韓國컴퓨터情報學會 論文誌
第16卷 第4號, 2011. 04.

2011-16-4-3-8

장르유사도와 선호장르를 이용한 협업필터링 설계

김 경 록*, 변 재 희*, 문 남 미*

Collaborative Filtering Design Using Genre Similarity and Preffered Genre

Kyung-Rog Kim*, Jaehee Byeon*, Nammee Moon*

요 약

전자상거래와 소셜미디어 서비스의 활성화에 따라, 집단지성을 개인 맞춤 서비스에 활용하는 추천시스템에 관한 연구가 활발히 진행되고 있다. 또한, 스마트폰의 발달과 모바일 환경의 발달에 따라 단말의 제약성에도 불구하고 개인화 서비스에 대한 연구가 가속화되고 있다. 대표적인 예로 위치기반 서비스와의 결합이다. 이에 본 연구에서는 영화의 장르유사도와 선호장르를 이용한 추천시스템을 제안한다. 영화 장르 유사도 프로파일을 생성하여 이를 모바일 실험 환경에서 서비스 될 수 있도록 설계하고 프로토 타이핑 한 후에 MovieLens 데이터를 적용하여 평가한다.

▸ 키워드 : 추천 시스템, 협업 필터링, 클러스터링, 다중 속성, 인터랙티브 서비스, 스마트폰, LBS

Abstract

As e-commerce and social media service evolves, studies on recommender systems advance, especially concerning the application of collective intelligence to personalized custom service. With the development of smartphones and mobile environment, studies on customized service are accelerated despite physical limitations of mobile devices. A typical example is combined with location-based services. In this study, we propose a recommender system using movie genre similarity and preferred genres. A profile of movie genre similarity is generated and designed to provide related service in mobile experimental environment before prototyping and testing with data from MovieLens.

▸ Keyword : recommender system, collaborative filtering, clustering, multi-property, interactive service, smart phone, LBS

• 제1저자 : 김경록 • 교신저자 : 문남미
• 투고일 : 2010-06-23, 심사일 : 2010-11-16, 게재확정일 : 2011-02-18
* 호서대학교벤처전문대학원 IT응용기술학과(Dept. of IT Application Tech., Hoseo Graduate School of Venture)
※ 본 연구는 한국콘텐츠진흥원의 2010년도 문화콘텐츠산업기술지원사업의 "웹 미디어의 IPTV 서비스 활용을 위한 콘텐츠 동적 결합과 콘텐츠 생성 기술" 과제 연구 결과로 수행된 결과입니다

Ⅰ. 서 론

Web2.0 서비스가 늘어나면서 개인맞춤서비스에 대한 관심과 중요성이 날로 증가하고 있다. 개인맞춤서비스는 고객이 필요로 하는 정보를 명시적으로 묻지 않고 제공하는 것이다[1]. 특히, Web2.0, UCC(User Create Contents) 등은 집단지성 특성을 이용하여 사용자들의 다양한 의견과 지식을 생성, 공유, 소비하는 지적 흐름을 나타내고 있다[2]. 예를 들어, 집단지성을 이루는 폭소노미(folksonomy)는 사람들이라는 "Folks"와 분류법이라는 "Taxonomy"의 합성어로 모두가 참여해서 분류한다는 의미를 지니고 있다[3]. 이는 날로 증가하는 다양한 정보의 생산과 공유를 위해 대중의 직접적인 참여를 의미하며, Social Network Service의 근간이 된다. 대표적인 예로는 사용자의 평가치(User Rating)가 있다.

더불어 최근 iPhone, Android, Blackberry 등 스마트폰의 발전에 따라 모바일 콘텐츠의 이용률이 급격히 증가하고 있다. 특히, 위치 기반 서비스(Location-based service, LBS)와의 결합 서비스가 증가하고 있다. 이는 이동 통신 단말기의 위치정보를 바탕으로 스마트폰 사용자에게 여러 가지 위치 관련 정보도 함께 제공하는 서비스이다. 이에 따라 보다 더 편리하게 이를 이용할 수 있도록 모바일이 지니고 있는 개인 정보를 바탕으로 개인화된 맞춤 정보 제공을 위한 추천시스템에 대한 요구가 증가하고 있다. 또한, 활용 분야도 다양하게 증가하고 있다[4-7].

이에 본 연구에서는 이러한 정보를 활용하여 개인맞춤서비스를 위한 추천시스템에 관한 연구를 진행하고자 한다.

Ⅱ. 관련 연구

1. Mobile2.0

무선인터넷 및 모바일 환경의 발달에 따라 iPhone, Android 등 스마트폰을 활용한 다양한 어플리케이션들이 날로 증가하고 있다[8]. 특히 풀 브라우징기술을 이용하여 다양한 콘텐츠를 Web 수준으로 소비할 수 있는 개인화 미디어로 빠르게 진화하고 있으며 개인이 항상 소지하면서 정보를 접하는 수단으로 사용할 뿐만 아니라, 정보 소비에 대한 자료를 서비스 제공자에게 제공함으로써 서비스 제공자는 수집된 정보를 이용하여 새로운 정보 추천에 유용하게 이용 할 수 있다[9]. 대표적인 서비스로는 위치기반 서비스인 googlemaps, 소셜네트워크 서비스인 twitter, facebook 등을 들 수 있다. 이는 무선 네트워크 기술의 발전과 모바일 단말의 성능 개선, 그리고 모바일 기기로의 컨버전스가 가속화 되고 있다는 것을 보여 주는 것이다. 이들의 큰 장점은 다수의 사용자 간의 활발한 상호작용이 가능하여 창조적 문화활동 및 사회적 네트워크를 쉽게 형성할 수 있다는 점이다. 즉, 모바일 방송 서비스, 위치서비스, 동영상 공유 서비스, 소셜미디어 서비스, 모바일 교수-학습 서비스 등 다양한 어플리케이션 이용과 사회적 상호작용의 확대에 있다[6].

좀 더 구체적으로 모바일 서비스의 특성을 살펴보면, 편재성(Ubiquity)은 실시간 정보를 받아볼 수 있는 속성, 도달성(Reachability) 언제 어디서나 접속할 수 있는 속성, 편의성(Convenience) 간소화된 통신 도구의 속성, 휴대성은 항상 지니고 다니면서 언제 어디서나 정보의 소비가 가능한 속성, 지역기반(Location)은 실시간으로 고객의 위치정보를 보여주는 속성, 즉시 연결성(Instant Connectivity)은 필요한 정보에 바로 접근이 가능한 속성, 개인화(Personalization)는 각 개인의 특성에 맞는 정보제공 및 서비스 차별화가 가능하다는 의미의 특성을 지닌다[8]. 이러한 서비스들의 특성에도 불구하고, 모바일 단말이 갖는 화면크기 제한, 불편한 입력 방식 등의 단점으로 인하여 이를 극복하면서 고객이 최소의 시간 투자로 원하는 정보를 쉽게 얻을 수 있도록 해 주기 위한 대안이 필요하다. 즉, 모바일 단말의 특징을 고려하고 서비스 특성을 살려 고객의 선호도를 간접적으로 파악하고 이를 활용하여 추천하는 추천시스템이 필요하다.

2. 군집화(Clustering)

군집화는 데이터의 집합을 분류하는 방법으로, 분할적 군집화(Partitional Clustering)와 계층적 군집화(Hierarchical Clustering)로 나눌 수 있다. 분할적 군집화는 주어진 목적함수를 최적화하기 위해 데이터 집합을 K개의 군집으로 나누는 것으로 K-평균(K-Means)이 대표적이다. 또한, 계층적 군집화는 가장 유사한 두 개체들을 선택하여 계속 병합해가는 병합적 계층 군집화가 대표적이다[10-11]. 즉, 데이터집합이 다중속성을 가지는 경우 내부에서 구조와 개별 그룹을 발견하는데 이용한다. 예를 들어 본 연구에서 사용하는 영화 데이터의 경우는 장르 속성에 따라 개별 그룹을 형성할 수 있으며, 분할적 방법을 이용하여 그룹화 한다.

3. 추천시스템(Recommender System)

추천시스템은 사용자의 과거 선호(Preference)도에 기반하여 예측한 후 새로운 아이템에 대한 평가치를 예측하여 제

공하는 시스템이다[12-13]. 즉, 전자상거래의 경우는, 고객 요구에 가장 적합한 상품을 추천하는 것으로 고객의 신상정보를 바탕으로 판매자가 소비자의 소비 패턴, 행동 패턴을 실시간으로 분석한 후 추천아이템을 생성, 전달하여 소비자의 구매 의사결정에 도움을 주는 방법이다[14].

추천방식은 크게 내용기반(Content-based), 협업기반(Collaborative), 혼합(Hybrid) 방식으로 나눌 수 있으며, 아마존닷컴과 Netflix, 그룹렌즈 등이 협업기반 방식을 성공적으로 상용화한 대표적인 예이다.

내용기반(Contents-based)은 소비한 아이템들에 대한 평가를 종합하여, 좋은 평가를 받은 아이템과 비슷한 특성을 보이는 새로운 아이템을 추천해주는 방식이다. 이 방식은 아이템의 명시적 특징만을 이용하여 추천하기 때문에 취향 및 선호도 등에 대한 소비자의 욕구를 만족시키기 어렵다. 또한, 소비자의 경험을 이용해야 하기 때문에 아이템에 대한 소비자의 경험이 없다면 아이템을 추천하지 못하는 문제점도 있다[15].

협업기반(Collaborative-based)은 집단지성을 활용한 것으로 고객들의 프로파일 정보를 바탕으로 아이템에 대한 목표고객의 평가치와 유사그룹 고객의 평가치를 바탕으로 목표고객이 선호할만한 아이템을 추천하는 기법이다. 주로 도서, 영화 추천 서비스 등에서 명시적 정보만을 가지고 분석 추천하기 힘든 아이템에 대해 좋은 추천 성능을 나타낸다[16-18].

Hybrid Method는 내용기반(Contents-based)과 협업기반(Collaborative-based)방식 혹은 아이템 기반(Item-Based)과 사용자 기반(User-Based) 결합 방식이 있다. 아이템기반(Item-based)은 각 항목 별로 가장 유사한 항목들을 미리 계산하여 사용자가 평가한 상위 항목들을 보고 유사한 항목들의 가중치 목록을 생성하여 추천하는 방식이고, 사용자기반(User-based)은 모든 사용자들의 평가치 데이터를 가지고 사용자간 유사도를 계산하여 추천하는 방식이다[15,19-21]. 아이템 기반은 데이터세트가 희박하거나, 큰 데이터세트인 경우에 유리하나 항목유사도 테이블을 유지해야 하는 부담이 있으며, 사용자기반은 대용량 데이터세트에서는 사용자간의 중첩이 발생할 경우가 흔치 않기 때문에 선호도 예측 능력이 감소하는 경향이 있다[19,21]. 이러한 각 방법의 특성을 정리하면 다음 표와 같다.

표 1. 추천 방식 비교
Table 1. Comparison of Recommendation Method

추천 접근 방식	아이디어와 관련 기술	한계성 및 주요이슈
내용 기반 방식	(고객의 과거 선호 아이템)의 유사 속성을 이용함 - User profiling - Clustering	새로운 사용자 문제 확장성 문제 성능 문제
협업 방식	집단지성을 바탕으로 유사 사용자 그룹의 의견을 이용함 - Nearest neighbor - Clustering	희박성 문제 (평가치가 없는 경우)새로운 아이템에 대한 문제 확장성 문제
혼합 방식	내용 기반과 협업기반 방식을 혼합 활용하거나, 사용자 기반 협업 방법과 내용기반 협업 방법을 혼합 활용함	

위에서 제시한 추천 방식 중, 협업기반(Collaborative) 방식은 다양한 비즈니스에서 가장 많이 적용되고 있으나, 근본적으로 입력 데이터집합의 희박성 문제, 신규 아이템에 추천 문제, 그리고 확장성 문제 등이 있다.

- 입력 데이터의 희박성(Sparsity) 문제 : 고객의 평가치 정보(선호도 데이터)를 많이 확보할수록 추천의 정확도를 높일 수 있으나, 고객이 직접 평가에 참여하지 않거나, 구매 정보 분석을 통하여 선호 데이터를 얻을 수 없는 경우 고객-아이템 행렬은 희박할 수 밖에 없으며, 이는 유사집단을 탐색하는 과정에서 아주 적은 선호도 데이터를 사용하므로 고객들 간의 유사도 측정시 신뢰도가 떨어지게 된다[19,22-23].
- 신규 아이템 추천 문제 : 협업필터링은 아이템에 대한 고객의 선호도를 기반으로 추천하므로, 신규 아이템의 경우는 어느 고객도 평가하지 않아 그 상품의 선호도를 알 수 없기 때문에 누군가가 선호도를 입력하거나 구매하기 전에는 그 상품을 추천할 수 없다[19,23-25].
- 확장성(Scalability) 문제 : 빠르게 늘어나는 고객과 아이템을 처리할 수 있는 확장성을 가져야 한다. 즉, 현재 서비스를 이용하고 있는 다수의 고객들에게 추천을 제공하는 동시에 그들의 선호도를 실시간으로 학습하여 유사 선호도 집단을 빠르게 탐색할 수 있어야 한다[22-23].

4. 성능측정(Performance Measurement)

추천시스템의 예측 성능을 측정하는 지표로는 MAE (Mean Absolute Error), MAPE(Mean Absolute Precent Error), RMSE(Root Mean Squared Error)가 보편적으로 이용된다. 이중 본 연구에서는 가장 많이 사용되는 MAE를 이용하여 고객이 실제 부여한 평가치(p_i)와 추천 알고리즘에 의해 예측된 평가치(r_i)의 차이로 성능을 측정한다[23-24].

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|p_i - r_i| \quad \cdots\cdots (1)$$

MAE 값이 작을수록 추천시스템의 예측 정확도가 높음을

의미한다.

기존 장르유사도를 이용한 아이템 기반 협업 필터링 방법은 아이템의 장르들을 바탕으로 이웃 아이템 후보를 선택한 후, 근접 이웃 그룹과 평점을 이용하여 목표아이템과 후보 아이템간의 유사도를 계산하여 추천하는 방법으로 그룹의 크기를 바탕으로 추천관계를 알아보고자 한 것이다.[25] 이를 개선하기 위해 본 연구에서는 장르유사도와 선호장르를 이용한 협업필터링(Collaborative Filtering Using Genre Similarity and User Preferred Genre : CF_GS_UP) 방법을 제안한다.

III. 장르유사도와 선호장르를 이용한 협업필터링 설계

1. 개요

본 연구에서 제안한 추천시스템은 고객의 과거 평가치 정보를 바탕으로 선호 영화 장르와 영화 장르 유사도를 도출하여 프로파일을 생성하고, 목표고객의 인구통계학적 정보인 연령, 성별, 직업을 바탕으로 유사고객 그룹을 형성한다.

목표고객과 유사고객이 공통으로 평가한 영화의 평가치를 바탕으로 피어슨 상관계수를 이용하여 최근접 유사그룹을 도출한다. 이를 바탕으로 추천 목록을 도출하여 모바일 환경에서 위치정보와 함께 개인화된 추천 정보를 제공한다.

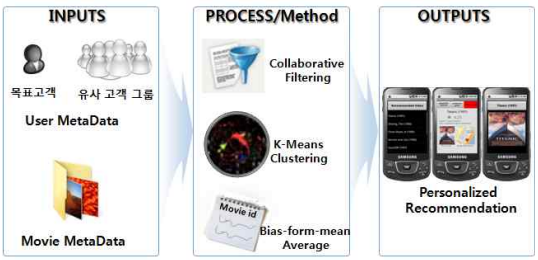

그림 1. 장르유사도와 선호장르를 이용한 추천 개요
Fig. 1. Overview of Recommended by genre similarity and genre preference

2. 필터링 과정

장르유사도와 선호장르에 기반하여 평가치 예측을 위한 필터링 과정은 아래 그림과 같으며 이를 세부적으로 살펴보면 다음과 같다.

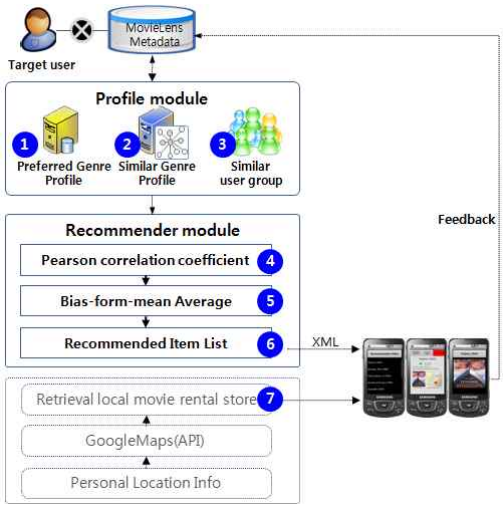

그림 2. 평가치 예측을 위한 필터링 과정
Fig. 2. Filtering process for the prediction rating

① 먼저, 기본 데이터 준비를 위하여, 고객이 평가한 아이템을 바탕으로 아이템별 장르 속성을 도출한 후 각 장르별 평점 평균에 따른 우선순위로, 고객 선호장르 프로파일을 <User, Item, Rating, user_ preference_genre1, user_preference_ genre2, user_preference_genre3>와 같은 Tuple형태로 생성한다.

② 고객의 이용정보 데이터를 바탕으로 아이템 장르 속성 데이터를 수집 분석한다. 아이템 장르 속성 사이 유사도 프로파일은, 아이템 속성들에 대한 피어슨 상관계수를 구한 후, 이들 관계를 파악하기 위해 K-Means Clustering으로 유사그룹을 도출한다. 그룹내 관계를 피어슨 상관계수로 결정한다. 즉, 장르유사도 프로파일은 <Item, Title, Genre, Similarity_Genre>와 같은 Tuple형태로 생성한다.

③ 이제, 목표고객의 인구통계학적 속성(성별, 연령대, 직업)을 바탕으로 전체 사용자 중에서 속성 값이 모두 같은 사용자에 대해 유사그룹을 도출한다.

④ 이제, 목표고객(u_a)의 선호장르(①)와 아이템 속성(장르) 유사도(②)를 바탕으로 새로운 선호장르를 도출한 후, 인구통계학적 유사그룹(③)에 대해, 목표고객(u_a)와 유사그룹내 고객(u_i)간의 유사도$P(u_a, u_i)$를 널리 사용되는 피어슨 상관계수(Pearson correlation coefficient)를 이용하여 계산한다[24,27].

$$P(u_a,u_i)=\frac{\sum_{j=1}^{m}(r_{a,j}-\overline{r_a})(r_{i,j}-\overline{r_i})}{\sqrt{\sum_{j=1}^{m}(r_{a,j}-\overline{r_a})^2}\sqrt{\sum_{j=1}^{m}(r_{i,j}-\overline{r_i})^2}} \quad \cdots\cdots (2)$$

(단, $r_{a,j}$: 아이템(j)에 대한 U_a의 평가치, $\overline{r_a}, \overline{r_i}$: U_a, U_i평가치 평균)

피어슨 상관계수 값이 1에 가까울수록 유사도가 높으며, -1에 가까울수록 선호도 차이가 크며, 0에 가까운 경우는 선호도간 관계가 성립하지 않는다.

⑤ 다음으로는, 최 근접 유사그룹 내 유사도와 이들의 평가치를 이용하여 목표고객이 평가하지 않은 아이템의 평가치를 예측한 후, 각 고객의 평점의 평균과 유사도를 가중치로 적용하여 목표고객(u_a)의 아이템(i)에 대한 평가 예측치($P_{a,i}$)를 구한다[24].

$$P_{a,i}=\overline{r_a}+\frac{\sum_{u=1}^{n}(r_{u,i}-\overline{r_u})\times P(u_a,u_u)}{\sum_{u=1}^{n}P(u_a,u_u)} \quad \cdots\cdots (3)$$

⑥ 마지막으로, Top-N기법으로 상위 10개 추천 목록을 생성하여 서비스 할 수 있도록 한다[27-28].

⑦ 다른 한편, 개인의 위치정보를 바탕으로 구글 지도 API를 연계하여 가까운 영화대여점 위치를 찾아 볼 수 있도록 연계 서비스한다.

IV. 구현 및 실험

1. 실험 데이터세트 구성

본 연구에서 데이터집합은 미네소타대학의 GroupLens Research Project에서 수집된 MovieLens Data Set로 943명의 사용자들이, 18개 장르, 1682개 영화에 대해 100,000건의 평점을 매긴 자료를 바탕으로 한다. 각 영화는 미국 영화 데이터베이스인 IMDB (Internet Movie Database, http://www.us.imdb.com)의 장르 기준에 따른 것으로 최소 1개 장르에서 최대 5개 장르까지 속한다[27-28].

주어진 데이터집합은 예상 평가치 1,586,126(943*1682)건의 6.3%에 불과하며, 희박성은 93.7%이다.

이제 본 연구에서 제안한 장르유사도와 선호장르를 이용하는 추천 방식을 위하여 기본 데이터집합을 바탕으로, 고객 선호장르 프로파일(User Preference Profile)과 장르간 유사도 프로파일(Genre Similarity Profile)을 생성하고, 우편번호코드를 재구성하여 지역 위치정보 코드를 생성한다. 이는 협업 추천 방식의 확장성 문제에 대체하고 모바일 실험 환경에서 활용하기 위한 것이다.

먼저, 고객 선호장르 프로파일을 생성한다. 각 고객(u_a)이 본 영화와 평점(UM_{Ri})을 바탕으로, 아래 식을 이용하여 장르 평점 평균을 구한다.

$$\text{장르}(G_j)\text{평점평균}=\frac{\sum_{i=1}^{5}(\text{평점}_i\times\text{평점}i\text{의 아이템(영화) 개수})}{\text{본 아이템(영화) 개수}} \quad \cdots\cdots (4)$$

이를 바탕으로 상위 3개 장르를 고객 선호장르 프로파일로 생성한다.

다음으로, 장르 유사도 프로파일을 생성한다. 이는 하나의 영화가 동시에 여러 장르의 특성을 가질 수도 있기 때문이다. 예를 들어, 스타워즈(Star Wars)는 액션, 어드벤처, 공상과학, 스릴러, 웨스턴의 5가지 장르 특성을 가지고 있다.

① 각 영화가 포함하는 장르의 집합을M_i라고 하면, 모든 영화에 대해 아래와 같이 나타낼 수 있다.

$$M_i=\{G_1,G_2,...,G_{19}\} \quad (\text{단, } G_i\text{: 장르}i\text{의 값으로, 0 or 1}) \quad \cdots\cdots (5)$$

② 이 M_i를 이용하여 피어슨 상관계수로 장르 사이 유사도 값을 구하면 아래 표와 같다.

표 2. 영화 장르 사이 피어슨상관계수
Table 2. Pearson correlation coefficient between genre

Division	Correlation between Genre of Movies																	
	Action	Adventure	Animation	Children's	Comedy	Crime	Documentary	Drama	Fantasy	Film-Noir	Horror	Musical	Mystery	Romance	Sci-Fi	Thriller	War	Western
Action	1.00	0.34	-0.04	-0.06	-0.16	0.05	-0.07	-0.21	0.01	-0.05	-0.01	-0.04	0.00	-0.06	0.27	0.22	0.09	0.05
Adventure	0.34	1.00	0.04	0.27	-0.11	-0.02	-0.05	-0.20	0.16	-0.04	-0.04	0.01	-0.03	-0.04	0.24	0.01	0.04	0.00
Animation	-0.04	0.04	1.00	0.44	-0.02	-0.04	-0.03	-0.14	0.05	-0.02	-0.02	0.33	-0.03	-0.05	0.02	-0.04	-0.02	-0.02
Children's	-0.06	0.27	0.44	1.00	0.02	-0.07	-0.05	-0.16	0.29	-0.03	-0.07	0.22	-0.02	-0.09	0.00	-0.11	-0.05	-0.02
Comedy	-0.16	-0.11	-0.02	0.02	1.00	-0.09	-0.11	-0.34	0.01	-0.08	-0.08	0.04	-0.05	0.08	-0.10	-0.22	-0.05	-0.01
Crime	0.05	-0.02	-0.04	-0.07	-0.09	1.00	-0.05	0.01	-0.01	0.17	-0.03	-0.05	0.08	-0.07	-0.05	0.13	-0.06	-0.03
Documentary	-0.07	-0.05	-0.03	-0.05	-0.11	-0.05	1.00	-0.13	-0.02	-0.02	-0.04	-0.03	-0.03	-0.07	-0.04	-0.07	-0.02	-0.02
Drama	-0.21	-0.20	-0.14	-0.16	-0.34	0.01	-0.13	1.00	-0.06	-0.08	-0.18	-0.10	-0.07	-0.03	-0.18	-0.16	0.04	-0.06
Fantasy	0.01	0.16	0.05	0.29	0.01	-0.01	-0.02	-0.06	1.00	-0.01	-0.03	-0.02	-0.02	-0.02	0.10	-0.03	-0.02	-0.02
Film-Noir	-0.05	-0.04	-0.02	-0.03	-0.08	0.17	-0.02	-0.08	-0.01	1.00	-0.03	-0.02	0.16	-0.04	0.01	0.15	-0.03	-0.02
Horror	-0.01	-0.04	-0.02	-0.07	-0.08	-0.03	-0.04	-0.18	-0.03	-0.03	1.00	-0.05	-0.02	-0.09	0.07	0.04	-0.05	-0.03
Musical	-0.04	0.01	0.33	0.22	0.04	-0.05	-0.03	-0.10	-0.02	-0.02	-0.05	1.00	-0.04	0.04	-0.03	-0.07	-0.02	-0.02
Mystery	0.00	-0.03	-0.03	-0.02	-0.05	0.08	-0.03	-0.07	-0.02	0.16	-0.02	-0.04	1.00	-0.01	0.01	0.20	-0.04	-0.03
Romance	-0.06	-0.04	-0.05	-0.09	0.08	-0.07	-0.07	-0.03	-0.02	-0.04	-0.09	0.04	-0.01	1.00	-0.07	-0.07	0.05	-0.04
Sci-Fi	0.27	0.24	0.02	0.00	-0.10	-0.05	-0.04	-0.18	0.10	0.01	0.07	-0.03	0.01	-0.07	1.00	0.13	0.06	-0.03
Thriller	0.22	0.01	-0.04	-0.11	-0.22	0.13	-0.07	-0.16	-0.03	0.15	0.04	-0.07	0.20	-0.07	0.13	1.00	-0.05	-0.05
War	0.09	0.04	-0.02	-0.05	-0.05	-0.06	-0.02	0.04	-0.02	-0.03	-0.05	-0.02	-0.04	0.05	0.06	-0.05	1.00	0.00
Western	0.05	0.00	-0.02	-0.02	-0.01	-0.03	-0.02	-0.06	-0.02	-0.02	-0.03	-0.02	-0.03	-0.04	-0.03	-0.05	0.00	1.00

③ 또한, 장르 사이의 그룹화를 위해 K-Means Clustering을 실시한다. 아래 그림은 상관계수의 변화를 나타내고 있으며, 여기서 변화가 큰 1678을 기준으로 그룹화를 할 수 있다. 즉, K=4 (K 값=1682-1678)로 결정한다. 이는 4개의 그룹으로 나누는 것이 가장 이상적이라는 의미다.

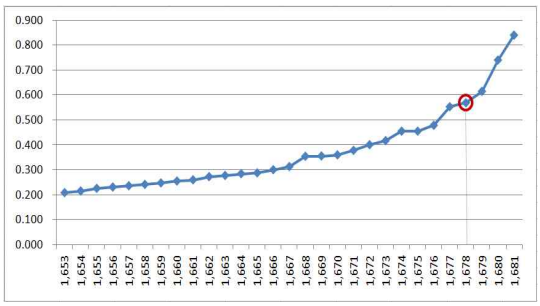

그림 3. 1-Coefficients 계산 단계 변화 그래프
Fig. 3. 1-Coefficients calculated phase change graph

④ 장르사이 유사도와 K-Means Clustering을 바탕으로 형성된 4개의 그룹은, 드라마 중심 가족 영화, 액션 중심의 모험영화, 아동 영화, 범죄 영화 그룹이며, 장르유사도는 4개의 그룹 내 장르사이로 한정한다.

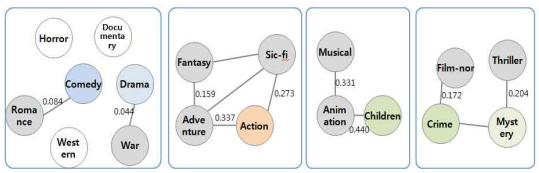

그림 4. Centroid Method를 이용한 군집 및 장르유사도
Fig. 4. Using Centroid method clustering and genre similarity

다음으로, 지역 위치정보 코드를 생성한다. 이는 사용자 우편번호 정보를 활용하여 지역 위치 정보를 생성한다. 즉, 우편번호의 앞 2자리를 이용하여 공통 그룹화 하여 10개의 그룹으로 재구성하고 이를 지역 위치 정보 코드화 한다.

지금까지 도출된 Movie Cluster, 고객 선호장르 프로파일, 장르유사도 프로파일, 위치정보 결과를 MovieLens Data Set에 추가하여 전체 Metadata를 아래와 같이 구성한다.

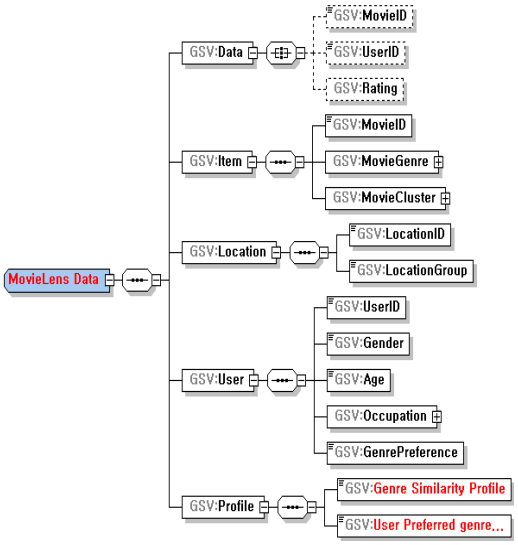

그림 5. MovieLensDataSet의 메타데이터
Fig. 5. Metadata of MovieLensDataSet

2. 실험

본 연구에서 제안한 장르유사도와 선호장르를 이용한 협업추천(CF_GS_UP : Collaborative Filtering Using Genre Similarity and User Preferred Genre) 방식에 대한 실험을 위하여, 943명의 고객이 시청한 영화의 빈도수를 기준으로 3그룹으로 나누었다. 즉, 고객이 가장 적게 본 영화의 개수는 20개였으며, 가장 많이 본 영화의 개수는 737개였다. 따라서 각 그룹은 최소값 20, 최대값 737 사이에서 3그룹으로 나뉜다.

그 후 각 그룹에 속한 고객 중 무작위로 4명씩을 추출하여 총 12명을 실험군 목표고객으로 선정하여 진행한다.

본 연구에서는 인구통계학적 유사그룹 방식과 선호장르가 반영된 유사그룹 방식을 추가하여 비교 실험이 되도록 한다.

먼저 인구통계학적 유사그룹을 바탕으로 협업필터링(Collaborative Filtering Using User Similarity: CF_US)방법을 적용한다. 다음으로는 목표고객의 선호장르가 반영된 인구통계학적 유사그룹을 바탕으로 협업필터링(collaborative Filtering Using Preference Genre Based User Similarity: CF_PG_US)방법을 적용한다. 끝으로, 본 연구에서 제안한 장르유사도와 선호장르 기반 협업필터링(Collaborative Filtering Using Genre Similarity and User Preferred Genre: CF_GS_UP)방법을 적용하여 상호 비교 분석 한다.

첫 번째, CF_US 모델은 목표고객과 인구통계학적 요소(성별, 나이, 직업)가 유사한 고객을 대상으로 한다. 즉, 이들이 평가한 각 영화의 평점을 기반으로 피어슨 상관계수(Pearson correlation coefficient)를 이용하여 유사도를 산출한 후 최근접 방법(K-nearest Method)으로 목표고객과 유사 상관관계를 나타내는 유사 그룹을 도출한다.

이제, 이 유사 그룹 고객들이 본 영화를 바탕으로 목표고객이 보지 않은 영화를 추천하기 위한 평가 예측치는 평균반영평균(Bias-form-mean Average) 방법을 이용하여 도출한다. 마지막으로 Top-N기법으로 추천 목록을 생성하여 서비스 할 수 있도록 한다[27-28].

아래 표는 CF_US 모델을 적용하여, 12명의 실험군 목표고객 중 498번 대한 추천 실험 결과표이다. 즉, 목표고객의 인구통계학적 유사 그룹에 대해, 유사 그룹이 시청한 영화의 평점을 바탕으로 평가 예측치를 구한 것이다.

목표고객이 보지 않은 영화를 대상으로 각 영화에 대해 유사 고객 그룹이 평가한 평점에서 목표고객의 평점 평균을 뺀 값에 유사 고객의 피어슨 값을 곱한 후에 각 영화별로 합한다. 이를 피어슨 합으로 나누면 평가 예측치가 도출된다. 이제 상위 10개를 추천 목록을 생성하여 추천 목록으로 사용한다.

표 3. CF_US를 통한 추천 결과
Table 3. The recommended results used CF_US

구분		평점평균	Pearson	상위 10개 추천영화 평가예측치									
사용자 구분	USER ID			313	200	408	285	298	216	48	157	196	201
목표 사용자	498	3.2886											
유사사용자	293	2.8350	0.460	0.54	0.54	0	1	0.54	0.54	1	1	0.54	0
	801	3.9091	0.408	0.45	0	0	0	0	0	0	0	0	0
	21	2.4314	0.354	0	0.91	0.91	0	0.91	0	0	0	0	0.91
	741	3.2535	0.321	0.24	0	0	0	0	0.24	0.24	0	0.56	0
	199	2.5938	0.270	0.38	0	0.65	0.38	0	0	0	0	0	0
	201	2.8697	0.216	0.46	0.46	0.24	0.24	0	0.24	0.03	0.24	0.24	0.24
	22	3.1122	0.197	0	0	0	0	0	0.18	0	0	0	0.18
	896	2.7972	0.117	0.14	0.14	0	0	0	0.26	0.14	0.14	0.02	0.02
	445	1.8400	0.094	0.02	0	0.11	0	0.02	0	0	0	0	0
사용자 평가치 합				2.22	2.05	1.91	1.62	1.46	1.45	1.41	1.38	1.36	1.35
사용자 평가치 합 / 피어슨 합				0.91	0.84	0.78	0.66	0.6	0.6	0.58	0.57	0.56	0.55
평가예측치				4.2	4.13	4.07	3.95	3.89	3.88	3.86	3.86	3.85	3.84
순위				1	2	3	4	5	6	7	8	9	10

두 번째, CF_PG_US 모델은 고객 선호장르 프로파일을 기반으로 목표고객의 선호장르가 반영된 인구통계학적 유사 고객을 대상으로 한다. 먼저, 인구통계학적 요소가 유사한 고객을 대상으로 선호장르가 2개 이상 유사한 고객을 유사 고객으로 재구성한다. 그 후 이들이 평가한 각 영화의 평점을 기반으로 피어슨 상관계수(Pearson correlation coefficient)를 이용하여 유사도를 산출한 후 최근접 방법(K-nearest Method)으로 목표고객과 유사 상관관계를 나타내는 유사 그룹을 도출한다. 그 이후는 첫 번째 방법을 적용하여 도출한다.

아래 표는 CF_PG_US 모델을 적용하여, 12명의 실험군 목표고객 중 498번 대한 추천 실험 결과표이다. 즉, 인구통계학적 유사그룹에 대해 목표고객의 선호장르를 반영하여 목표고객의 유사그룹을 재도출한 후, 유사 그룹이 시청한 영화의 평점을 바탕으로 평가 예측치를 구한 것이다.

이는 목표고객이 보지 않은 영화를 대상으로 각 영화에 대해 유사 고객 그룹이 평가한 평점에서 목표고객의 평점 평균을 뺀 값에 유사 고객의 피어슨 값을 곱한 후에 각 영화별로 합한다. 이를 피어슨 합으로 나누면 평가 예측치가 도출된다. 이제 상위 10개를 추천 목록을 생성하여 추천 목록으로 사용한다.

표 4. CF_PG_US를 통한 추천 결과
Table 4. The recommended results used CF_PG_US

구분		평점평균	Pearson	상위 10개 추천영화 평가예측치									
사용자 구분	USER ID			313	200	408	285	298	48	157	196	357	705
목표 사용자	498	3.2886											
유사 사용자	293	2.835	0.4603	0.54	0.54	0	1	0.54	1	1	0.54	0.54	1
	801	3.9091	0.4082	0.45	0	0	0	0	0	0	0	0	0
	21	2.4314	0.3541	0	0.91	0.91	0	0.91	0	0	0	0	0
	741	3.2535	0.3208	0.24	0	0	0	0	0.24	0	0.56	0.56	0
	199	2.5938	0.2697	0.38	0	0.65	0.38	0	0	0	0	0	0
	201	2.8697	0.2162	0.46	0.46	0.24	0.24	0	0.03	0.24	0.24	0.24	0.03
	896	2.7972	0.1174	0.14	0.14	0	0	0	0.14	0.14	0.02	0	0.26
	445	1.84	0.0945	0.02	0	0.11	0	0.02	0	0	0	0	0
사용자 평가치 합				2.22	2.05	1.91	1.62	1.46	1.41	1.38	1.36	1.34	1.28
사용자 평가치 합 / 피어슨 합				0.99	0.91	0.85	0.72	0.65	0.63	0.62	0.61	0.6	0.57
평가예측치				4.28	4.2	4.14	4.01	3.94	3.92	3.91	3.9	3.89	3.86
순위				1	2	3	4	5	6	7	8	9	10

세 번째, CF_GS_UP 모델은 장르유사도와 선호장르가 반영된 유사 고객을 대상으로 한다. 즉, 인구 통계학적 유사 고객에 대해 목표고객의 선호장르를 장르유사도를 사용하여 재구성하고 이 선호장르가 반영된 유사 고객 그룹을 도출한다. 이는 장르유사도 프로파일을 기반으로, 목표고객의 선호장르를 재구성하는 것이다. 그 후에는 두 번째 모델을 적용하여 구한다.

아래 표는 CF_GS_UP 모델을 적용하여, 12명의 실험군 목표고객 중 498번 대한 추천 실험 결과표이다. 즉, 목표고객의 선호장르를 장르유사도 프로파일을 이용하여 재구성한 후 유사 그룹을 도출하여 이들이 시청한 영화의 평점을 바탕으로 평가 예측치를 구한 것이다.

이는 목표고객이 보지 않은 영화를 대상으로 각 영화에 대해 유사 고객 그룹이 평가한 평점에서 목표고객의 평점 평균을 뺀 값에 유사 고객의 피어슨 값을 곱한 후에 각 영화별로 합한다. 이를 피어슨 합으로 나누면 평가 예측치가 도출된다. 이제 상위 10개를 추천 목록을 생성하여 추천 목록으로 사용한다.

표 5. CF_GS_UP를 통한 추천 결과
Table 5. The recommended results used CF_GS_UP

구분		평점평균	Pearson	상위 10개 추천영화 평가예측치								
사용자 구분	USER ID			313	96	200	318	276	346	216	588	682	1135
목표 사용자	498	3.2886											
유사 사용자	801	3.9091	0.4082	0.45	0	0	0	0	0	0	0	0.45	0
	201	2.8697	0.2162	0.46	0.24	0.46	0.46	0.46	0.24	0.24	0.24	0.03	0.46
	896	2.7972	0.1174	0.14	0.26	0.14	0.14	0	0	0.26	0.26	0	0
	445	1.84	0.0945	0.02	0.2	0	0	0.11	0.3	0	0	0	0
사용자 평가치 합				1.06	0.71	0.6	0.6	0.57	0.54	0.5	0.5	0.47	0.46
사용자 평가치 합 / 피어슨 합				1.27	0.85	0.72	0.72	0.68	0.65	0.6	0.6	0.57	0.55
평가예측치				**4.56**	**4.13**	**4.01**	**4.01**	**3.97**	**3.94**	**3.89**	**3.89**	**3.85**	**3.84**
순위				1	2	3	4	5	6	7	8	9	10

지금까지 실험을 바탕으로, 실험 목표고객 12명에 대한 각 모델별 실험을 통하여 얻은 추천 결과에 대한 MAE(Mean Absolute Error) 값을 그래프로 나타내면 아래와 같다.

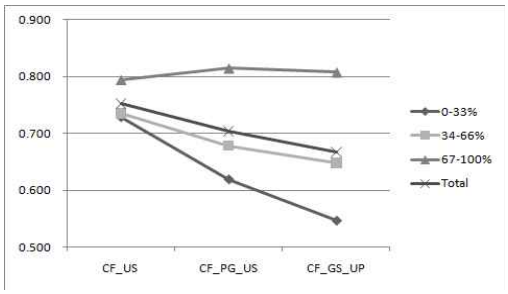

그림 6. MAE 그래프
Fig. 6. MAE Graph

본 실험에서는 영화 시청 빈도가 가장 적은 0~33% 그룹의 MAE값이 가장 적다. 이는 아이템에 대한 평가치가 적은 경우에도 본 연구에서 제안한 기법이 효과가 있음을 나타내고 있다.

각 모델별 실험을 통해 얻은 추천 결과값에 대한 MAE에 대해 Paired-Samples T-test를 실시하여 실험 결과를 검증한다.

표 6. 대응표본 T 검증
Table 6. Paired-Samples T test
(P=0.95, df=120 α=1.658)

Division	T- test
CF_US - CF_PG_US	1.659
CF_US - CF_GS_UP	3.397
CF_PG_US - CF_GS_UP	2.527

지금까지 협업필터링 모델에 대한 실험 결과, 선호장르를 이용한 ②(CF_PG_US), ③(CF_GS_UP)모델이 ①(CF_US) 모델에 비해 평균 오차값이 감소한 것을 관찰 할 수 있다.

또한, ③(CF_GS_UP) 모델이 ②(CF_PG_US) 모델보다 평균 오차값이 감소한 것을 관찰 할 수 있다. 이는 본 연구에서 제안한 장르유사도와 선호장르를 이용한 협업필터링 모델이 성능 향상에 도움이 된다는 것이다.

3. 모바일 환경 적용

각 협업필터링모델에서 도출된 최종 추천목록을 모바일 단말을 통해 서비스 할 수 있도록 실험환경을 구한다. 이를 위한 개발환경은 MySQL5.1, JDK1.6, Eclipse, Android Build Target Google APIs 2.1, XML Parser1.0, SDK으로 구성한다.

도출된 추천목록은 안드로이드 모바일 실험 환경에서 XML파일로 구성하여 안드로이드의 XMLPullParser를 이용하여 ListView형식과 ImageView형식으로 구성된 목표고객에게 서비스 할 수 있도록 한다.

즉, 아래 그림에서와 같이 도출된 추천 목록, 이에 대한 세부 정보 및 위치정보, 샘플 영화로 구성한다.

그림 7. 안드로이드 플랫폼에서의 추천목록 화면
Fig. 7. Recommendation list screen in the Android platform

V. 결론

본 연구에서는 MovieLens Data Set를 기반으로 고객 선호 장르 프로파일과 장르 유사도 프로파일을 추가 확장하여 추천목록을 도출하여 모바일 실험 환경에서 서비스 될 수 있도록 구현하였다. 이는 협업필터링의 희박성(Sparsity)의 한계점을 극복하고자 한 것과 안드로이드(Android)폰을 이용하여 모바일 서비스에 적용한 것의 의미가 있다. 더 나아가, 본 제안 모델은 모바일의 특성과 장르유사도의 특성을 반영하여, 신규 아이템이나 신규 고객에게도 최소한의 추천을 해 줄 수 있다는 것을 의미한다. 또한, 협업필터링 확장성(Scalability)

의 한계점을 극복할 수 있도록 장르기반 프로파일과 고객 선호장르 프로파일을 추가하여 추천함으로써 추천 성능을 높인 것 이다. 하지만, 실험데이터의 한계로 정확한 근접 무비샵 위치 정보에 대한 추천을 함께 제공하지 못하고 실험에서 설정한 위치 MapView 서비스로 대체한 점이 아쉽다.

다른 한편, MovieLens Data Set에서는 장르라는 대표 속성만 제공하고 있지만, 유사 속성으로 배우, 감독, 음악 등의 속성을 더 포함할 수 있으며, 이들을 추가 클러스터링하여 추천한다면, 보다 다양한 고객들의 기호를 충족시킬 수 있을 것이다.

또한, 콘텐츠에 대한 선호기호를 반영하고 있는 교육 서비스 부분 등으로 대상 범위를 확대할 수도 있고, 더 나아가 추천된 결과에 대한 고객의 피드백(feedback)을 되받아 적용할 수 있도록 연구를 확대 하고자 한다.

참고문헌

[1] M.D.Mulvenna and S.S.Anand and A.G.Buchner, "Personalization on the Net Using Web Mining : Introduction," Communications of the ACM, Vol. 43, No. 8, pp. 122~125, 2000.

[2] H.J. Kwon and D.K.Chung, K.S.Hong, "A Multimedia Recommender System Using User Playback Time," Korean Society for Internet Information, Vol. 10, No. 1, pp.111-121, 2009.

[3] J.W.Choi, "An Application for Calculation and Visualization of Narrative Relevance of Films Using Keyword Tags," Korea Advanced Institute of Science and Technlogy : Master's Thesis, pp.19-21, 2007.

[4] K.Chorianopoulos, "Personalized and mobile digital TV applications," Multimedia tools and Applications, Vol. 36, pp.1-10, 2008.

[5] T.Q.Lee, Y.Park, Y.T.Park, "A time-based approach to effective recommender systems using implicit feedback," Expert Systems with Applications, Vol. 34, Issue. 4, pp.3055-3062, May 2008.

[6] K.R.Kim, J.H.Lee, J.H.Byeon, N.M. Moon, "Recommender System Using the Movie Genre Similarity in Mobile Service," The 4th International Conference on Multimedia and Ubiquitous Engineering, 2010.

[7] J.M.Oh, J.H.Song, N.M.Moon, "Preference Element Selectable Interactive Recommender System by Employing Collaborative Filtering," The 4th International Conference on Multimedia and Ubiquitous Engineering, 2010.

[8] J.Y.Park and B.S. Chon, "Analysis on Mobile Content Services of the Demestic Media Comoanies," The Journal of the Korea Contents Association, Vol. 10, No. 1, pp.160-169, 2009.

[9] H.S.Park and M.H.Park, S.B.Cho, "LBS Information Recommendation in Mobile Environment Using Multi-Criteria Decision Making," Journal of Computing Science and Engineerin, Vol. 14, No. 3, pp. 306-310, 2008.

[10] B.J.An and E.J.Kim, Y.B.Lee, "A Hiererchical Representatives Clustering Technique for Data Mining," Korean Institute of Information Scientists and Engineers, Vol. 27, No. 2, pp.69-71, 2000.

[11] M.H.Huh and Y.G.Lee, "Reproducibility estimation and Application of K-means clustering," The Korean Journal of Applied Statistics, Vol. 17, No. 1, pp.135-144, 2004.

[12] D.J.Lee and S.K.Lee, S.G.Lee, "Considering temporal context in music recommendation based on collaborative filtering," Proceedings of Korea computer congress, Vol. 36, No. 1, 2009.

[13] G.Adomavicius and A.Tuzhilin, "Toward the Next Generation of Recommender Systems : A Survey of the State-of-the-Art and Possible Extensions," IEEE Transactions on Knowledge and Data Engineering, Vol. 17, No. 6, pp. 734-749, June 2005.

[14] H.C. Lee, "Enhancement of Collaborative Filtering in Electronic Commerce Recommender System," Kangwon University Graduation School : Doctoral thesis, 2009.

[15] S.H.Jo, "Weight Recommendation Technique Based on Item Quality To Improve Performance of New User Recommendation and Recommendation on The Web, " Hannam University Graduation School : Doctoral thesis, 2008.

[16] G.Lekakos and G.M.Giaglis, "Improving the Prediction Accuracy of Recommendation Algorithms : Approaches Anchored on Human Factors," Interacting with Computers, Vol. 18, pp. 410-431. 2006.

[17] S.J.Lee and T.R.Jeon, G.D,Baek, S.S.Kim, "A Movie Rating Prediction System of User Propensity Analysis based on Collaborative Filtering and Fuzzy System," Journal of Korean institute of intelligent systems, Vol. 19, No. 2, pp.242-247, 2009.

[18] K.C.Park, "Collaborative Filtering Method Considering Purchase Interval for Mobile Multimedia Contents Recommendation," Hanyang University : Master's Thesis, 2008.

[19] J.S.Lee and S.D.Park, "Performance Improvement of a Movie Recommendation System using Genre-wise Collaborative Filtering," Journal of intelligent information systems, Vol.13, No. 4, pp.65-78, 2007.

[20] B.M.Sarwar and G.Karypis, J.Konstan, J.Riedl, "Item-Based Collaborative Filtering Recommendation Algorithms," ,Proceedings of the 10th international conference on World Wide Web, pp.285-295, 2001.

[21] H.S.Lee, J.H.Kwon, "Collaborative Filtering Mobile Contents Recommender Application Using Context and Folksonomy," Journal of Korean Institute of Information Technology, Vol. 7, No.2, pp.132-140, April 2009.

[22] Y.H.Cho and J.K.Kim, "Application of Web usage mining and product taxonomy to collaborative recommendations in e-commerce," Expert System with Applications, Vol 26(1), pp.233-246, 2004.

[23] J.K.Kim and Y.H.Cho, S.T.Kim, H.K.Kim, "A Personalized Recommender System for Mobile Commerce Application," The Korea Society of Management Information Systems, Vol. 15, No. 3, pp.223-240, 2005.

[24] P.Melville, R.J.Mooney, R.Nagarajan, "Content-Boosted Collaborative Filtering for Improved Recommendations," AAAI-02, pp.187-192, 2002.

[25] Y.Zhang, W.Song, "A Collaborative Filtering Recommendation Algorithm Based on Item Genre and Rating Similarity," 2009 International Conference on Computational Intelligence and Natural Computing, pp. 72-75, 2009.

[26] B.I.Kwon, N.M.Moon, "Recommendation system for supporting self-directed learning on e-learning marketplace," Journal of The Korea Society of Computer and Information, Vol. 15, No. 2, pp.135-146, 2010.

[27] B.M. Sarwar and G.Karypis, J.Konstan, J.Riedl, "Application of Dimensionality Reduction in Recommender System : A Case Study," WebKDD-2000 Workshop, 2000.

[28] B.M. Sarwar and G.Karypis, J.Konstan, J.Riedl, "Recommender Systems for Large-scale E-Commerce : Scalable Neighborhood Formation Using Clustering," 5th International Conference on Computer and Information Technology, 2002.

저 자 소 개

김 경 록

1999: 아주대학교 공학사
2006: 서울벤처정보대학원대학교 공학석사
현재: 차세대학습산업기반센터 사무국장
관심분야: 양방향 서비스, 이러닝, Information Learing, 컨설팅
Email : it4all@naver.com

변 재 희

2010: 덕성여자대학교 공학사
현재: 호서대학교 벤처전문대학원 IT응용기술학과 석사과정
관심분야: 모바일콘텐츠, 추천시스템 등
Email : bjaeh9188@gmail.com

문 남 미

1985: 이화여자대학교 컴퓨터학과 공학사
1987: 이화여자대학교 공학석사
1998: 이화여자대학교 공학박사
1999: 이화여자대학교 조교수
2004: 서울벤처정보대학원대학교 디지털미디어학과 교수
현재: 호서대학교 벤처전문대학원 IT응용기술학과 교수
관심분야: 디지털데이터방송비즈니스 모델, HCI, T-Commerce, 메타데이터 등
Email : mnm@hoseo.edu