_{cx}}{\sum_{i=1}^m u_{cx}} \right) // \left(\frac{u_{cx}}{\sum_{j=1}^n u_c} + \frac{u_{cx}}{\sum_{i=1}^m u_{cx}} \right) \quad (11)$$

u_{cx} : 사용자 정보 중 속성에서 개별 요소에 속한 개수

u_c : 사용자 정보 중 속성의 전체 개수

j : 사용자

n : 전체 사용자

i : 아이템

m : 전체 아이템

$R_{i,u_{cx}}$ 는 아이템에 대한 사용자 정보 중 특정 속성의 개별 요소들의 비율을 의미한다. 식 (11)을 사용하여 사용자 정보 속성인 성별, 연령, 직업의 각 요소들의 비율을 계산 한 후 가장 비율이 큰 요소를 해당 속성의 특성으로 선택한다.

예를 들어 Action 장르의 아이템 2의 사용자 정보를 사용하여 속성에서 특성 추출을 위해 성별, 연령별, 직업별 비율을 계산하면 다음과 같다.

Action 장르 중 아이템 2를 4이상의 평가값으로 선택한 사용자의 수는 43명이다. 표 9, 표 10, 표 11을 통해 직업별, 성별, 연령별 인원수를 확인할 수 있다. 표 12는 Action 장르 중 아이템 2를 선택한 사용자들의 정보를 나타낸다. Action 장르 중 아이템 2의 직업별, 성별, 연령별 인원은 표 13, 표 14, 표 15와 같다.

<표 12> Action 장르 중에서 아이템 2
 <Table 12> Item 2 of Action genre

item	user	genre	rating	gender	age	job
2	42	AC-AC-TH	5	M	A4	J1
2	87	AC-AC-TH	4	M	A5	J1
2	130	AC-AC-TH	4	M	A3	J13
2	200	AC-AC-TH	4	M	A5	J15
2	213	AC-AC-TH	4	M	A4	J7
2	250	AC-AC-TH	4	M	A3	J7
2	256	AC-AC-TH	5	F	A4	J13
2	276	AC-AC-TH	4	M	A3	J19
2	279	AC-AC-TH	4	M	A4	J15
⋮						
2	751	AC-AC-TH	4	F	A3	J14
2	796	AC-AC-TH	5	F	A4	J21
2	798	AC-AC-TH	4	F	A5	J21
2	804	AC-AC-TH	4	M	A4	J4
2	807	AC-AC-TH	4	F	A5	J8
2	844	AC-AC-TH	4	M	A3	J5
2	846	AC-AC-TH	5	M	A3	J10
2	864	AC-AC-TH	4	M	A3	J15
2	886	AC-AC-TH	4	M	A3	J19
2	892	AC-AC-TH	4	M	A4	J14
2	934	AC-AC-TH	4	M	A7	J5
2	943	AC-AC-TH	5	M	A3	J19

<표 13> 직업별 인원

<Table 13> The number of people by job

(단위 : 명)

	J1	J2	J3	J4	J5	J6	J7
아이템2	2	0	0	2	4	1	3
	J8	J9	J10	J11	J12	J13	J14
아이템2	1	0	1	0	0	2	5
	J15	J16	J17	J18	J19	J20	J21
아이템2	6	0	0	0	14	0	2

<표 14> 성별 인원

<Table 14> The number of people by gender

(단위 : 명)

	M	F
아이템2	33	10

<표 15> 연령별 인원

<Table 15> The number of people by age

(단위 : 명)

	A1	A2	A3	A4	A5	A6	A7
아이템2	0	3	21	12	6	0	1

Action 장르 중 아이템 2의 각 사용자 정보별 요소들의 비율을 식 (11)을 사용하여 계산하면 표 16, 표 17, 표 18과 같다. 따라서 Action 장르 아이템 2의 아이템 특성은 성별은 M(남자), 연령은 A3(20~29세), 직업은 J19(student)라는 결과를 추출할 수 있다.

<표 16> 장르별 비율
 <Table 16> The ratio of genre

(단위 : 비율)

	J1	J2	J3	J4	J5	J6	J7
아이템2	0.035	0.0	0.0	0.031	0.074	0.033	0.08
	J8	J9	J10	J11	J12	J13	J14
아이템2	0.034	0.0	0.038	0.0	0.0	0.078	0.072
	J15	J16	J17	J18	J19	J20	J21
아이템2	0.111	0.0	0.0	0.0	0.12	0.0	0.047

<표 17> 성별 비율
 <Table 17> The ratio of males to females
 (단위 : 비율)

	M	F
아이템2	0.097	0.068

<표 18> 연령별 비율
 <Table 18> The ratio of age

(단위 : 비율)

	A1	A2	A3	A4	A5	A6	A7
아이템2	0.0	0.054	0.116	0.089	0.06	0.0	0.029

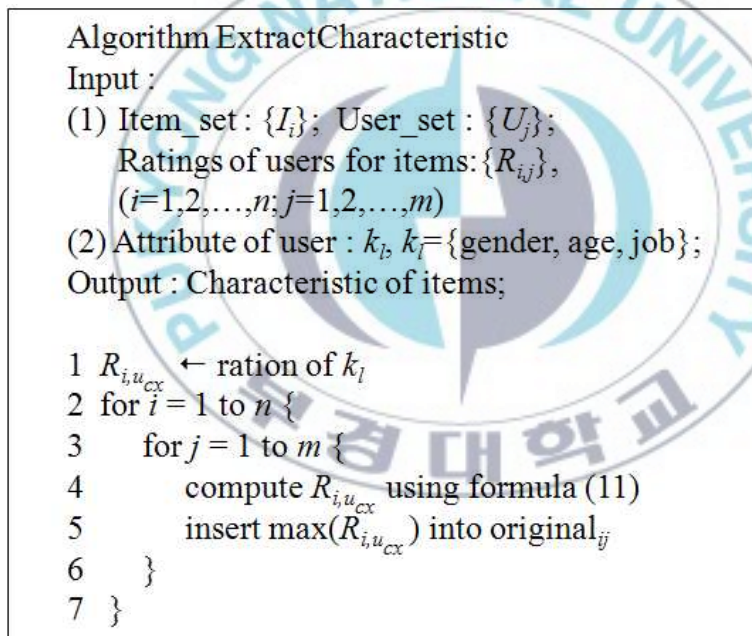
식 (11)을 사용하여 Action 장르의 나머지 아이템들의 사용자 정보 특성을 추출하면 표 19와 같다.

<표 19> Action 장르의 아이템들의 특성
 <Table 19> The characteristic of items of Action genre

item	genre	gender	age	job
2	AC-AD-TH	M	A3	J19
4	AC-CO	M	A3	J19
17	AC-CO	M	A2	J19
21	AC-AD-CO	F	A2	J13
22	AC-DR	M	A3	J19
24	AC-AD-CR	M	A3	J19
27	AC	M	A2	J19
28	AC-DR-TH	M	A3	J19
29	AC-AD-CO	F	A3	J8
33	AC-RO-TH	M	A3	J15
39	AC-CR	M	A3	J6
		⋮		
1303	AC	F	A4	J21
1372	AC	M	A4	J2
1413	AC	F	A2	J15
1415	AC-CH	M	A3	J15
1478	AC-CR-DR	F	A3	J10
1483	AC-DR-RO	F	A3	J8
1484	AC-CO	M	A3	J1
1491	AC-DR-TH	M	A4	J15
1523	AC-AD	M	A5	J20
1597	AC-DR	M	A4	J21
1613	AC-DR	M	A6	J14
1615	AC-AD-CH	F	A4	J8
1646	AC-DR	M	A3	J11

11개의 장르별로 아이템들의 특성을 추출한 후 아이템 특성과 평가값을 사용하여 $\text{아이템} \times \text{사용자}$ 매트릭스를 생성한다.

아이템의 평가값이 4이상인 데이터가 아닌 경우는 장르 특성만을 표시하고 사용자 정보 특성은 비워둔다. 이런 아이템들은 비록 사용자 정보 특성 가중치 값은 지정되지 않지만 같은 장르인 아이템과는 장르 가중치 적용이 가능하다. 그림 4는 장르와 사용자 정보를 사용하여 아이템의 특성을 추출하는 알고리즘이고, 그림 5는 알고리즘에 의해 추출된 특성을 포함하여 새로이 생성된 매트릭스 중 Action 장르의 $\text{아이템} \times \text{사용자}$ 매트릭스를 나타낸 것이다.



```

Algorithm ExtractCharacteristic
Input :
(1) Item_set :  $\{I_i\}$ ; User_set :  $\{U_j\}$ ;
    Ratings of users for items:  $\{R_{ij}\}$ ,
    ( $i=1,2,\dots,n; j=1,2,\dots,m$ )
(2) Attribute of user :  $k_i, k_i = \{\text{gender, age, job}\}$ ;
Output : Characteristic of items;

1  $R_{i,u_{cx}} \leftarrow \text{ration of } k_i$ 
2 for  $i = 1$  to  $n$  {
3   for  $j = 1$  to  $m$  {
4     compute  $R_{i,u_{cx}}$  using formula (11)
5     insert  $\max(R_{i,u_{cx}})$  into original  $i_j$ 
6   }
7 }
    
```

(그림 4) 아이템의 특성 추출 알고리즘
(Figure 4) The algorithm to extract characteristic of items

	genre	gender	age	job	1	2	3	4	5	936	937	938	939	940	941	942	943
2	AC-AD-TH	M	A3	J19					3								5
4	AC-CO	M	A3	J19	3									2			
17	AC-CO	M	A2	J19	3				4								
21	AC-AD-CO	F	A2	J13													
22	AC-DR	M	A3	J19													4
24	AC-AD-CR	M	A3	J19	3				4	4							4
27	AC	M	A2	J19	2												4
28	AC-DR-TH	M	A3	J19	4												4
29	AC-AD-CO	F	A3	J8	1												
33	AC-RO-TH	M	A3	J15	4												
39	AC-CR	M	A3	J6	4												
50	AC-AD-RO	M	A3	J19	5	5		5		4	5			4		5	4
53	AC-TH	M	A3	J19	3												3
54	AC-DR-TH	M	A2	J15	3												4
62	AC-AD-SC	M	A2	J15	3				4								3
68	AC-RO-TH	M	A3	J19													4
73	AC-CO	M	A4	J19	3												3
74	AC-CO	F	A7	J10	1												
79	AC-TH	M	A4	J19	4				3							5	5
80	AC-CO	M	A2	J10	4				2								2
⋮																	
1414	AC																
1415	AC-CH	M	A3	J15													
1416	AC-SC																
1419	AC-SC																
1433	AC-DR																
1478	AC-CR-DR	F	A3	J10													
1483	AC-DR-RO	F	A3	J8													
1484	AC-CO	M	A3	J1													
1491	AC-DR-TH	M	A4	J15													
1523	AC-AD	M	A5	J20													
1548	AC-TH																
1552	AC																
1556	AC-DR-TH																
1586	AC-CR-DR																
1595	AC-TH																
1597	AC-DR	M	A4	J21													
1610	AC-CR																
1613	AC-DR	M	A6	J14													
1615	AC-AD-CH	F	A4	J8													
1618	AC-CR																
1646	AC-DR	M	A3	J11													

(그림 5) 새로운 아이템 × 사용자 매트릭스
 (Figure 5) New item × user matrix

3.2 아이템 추천 단계

이 단계에서는 생성된 장르별 매트릭스들을 기반으로 유사도를 구하고 유사도가 임계값 이상인 아이템들만을 최근접 이웃으로 지정한다. 지정된 최근접 이웃들의 평가값을 바탕으로 아이템의 예측값을 계산하고 상위 N 개의 예측값만을 가지고 평가값이 없는 아이템의 빈 셀을 채운다.

3.2.1 아이템 간 유사도 측정

새로이 생성된 매트릭스에서 아이템간의 유사도는 피어슨 상관관계수에 각 아이템을 선호하는 성별, 연령, 직업과 아이템의 장르를 가중치로 적용하여 계산한다. 식 (12-1)에서 식 (12-4)는 각 속성별 가중치 계산식이다.

$$w_{gen} = \begin{cases} \alpha, & \text{if } i_{a,gen} = i_{b,gen} \\ 0, & \text{otherwise} \end{cases} \quad (12-1), \quad w_{age} = \begin{cases} \beta, & \text{if } i_{a,age} = i_{b,age} \\ 0, & \text{otherwise} \end{cases} \quad (12-2)$$

$$w_{job} = \begin{cases} \gamma, & \text{if } i_{a,job} = i_{b,job} \\ 0, & \text{otherwise} \end{cases} \quad (12-3), \quad w_{grn} = \begin{cases} \lambda, & \text{if } i_{a,grn} = i_{b,grn} \\ \epsilon, & \text{elseif } i_{a,lgrn} = i_{b,lgrn} \\ \vartheta, & \text{elseif } i_{a,sgrn} = i_{b,sgrn} \\ 0, & \text{otherwise} \end{cases} \quad (12-4)$$

$w_{gen}, w_{age}, w_{job}, w_{grn}$ 은 두 아이템 사이에서 각 아이템을 선호하는 성별, 연령, 직업, 장르간의 가중치이다. 장르 가중치는 세 개의 분류를 기준으로 완전히 일치하는 경우 대분류까지만 일치하는 경우, 중분류까지만 일치하는 경우로 나누어 그 값을 다르게 적용한다. 각 속성의 가중치는 $[0, 0.4]$ 사이의 값을 가진다.

식 (13)은 각 속성별 가중치의 합을 구하는 식으로 각 속성의 합인 w_i 가 아이템의 특성 가중치 값이 된다.

$$w_i = 1 + w_{gen} + w_{age} + w_{job} + w_{grn} \quad (13)$$

식 (14)는 아이템 가중치 w_i 와 피어슨 상관계수를 사용하여 유사도를 구하는 식이다.

$$Sim(i, j) = \frac{\sum_{u \in U_{i,j}} w_i^2 (r_{u,i} - \bar{r}_i) \cdot (r_{u,j} - \bar{r}_j)}{\sqrt{\sum_{u \in U_{i,j}} (r_{u,i} - \bar{r}_i)^2} \cdot \sqrt{\sum_{u \in U_{i,j}} w_i^2 (r_{u,j} - \bar{r}_j)^2}} \quad (14)$$

$U_{i,j}$:아이템 i 와 j 가 공통으로 평점을 부여한 사용자의 집합

$r_{u,i}$:사용자 u 가 아이템 i 를 평가한 평점

\bar{r}_i : i 번째 아이템의 전체 평점 평균

표 20은 Action 장르의 목표 아이템 2와 비교 아이템 간의 아이템 특성

가중치를 사용한 유사도 계산 결과이다.

<표 20> 아이템 2와 비교 아이템간의 유사도
 <Table 20> Similarity between item 2 and comparative items

아이템	가중치 유사도	아이템	가중치 유사도	아이템	가중치 유사도
4	0	96	0.48	1181	0
17	0.71	101	0.59	1183	0
21	0.01	110	0	1188	0.66
22	0.61	117	0.76	1215	0.79
24	0.28	118	0.83	1222	0
27	0.47	121	0.95	1228	0.53
28	0.45	127	-0.04	1231	0
29	0.62	128	0.17	1239	0
33	0.43	144	0.43	1244	0
39	0.21	145	0.48	1250	0
50	0.18	147	0.75	1277	0
53	-0.17	148	0.64	1303	0
54	0.68	161	0.99	1314	0
62	0.52	172	0.91	1393	0
68	0.42			1407	0
73	0.65	1089	0	1413	0
74	0	1105	0	1415	1
79	0.75	1110	0.85	1419	0
80	0.32	1139	0	1478	0
82	0.51	1161	1	1483	0
92	-0.23	1180	0	1615	0

3.2.2 아이템 예측값 생성

유사도를 계산한 후 다음의 2가지 조건을 만족하는 아이템들만을 사용하여 예측값을 계산한다. 첫 번째 조건은 목표 아이템과 유사도가 높은 아이템들을 최근접 이웃으로 지정하기 위해 유사도의 값이 지정된 임계값 이상인 아이템들만을 대상으로 한다는 것이다. 두 번째 조건은 유사도는 높지만 목표 아이템과 비교 아이템을 공통으로 평가한 사용자들이 극소수인 경우는 정확하게 두 아이템이 매우 유사하다고 할 수 없으므로 목표 아이템과 비교 아이템을 공통 평가한 사용자들이 해당 장르의 전체 사용자들의 일정 비율 이상인 아이템들만을 최근접 이웃의 대상으로 한다는 것이다.

식 (15)는 식 (14)를 이용하여 계산된 유사도를 기준으로 위의 2가지 조건을 만족한 아이템들만을 사용하여 예측값을 계산하는 식이다.

$$P_{u,i} = \frac{\sum_{j=1}^n Sim_{i,j} \times r_{u,j}}{\sum_{j=1}^n Sim_{i,j}} \quad (15)$$

$r_{u,j}$: 사용자 u 의 아이템 i 의 이웃 아이템 j 에 대한 평가값

$Sim_{i,j}$: 목표 아이템 i 와 이웃 아이템 j 의 유사도

그림 6은 아이템 특성 가중치를 사용하여 유사도를 구하고 임계값 이상의 아이템만을 사용하여 예측값을 구하는 알고리즘이다.

```
Algorithm CreateItemPrediction
Input : item user matrix; similarity threshold :  $s-th$ ;
         $co_{ij}$  threshold :  $c-th$ ;
Output : Predicted Rating Value  $P_{u,i}$ 

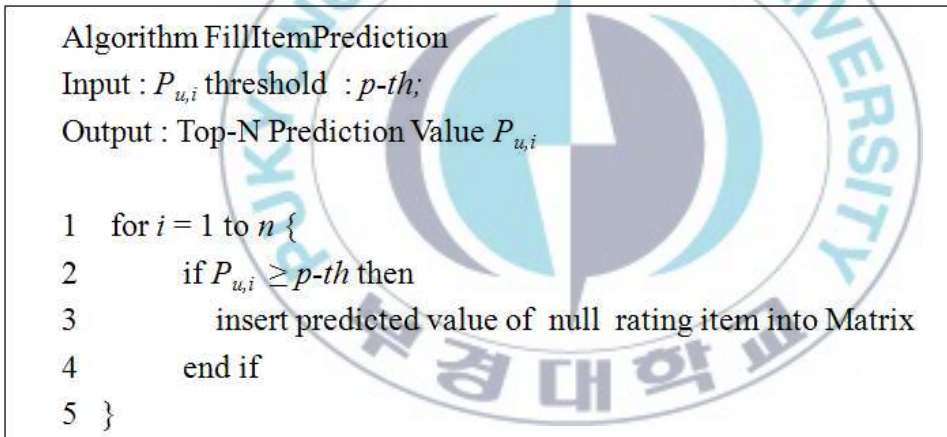
1 for  $i = 1$  to  $n$  {
2   compute the similarity of every two items using formular (14)
3   if  $sim(i,j) \geq s-th$  and  $co_{ij} \geq c-th$  then
4     select neighbors of each items;
5   end if
6   compute the  $P_{u,i}$  using formular (15)
7 }
```

(그림 6) 아이템 유사도 계산과 예측값 계산 알고리즘
(Figure 6) The algorithm to compute of item similarity and predictive value

3.2.3 아이템 추천

11개의 장르별 매트릭스에 모두 유사도를 계산하고 평가되지 않은 아이템의 예측값을 계산한 후 계산된 예측값 중에서 추천의 정확성을 위해 모든 예측값을 사용하지 않고 Top-N 개의 아이템 예측값만을 사용하여 평가되지 않은 아이템들에 대한 예측값으로 지정한다.

그림 7은 계산된 예측값 중에서 Top-N 개의 예측값만으로 평가되지 않은 아이템의 값을 채우는 알고리즘이고, 그림 8은 이 알고리즘을 사용하여 Action 장르 매트릭스의 예측값 중에서 Top-N 개의 예측값을 지정한 결과를 나타낸 것이다.



```
Algorithm FillItemPrediction
Input :  $P_{u,i}$  threshold :  $p-th$ ;
Output : Top-N Prediction Value  $P_{u,i}$ 

1  for  $i = 1$  to  $n$  {
2      if  $P_{u,i} \geq p-th$  then
3          insert predicted value of null rating item into Matrix
4      end if
5  }
```

(그림 7) 아이템 예측값 채우기 알고리즘
(Figure 7) The algorithm to fill predictive value of item

	genre	gender	age	job	1	2	3	4	5	936	937	938	939	940	941	942	943	
2	AC-AD-TH	M	A3	J19								1.2						
4	AC-CO	M	A3	J19														
17	AC-CO	M	A2	J19													2.2	
21	AC-AD-CO	F	A2	J13														
22	AC-DR	M	A3	J19	2													
24	AC-AD-CR	M	A3	J19														
27	AC	M	A2	J19														
28	AC-DR-TH	M	A3	J19														
29	AC-AD-CO	F	A3	J8													1	
33	AC-RO-TH	M	A3	J15					...									
39	AC-CR	M	A3	J6													2.7	
50	AC-AD-RO	M	A3	J19														
53	AC-TH	M	A3	J19														
54	AC-DR-TH	M	A2	J15														
62	AC-AD-SC	M	A2	J15														
68	AC-RO-TH	M	A3	J19							1.4							
73	AC-CO	M	A4	J19														
74	AC-CO	F	A7	J10														
79	AC-TH	M	A4	J19														
80	AC-CO	M	A2	J10														
⋮																		
1414	AC																	
1415	AC-CH	M	A3	J15														
1416	AC-SC																	
1419	AC-SC																	
1433	AC-DR																	
1478	AC-CR-DR	F	A3	J10														
1483	AC-DR-RO	F	A3	J8					...									
1484	AC-CO	M	A3	J1														
1491	AC-DR-TH	M	A4	J15														
1523	AC-AD	M	A5	J20														
1548	AC-TH																	
1552	AC																	
1556	AC-DR-TH																	
1586	AC-CR-DR																	
1595	AC-TH																	
1597	AC-DR	M	A4	J21														
1610	AC-CR																	
1613	AC-DR	M	A6	J14														
1615	AC-AD-CH	F	A4	J8														
1618	AC-CR																	
1646	AC-DR	M	A3	J11														

(그림 8) Action 장르의 상위 N개의 예측값
 (Figure 8) Predictive values of Top N of Action genre

제안하는 기법은 아이템의 장르와 사용자 정보를 이용하여 아이템의 특성을 추출한 후 그 특성을 가중치로 사용하여 목표 아이템과 이웃 아이템 간의 유사도를 계산하고 조건을 만족하는 최근접 이웃의 평가값들을 사용하여 예측값을 생성하는 방법이다.

이러한 방법을 통해 평가값이 없는 비어 있는 셀에 예측값 중 Top-N 개의 값을 채움으로써 희소성의 문제를 줄일 수 있다. 아이템을 장르별 매트릭스로 분류하여 사용하기 때문에 새로운 아이템이 입력되었을 때 해당 장르의 매트릭스에 입력되어 같은 특성을 가진 아이템을 선호하는 사용자에게 추천이 가능하여 초기 평가 문제 또한 줄일 수 있어 추천의 정확성도 높일 수 있다.

사용자에게 아이템을 추천할 때는 예측값을 모두 사용하는 것이 아니라 3점(보통) 이상의 값으로 예측된 아이템들만 사용자에게 추천한다. 그 이유는 사용자가 아이템을 선택할 때는 이웃 사용자들의 상품에 대한 만족도가 최소 보통 이상인 아이템을 추천받았을 경우 선택할 가능성도 높으므로 예측값 중에서 실제 추천이 이루어지는 아이템들은 평가값이 3점 이상인 아이템들이다.

<표 21> 제안하는 기법의 추천 아이템

<Table 21> Recommendation item of suggested method

item	187	313	184
예측값	4.29	4.17	3.96

<표 22> 아이템 기반 기법의 추천 아이템

<Table 22> Recommendation item of item-based method

item	313	187	838
예측값	4.55	3.74	3.32

<표 23> 장르 기반 기법의 추천 아이템

<Table 23> Recommendation item of genre-based method

item	187	250	323
예측값	5	5	3.67

<표 24> 사용자 아이디 7이 선택한 아이템들

<Table 24> Selected items by User ID 7

item	평가값	item	평가값
27	4	380	3
29	3	391	4
62	3	403	3
80	4	405	3
187	4	449	3
201	2	572	3
229	3	578	2
231	3	590	3
234	5	597	4
380	4	665	5

표 21, 표 22, 표 23은 각각 제안하는 기법, 아이템 기반 기법, 장르 기반 기법들을 사용해 예측값을 구한 후 그 중 사용자 아이디 7번에게 실제

추천이 되는 예측값이 3이상인 아이템들 중 상위 3개를 나타낸 것이다.
표 24는 사용자 아이디 7번이 실제 선택한 아이템들과 평가값을 나타낸 것이다.

각 기법에서 추천한 아이템 중 아이템 187이 사용자가 선택한 아이템이다. 제안하는 기법과 장르 기반 기법은 아이템 187을 예측값이 가장 큰 1순위로 추천하고 있고 아이템 기반 기법은 예측값이 2번째로 큰 2순위로 추천하고 있다. 추천을 받은 사용자의 아이템 선택에서도 아이템 187은 높은 순위에 있다.



4. 실험 및 평가

본 장에서는 본 논문에서 제안한 아이템 특성을 이용한 추천 기법과 기존 기법들을 비교하여 제안하는 기법이 더 효율적임을 입증하기 위한 실험 및 평가 방법에 대해 기술하고 그 결과를 분석한다.

4.1 실험 환경

본 논문에서 제안한 아이템 특성을 가중치로 이용한 추천 기법의 목적은 사용자에게 정확한 상품을 추천하고 초기 평가 문제를 해결하는데 있다. 따라서 추천 기법의 성능 측정은 가상의 데이터를 사용하는 것보다 실제 추천 시스템을 통해 일정량 이상으로 수집된 데이터를 사용하는 것이 더 적절하다. 본 논문에서는 미네소타 대학의 GroupLens Research Project의 일환으로 진행된 MovieLens를 통해 수집된 사용자 평가 데이터를 실험 데이터로 사용하였다.

실험에 사용된 MovieLens 데이터는 1997년 9월부터 1998년 4월까지 MovieLens 웹 사이트에서 수집되었다. 이 데이터에는 943명의 사용자가 1,682개의 영화에 대해 1부터 5까지의 점수로 평가한 100,000개의 평가 값이 저장되어 있으며 각 사용자는 최소 20편의 영화에 대해 평가를 하였다[9]. MovieLens 데이터는 u.data, u.item, u.user, u1.base~u5.base,

u1.test~u5.test 등으로 구성된다. u.data는 100,000개의 평가값이 들어 있는 데이터 집합으로 user id, item id, rating, timestamp로 구성된다.

u.item은 영화 아이টে에 대한 정보가 저장되어 있으며 영화는 19개의 장르로 분류되어 있다. u.user는 사용자에게 대한 연령, 성별, 직업, 우편번호 정보를 저장하고 있다. u1.base~u5.base와 u1.test~u5.test는 실험을 위해 u.data를 80%의 훈련 집합(training dataset)과 20%의 테스트 집합(test dataset)으로 나누어 놓은 데이터들이다.

실험은 2GB의 메모리와 Intel Core 2 Duo 3.16GHz, Windows XP의 환경에서 수행하였다. 실험을 위한 제안 기법과 비교 기법들의 프로그램은 Microsoft Visual C++ 6.0으로 구현하였다

4.2 실험 방법

본 논문에서 제안하는 기법의 성능을 평가하기 위해 아이টে 유사도는 피어슨 상관계수에 아이টে 특성을 가중치로 사용하여 측정하였고, 테스트 아이টে의 선호도 예측값은 임계값을 적용하여 생성하였다. 생성된 예측값 중 Top-N 개의 값만을 사용하였다. 실험 데이터는 80%의 훈련 집합과 20%의 테스트 집합으로 나누어 실험을 하였다.

유사도를 계산한 후 그 결과 값이 임계값 이상이며 목표 아이টে와 비교 아이টে를 공통으로 평가한 사용자의 수가 일정 수 이상인 아이টে들만을 최근접 이웃으로 지정하고 이 이웃들의 평가값을 사용하여 예측값을 생성하였다. 실험에서는 최근접 이웃 지정을 위해 유사도 임계값을 0.3에서

0.8까지 0.1씩 변화를 주어 실험하였다.

예측값을 생성한 후 Top-N 개의 값만을 비어있는 셀의 실제 예측값으로 사용하였다. 추천의 성능을 측정하기 위해 N의 개수를 10~50으로 10씩 변화를 주어 실험을 하였다. N의 개수는 일반적으로 일정범위까지는 N의 크기가 커질수록 측정범위가 확장됨으로써 추천 성능이 높게 나타나고 N의 크기가 작아지면 추천 성능이 낮게 나타난다. 따라서 적절한 N의 개수를 구하기 위해 유사도 임계값과 같이 변화를 주어 실험을 하고 이를 통해 적절한 임계값과 N의 개수를 구하였다.

예측의 정확성을 평가하기 위해서 통계적 정확성 측정기법(statistical accuracy metrics)과 의사결정지원 측정기법(decision-support metrics)이 사용된다. 통계적 정확성 측정기법은 사용자가 평가한 값들과 예측된 값들을 비교하여 예측의 정확성을 측정하는 기법이다. 의사결정지원 측정기법은 사용자가 양질의 아이템들을 선택하는 것을 얼마나 잘 지원하는가를 측정하는 기법이다[39].

MAE(Mean Absolute Error)는 통계적 정확성 측정기법으로 아이템에 대한 사용자의 실제 평가값과 추천 시스템의 예측값의 차이에 대한 절대 평균으로 추천의 성능을 평가한다.

정확율, 재현율은 의사결정지원 측정기법으로 정확율은 추천된 아이템이 적합할 확률[74]로서 추천된 Top-N개의 집합에 대한 히트 집합의 비율이다. 재현율은 적합한 아이템이 선택될 확률[74]로서 테스트 집합에 대한 히트 집합의 비율이다. 정확율과 재현율은 정보검색 분야에서 문서 추천 성능을 분석하기 위해서 많이 사용된다.

그러나 정확율과 재현율은 상반되는 경향이 있어 추천목록의 개수가 증

가하면 재현율을 커지지만 정확율은 감소하기 때문에 정확율과 재현율을 결합하여 하나의 측정 기준으로 채택하게 될 경우 극단적인 값에 종속적이게 되어 정확한 추천 성능 분석이 어렵다. 따라서 이러한 문제를 해결하기 위해 F1 척도를 사용한다. F1 척도는 정확율과 재현율의 상충관계를 해소하기 위해 조화평균을 활용하여 성능 측정을 한다.

본 논문에서는 예측 성능 평가를 위해 MAE 기법을 사용하였고, 사용자가 양질의 아이템을 추천받는가를 평가하기 위해 정확율, 재현율, F1 척도를 사용하였다. 마지막으로 추천 기법의 효율성을 평가하기 위해 아이템 변화에 따른 세 기법의 예측값 생성 속도를 측정하여 확장성을 비교하였다.

4.3 성능 평가

성능 평가 실험은 3가지로 이루어졌다. 실험 1에서는 각각의 추천기법에 대하여 예측의 정확성을 측정하기 위해 각 기법의 MAE 값을 비교하였고, 실험 2에서는 아이템 기반 추천 기법, 장르 기반 추천 기법과 본 논문에서 제안하는 아이템 특성을 이용한 추천 기법들의 추천의 적합성을 측정하기 위해 정확율, 재현율, F1 값을 비교하였다. 실험 3에서는 추천 기법의 효율성에 영향을 주는 확장성 문제를 평가하기 위해 각 기법의 아이템 증가에 따른 예측값 생성 속도를 비교하였다.

4.3.1 예측 성능 평가

본 논문에서는 예측 성능 평가를 위해 훈련 데이터의 유사도를 계산한 후 최근접 이웃을 지정할 때 유사도 임계값을 0.3에서 0.8까지 0.1씩 변화를 주어 예측값을 계산하였다. 그리고 각각의 계산에서 얻어진 예측값 중 Top-N 개의 예측값을 각 기법 간의 성능 비교를 위해 사용하였다.

예측 성능은 아이템 기반 기법, 장르 기반 기법, 제안하는 기법의 예측값과 테스트 집합의 평가값을 MAE 기법을 사용하여 비교 측정하였다.

표 25와 그림 9는 각각의 기법들에 대해 유사도 임계값에 변화를 주어 얻은 결과들과 테스트 집합을 비교하여 MAE 값을 측정한 결과를 나타낸 것이다.

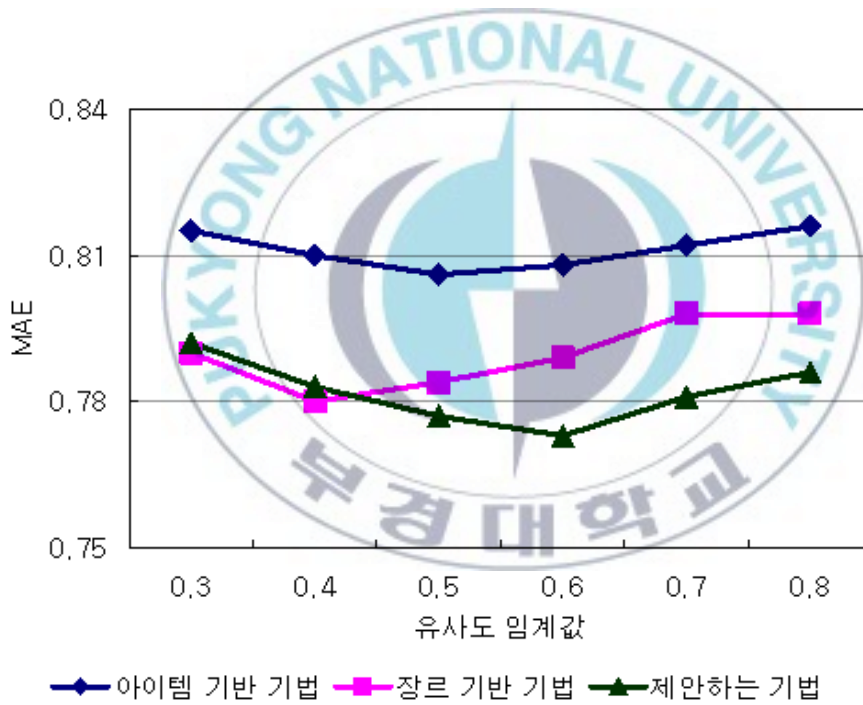
표 25와 그림 9에서 볼 수 있듯이 제안하는 기법은 임계값이 0.3, 0.4 일 때는 장르 기반 기법과 MAE 값에 거의 차이가 없지만 0.5 이상일 때는 제안하는 기법이 비교 기법들보다 MAE 값이 낮아 예측 성능이 우수함을 알 수 있다.

제안하는 기법은 임계값이 0.6일 때 비교 기법들의 MAE 값들과 가장 큰 차이를 보인다. 유사도 임계값이 0.6일 때 제안하는 기법이 아이템 기반 기법 보다 3.6%, 장르 기반 기법보다 1.0% 예측 성능이 향상됨을 알 수 있다. 따라서 제안하는 기법은 0.6을 유사도 임계값으로 지정하는 것이 예측 성능을 가장 많이 향상시킬 수 있다.

<표 25> 유사도 임계값 변화에 따른 예측 성능 비교

<Table 25> The comparison of prediction quality based on changing threshold of similarity

추천기법 임계값	아이템 기반 기법	장르 기반 기법	제안하는 기법
0.3	0.815	0.79	0.792
0.4	0.81	0.78	0.783
0.5	0.806	0.784	0.777
0.6	0.808	0.789	0.773
0.7	0.812	0.798	0.781
0.8	0.816	0.798	0.786



(그림 9) 유사도 임계값 변화에 따른 예측성능 비교

(Figure 9) The comparison of prediction quality based on changing threshold of similarity

4.3.2 추천 성능 평가

본 논문에서는 추천 성능 평가를 위해 훈련 데이터를 사용하여 유사도를 계산하고 예측값을 계산한 후 예측값 결과 중 빈 셀에 채워지는 Top-N의 개수를 10에서 50까지 10씩 변화를 주어 예측값으로 채웠다. 채워진 예측값을 모두 사용하지 않고 예측값이 3이상인 데이터들에 한해 아이템 기반 기법, 장르 기반 기법과 제안하는 기법의 추천의 적합성을 정확율, 재현율, F1 값으로 비교하였다.

예측값이 3이상인 데이터만을 사용한 것은 3.2절에 기술했듯이 사용자들의 아이템에 대한 만족도가 보통 이상은 되어야 추천을 했을 때 선택될 가능성이 높기 때문이다.

<표 26> 아이템 기반 기법의 정확율과 재현율
 <Table 26> Precision and Recall of item-based method
 <단위 : %>

임계값		0.3	0.4	0.5	0.6	0.7	0.8
N값							
10	정확율	32.79	30.29	28.59	26.36	25.09	24.00
	재현율	11.60	10.67	9.89	9.29	8.19	7.26
20	정확율	30.14	28.23	27.60	24.14	23.04	22.92
	재현율	14.10	12.83	11.97	10.76	9.89	8.50
30	정확율	28.45	26.43	25.64	22.17	21.40	20.86
	재현율	17.06	15.35	13.49	12.13	11.07	9.12
40	정확율	26.84	25.32	23.32	21.13	20.07	19.35
	재현율	18.95	16.87	15.38	14.01	11.93	10.28
50	정확율	25.03	23.51	22.03	20.03	18.37	17.93
	재현율	20.12	17.97	16.92	15.26	13.04	11.05

<표 27> 장르 기반 기법의 정확율과 재현율

<Table 27> Precision and Recall of genre-based method

<단위 : %>

임계값 N값		0.3	0.4	0.5	0.6	0.7	0.8
		10	정확율	33.96	30.15	28.76	26.93
	재현율	12.30	11.41	10.87	9.81	9.28	8.75
20	정확율	31.67	28.34	27.67	24.32	23.36	23.03
	재현율	14.55	13.53	13.00	12.20	11.14	10.61
30	정확율	30.23	26.95	25.82	23.05	21.87	20.98
	재현율	17.44	16.31	15.35	14.06	13.00	12.47
40	정확율	28.83	25.19	24.02	21.36	20.27	19.62
	재현율	19.33	18.85	16.79	15.42	13.79	13.00
50	정확율	27.53	23.68	22.35	20.23	18.97	18.21
	재현율	20.87	19.57	18.51	16.97	14.53	14.06

<표 28> 제안하는 기법의 정확율과 재현율

<Table 28> Precision and Recall of proposed method

<단위 : %>

임계값 N값		0.3	0.4	0.5	0.6	0.7	0.8
		10	정확율	38.33	35.17	32.12	30.18
	재현율	14.10	12.89	11.69	10.57	9.46	8.93
20	정확율	36.33	33.00	30.88	28.53	27.30	26.00
	재현율	17.87	15.67	14.84	13.91	12.61	11.65
30	정확율	34.78	31.89	28.85	26.92	26.04	24.21
	재현율	22.18	18.96	17.83	16.98	14.85	13.11
40	정확율	32.73	29.25	27.06	24.57	24.23	22.98
	재현율	23.37	20.97	19.18	18.03	15.21	13.98
50	정확율	31.25	27.89	25.06	23.26	22.04	20.87
	재현율	24.36	22.27	20.96	19.35	15.77	14.58

표 26, 표 27, 표 28은 각각 아이템 기반 기법, 장르 기반 기법, 제안하는 기법의 유사도 임계값과 Top-N의 개수의 변화에 따른 정확율과 재현율의 실험 결과를 나타낸 것이다.

표 26, 표 27, 표 28에서 볼 수 있듯이 모든 기법들은 유사도 임계값이 낮을 때 정확율과 재현율이 높게 나온다. 그 이유는 유사도 임계값이 낮으면 최근접 이웃 아이템들의 수가 많아 예측되는 아이템들의 수도 많기 때문에 정확율, 재현율이 높게 나타난다.

아이템 기반 기법, 장르 기반 기법, 제안하는 기법을 비교하면 제안하는 기법이 비교 기법들보다 정확율과 재현율이 높게 나오는 것을 알 수 있다.

제안하는 기법과 아이템 기반 기법을 비교하면 제안하는 기법이 유사도 임계값 변화와 이웃 사용자 수의 변화에 따른 정확율과 재현율이 모두 높게 나타난다. 이는 제안하는 기법이 목표 사용자에게 추천한 아이템들 중 사용자에게 선택되는 아이템들이 많고 목표 사용자가 만족도를 높게 평가한 아이템들 중에 제안하는 기법이 추천한 아이템들이 많이 포함되어 실제로 사용자들의 아이템 선택을 더 효과적으로 도울 수 있다는 것을 의미한다.

장르 기반 기법과 비교해서는 유사도 임계값 0.6까지는 정확율과 재현율 모두 차이가 있지만 유사도 임계값 0.7 이상이 되면 재현율의 차이가 많이 줄어든다. 이는 두 기법 모두 장르 속성을 사용하기 때문에 유사도 임계값이 높을 경우 추천되는 아이템의 수에 차이가 많지 않기 때문이다.

표 29, 표 30, 표 31은 아이템 기반 기법, 장르 기반 기법, 제안하는 기법의 유사도 임계값과 Top-N의 개수의 변화에 따른 정확율과 재현율을 사용한 F1 값을 나타낸 것이다.

제안하는 기법과 아이템 기반 기법의 F1 측정 결과를 비교하면 정확율

과 재현을 결과와 같이 제안하는 기법의 F1 값이 높아 추천 성능이 향상된 것을 알 수 있다. 장르 기반 기법과의 비교에서도 제안하는 기법의 F1 값이 높음을 알 수 있다.

<표 29> 아이템 기반 기법의 F1
 <Table 29> F1 of item-based method

<단위 : %>

N값 임계값	10	20	30	40	50
0.3	17.14	19.21	21.33	22.22	22.31
0.4	15.78	17.64	19.42	20.25	20.37
0.5	14.70	16.70	17.68	18.54	19.14
0.6	13.74	14.89	15.88	16.85	17.32
0.7	12.35	13.84	14.59	14.96	15.25
0.8	11.15	12.40	12.69	13.43	13.67

<표 30> 장르 기반 기법의 F1
 <Table 30> F1 of genre-based method

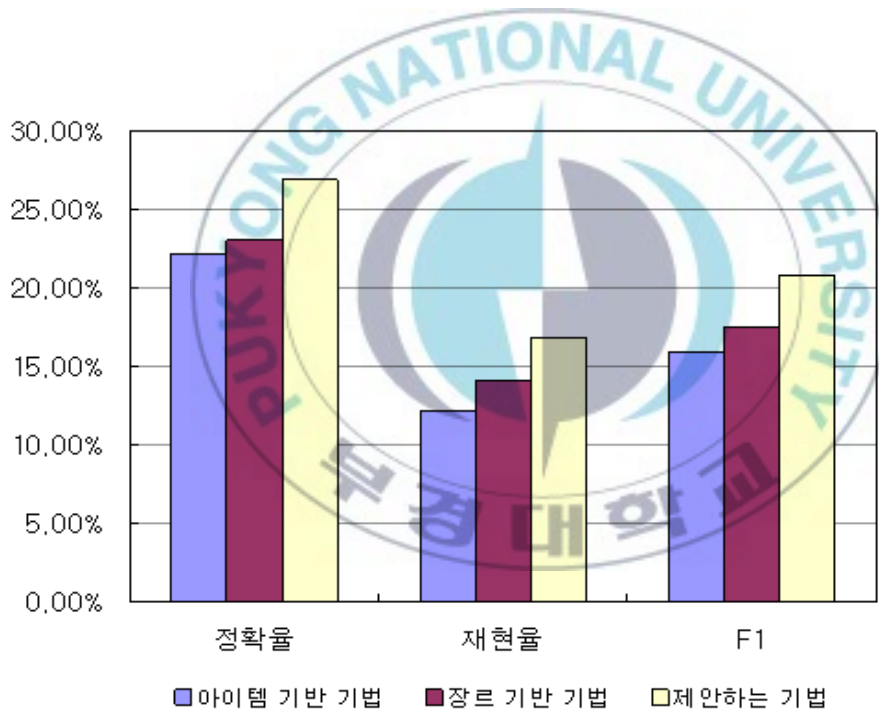
<단위 : %>

N값 임계값	10	20	30	40	50
0.3	18.06	19.94	22.12	23.14	23.74
0.4	16.55	18.32	20.32	21.56	21.43
0.5	15.78	17.69	19.25	19.76	20.25
0.6	14.38	16.25	17.47	17.91	18.46
0.7	13.63	15.09	16.31	16.41	16.46
0.8	12.84	14.53	15.64	15.64	15.87

<표 31> 제안하는 기법의 F1
 <Table 31> F1 of proposed method

<단위 : %>

N값 임계값	10	20	30	40	50
0.3	20.62	23.96	27.09	27.27	27.38
0.4	18.87	21.25	23.78	24.43	24.77
0.5	17.14	20.05	22.04	22.45	22.83
0.6	15.66	18.70	20.82	20.80	21.13
0.7	14.25	17.25	18.91	18.69	18.39
0.8	13.52	16.09	17.01	17.38	17.17



(그림 10) 정확율, 재현율, F1의 비교

(Figure 10) The comparison of a precision, a recall and a F1

그림 10은 유사도 임계값이 0.6이고 Top-N의 개수가 30일 때 아이템 기반 기법, 장르 기반 기법, 기법, 제안하는 기법의 정확율, 재현율, F1값을 비교한 것이다.

정확율, 재현율, F1값을 사용한 추천 성능 평가 결과 제안하는 기법이 아이템 기반 기법과 장르 기반 기법보다 추천 성능이 향상되어 추천의 적합성이 더 높음을 알 수 있다.

4.3.3 추천 효율성 평가

본 논문에서는 추천 기법의 효율성을 평가하기 위해 아이템 수의 확장에 따른 변화를 비교하였다. 각 기법의 아이템 수를 100에서 700까지 100씩 변화를 주고 이에 따른 아이템 기반 기법, 장르 기반 기법, 제안하는 기법의 예측값 생성 속도를 측정하였다. 아이템의 수를 700까지로 제한한 이유는 제안하는 기법이 장르별로 그룹을 나누어 예측값을 구하기 때문에 장르별 아이템 수가 가장 많은 Drama 장르의 아이템으로 비교하기 위해서이다. 실험은 유사도 임계값을 0.6으로 N의 개수를 30으로 지정하여 5회에 걸쳐 실시하였으며 결과는 그 평균으로 나타냈다.

제안하는 기법은 아이템 특성을 추출하기 위한 추가적인 수행시간이 필요하며, 표 32는 아이템 특성을 추출하기 위한 수행시간을 장르별로 측정한 것이다. 표 32에 나타난 것처럼 가장 아이템이 많은 Drama의 아이템 특성 추출 시간이 0.23초 정도 밖에 소요되지 않으므로 아이템 특성 추출 시간이 제안하는 기법의 전체 수행시간에는 큰 영향을 주지는 않는다.

<표 32> 아이템 특성 추출 시간

<Table 32> An extraction time for characteristic of item

<단위 : sec>

장르	아이템 수	소요시간	장르	아이템 수	소요시간
Action	216	0.109	Horror	72	0.016
Adventure	130	0.060	Romance	207	0.100
Children	111	0.034	Sci-fi	68	0.019
Comedy	435	0.159	Thriller	190	0.088
Crime	86	0.036	Etc	64	0.015
Drama	584	0.225			

아이템을 추출한 후 $\text{아이템} \times \text{사용자}$ 매트릭스를 생성하는데 걸리는 시간을 트랜잭션 수에 변화를 주어 비교하였다. 먼저, 제안하는 기법과의 비교를 위해 11개의 장르별 트랜잭션 데이터베이스에서 새로운 매트릭스를 생성하는데 걸리는 시간과 같은 수의 트랜잭션 데이터베이스에서 비교 기법들의 매트릭스를 생성하는데 걸리는 시간을 측정하였다. 다음으로 제안하는 기법은 Drama의 트랜잭션이 19,021개로 가장 많지만 아이템 기반 기법과 장르 기반 기법은 전체 트랜잭션 80,000개를 대상으로 하기 때문에 트랜잭션 수가 Drama보다 많을 경우 매트릭스의 생성 시간도 비교를 위해 측정하였다.

표 33은 각 기법별 매트릭스 생성 시간을 나타낸 것으로 제안하는 기법은 아이템 특성을 추출하는 시간이 포함된 것이다. 표 34는 트랜잭션 수를 20,000개부터 10,000개씩 증가시켜 80,000개까지 변화를 주어 아이템 기반 기법과 장르 기반 기법의 매트릭스 생성 시간을 측정한 것이다.

<표 33> 트랜잭션 변화에 따른 매트릭스 생성 시간 비교
 <Table 33> The comparison of creation time of matrix based on
 changing transaction

<단위 : sec>

추천기법 트랜잭션 수	아이템 기반 기법	장르 기반 기법	제안하는 기법
1100	0.005	0.005	0.011
2700	0.010	0.012	0.021
3800	0.014	0.015	0.026
5900	0.016	0.020	0.031
8000	0.021	0.026	0.047
11000	0.030	0.035	0.068
19000	0.057	0.063	0.156

<표 34> 비교 기법들의 트랜잭션 변화에 따른 매트릭스 생성 시간 비교
 <Table 34> The comparison of creation time of matrix based on changing
 transaction of comparison method

<단위 : sec>

추천기법 트랜잭션 수	아이템 기반 기법	장르 기반 기법
20000	0.063	0.079
30000	0.110	0.123
40000	0.156	0.172
50000	0.234	0.250
60000	0.328	0.344
70000	1.094	1.141
80000	1.917	2.181

표 33에서 제안하는 기법의 매트릭스 생성 시간이 아이템 기반 기법보다 평균 2.2배, 장르 기반 기법보다 1.9배 정도 더 소요되는 것을 볼 수 있

다. 이는 제안하는 기법은 아이템 특성을 추출하기 위한 시간이 더 소요되기 때문이다. 그러나 표 34에 나타난 것처럼 제안하는 기법은 장르별로 11개 그룹으로 나뉘어 있기 때문에 최대 트랜잭션의 수가 Drama의 트랜잭션 수 19,000여개 정도이지만 비교 기법들은 트랜잭션 수가 4배 정도 많아 80,000개의 트랜잭션을 매트릭스로 생성하기 위해 제안하는 기법보다 많은 시간을 소요한다. 11개의 그룹을 모두 매트릭스로 생성하는데 0.55초 정도의 시간이 소요되는데 아이템 기반 기법은 약 3.5배, 장르 기반 기법은 약 4배 정도의 시간이 더 소요된다.

표 35는 제안하는 기법의 각 장르의 예측값 생성 시간을 나타낸 것이다. 표 36과 그림 11은 각각의 기법들에 대해 아이템 수에 변화를 주어 예측값 생성에 소요된 시간을 나타낸 것이다.

<표 35> 장르별 예측값 생성 시간
 <Table 35> A creation time of predictive value by the genre
 <단위 : sec>

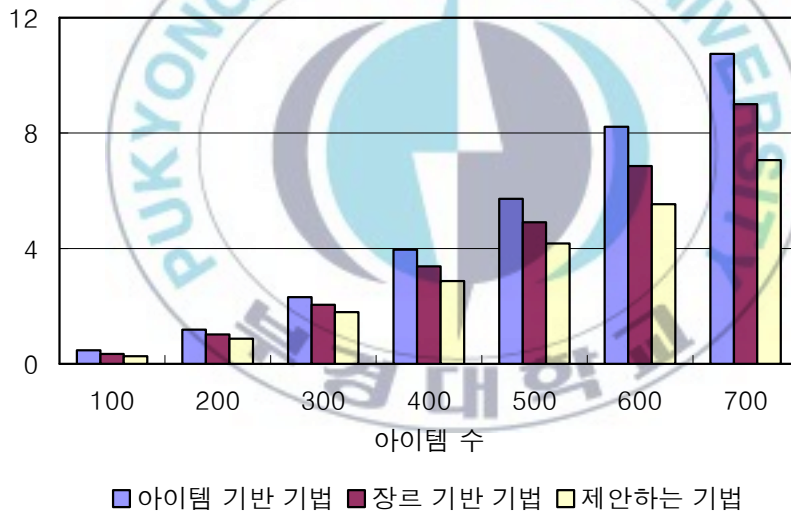
장르	아이템	사용자	생성 시간
Action	249	937	1.375
Adventure	135	882	0.466
Children	122	772	0.316
Comedy	497	939	3.940
Crime	102	898	0.301
Drama	698	943	7.032
Horror	86	714	0.156
Romance	234	935	1.116
Sci-fi	72	863	0.175
Thriller	213	932	0.995
Etc	76	544	0.113

<표 36> 아이템 변화에 따른 예측값 생성 시간 비교

<Table 36> The comparison of creation time of predictive value based on changing item

<단위 : sec>

추천기법 아이템	아이템 기반 기법	장르 기반 기법	제안하는 기법
100	0.469	0.344	0.265
200	1.188	1.015	0.875
300	2.312	2.047	1.797
400	3.953	3.375	2.875
500	5.719	4.907	4.171
600	8.219	6.859	5.531
700	10.75	9.000	7.062



(그림 11) 아이템 변화에 따른 예측값 생성 시간 비교

(Figure 11) The comparison of creation time of predictive value based on changing item

표 36과 그림 11에서 볼 수 있듯이 제안하는 기법과 장르 기반 기법 모두 아이템 기반 기법보다 예측값 생성에 소요되는 시간이 짧다. 이는 두 기법 모두 유사도를 계산할 때 조건을 주어 해당하는 조건에 만족하는 아이템을 최근접 이웃으로 지정하기 때문에 단순한 임계값만을 지정한 아이템 기반 기법보다 계산해야 하는 아이템의 수가 감소하기 때문이다.

제안하는 기법과 장르 기반 기법을 비교해 보면 아이템의 수가 300개 미만일 때는 장르 기반 기법과 제안하는 기법의 예측값 생성시간에 차이가 별로 없지만 아이템의 수가 300개를 넘어가면서 제안하는 기법과 장르 기반 기법 사이의 예측값 생성시간에 차이가 점점 커짐을 볼 수 있다. 아이템 수가 작을 때는 두 기법 모두 예측값 생성시간이 짧고 두 기법 간에 시간 차이가 별로 나지 않는다. 그러나 제안하는 기법은 아이템이 추가되어도 추가된 아이템이 속한 장르 그룹만 계산 작업을 수행하지만 장르 기반 기법은 아이템이 추가될 때마다 전체 데이터를 대상으로 계산 작업을 수행해야 하기 때문에 아이템 수가 많아질수록 속도의 차이가 커진다.

아이템의 수와 함께 사용자의 수 또한 예측값 생성 시간에 영향을 미친다. 이 실험에 사용한 Drama 장르는 사용자가 943명으로 비교 기법들과 사용자 수가 같지만 제안하는 기법은 장르별로 나뉘어져있기 때문에 장르에 따라 사용자의 수에 차이가 있다. 표 35에서 Etc와 Sci-fi 장르를 보면 아이템의 수는 Sci-fi 장르가 작지만 사용자의 수가 많아 예측값 생성에서 더 많은 시간이 소요된 것을 볼 수 있다. 따라서 장르 기반 기법과 아이템 기반 기법은 항상 전체 사용자를 대상으로 예측값을 계산해야 하지만 제안하는 기법은 항상 전체 사용자가 대상이 되지 않으므로 예측값 생성 속도가 비교 기법들보다 단축되어 효율성이 향상된다.

5. 결 론

인터넷의 확산과 다양한 정보기기들의 발달로 각종 정보들이 급격히 증가하였다. 이로 인해 사용자들은 수많은 정보들 속에서 자신이 원하는 정보를 찾기 위해 많은 노력을 해야만 한다.

기업들은 사용자들이 원하는 정보를 빠르게 제공하여 더 많은 고객을 확보하기 위해 추천 시스템을 도입하고 있다. 추천 시스템은 사용자들이 관심 있고 좋아할 만한 아이템이나 서비스를 추천해 주어 원하는 아이템이나 서비스를 쉽고 빠르게 찾을 수 있도록 돕는 시스템으로 협업 필터링이 가장 널리 사용되고 있다.

협업 필터링이 e-commerce에서 성공적으로 널리 사용되고 있지만 희소성, 확장성, 초기 평가 등의 문제점을 가진다. 본 논문에서는 추천의 정확성에 심각한 문제를 발생시키는 데이터의 희소성을 줄이고 초기 평가문제를 줄이기 위해 아이템의 장르 정보와 사용자들의 정보를 가중치로 사용한 기법을 제안하였다.

제안하는 기법은 데이터베이스를 이용하여 아이템을 11개의 장르별로 분류하고 각 장르별로 평가값이 4이상인 아이템들의 사용자 정보를 사용하여 아이템 특성을 추출한 후 11개의 새로운 매트릭스를 생성한다. 11개의 매트릭스별로 아이템 특성을 가중치로 이용하여 유사도와 예측값을 계산한다. 계산된 예측값 중에서 Top-N 개의 예측값만으로 평가되지 않은 아이템들의 빈 셀을 채운다. 예측값을 계산하여 평가값이 없는 빈 셀을 채움으로써 데이터의 희소성을 줄일 수 있다. 사용자에게 아이템을 추천할

때는 정확성을 높이기 위해 채워진 평가 예측값 중 값이 3이상인 아이템만을 추천한다.

실험 평가를 통해 제안하는 기법과 비교 기법들 간의 예측의 정확성, 추천의 적합성, 추천의 효율성을 비교하였다. 예측 성능의 정확성은 MAE 값으로 비교하였으며 아이템 기반 기법, 장르 기반 기법보다 제안하는 기법이 MAE 값이 낮아 예측의 정확성이 높음을 알 수 있었다. 제안하는 기법은 유사도가 0.6 일 때 아이템 기반 기법보다 3.6%, 장르 기반 기법보다 1.0% 예측의 정확성이 향상되었다.

이웃의 크기 변화와 유사도 임계값 변화에 따른 추천의 적합성 비교에서도 제안하는 기법이 아이템 기반 기법, 장르 기반 기법보다 정확율, 재현율, F1값이 높아 추천의 적합성이 높았다.

아이템 수의 변화에 따른 추천의 효율성 비교에 있어서도 제안하는 기법이 아이템 기반 기법, 장르 기반 기법보다 예측값 생성까지의 연산 시간이 단축됨을 알 수 있었다. 제안하는 기법은 유사도가 0.6이고 Top-N의 개수가 30일 때, 예측값 생성을 위한 연산 시간이 아이템 기반 기법보다 약 30%, 장르 기반 기법보다 약 17% 단축되어 확장성 문제를 감소시켜 효율성이 향상되었다.

아이템 기반 기법과 장르 기반 기법은 새로운 아이템이 추가되었을 때 전체 데이터를 대상으로 유사도를 계산하는 작업이 필요하며 새로운 아이템은 평가값이 없어 추천이 이루어지기 힘든 초기화 문제를 가진다.

제안하는 기법은 아이템의 장르 특성과 사용자 정보 특성을 추출한 11개의 장르별 매트릭스를 사용하므로 아이템의 특성을 추출하기 위한 시간이 추가되긴 하지만 그 시간이 짧아 확장성과 초기 평가 문제를 감소시킬

수 있다.

새로운 아이템이 추가되었을 때 아이템의 장르 정보에 해당하는 그룹으로 추가되어 같은 장르 속성을 가진 아이템을 선호하는 사용자에게 빠른 추천이 가능하며, 새로운 사용자가 추가되었을 때도 가입 시 입력한 사용자 정보를 바탕으로 사용자가 선호하는 장르에서 추출된 특성을 기준으로 빠른 추천이 가능하다.

제안하는 기법은 아이템의 특성으로 사용자의 정보를 분석하여 사용하는 기법이기에 때문에 사용자의 수가 너무 작을 경우 아이템 특성으로 사용하기에는 정확성이 떨어진다. 따라서 추천 시스템이 일정 수의 사용자가 확보되지 않은 초기 상태에는 적합하지 않을 수 있다. 그러나 이 문제는 e-commerce의 이용이 보편화되고 있기 때문에 빠르게 해결될 수 있는 문제이다.

향후 연구과제는 스마트 폰, Tablet PC, PDA 등 다양한 종류의 모바일 기기의 보급으로 m-commerce 이용자들이 증가하고 있으므로 모바일 환경에서도 사용자가 원하는 아이템을 보다 빠르고 정확하게 추천할 수 있는 방법을 연구하는 것이다.

[참고문헌]

- [1] T. HengSong, Y. HongWu, "A Collaborative Filtering Recommendation Algorithm Based On Item Classification," *Pacific-Asia Conference on Circuits, Communications and System.*, pp. 694-697, May 2009.
- [2] U. Shardanand, P. Maes, "Social Information Filtering : Algorithms for Automating 'Word of Mouth'," *CHI '95 Proceedings of the SIGCHI Conference on Human factors in computing systems*, pp. 210-217, 1995.
- [3] Manow Papagelisa, Dimitris Plexoosakis, "Qualotative analysis of user-based and item-based prediction algorithms for recommendation agents," *ACM Transactions on Information 22, vol 1*, pp. 116-142, 2004.
- [4] B. Sarwar, G. Karypis, J. Konstan, J. Riedl, "Analysis of Recommendation Algorithms for E-Commerce," *In Processing of the 2nd ACM Conference on Electronic Commerce*, pp. 158-167, Oct 2000.
- [5] M. Pazzani, " A framework for collaborative, content-based and demographic filtering," *Artificial Intelligence Review, Vol. 13, No 5*, pp. 393-408, 1999.
- [6] P. Melville, R. Mooney and R. Nagarajan, "Content-Boosted Collaborative Filtering for Improved Recommendations,"

Proceedings of the eighteenth national Conference on Artificial Intelligence, pp. 187-192, 2002.

- [7] 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*, 1999.
- [8] M. Pazzani, D. Billsus, "Learning and revising user profiles: the identification of interesting Web sites," *Machine Learning, Vol. 27, No.3*, pp. 313-331, 1997.
- [9] M. Balabanovic, Y. Shoham, "Fab: content-base, collaborative recommendation," *Communication of the ACM, Vol. 40, Issue 3*, pp. 66-72, 1997.
- [10] Ye Zhang, Wei Song, "A Collaborative Filtering Recommendation Algorithm Base on Item Genre and Rating Similarity," *International Conference on Computational Intelligence and Natural Computing*, pp. 72-74, 2009.
- [11] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, J. Riedl, "GroupLens: an open architecture for collaborative filtering of netnews," *Proceedings of the ACM Conference on Computer Supported Cooperative Work*, pp. 175-186, 1994.
- [12] D. Goldberg, D. Nichols, B. Oki, D. Terry, "Using collaborative filtering to weave an information tapestry," *In Communications of the ACM, Vol. 35, No. 12*, pp. 61-70, 1992.

- [13] J. Breese, D. Heckerman, C. Kadie, "Empirical Analysis of Predictive Algorithms for Collaborative Filtering," *Proceedings of the 14th Annual Conference on Uncertainty in Artificial Intelligence*, pp. 43-52, 1998.
- [14] J. B. Schafer, J. A. Konstan and J. Reidle, "E-Commerce Recommendation Application," *Data Mining and Knowledge Discovery, Vol. 5, Issue 1-2*, pp. 115-153, 2001.
- [15] J. B. Schafer, J. A. Konstan and J. Reidle, "Recommender systems in e-commerce," *ACM Conference on Electronic Commerce*, pp. 158-166, 1999.
- [16] K. Swearingen, R. Sinha, Interaction design for recommender systems, *In DIS2002, ACM*, 2002.
- [17] P. Renick, H.R. Varian, "Recommender systems," *Communications of the ACM Vol. 40, No 3*, pp. 56-58, 1997.
- [18] R. J. Mooney, L. Roy, "Content-Based Book Recommending Using Learning for Text Categorization," *Proceedings of the fifth ACM conference on ACM 2000 digital libraries*, pp. 195-204, 2000.
- [19] B. J. Pine, *Mass Customization*, Harvard Business School Press. Boston, Massachusetts, 1993.
- [20] R. A. Golding, "Improving Rule-Based System through Case-Based Reasoning," *Proc. of 9th Conference of A.I.*, pp. 22-27, 1991.

- [21] Christopher Reisbeck, Roger Schank, *Inside Case-Based Reasoning*, Psychology Press, 1989.
- [22] P. C. Chang, C. Y. Lai, "A Hybrid system Combining Self-organizing Maps with Case-based Reasoning in Wholesaler's New-release Book Forecasting", *Expert Systems with Applications*, Vol.29, pp. 183-192, 2005.
- [23] R. J. Kuo, Y. P. Kuo and K. Y. Chen, "Developing a Diagnostic of Fuzzy Case-Based Reasoning and Fuzzy Ant Colony System," *Expert System with Applications*, vol. 28, pp. 783-797, 2005
- [24] A. Varman, N. Robby, "ICARUS: Design and Development of a Case-Based Reasoning System for Locomotive Diagnostics," *Engineering Applications of Artificial Intelligence*, Vol. 12, pp. 681-690, 1999.
- [25] H. C. Wang, H. S. Wang, "A Hybrid Expert System for Equipment Failure Analysis," *Expert Systems with Applications*, Vol. 28, pp. 615-622, 2005.
- [26] Mehmet Goker, Thomas Roth-Berghofer, "The Development and Utilization of the Case-Based Help-Desk Support System HOMER," *Engineering Applications of Artificial Intelligence*, Vol. 12, pp. 664-680, 1999.
- [27] S. Wesley Chanchien, Ming-Chin Lin, "Design and Implementation of a Case-based Reasoning System for Marketing Plans," *Expert Systems with Applications*, Vol. 28, pp. 43-53, 2005.

- [28] David leake, Ana Maguitman, Thomas Reichherzer, "Cases, Context, and comfort : opportunities for case-based reasoning in smart homes", *Lecture Notes in Artificial Intelligence, Vol. 4008*, pp. 109-131, 2006.
- [29] A. Aamodt, E. Plaza, "Case-nased Resoning: Fundamental Issues, Methodological Variations, and System Approaches," *Artificial Intelligence Communication, Vol. 7, No. 1*, pp. 39-59, 1994
- [30] J. A. Konstan, B. Miller, D. Maltz, J. L. Herlocker, L. R. Gordon and J. Riedl, "GroupLens: Applying Collaborative Filtering to Usenet News," *Communications of the ACM, Vol. 40, No. 3*, pp. 77-87, 1997
- [31] Y. Z. Wei, L. Moreau and N. R. Jennings, "Learning users' interests by quality classification in market-based recommender systems," *IEEE Trans on Knowledge and Data Engineering, Vol.17, No. 12*, pp. 1678-1688. 2005.
- [32] B. Sarwar, G. Karypis, J. Konstan and J. Riedl, "Item-based Collaborative Filtering Recommendation Algorithms," *Processing of the 10th International World Wide Web Conference*, ACM Press, pp. 285-295, 2001.
- [33] G. R. Lin, Q. Yang, W. Xi, H. J. Zeng, Y. Yu and Z. Chen, "Scalable Collaborative Filtering Using Cluster-based Smoothing," *In Proceedings of the 28th annual international*

ACM SIGIR conference on Research and development in information retrieval, ACM Press, pp. 114-121, 2005.

- [34] T. Hofmann, J. Puzicha, "Latent Class Models for Collaborative Filtering," *In Proceedings of the 16th International Joint Conference on Artificial Intelligence*, pp. 688-693, 1999
- [35] A. Kohrs, B. Merialdo, "Clustering for Collaborative Filtering Application," *In proceedings of CIMCA'99*, IOS Press, 1999.
- [36] Z. Huang, H. Chen, D. Zeng, "Applying associative retrieval techniques to alleviate the sparsity problem in collaborative filtering," *ACM Trans. Inf. Syst.*, 22(1), pp. 116-142, 2004.
- [37] Liang Zhang, Bo Xiao, Jun Guo, Chen Zhu, "A Scalable Collaborative Filtering Algorithm Based on Localized Preference," *Proceedings of the 7th International Conference on machine Learning and Cybernetics, Kunming*, pp. 160-167, July 2008.
- [38] J. L. Herlocker, J. A. Konstan, A. Borchers and J. Riedl, "An Algorithmic Framework for Performing Collaborative Filtering," *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 230-237, 1999.
- [39] T. Hofmann, "Collaborative Filtering via Gaussian Probabilistic Latent Semantic Analysis," *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 259-266, 2003.

- [40] B. Sarwar, *Sparsity Scalability and Distribution in Recommendation Systems*, University of Minnesota, 2001.
- [41] G. Salton, *Automatic Text Processing*. Addison-Wesley, 1989.
- [42] G. Kowalski, *Information Retrieval Systems : Theory and Implementation*, Kluwer Academic Publishers, 1997.
- [43] U. Shardanand, "Social Information Filtering for Music Recommendation," *Learning and Common Sense Group, TR-94-04, MIT Media Laboratory*, 1994.
- [44] B. Sheth, *A learning approach to personalized information filtering*, MIT, 1994.
- [45] B. Buchanan and E. Shortliffe, *Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project*, Addison-Wesley, 1984.
- [46] G. Karypis, "Evaluation of Item-Based Top-N Recommendation Algorithm," *Technical Report CS-TR-00-46*, University of Minnesota, 2000.
- [47] J. Riedl and J. Konstan, *Word of Mouse: The Marketing Power of Collaborative Filtering*, Warner Books, 2002.
- [48] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*, Morgan Kaufmann, 2000
- [49] P. N. Tan, M. Steinbach, V. Kumar, *Introduction to Data Mining*, Addison-Wesley, 2006.
- [50] G. Salton and M. J. McGill, *Introduction to Modern Information*

Retrieval, McGraw-Hill, New York, 1983

- [51] B. Mobasher, X. Jin, Y. Zhou, "Semantically enhanced collaborative filtering on the web," *In web Mining: From Web to Semantic Web, Lecture Notes in Computer Science, volume 3209*, pp. 57-76 , Springer, 2004.
- [52] HengSong Tan, HongWu Ye, "A Collaborative Filtering Recommendation Algorithm Based on Item Classification," *Pacific-Asia Conference on Circuits, Communication and System*, pp. 694-697, 2009.
- [53] Sutheera Puntheeranurak, Hidekazu Tsuji, "A Multi-Clustering Hybrid Recommender System," *7th International Conference on Computer and Information Technology*, pp. 223-228, 2007.
- [54] Y. Dei, H. Ye, S. Gong, "Personalized Recommendation Algorithm Using Demography Information," *Second International Workshop on Knowledge Discovery and Data Mining*, pp. 100-103, 2009.
- [55] F. Wu, L. He, L. Ren and W. Xia, "An Effective Similarity Measure for Collaborative Filtering," *In Proceedings of GrC*, pp. 659-664, 2008.
- [56] David leake, Ana Maguitman, Thomas Reichherzer, "Cases, Context, and comfort : opportunities for case-based reasoning in smart homes", *Lecture Notes in Artificial Intelligence, Vol.4008*, pp. 109-131, 2006.

- [57] R. J. Mooney, P. N. Bennet and L. Roy, "Book Recommending Using Text Categorization with Extracted Information," *Proc. Recommender Systems Papers from 1998 Workshop, Technical Report WS-98-08*, pp. 49-54, 1998.
- [58] O. Bora Fikir, İlker O. Yaz and Tansel Özyer, "A Movie Rating Prediction Algorithm with Collaborative Filtering," *2010 International Conference on Advances in Social Networks Analysis and Mining*, pp. 321-325, 2010.
- [59] Y.-F. Kuo, L.-S. Chen, " Personalization technology application to Internet content provider," *Expert Systems with Applications, Vol 21, Issue 4*, pp. 203-215, 2001.
- [60] C. Stanfill, D. Waltz, "Toward memory-based reasoning", *Communications of the ACM , Volume 29 Issue 12*, pp. 1213-1228, 1986.
- [61] Wu Jian-wei, Yu Jiu-hua, "An Item-based Weighted Collaborative Recommended Algorithm Based on Screening User's Preferences," *2010 International Conference on Computer Application and System Modeling*, pp. 161-164, 2010.
- [62] Xiangwei Mu, Yan Chen and Jinsong Zhang, "Improvement of Collaborative Filtering Algorithm Based on Hesitation Degree," *Computer, Mechatronics, Control and Electronic Engineering*, pp. 302-305, 2010.
- [63] 이재식, 박석두, "장르별 협업필터링을 이용한 영화 추천시스템의 성

- 능 향상,” 한국지능정보시스템학회논문지, 제13권 제4호, pp. 65-77, 2007.
- [64] 이회춘, “전자상거래 추천시스템에서 협력적 필터링의 성능향상”, 강원대학교 컴퓨터과학과, 박사학위논문, pp. 7-11, 2008.
- [65] 유상원, 이홍래, 이형동, 김형주, “TV 프로그램을 위한 내용기반 추천 시스템”, 정보과학회논문지, 컴퓨팅의 실제 제9권 제6호, pp. 683-692, 2003.
- [66] 박진희, 허철희, 정환목, “전자상거래를 위한 규칙 및 사례기반 추론 에이전트”, 한국전자거래학회지, 제8권 제1호, pp. 55-70, 2003.
- [67] 이재필, 이말레, 이헌주, 김기태, “사례기반 추론을 사용한 규칙기반 시스템의 예외 처리,” 정보과학회논문지(B) 제25권 제1호, pp. 132-140, 1998.
- [68] 이재식, 명훈식, “사례기반 추론을 이용한 인터넷 서점의 서적 추천 시스템 개발,” 한국전자상거래학회지 제13권 제4호, pp. 173-190, 2008.
- [69] 김영지, 문현정, 옥수호, 우용태, “사례기반추론기법을 이용한 개인화된 추천시스템 설계 및 구현,” 정보처리학회논문지 D 제9-D권 제6호, pp. 1009-1016, 2002.
- [70] 이재식, 이진천, “유사도 임계치에 근거한 최근접 이웃 집합의 구성”, 한국지능정보시스템학회논문지 제13권 제2호, pp. 1-14, 2007.
- [71] 이건호, 이동훈, “사례기반추론과 규칙기반추론을 이용한 e-쇼핑몰의 상품추천 시스템,” 정보처리학회논문지 D, 제11-D권 제5호, pp. 1189-1196, 2004.

- [72] 홍태호, 이회정, 서보밀, “클러스터링 기반 사례기반추론을 이용한 웹 개인화 추천시스템,” 한국지능정보시스템학회논문지, 제 11권 제1호, pp. 107-121, 2005.
- [73] 이준규, “인터넷 개인화 아이템 추천 알고리즘에 대한 연구,” 연세대학교 응용통계학과, 석사학위논문, pp. 3-14, 2000.
- [74] 고수정, “연관 아이템 트리를 이용한 추천 에이전트,” 정보과학회논문지 : 소프트웨어 및 응용 제 36권 제4호, pp. 298-305, 2009.
- [75] 이용준, 이세훈, 왕창중, “인구 통계 정보를 이용한 협업 여과 추천의 유사도 개선 기법,” 정보과학회논문지 : 컴퓨팅의 실제 제 9권 제 5호, pp. 521-529, 2003.
- [76] 김병만, 이경, “항목 속성과 평가 정보를 이용한 혼합 추천 방법,” 정보과학회논문지 : 소프트웨어 및 응용 제 31권 제 12호, pp. 1672-1683, 2004.
- [77] 고수정, “협력적 여과 시스템에서 귀납 추리를 이용한 순위 결정,” 정보과학회 논문지 : 소프트웨어 및 응용 제 37권 제 9호, pp. 659-721, 2010.
- [78] 이형동, 김형주, “협업 필터링 추천시스템에서의 취향 공간을 이용한 평가 예측 기법,” 정보과학회논문지: 데이터베이스 제 34권, 제 5호, pp. 389-395. 2007.
- [79] 최인복, 이재동, “이웃크기를 이용한 사용자기반과 아이템기반 협업 여과의 결합예측 기법,” 정보과학회논문지B 제16-B권 제1호, pp. 55-62, 2009.
- [80] 고수정, 임기욱, 이정현, “협력적 여과 시스템을 위한 효과적인 사용

자 군집 알고리즘,” 한국정보처리학회 논문지B, 제8-B권 제2호, pp. 144-154, 2001.

[81] 김형일, “협동적 필터링을 위한 데이터 블러링 기법,” 동국대학교 컴퓨터 공학과, 박사학위논문, pp. 59-61, 2003.

[82] 오정민, 문남미, “확장된 협업 필터링을 활용한 선호 요소 가변 추천 시스템,” 전자공학회 논문지 제47권 CI편 제4호, pp. 18-24, 2010.



감사의 글

박사과정을 시작해 본 논문이 완성되기까지 4년 반이란 시간은 힘들면서도 알찬시간이었던 것 같습니다. 박사과정 동안 포기하지 않고 열심히 연구할 수 있도록 도와주신 분들께 이렇게 결과를 내고 감사의 마음을 전할 수 있게 되어 너무도 행복합니다.

학문적, 인성적으로 많은 지도와 가르침을 주시고 이끌어주신 존경하는 윤성대 교수님께 깊이 감사드립니다. 너무도 부족한 제가 논문을 쓸 수 있도록 박사과정 동안 많은 국내외 논문을 읽고 분석하고 발표하는 과정을 통해 기초를 탄탄하게 닦을 수 있게 해주신 점 너무도 감사드립니다. 당시에 너무 힘들었지만 그러한 과정이 없었다면 아마 제 자신과 적당히 타협하고 자기 합리화만 시키며 논문은 쓰지도 못했을 것입니다. 다시 한 번 교수님께 깊이 감사드리며 앞으로도 모든 일에 항상 최선을 다하겠습니다.

바쁘신 와중에도 항상 인자한 웃음으로 본 논문이 좀 더 나은 논문이 될 수 있도록 많은 조언을 주시고 세심히 살펴 주신 김종진 교수님, 논문의 표현까지 세심히 지도해주시며 연구실이 가까우니 언제든 오라시며 살펴 주신 조우현 교수님, 논문의 장점을 보다 잘 부각시킬 수 있도록 지도해주시고 꼼꼼히 살펴 주신 김영봉 교수님 감사드립니다. 허리를 다치셔서 의자에 앉아 계시는 것도 불편하신데 제 논문의 완성을 위해 조언해주신 동서대학교 이재욱 교수님 감사드립니다. 그리고 전자상거래협동과정의 모든 교수님들께 감사드립니다.

지금은 졸업했지만 늦게 까지 같이 논문도 쓰고 수다도 떨었던 후배 김나희, 묵묵히 연구실 잘 챙기고 있는 것만으로도 든든했던 후배 이영석에

게도 감사드립니다. 볼 때마다 늘 걱정해주던 홍진숙 선생님, 연구실의 선배님들 모두 감사드립니다.

박사과정 시작해서 지금까지 항상 늦게 오는 저를 위해 밤잠을 설치시키고 입 짧은 막내딸 아침에 뭐라도 먹고 가게 하려고 일찍 일어나 챙겨주시며 함께 고생하신 부모님 너무 너무 감사드리고 사랑합니다. 그리고 연구실 선배이자 지금까지 나태해지지 않게 항상 격려하고 이끌어준 든직한 우리 오빠. 겁 많은 동생 때문에 황금 같은 주말에도 연구실에 같이 있어주고 논문이 잘 안 풀릴 때마다 함께 고민해주고 짜증 다 받아줘서 너무 고마워. 오빠가 없었으면 논문 쓸 때 엄청 힘들었을 거야. 항상 걱정해 주던 언니들과 형부들, 이제는 이모일도 척척 도와주는 민정이, 주현이, 성현이, 상현이도 모두 고마워.

멀리 있어 자주 보지는 못해도 공부하는 친구를 항상 격려해주던 은주, 친구들 모임에 번번히 빠져도 항상 챙겨주던 종욱이, 종성이, 석용이도 고마워. 같은 과는 아니지만 박사과정의 힘든 점을 같이 얘기하며 용기주신 전섭태 대표님과 볼 때 마다 항상 응원해준 석사 동기들도 너무 감사드립니다.

마지막으로 본 논문이 완성될 수 있도록 응원해주시고 애써 주신 모든 분들께 다시 한 번 감사드립니다.

2011년 6월 21일 연구실에서