멀티미디어 콘텐츠를 위한 이용빈도 기반 하이브리드 추천시스템에 관한 연구*

A Study on Hybrid Recommendation System Based on Usage Frequency for Multimedia Contents

김 용(Yong Kim)**, 문성빈(Sung-Been Moon)***

초 록

정보기술과 인터넷의 발전에 따른 정보의 폭발적인 증가로 인하여 정보과잉에 따른 적절한 정보의 선택이 필요하게 되었다. 이를 위하여 이용자가 정보를 효율적으로 이용할 수 있도록 검색 또는 여과하는 일을 수행하기 위하여 정보검색 및 정보여과 시스템이 등장하게 되었다. 이러한 일련의 정보환경의 변화에 대한 보다 적극적인 대응방법으로서 도서관 및 정보센터에서는 이용자가 원하는 정보를 정확하고 효율적으로 제공하기 위한 노력의 일환으로서 이용자에게 맞춤화된 정보추천서비스 제공이 요구된다. 본 연구에서는 도서관 및 정보센터에서 적극적인 정보서비스를 위한 방법으로 이용자에게 맞춤화된 정보를 제공할 수 있는 개인화 추천시스템을 구축하기 위한 방안을 제안하였다. 이를 위하여 기존의 추천방법에 대한 장단점을 분석하고 기존 추천방법에 대한 문제점을 해결하기 위한 방법으로서 대용량 콘텐츠 및 이용자 환경에서 이용자의 콘텐츠 이용빈도를 기준으로 멀티미디어 콘텐츠를 위한 개인화된 하이브리드 추천방법을 제안하였다. 이를 위하여 이용빈도에 있어서 상위 이용자 및 콘텐츠를 분리하고 적절한 추천방법에 적용하기 위한 새로운 형태의 추천방법 및 대용량 추천시스템에 적합한 연관규칙과 협업여과방법에 대한 조합방법을 제안하였다.

ABSTRACT

Recent advancements in information technology and the Internet have caused an explosive increase in the information available and the means to distribute it. However, such information overflow has made the efficient and accurate search of information a difficulty for most users. To solve this problem, an information retrieval and filtering system was developed as an important tool for users. Libraries and information centers have been in the forefront to provide customized services to satisfy the user' s information needs under the changing information environment of today. The aim of this study is to propose an efficient information service for libraries and information centers to provide a personalized recommendation system to the user. The proposed method overcomes the weaknesses of existing systems, by providing a personalized hybrid recommendation method for multimedia contents that works in a large-scaled data and user environment. The system based on the proposed hybrid method uses an effective framework to combine Association Rule with Collaborative Filtering Method.

키워드 : 개인화, 추천, 협업여과, 연관규칙, 선호도, 정보여과, 하이브리드
personalization, recommendation, hybrid, collaborative filtering, association rule, information filtering, user preference

---

* 본 연구는 연세대학교 대학원 박사학위논문의 일부를 요약한 것임.
** KT BcN 본부 책임연구원 (yongkim@kt.co.kr)
*** 연세대학교 문헌정보학과 교수 (sbmoon@yonsei.ac.kr)
■ 논문접수일자 : 2006년 8월 10일
■ 게재확정일자 : 2006년 9월 15일

1. 서 론

인터넷의 발전에 따라 대부분의 도서관 및 정보센터를 포함한 정보제공자들은 인터넷을 통하여 정보를 제공함으로써 자관의 서비스 이용자 및 모든 정보이용자들의 요구를 충족시킬 수 있을 것이라는 기대 하에 전자도서관과 같은 정보저장소를 구축하였으나 정보생산을 위한 다양한 저작도구(authoring tool)의 발달과 인터넷의 확산에 따른 정보의 폭발적인 증가에 의한 정보과잉(information overflow)은 이용자에게 자신이 원하는 정보를 찾기 위하여 엄청난 시간과 노력을 요구하는 결과를 초래하였다. 이러한 문제를 해결하기 위한 다양한 방법들이 도서관 및 정보센터를 중심으로 시도되었으며 이에 대한 대표적인 방법으로서 검색엔진의 등장이라고 할 수 있다. 그러나 단순히 검색엔진의 등장만으로는 이러한 문제를 근본적으로 해결할 수 없었으며 이를 위하여 전통적으로 도서관이나 정보센터에서 제공되던 선택적 정보배포(SDI : Selective Dissemination of Information) 서비스가 중요한 대안으로서 제시되었다. 선택적 정보배포 서비스를 기반으로 보다 능동적이고 발전된 형태로서 이용자의 프로파일을 기반으로 하는 푸시기술(push technology)을 적용시킨 맞춤정보서비스(customized information service)가 등장하게 되었다. 이러한 푸시기술을 적용한 맞춤정보서비스는 전통적으로 도서관 및 정보센터에서 이용자의 요구에 맞추어 주기적으로 정보를 제공하는 선택적 정보배포 서비스와 그 기능이 유사하다고 할 수 있다. 그러나 맞춤 정보 서비스는 단순히 이용자의 프로파일 정보와 일치되는 모든 정보를 제공함으로써 여전히 정보과잉의 한계점을 극복할 수는 없었으며, 특히 이용자의 정보요구 행태의 변화에 적절하게 대응할 수 없는 제약점이 있다. 이러한 사회, 문화적인 요구에 따라 이용자요구에 적합한 정보의 추출과 제공을 위한 방법으로서 대용량의 정보에서 개인별 맞춤화된 정보와 서비스를 제공할 수 있는 개인화 서비스에 대한 관심은 더욱 높아지고 있다. 개인화 서비스는 도서관 및 정보센터의 관점에서 정보이용자의 요구를 보다 정확하게 분석하고 이를 기반으로 이용자의 정보요구에 적합한 정보를 제공한다는 측면에서 매우 중요하고 필수적인 정보서비스로 고려되고 있다.

정보제공의 맞춤화(customization)는 최근 다양한 분야에서 응용되고 있는 정보검색기술, 푸시기술, 웹 에이전트(web agent)기술 등을 기반으로 하는 관련 기술들의 결합으로 구현될 수 있다. 맞춤정보서비스는 사이버공간상에서 이용자가 원하는 정보만을 효과적으로 검색하여 제공함으로써 이용자의 정보에 대한 맞춤화 요구를 충족시키고자 개발되었다. 이러한 일련의 노력으로 웹상에서 개인별 프로파일 정보에 따른 "MyLibrary" 서비스가 제공되고 있으며 그 효과에 대해서는 약 75% 이상의 이용자들이 해당 서비스에 만족을 표시하고 있다(김현희 2002). 한편, 이전의 맞춤정보서비스가 명시적으로 이용자가 입력한 인구통계학적 정보 및 선호분야에 대한 키워드와 일치되는 정보를 제공하는데 비하여 근래에는 보다 발전되고 지능화된 기술과 서비스 방법을 기반으로 이용자 개개인의 묵시적인 행위정보 및 보다 세부적인 이용자 정보를 분석하여 적합한 정보

서비스를 제공함에 따라 맞춤정보서비스라는 용어를 대신하여 개인화(personalization)라는 용어가 보다 광범위하게 사용되고 있다. 이러한 개인화 추천서비스를 위하여 본 연구에서는 다양한 추천방법의 장단점에 대해서 알아보고, 보다 대용량의 멀티미디어 콘텐츠 환경에서 효율적이면서 추천의 정확성을 높일 수 있는 추천방법을 제안하여 이를 적용한 개인화 추천시스템을 설계 및 구현하고 제안한 추천방법에 대한 평가를 하고자 한다.

1.2 연구범위 및 구성

본 연구에서는 대표적인 멀티미디어 자료인 음악 콘텐츠를 대상으로 개인에게 맞춤화된 형식으로 추천서비스를 제공하기 위한 후보 콘텐츠 추천 및 결합방법의 제안 및 구현을 목표로 하고 있다. 이를 위하여 기본적으로 요구되는 웹상에서의 이용자 콘텐츠 이용행태에 대한 분석 및 데이터베이스 구축을 위한 데이터의 전처리방법, 후보 콘텐츠 추천 및 결합방법을 제안하고 추천시스템을 구현함으로써 이에 대한 실질적인 평가를 하고자 한다. 본 연구에서 목표하고 있는 구체적인 연구 범위와 내용은 다음과 같이 요약될 수 있다.

첫째, 기존 추천시스템의 구성과 특징을 비교 분석한다.

둘째, 다양한 추천방법에 대한 기존 연구들에 대한 분석을 통하여 본 연구에서 제안한 모형과 기반구조의 설계를 포함하는 기본 방향을 설정한다.

셋째, 기존 추천방법의 장단점을 분석하고 기존의 제한점을 극복하기 위한 후보 콘텐츠 추천 및 결합방법을 적용한 하이브리드 기반의 추천 모형을 정의한다.

넷째, 대용량 콘텐츠 및 이용자 환경에서 이용자 및 콘텐츠간의 연관성을 반영하기 위하여 이용자의 웹 로그를 이용하여 행동 데이터를 분석하고 이를 기반으로 콘텐츠 이용빈도 및 이용자별 콘텐츠 이용빈도에 대한 분포를 분석한다. 또한, 분석된 결과를 기반으로 실험을 통하여 추출된 임계값을 기준으로 콘텐츠와 이용자를 분할하고 이를 통하여 협업여과 추천방법과 연관규칙을 기준으로 추천 콘텐츠를 선정하는 방법을 제안한다.

다섯째, 콘텐츠 이용빈도에 있어서 다이용 이용자와 다이용 콘텐츠를 적절히 분리하고 활용하는 새로운 형태의 추천방법과 함께, 대용량 추천시스템에 적합한 연관규칙과 협업여과 추천방법에 대한 결합방법의 제안을 통하여 추천의 정확도와 확장성 향상을 유도한다.

여섯째, 특정 주제나 장르에 관계없이 다수의 이용자에게 이용빈도가 높은 콘텐츠를 추천서비스에 적용하기 위한 방법론을 제안한다.

일곱째, 제안된 추천 모형을 동적으로 반영할 수 있는 응용 개발 환경을 설계 및 구현한다.

여덟째, 실험을 통하여 추천시스템의 효율성을 기존의 추천방법과 비교하여 성능의 우수성을 검증한다.

2. 이론적 배경

본 장에서는 개인화 추천서비스의 유형 및 특징을 알아보고, 추천시스템에서 사용되는 다양한 추천방법에 대한 비교를 위하여 각 추천

방법에 대한 장단점분석과 함께, 관련된 선행 연구를 알아본다. 특히 본 연구에서 적용하고 있는 협업여과 추천방법 및 연관규칙에 대해서 보다 구체적으로 알아보며 해당 추천방법에 대한 장단점 분석을 기반으로 추천의 정확성 및 확장성 향상을 위하여 본 연구에서 제안하고 있는 방법론에 대하여 간략히 기술한다.

2.1 개인화 추천서비스의 정의

웹사이트 상에서의 개인화는 웹사이트에 들어오는 이용자를 각 이용자의 성향과 행태별로 세분화하여, 이용자가 선호할 수 있는 적합한 정보를 보여주거나 서비스를 제공하는 것을 의미한다. 이는 이용자의 요구를 만족시킴으로서 해당 웹사이트에 대한 이용자의 충성도(loyalty)를 높여줄 뿐 아니라 타켓 마케팅(target marketing)과 일대일 마케팅(one-to-one marketing)을 가능하게 해준다는 점에서 의의가 크다. 개인화라는 용어의 범위는 상당히 넓게 이해되며 개인화 서비스의 종류도 다양한 형태를 보여주고 있다. Yahoo의 My Yahoo!와 같이 이용자별로 이용자가 원하는 스타일과 내용들을 선별해서 볼 수 있도록 메뉴를 구성해주는 방법, 전자 상거래 업체들에서 적용되는 이용자의 개인적 취향에 따르는 적절한 제품을 추천해 주는 방법, 더 나아가 이용자가 가장 관심을 가질 수 있는 배너광고를 선택해서 회전식으로 보여주는 것까지 모두 포함될 수 있다. 개인화 추천이라는 용어는 최근 인터넷의 대중화에 따라 협의의 의미로 웹사이트 개인화라는 의미로 많이 사용되고 있다.

2.2 개인화 추천방법

개인화 추천서비스에 대한 문제가 학술적으로 처음 발표된 것은 90년대 중반부터 라고 할 수 있다(Hill et al. 1995 ; Rensnick et al. 1994 ; Shardanand and Maes 1995). 초기 추천에 관한 연구를 시작으로 추천 문제는 지금까지 광범위하게 연구되어 왔으며 정보검색, 데이터마이닝 분야의 다양한 방법을 기반으로 다양한 추천방법들이 제안되었으며 실제 추천시스템의 구현에 적용되어 연구되어져 왔다. 대표적인 추천방법은 추천 과정에 따라 〈표 1〉과 같은 네 가지 범주로 분류할 수 있다(Burke 2002). 이러한 추천방법 중에서 최근에는 협업여과 추천방법이 가장 대중적으로 이용되고 있다.

2.2.1 협업여과 추천방법

협업여과 추천방법은 이용자의 선호도에 대한 데이터를 기반으로 새로운 이용자가 관심을 가질 것으로 생각되는 항목을 추천해 주는 방법이다. 규칙기반 추천이나 내용기반 추천 등의 방법이 항목 자체의 속성정보를 사용해서 이용자에게 추천하는 것과는 달리 협업여과 추천방법은 항목에 대한 다른 이용자들의 선호를 기반으로 하기 때문에 협업(collaborative)이라는 용어를 사용하게 된다. 협업여과라는 용어는 1990년 초에 제록스(Xerox)의 PARC 연구소에서 개발한 Tapestry에서 처음 사용되었으며, 이 시스템은 전자우편, 넷뉴스 등에 이용자의 평가 또는 주석을 붙임으로서 문서의 참조시에 이를 참고하도록 하였다(Goldberg et al. 1992). 한편, 이전의 추천시스템들이

〈표 1〉 추천방법의 분류 및 장단점

분류	설명	장점	단점
협업여과 추천방법	- 이용자와 유사한 취향이나 정보요구를 갖는 이웃들의 취향에 기반한 추천	- 다양한 형태의 정보에 적용 가능 - 초기 이용자 문제 부분적으로 해결 - 데이터가 충분하면 예측력 높음	- 충분한 트랜잭션 데이터 필요 - 이용자 및 콘텐츠 규모가 클수록 많은 연산량 요구
규칙기반 추천방법	- 자료를 통하여 규칙을 형성하고, 규칙에 따라 추천 - 데이터마이닝 기법들이 주로 이용됨	- 추천 시간이 짧음 - 확장성 및 희소성문제 일부 해결	- 다양하고 대용량의 콘텐츠에 체계적 적용이 어려움 - 개인성향 반영의 어려움
인구통계 기반 추천방법	- 개인의 특성을 기반으로 이용자를 분류하는 것을 목적으로 하며, 인구 통계 분류를 기반으로 추천	- 구축 용이성	- 정확성에 한계 - 공학적인 추천에는 부적절
내용기반 추천방법	- 정보검색에 뿌리를 두고 있으며 텍스트 정보 적용 - 콘텐츠의 특징들에 의해 이용자의 프로파일을 구성하고 유사한 특징들을 가진 콘텐츠를 추천	- 추천대상의 속성 및 이용자의 성향을 반영 가능 - 초기평가와 희소성 문제를 부분적으로 해결	- 멀티미디어 자료에 적용이 어려움 - 계산의 복잡성 - 신규이용자 문제 - 용어의 중요성

대부분 텍스트 기반의 자료를 대상으로 하였으나 협업여과 추천방법을 이용하여 다양한 멀티미디어 콘텐츠를 추천하는데 이용한 시도들이 있었다. 대표적으로 Ringo 시스템(Shandanand and Maes 1995)은 음악 앨범 추천용으로, 제스터 시스템(Gupta et al. 1999)은 유머 추천용으로, 비디오 추천용으로 Bellcore(Hill et al. 1995), 그리고 플라이캐스팅은 온라인 라디오 추천용으로 개발된 협력여과방법에 기반을 둔 추천시스템들이다. 최근에는 인터넷 서점인 아마존이나 인터넷 CD 상점인 CDNow, 그리고 인터넷 영화 추천 사이트인 MovieFinder 등에서 협력여과 추천방법을 적용시켜 성공을 거두고 있다.

이러한 협업여과 추천방법의 대표적 특징은 다음과 같다. 첫째, 협업여과 추천방법을 사용해서 항목의 선호도를 예측할 때 예측 대상이 되는 항목이 다양한 분야에 속해 있다고 가정하면, 이용자 A가 선호하는 분야와 같은 분야의 항목을 선호했던 이웃이 새로운 분야를 선호하였을 경우 특정 이용자에게 새로운 분야의 항목을 추천할 수 있다. 둘째, 협업여과 추천방법은 동질적(homeogenous)인 선호도를 가진 이용자 집단에서 정확한 예측이 가능하다. 셋째, 협업여과 추천방법은 많은 계산량을 요구하지 않은 중소규모의 환경에서 실시간 추천(Real-time recommendation)을 위해서 고안되었기 때문에 개인화된 추천을 위한 계산은 복잡하지 않고 속도도 빠르다. 그러나 대용량 환경에서는 이용자 간의 유사도 및 콘텐츠 선정을 위하여 훨씬 복잡하고 많은 계산량을 요구한다. 넷째, 이용자의 선호도 정보가 대부분

누락값인 경우에 정확한 예측이 불가능하다. 특히, 대용량 콘텐츠 환경에서 콘텐츠에 대한 이용자의 구매 또는 이용은 제한적이며 따라서 많은 콘텐츠에 있어서 이용자의 선호도 정보가 누락되는 경우가 빈번하게 발생한다.

2.2.2 규칙기반 추천방법

규칙기반 추천방법은 이용자의 행동패턴은 일정한 규칙을 가진다는 가정 하에 유용한 규칙을 찾아내는 방법으로서 대체로 이용자의 구매 또는 정보이용 경향을 파악하려는 경우에 많이 쓰이는 방법이다. 규칙기반 추천기법은 이용자의 프로파일 데이터, 구매 데이터, 웹 로그 데이터 등에 근거하여 조건문 형식의 규칙을 이용한 개인화된 추천을 제공하는 방법으로서 이용자에 의해서 입력되고 이용자에 의해서 생성된 데이터를 활용해서 세밀한 분석과 추론과정을 통해 규칙을 생성하게 된다. 규칙기반 추천을 위하여 데이터마이닝 방법이 주로 사용되는데 대표적으로 트랜잭션 데이터가 누적된 데이터베이스에서 각 트랜잭션 간의 상호관계를 통계적 방법에 의해 연관성이 있는 항목들 사이의 규칙성을 추출하는 연관규칙이 많이 사용된다. 연관규칙은 대형 데이터베이스에서의 항목 간의 규칙을 추출하는데 있어서 효과적으로 수행방법론으로서 예전에 미처 알지 못했거나 생각지도 않았던, 또는 규칙으로 표현될 가능성이 있는 조합이 너무 많아서 일일이 통계적 방법으로 처리하지 못했던 연관성을 찾아 주는 유용한 방법이다. 그러나 연관규칙의 응용에 가장 큰 문제점은 수행 시간이 오래 걸릴 수 있다는 점과 적절한 수행을 위한 최소 신뢰도와 최소 지지도와 같은 매개변수(parameter) 설정이 어렵다는 점이다. 이러한 연관규칙의 문제점에도 불구하고 다음과 같은 장점으로 인하여 연관규칙방법이 근래에 많은 추천시스템에 적용되고 있다. 첫째, 〈선행부→결과부〉로 표현되는 연관성 분석 결과는 이해하기 쉽고 또 이를 즉각적으로 실제에 적용하기 용이하다. 둘째, 서술적 모델(descriptive model)이라고도 불리는 비지도 학습(unsupervised learning) 분석기법으로 대부분의 데이터마이닝 기법들은 예측적 모델(predictive model)로서 지도 학습적(supervised learning) 특성을 가지고 있으므로, 목적변수가 뚜렷이 없는 경우 적용하기 쉽지 않은 데 비하여 연관성 분석의 경우에는 이런 상황에서 확실한 해결 방법이 될 수 있다. 마지막으로, 사용이 편리한 분석 데이터의 형태라는 점이다. 이는 트랜잭션내용에 대한 데이터를 변환 없이 그 자체로 이용할 수 있는 간단한 자료구조를 갖는 분석방법이다. 이러한 연관규칙에 있어서의 성능의 척도로서 지지도, 신뢰도 및 향상도가 이용된다, 연관규칙을 추출하는 대표적인 방법으로서 본 연구에서 적용하고 있는 Apriori 알고리즘이 있다. Apriori 알고리즘은 이진연관규칙에 대한 빈발항목집합을 찾아내는데 유용한 알고리즘으로서 Apriori라는 명칭은 알고리즘이 빈발항목집합 특성인 사전지식(prior knowledge)을 사용한다는 데에서 비롯되었다(박우창 외 2003). Apriori 알고리즘은 해당 데이터베이스에 있는 특정한 길이(항목의 수)를 가지는 항목집합을 선택하는 반복적인 알고리즘으로서 그 과정은 특정 데이터베이스에서 발생하는 모든 트랜잭션을 분석하고 항목에 대한 지지도를 이용하

여 동시에 자주 나타나는 항목들을 정제하고 빈발항목집합에서 생성된 규칙들을 신뢰도를 이용하여 정제하는 방식으로 후보항목집합에서 각각의 지지도를 계산한 후 이용자가 정의한 지지도보다 크거나 같은 조건을 만족하는 데이터로 빈발항목집합을 구성한다(하단심, 황부현 2000).

2.2.3 내용기반 추천방법

내용기반 추천방법은 정보검색과 정보여과 연구에 뿌리를 두고 있으며, 과거에 연구되었던 방법들을 적용하였다. 대부분의 내용기반 추천방법은 문서, 웹사이트(URLs), 유즈넷 뉴스 메시지와 같은 정보 추천에 사용되고 있다. 내용기반 추천방법은 이용자의 항목에 대한 평가 정보 혹은 구매내역을 바탕으로, 미리 정의된 항목에 대한 특징들(features)에 의해 이용자의 프로파일(profile)을 구성하며 생성된 프로파일과 유사한 특징들을 가진 항목을 추천하는 방식으로서 항목과 이용자의 정보요구간의 유사도를 측정하고, 그 결과를 순위화하여 보여주며 원칙적으로 내용기반 추천방법은 과거에 이용자가 선호하는 항목과 유사한 것을 추천한다(Lang 1995 ; Billsus and Pazzani 1998). 내용기반 추천방법에서 다양한 후보 항목은 이용자에 의해서 이전에 점수가 주어진 항목과 비교하여 가장 잘 일치하는 항목을 추천한다. 따라서 내용기반 추천방법은 이용자의 이전 경험을 바탕으로 항목의 내용을 중심으로 분석하여 추천하는 방법이라고 할 수 있다. 한편, 내용기반 추천방법은 정보검색의 주요 방법인 불리안 모델, 벡터공간 모델, 확률 모델, 인공신경망 모델, 퍼지집합 모델과 같은 기법을 사용하여 주제에 적절한 텍스트 정보를 찾아 주는데 아주 효과적인 것이 증명되었다(Balabanovic and Shoham 1997).

추가적으로 현재 웹상에서는 위에서 알아본 추천방법들 중에서 특정한 방법만을 사용하여 개인화 추천서비스를 구현하지 않고 다양한 방법들을 함께 적용하여 구현하고 있다. 이러한 추천방법을 혼합형 방법(hybrid)이라고 할 수 있다. 즉, 개별적인 추천방법의 장점을 추천단계 및 분야에 따라 적절하게 적용하여 보다 정확한 항목을 추천하고자 하는 것이다. 대표적으로 협업여과 추천방법과 내용기반 추천방법을 혼합한 방법, 협업여과 추천방법과 데이터마이닝의 연관규칙, 유전자 알고리즘, 신경망 등의 방법을 결합한 방식이 있다. 예를 들어, 협력여과방법과 내용기반 방법의 결과를 선형적으로 결합시키는 방법으로 Claypool 등(1999)과 Wasfi(1999)가 제안한 시스템이 이러한 유형에 속한다. Wasfi에 의해 제안된 ProBuilder는 내용기반 추천방법과 협업여과 추천방법 모두를 사용하여 웹 페이지를 추천하였다. 이는 내용기반 필터링과 협력필터링을 순차적으로 결합시키는 방법이다. 이러한 유형에서는 먼저 유사한 취미를 가지는 이용자를 찾기 위해 내용기반 여과기법을 적용하고 그 다음 그 결과에 협업여과기법을 적용하는 방법으로 RAPP 시스템과 Fab 시스템 등이 있다. 또한 항목의 내용과 평가 정보를 혼합시키는 방법으로 Popescul (2001)의 확률모델, Basu(1998)의 Ripper 시스템, Good 등(1999)의 에이전트기반 방법, 김병만 외(2004)의 UCHM이 있다.

3. 제안된 하이브리드 추천방법

3.1 추천방법의 개요

최근의 상품 또는 콘텐츠 추천분야의 연구흐름과 함께, 기존의 협업여과 추천방법에서의 문제점인 희소성과 초기 이용자 문제 및 초기 평가문제를 해결하고 동시에 대용량의 멀티미디어 데이터 처리에 있어서의 확장성 문제를 해결하기 위하여 본 연구에서는 콘텐츠 이용빈도에 기반한 개선된 협업여과 추천방법과 데이터마이닝의 연관규칙(Association rule)을 혼합한 새로운 하이브리드 추천방법을 제안하고 있다.

3.1.1 제안된 하이브리드 추천방법의 특징

현재 멀티미디어 콘텐츠를 기반으로 서비스를 제공하는 사이트들은 제공되는 콘텐츠와 이용자 수에 있어서 이전의 규모와는 비교할 수 없을 만큼 대규모로써 수백만 건의 음악이나 영화 콘텐츠를 보유하고 수백만 명의 이용자에게 서비스를 제공하고 있다. 따라서 기존의 중소규모의 텍스트 자료를 주요한 대상으로 구축된 추천시스템을 적용하기에는 추천의 정확성과 확장성 측면에서 많은 문제점을 내포하고 있다. 또한, 기존의 대부분의 추천방법에서는 이용자의 선호도를 분석하기 위하여 이용자의 명시적 피드백을 사용하고 있다. 그러나 이러한 가정은 현재의 웹 및 정보환경에 비추어 보았을 때 현실적으로 적용하기가 어렵다. 일반적으로 이용자들은 웹상에서 항목에 대한 이용행위 또는 구매행위를 완료하고 항목에 대한 평가와 같은 추가적인 행위를 수행하지 않거나 성실하게 평가를 수행하지 않는 경향이 있다. 비록 이용자가 추가적인 행위로서 평가를 제공한다고 하더라도 평가에 대한 기준이 개인적으로 상이하기 때문에 서로 다른 기준에서 평가된 결과를 입력값으로 적용한다는 것은 추천의 정확성뿐만 아니라 추천시스템의 신뢰성에 대한 많은 문제점을 안고 있다. 따라서 본 연구에서는 이러한 이용자의 경향을 반영하여 보다 정확한 추천을 위하여 이용자의 실질적인 콘텐츠에 대한 선호도를 묵시적이면서 가장 정확하게 보여주는 콘텐츠에 대한 이용여부를 이용자의 선호도 정보로 적용하였다. 따라서 이용빈도에 따른 이용자의 콘텐츠에 대한 선호도 분석은 단순하면서도 가장 효과적인 방법이라고 할 수 있다.

1) 멀티미디어 콘텐츠의 특징

텍스트 자료와는 달리 멀티미디어 자료는 추천을 위하여 고려해야 할 부분들이 많다. 콘텐츠 자료는 자료의 내용을 대표하는 색인어로서 키워드를 추출할 수 있으며 따라서 2장에서 언급한 내용기반 추천방법을 이용하여 보다 정확한 추천이 가능하다. 그러나 멀티미디어 자료의 내용에 대한 분석이 어렵기 때문에 항목이나 콘텐츠에 대한 내용을 기반으로 하는 추천방법을 적용하는데 있어서는 다음과 같은 한계점이 있다. 첫째, 텍스트 문서와는 달리 콘텐츠에 대한 특성을 표현하는 도구로서 키워드를 적용하기 어렵다. 둘째, 유사도를 계산하기 위하여 속성 필드별 용어를 키워드로 사용하는 경우에 있어서 비슷한 용어들이 대부분이기 때문에 대용량의 콘텐츠의 추천을 위한 후보 콘

텐츠 추천과정에서 키워드 간의 관계성이 높아져서 계산이 불가능할 정도로 복잡해질 수 있다. 셋째, 협업여과 추천방법과 내용기반 추천방법에서 주로 사용하는 가중치 기법의 적용이 어렵다. 마지막으로 텍스트 자료에서도 문제점으로 지적되고 있는 용어의 중의성 문제는 멀티미디어 자료의 경우 필드별로 같은 용어가 키워드로 중복될 수 있으며 따라서 추천에 있어서 정확성을 낮추게 될 수 있다.

2) 제안된 하이브리드 추천방법의 특징

추천서비스를 둘러싸고 있는 새로운 환경적인 변화에 따른 문제점을 해결하고 보다 높은 추천의 정확성과 대규모의 이용자 및 콘텐츠의 추천을 위하여 본 연구에서는 이용자의 개인정보 및 콘텐츠 속성을 활용하지 않고 이용자 및 콘텐츠 이용빈도를 주요한 고려요소로서 설정하고 있다. 세부적인 방법론적 측면에서는 기존의 협업여과 추천방법을 개선하고 데이터마이닝에서 이용되는 연관규칙의 변형된 방법과 함께, 전체 이용자의 콘텐츠에 대한 선호도를 반영하고 있는 많은 이용자로부터 이용빈도가 높은 콘텐츠에 대한 이용빈도를 동시에 반영하고 있는 하이브리드 추천방법을 제안하고 있다. 본 연구에서 제안하고 있는 하이브리드 추천방법의 특징은 다음과 같이 정리할 수 있다.

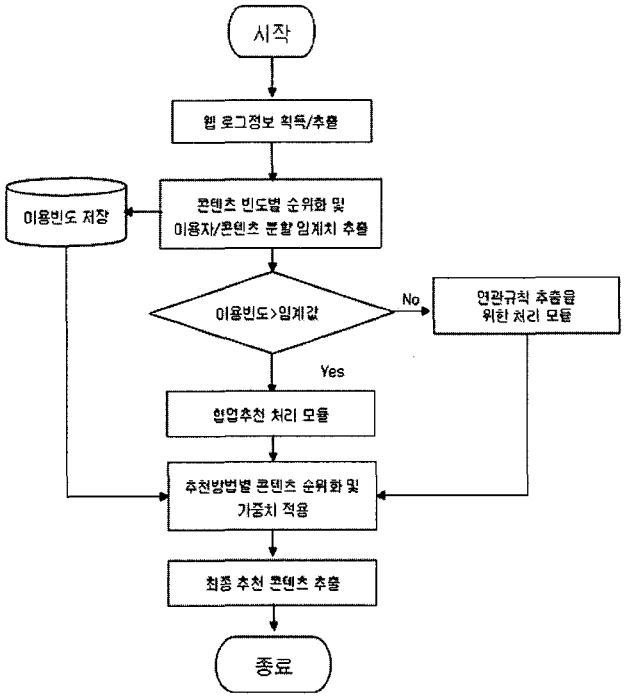
- 대용량의 멀티미디어 콘텐츠 및 이용자를 대상으로 추천서비스를 제공한다.
- 이용자 정보 및 콘텐츠 속성을 활용하지 않고 이용자의 콘텐츠에 대한 이용행위로서 이용빈도를 고려요소로 설정하였다.
- 연관규칙 추출을 위하여 콘텐츠의 이용빈도에서 특정 임계값 이상에 포함되는 일부 콘텐츠를 제거하고 나머지 콘텐츠만을 대상으로 활용하였다.
- 협업여과 추천방법을 위하여 연관규칙에 적용되지 않은 상위 이용빈도를 가지는 콘텐츠를 기반으로 유사이용자의 추출에 적용하였다.
- 많은 이용자에게 가장 많이 이용된 콘텐츠를 후보 추천 콘텐츠로 고려하였다.
- 최종 콘텐츠를 선정하는데 있어서 적용하고 있는 세 가지 방법에 대한 각각의 가중치를 실험을 통하여 추출하고 후보 추천 콘텐츠에 적용하여 순위별로 최종 추천 콘텐츠를 선정하였다.

3.1.2 콘텐츠 추천을 위한 단계

콘텐츠 추천을 위한 단계는 크게 협업여과 추천방법을 적용하여 추천 콘텐츠를 선정하는 단계와 연관규칙을 적용하여 콘텐츠를 추천하는 단계로 구분할 수 있으며 각각의 방법은 순차적이 아닌 병렬적(parallel)으로 진행된다. 단계별 방법과 각 단계별로 생성되는 결과물을 알아보면 다음과 같다.

먼저 이용자의 콘텐츠 이용빈도를 계산하기 위하여 이용자의 웹상의 행위 정보를 담고 있는 웹 로그를 전처리 과정을 통하여 분석하여 이용자 식별번호(user identification number), 콘텐츠 식별번호(contents identification number), 이용빈도(usage frequency) 등의 정보를 추출하고 이를 본 연구에서 필요한 형식으로 변환하여 데이터베이스에 저장하는 전처리 단계.

두 번째, 후보 콘텐츠 추천을 위하여 전체



〈그림 1〉 제안된 하이브리드 추천방법

실험데이터에서 각각의 콘텐츠에 대한 빈도수를 계산하여 이를 이용빈도에 따라 정렬하고 콘텐츠 및 이용자의 분할을 위한 임계값을 구하는 단계.

세 번째, 두 번째 단계에서 얻은 분할 임계값을 기준으로 전체 대상이 되는 콘텐츠와 이용자를 협업여과 추천방법과 연관규칙에 적용하기 위하여 분할하고 분할된 내용을 개별적으로 저장하는 단계.

네 번째, 연관규칙을 적용하여 후보 콘텐츠를 추출하는 과정으로서 임계값 이하 범위에 포함되는 콘텐츠를 통하여 빈발항목집합을 생성하고 이를 통하여 규칙을 추출하고 추출된 규칙을 통하여 추천 콘텐츠를 선정하는 단계.

다섯 번째, 세 번째 단계에서 추출된 임계값 이상에 포함되는 콘텐츠와 이용자를 분석하여 콘텐츠에 대한 이용자별 콘텐츠 이용빈도를 통하여 이용비율에 대한 벡터값을 추출하고 이를 통하여 이용자 유사도의 계산에 따른 유사이용자를 선정한다. 선정된 유사이용자에 대한 콘텐츠 이용성향을 기반으로 협업여과 추천방법을 통한 추천 콘텐츠를 선정하는 단계.

마지막 단계로서, 각 단계별로 선정된 후보 콘텐츠를 결합하여 최종 추천 콘텐츠를 선정하게 된다. 즉, 특정 임계값 이상의 최상위 이용빈도를 가지는 콘텐츠, 빈발항목집합과 함께 생성된 연관규칙을 통하여 추출된 콘텐츠, 협업여과 추천방법에 의하여 추출된 콘텐츠의 세 가지 방법을 결합하며 결합된 후보 추천 콘텐츠와 함께, 각각 적용된 추천방법에 적용하기

위한 가중치를 실험을 통하여 추출하고 이를 각각의 후보 추천 콘텐츠에 적용하여 정렬함으로써 최종의 추천 콘텐츠를 선정하는 단계로 구분할 수 있다.

〈그림 1〉은 본 연구에서 제안하고 있는 방법을 통하여 최종의 추천 콘텐츠를 선정하는 일련의 과정을 도식화하여 보여주고 있다.

3.2 후보 추천 콘텐츠 선정방법

후보 콘텐츠 추천과정에서는 최종 추천 콘텐츠의 선정을 위하여 본 연구에서 적용하고 있는 협업여과 추천방법, 연관규칙, 및 이용빈도별 통계적 추천의 세 가지 방법을 수행하고 각 방법에 따라 후보 추천 콘텐츠를 선정하는 단계라고 할 수 있다.

3.2.1 콘텐츠 및 이용자 집합의 분할

콘텐츠에 대한 이용자의 선호도를 측정하기 위하여 이용자의 콘텐츠 이용빈도에 대한 분석은 제안된 하이브리드 추천에 있어서 매우 중요한 방법론을 제공하고 있다. 이를 위하여 본 연구에서는 콘텐츠에 대한 이용빈도에 따른 분포를 분석하고 분할 임계값을 추출하여 협업여과 추천방법과 연관규칙 추천방법의 대상이 되는 콘텐츠와 이용자를 분할하였다. 콘텐츠와 이용자 분할을 위한 임계값의 추출은 콘텐츠 이용분포에서 급격한 변화를 보이는 구간을 선정하고 실험을 통하여 적합한 임계값을 선정한다.

제안된 하이브리드 추천방법에서는 이용량을 기준으로 이용자를 분리하고 이용빈도를 기준으로 콘텐츠를 분리하여 추천에 활용하였다. 이를 위하여 먼저 이용자를 분할하는데 있어서

HC(n)=이용량 상위 n%에 포함되는 콘텐츠 집합

HU(n)=전체 이용자 중에서 이용량이 상위 n%에 포함되는 이용자

전체 대상 이용자 중에서 콘텐츠에 대한 이용량이 많은 이용자(Heavy User)와 일반 이용량을 가지는 이용자(NU: Normal User)로 분리하였으며 다이용 이용자(HU)는 다음과 같이 정의할 수 있다.

예를 들어, HU(1)은 이용량 기준으로 상위 1%인 이용자를 의미하며 이 기준은 콘텐츠가 이용된 횟수가 아닌 콘텐츠를 이용한 이용자 수를 의미한다.

이용자 분할과 함께, 전체 대상이 되는 콘텐츠 집합에서 많은 이용자에 의해 이용량이 높은 콘텐츠와 일반 이용량을 가지는 일반 콘텐츠(NC: Normal Contents)로 분리할 수 있으며 다이용 콘텐츠(HC)는 다음과 같이 정의한다.

예를 들어, HC(1)는 이용량 기준으로 상위 1%의 범위에 포함되는 콘텐츠 집합이라고 할 수 있다.

이와 같이 이용자와 콘텐츠의 분할을 통하여 본 연구에서 제안하고 있는 하이브리드 추천방법을 통하여 개별 추천방법에 가장 적합한 대상을 한정하여 적용함으로써 해당 추천방법이 가지는 장점을 극대화하는 동시에 단점은 해결하고자 하는 목적을 달성할 수 있다. 한편, 추천서비스가 개인의 성향에 따라 적절한 콘텐츠

를 추천하는 목표를 가지고 있으나 전체 이용자 중에서 대다수의 이용자에게 많이 이용되는 콘텐츠의 경우 전체 이용자 그룹의 선호도를 반영한다고 할 수 있다. 따라서 본 연구에서는 최종 추천 콘텐츠 선정을 위하여 후보 추천 콘텐츠로서 전체 이용자 그룹의 선호도를 반영하는 다이용 이용자의 인기 콘텐츠를 추천에 반영하였다.

3.2.2 통계적 이용빈도 기반의 추천 콘텐츠 선정

본 연구에서는 특정 콘텐츠 분할 임계값 이상의 이용빈도를 가지는 콘텐츠를 다이용 콘텐츠로서 정의하고 있으며 이러한 다이용 콘텐츠는 통계적으로 이용자에게 많이 이용된 콘텐츠를 의미하는 것으로 전체 이용자 집합의 선호도를 반영한다고 할 수 있다. 그러나 추천의 목적이 개인별 성향에 적합한 콘텐츠를 추천하는 것이기 때문에 전체 이용자 집합의 선호도를 반영한다는 것은 동전의 양면과 같은 특징을 가지고 있다고 할 수 있다. 즉, 다이용 콘텐츠 중에서 많은 부분이 특정 기간에만 일시적으로 이용자에게 이용되는 경향이 있다. 그러나 이러한 다이용 콘텐츠는 전체 이용자 집합의 선호도를 보여주고 있다는 측면에서 분명한 추천서비스의 효과를 가지고 있다. 따라서 본 연구에서는 이러한 다이용 콘텐츠가 가지고 있는 특성을 고려하여 제안된 하이브리드 추천방법에서 후보 추천 콘텐츠를 추출하는 방법 중의 하나로 고려하는 동시에 최종 콘텐츠를 선정하는 결합추천과정에서 다이용 콘텐츠에 대한 가중치를 조절함으로써 이를 해결하고자 하였다. 한편, 제안된 하이브리드 추천방법에서 통계적 이용빈도에 기반한 후보 추천 콘텐츠를 추출하는데 있어서 이용자 선호도는 단순 이용빈도가 아닌 최상위 이용빈도를 가지는 콘텐츠에 대한 상대비율로 표현한다.

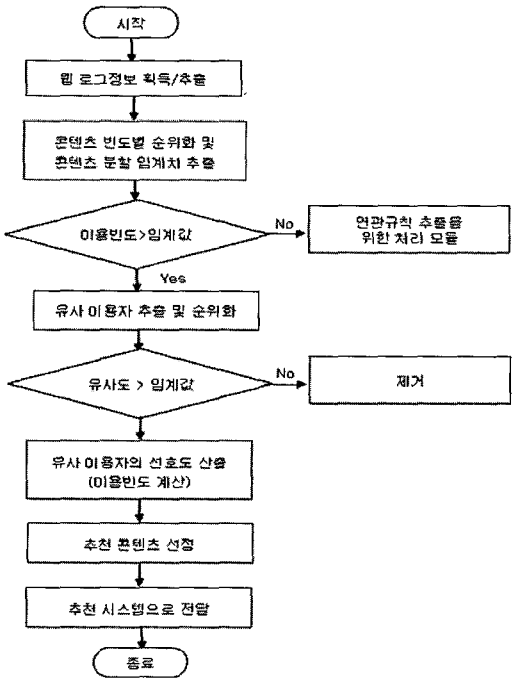
$$Popilarity(c)=\frac{\text{이용빈도}}{\text{최상위 콘텐츠의 이용빈도}} \quad (1)$$

3.2.3 협업여과 추천방법 기반의 추천 콘텐츠 선정

협업여과 추천방법은 콘텐츠 이용에 있어 해당 이용자와 비슷한 경향을 보여주는 유사이용자를 추출하여 유사이용자의 성향에 따라 추천 콘텐츠를 선정하는 방법으로 추천 콘텐츠를 선정하기 위하여 이용자에 대한 유사이용자 선정을 위한 이용자 간의 유사도 측정이 선행된다. 이렇게 선정된 유사이용자는 추천대상이 되는 이용자와 콘텐츠 이용경향에 있어서 비슷한 유형을 보인다는 가정 하에 유사이용자의 콘텐츠에 대한 선호도를 분석하고 이를 통하여 해당 이용자의 선호도 값에 대한 예측이 이루어진다. 〈그림 2〉에서는 본 연구에서 제안하고 있는 협업여과 추천방법에 따른 후보 콘텐츠 추천 흐름을 보여주고 있다.

1) 유사도 계산

이용자 간의 유사도를 계산하기 위해서는 이용자의 콘텐츠에 대한 선호도를 알아야 한다. 이를 위하여 본 연구에서는 연구의 대상으로 고려하고 있는 음악 콘텐츠에 대한 이용자의 선호도를 이용자의 콘텐츠 이용빈도로 고려하고 있다. 이러한 이용자 선호도를 기반으로 다이용 콘텐츠 벡터(HCV)는 다이용 콘텐츠(HC)에 대한 이용자의 선호도를 정규화된 벡



〈그림 2〉 제안된 협업여과 추천방법을 통한 추천 콘텐츠 선정

터(normalized vector)로 표현한 것으로 다이용 콘텐츠의 수를 "k"라 정의하고, 다이용 콘텐츠(HC)에 대하여 이용자의 선호도를 표현하는 이용빈도를 기준으로 순위를 정한다. 이때, 순위는 해당 콘텐츠에 대한 이용자의 선호도를 기준으로 하여 내림차순으로 정렬한다. 한편 정규화하지 않은 다이용 콘텐츠 벡터(HCV_bn : HCV before normalization)는 다음과 같이 정의할 수 있다.

$HCV_{bn}(u)=\{w_1,w_2,w_3,...,w_n\}$
w_k=(해당 이용자의 i번째 다이용 콘텐츠 이용 횟수)

위의 정의를 기준으로 벡터의 길이는 다이용 콘텐츠의 수와 같으며 HCV_{bn}을 정규화를 하게 되면 모든 이용자에 대해서 다음을 만족한다.

$$HCV=\frac{HCV_{bn}}{\|HCV_{bn}\|}=\frac{1}{norm(HCV_{bn})}(HCV_{bn}) \quad (2)$$
$$=\left(\frac{w_1}{\sqrt{w_1^2,w_2^2,\ldots,w_n^2}}+\frac{w_2}{\sqrt{w_1^2,w_2^2,\ldots,w_n^2}}+\ldots+\frac{w_i}{\sqrt{w_1^2,w_2^2,\ldots,w_n^2}}\right)$$

따라서 항상 $\|HCV\|=1$이 된다.

이러한 사전과정을 수행한 다음 다이용 콘텐츠에 대한 상대적 이용벡터(HCV)를 통하여 아래와 같이 내적계수(inner product) 방법을 통하여 쉽게 유사도를 계산할 수 있다.

$$Sim(u_1,u_2)=HCV(u_1)\cdot HCV(u_2) \quad (3)$$

2) 유사이용자 추출

본 연구에서는 이용자를 표현하는데 있어서 특정 기간 동안에 이용자가 이용한 콘텐츠 집합과 해당 콘텐츠에 대한 이용빈도로써 표현하고 있다. 이를 위하여 특정 이용자(u)에 대한 이용내역(usage)은 다음과 같이 표현된다.

$$Usage_{bn}(u)=\{c_1,c_2,c_3,...,c_n\}$$

위의 정의에서 는 콘텐츠의 인식번호를 의미한다. 한편, 이용자(u)가 콘텐츠를 이용하는 경우 해당 이용자가 이용한 콘텐츠는 다이용 콘텐츠에 대한 상대적 이용벡터(HCV : Heavy Contents Vector)로서 아래와 같이 표현할 수 있으며 는 해당 콘텐츠의 이용빈도에 대한 정규화된 이용벡터를 의미한다.

$$HCV(u)=\{u_1,u_2,u_3,...,u_n\}$$

위에서 정의하고 있는 다이용 콘텐츠 벡터(HCV)를 기준으로 유사이용자를 추출한다. 그런데, 기존 연구에서는 유사이용자를 추출의 대상을 전체 이용자를 대상으로 유사이용자를 선택하였으나 본 연구에서는 이용빈도가 높은 이용자(HU) 중에서만 선택하게 된다.

위의 방법을 통하여 추출한 이용자(u)에 대한 유사이용자 목록(RelUserList(u))은 아래와 같이 정의할 수 있다.

$RelUserList\ (u)=\{(u_1,s),(u_2,s),(u_3,s),...,(u_n,s)\}$
u_k=다이용 이용자중 상위 유사도를 갖는 k명
S=해당 다이용 이용자와의 유사도

$COL\ (u)=\{(c_1,s),(c_2,s),(c_3,s),...,(c_n,s)\}$
c_k=유사이용자의 후보 추천 콘텐츠
S=해당 콘텐츠의 선호도

3) 후보 추천 콘텐츠 추출

유사이용자 집합에 포함되는 이용자들이 이용한 콘텐츠 목록을 결합하여 후보 추천 콘텐츠를 추출한다. 이때 대상 콘텐츠는 다이용 이용자들이 이용한 모든 콘텐츠가 아닌 일반 이용빈도를 가지는 콘텐츠(NC) 중에서만 고려한다. 위의 과정을 통하여 유사이용자 성향에 기반한 후보 추천 콘텐츠(COL)는 다음과 같이 정의된다.

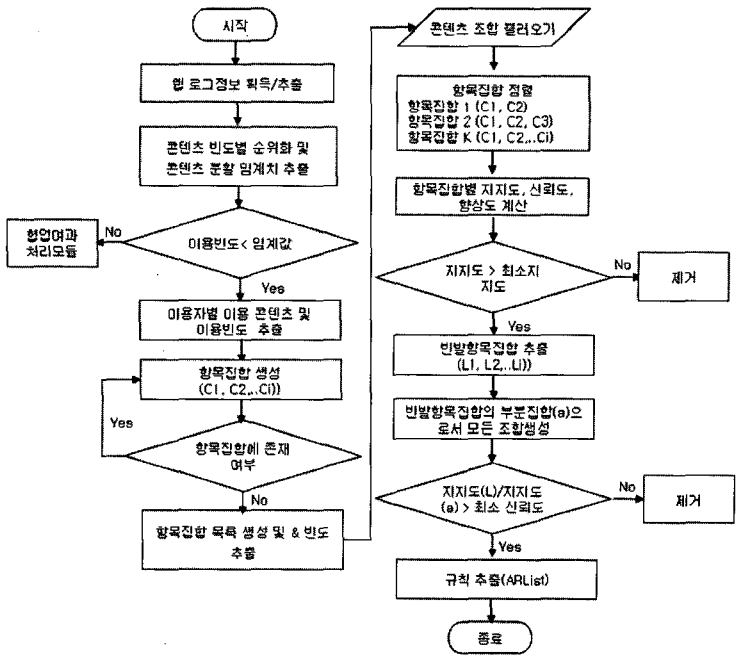
〈그림 3〉은 유사이용자의 성향을 기반으로 추천 콘텐츠를 선정하는 협업여과 추천방법에 기반한 후보 추천 콘텐츠(COL)에 대한 선호도를 계산하기 위한 방법을 설명하고 있다.

3.2.4 연관규칙 기반의 추천 콘텐츠 선정

본 연구에서는 연관규칙의 장점을 살리면서 대용량 환경에서 적합하게 적용하기 위하여 대상 콘텐츠에 대한 제한을 통한 개선된 방법을 적용하고 있다. 제안된 방법을 통하여 규칙을 추출하고 이를 기반으로 후보 추천 콘텐츠를 선정하는 방법은 다음과 같다. 먼저, 전체 이용 로그에서 이용빈도가 높은 콘텐츠 집합(HC)을 제외한 로그를 준비한다. 둘째, 해당 로그에서 Apriori 알고리즘을 적용하여 각각의 콘텐츠 집합에서 최소 지지도를 만족하는 빈발 항목집합을 추출한다. 셋째, 이렇게 추출된 빈

```
Algorithm Calculate_COL
    COL(u) = {}
    RelUserList(u)에 포함된 유사이용자(u) 및 유사이용자의 이용내역
     (Usage(u))내의 모든 콘텐츠(c_k)에 대해서 다음을 반복한다.
        COL 내에 c_n가 없으면, (c_n, s)를 추가한다.
        COL 내에 c_n가 포함되어 있으면, 기존의 값에 s를 더한 값으로
       변경한다.
    Return COL
```

〈그림 3〉 협업여과 추천방법에서의 후보 추천 콘텐츠 추출방법



〈그림 4〉 제안된 연관규칙 기반의 콘텐츠 추천흐름도

발항목집합을 기준으로 연관규칙을 추출하고, 추출된 연관규칙을 저장한다. 추출된 연관규칙을 통하여 해당 콘텐츠(c)에 대한 관계성이 높은 후보 추천 콘텐츠(Related(c))가 추출된다. 그러나 연관 콘텐츠(Related(c))는 공집합일 수 있다. 한편, 연관 콘텐츠(Related(c))는 다음과 같이 정의된다.

$Related(c)=(c_1,m),(c_2,m),\ldots,(c_n,m)$

여기서 "c_k"는 콘텐츠에 부여된 식별번호(identification number)를 의미하며 "m"은 해당 콘텐츠와의 연관성을 의미하는 수치값(measures of relation)으로서 본 연구에서는 연관성에 대한 수치값으로서 항목에 대한 지지도를 적용하였다. 〈그림 4〉는 본 연구에서의 연관규칙을 추출하는 과정을 보여주고 있으며 생성된 규칙을 통하여 콘텐츠를 선정한다.

1) 연관성 및 규칙 추출

위에서 언급한 바와 같이 이전의 연구에서는 콘텐츠간의 연관성 추출을 위하여 데이터마이닝의 연관성 추출기법을 그대로 활용했으나, 본 연구에서는 효과적인 추천을 위하여 다음과 같이 변형된 연관성 추출기법을 사용한다.

하나, 선행부와 결과부에는 각각 하나의 항목집합만을 가지는 규칙이 생성된다. 그러나 결과부 또는 선행부에 두 개 이상의 항목을 가지는 규칙은 허용되지 않는다.

둘, 규칙을 추출하는데 있어서 향상도가 1 이상인 것만 추출한다. 이를 통하여 통계적으로 의미가 있는 규칙만을 추출할 수 있다.

셋, 추출된 연관규칙 중에서 하나의 동일한 선행부에 대응되는 결과부의 경우의 수를 제한한다. 여기서 결과부에 나타날 수 있는 항목에 대한 경우의 수를 결과부 제한수(consequence restriction number)라고 정의한다. 즉, "CRN = 3"인 경우에 있어서는 동일한 선행부에 대하여 4개의 규칙이 추출될 수 없다.

한편, 연관규칙의 추출을 위하여 먼저 빈발항목집합을 추출하여야 하며 해당 빈발항목집합을 추출하는데 있어서 기준으로 연관규칙의 측정도구 중에서 신뢰도(support)를 적용한다. 따라서 위의 네 개의 규칙들 중에서 세 개만이 선택되어야 하며 신뢰도가 가장 작은 규칙은 제외된다.

2) 연관규칙에 기반한 후보 추천 콘텐츠 추출

연관규칙 목록(ARList)을 통하여 콘텐츠 간의 상호 연관성을 통하여 추천 콘텐츠를 선정하는 과정으로서 추천대상 이용자가 과거에 이용한 콘텐츠 이용내역을 기반으로 해당 이용자의 콘텐츠에 대한 선호도를 콘텐츠 간의 연관성을 통하여 해당 이용자에게 추천하는 방법을 이용한다. 본 연구에서는 추천대상 콘텐츠를 일반 이용빈도를 가지는 콘텐츠(NC)를 대상으로 한다. 콘텐츠 간의 연관성에 대한 기준은 연관규칙에서의 지지도가 된다. 이를 위하여 콘텐츠 간의 상호 연관성을 분석하여 추천하는 연관규칙 기반의 후보 추천 콘텐츠(REL)는 다음과 같이 정의된다.

$REL(u)=\{(c_1,s),(c_2,s),(c_3,s),\ldots,(c_n,s)\}$
c_k=연관규칙 기반의 후보 추천 콘텐츠
S=해당 콘텐츠의 선호도

추천 콘텐츠를 선정하기 위하여 해당 이용자가 이용한 콘텐츠들의 목록(Usage(u))과 추출된 연관규칙을 통하여 생성된 연관규칙 목록(ARList)을 기반으로 〈그림 5〉의 과정을 반복하면서 후보 추천 콘텐츠(REL)에 대한 선

```
Algorithm Calculate_REL
    REL = { }
    Usage(u)에 포함된 각 콘텐츠(c_k) 및 연관 콘텐츠 (Related(c_j)) 내의
    (c_n, m) 각 항목에 대해 다음을 반영한다.
        REL내에 c_j가 없으면, (c_n, m)을 추가한다.
        REL내에 c_j가 포함되어 있으면, 기존의 값과 비교하여 최대값
      (max)을 취한 m 으로 변경한다.
    Return REL
```

〈그림 5〉 연관규칙 기반 후보 콘텐츠 추출방법

호도를 생성한다.

3.3 최종 추천 콘텐츠 선정 방법

본 연구에서는 최종적으로 이용자에게 추천될 추천 콘텐츠를 선정하기 위하여 협업여과 추천방법, 연관규칙 및 통계적 이용빈도의 세 가지 방법을 혼합한 하이브리드 추천방법을 제안하고 있다.

3.3.1 추천 정책

최종적으로 이용자에게 추천되는 콘텐츠를 선정을 위하여 먼저 후보 콘텐츠 추천과정을 통하여 추출된 후보 추천 콘텐츠는 결합과 순위화 과정을 통하여 최종의 추천 콘텐츠를 선정하게 된다. 최종 콘텐츠를 선정하는데 있어서 각 후보 추천 콘텐츠에 할당된 선호도가 동일한 기준에서 생성된 경우에는 선호도에 따른 순위화가 가능하다. 그러나 본 연구에서와 같이 개별 추천방법별 선호도에 대한 기준값은 서로 상이하고, 또한 선호도를 표현하는 수치값에 대한 표현도 다르기 때문에 각각의 추천방법을 통하여 선정된 후보 추천 콘텐츠의 결합과 순위화를 위한 방법이 필요하다. 이를 해결하기 위하여 본 연구에서는 결합추천방법(weighted combination recommendation method)을 제안하고 있다. 즉, 최종의 추천 콘텐츠를 선정하기 위하여 현실적으로 개별 추천방법별로 추출된 후보 추천 콘텐츠에 대한 순위화가 요구된다. 예를 들어, 협업여과 추천방법에만 나타나는 콘텐츠가 있을 수가 있고, 협업여과 추천방법과 연관규칙에 동시에 나타나는 콘텐츠도 있을 수가 있다, 두 가지 방법에 동시에 나타나는 콘텐츠의 경우에는 집합 내부에서 순위화 작업이 필수적이라고 할 수 있다. 따라서 결합추천과정에서 콘텐츠에 대한 순위화의 과정은 실질적인 추천 콘텐츠 선정을 위한 과정이라고 할 수 있다.

3.3.2 결합추천방법

결합추천방법은 후보 추천 콘텐츠 선정과정을 통하여 추출된 후보 콘텐츠를 추천방법별

가중치와 순위화를 통하여 결합하는 방법이라고 할 수 있다. 이러한 결합추천과정을 정리하면 다음과 같다.

첫째, 세 가지의 후보 추천 콘텐츠를 선정하는 방법에 따라 동일 개수의 후보 추천 콘텐츠를 선정한다.

둘째, 개별 방법별로 선정된 후보 추천 콘텐츠는 각각 선호도 추천의 기준값에 따른 선호도를 가지고 있으며 이를 선호도에 따라 적용된 추천방법별로 정렬한 다음, 해당 선호도에 대한 순위를 합산한다.

셋째, 실험을 통하여 세 가지의 추천방법별로 추천에서의 중요도에 따라 가중치를 선정하고, 선정된 가중치를 산출된 순위에 곱연산을 수행한다.

마지막으로 산출된 값을 기준으로 콘텐츠를 내림차순으로 정렬하고 순위에 따라 최종 추천 콘텐츠를 선정한다.

한편, 후보 콘텐츠 추천과정을 통하여 얻은 정보에는 연관규칙 목록(ARList), 최상위 인기콘텐츠 목록(PopList), 유사이용자 목록(RelUserList)이 있다. 이러한 세 가지의 목록을 통하여 추출된 후보 추천 콘텐츠는 아래와 같다.

- 이용자의 콘텐츠 이용빈도에 있어서 특정 임계값 이상에 포함되는 상위 이용빈도를 가지는 다이용 콘텐츠(POP)
- 연관규칙 처리모듈에서 선정된 콘텐츠로서 해당 이용자가 이용한 콘텐츠와 연관성이 있는 콘텐츠(REL)
- 협업여과 처리모듈에서 선정된 콘텐츠로서 해당 이용자의 관련 다이용 이용자가 이용한 콘텐츠(COL)

최종 추천콘텐츠를 선정하기 위한 결합추천 방법은 〈그림 6〉과 같다.

한편, 본 연구에서 최종 추천 콘텐츠를 선정하는 것은 후보 콘텐츠 추천과정에서 개별 추천방법에 따라 추출된 콘텐츠 집합 "POP", "REL", "COL"을 결합하는 것이다. 이를 위하여 개별 콘텐츠 집합별로 순위화를 수행한 다음에 개별 콘텐츠 집합에 대하여 추천방법별 가중치를 적용해야 한다. 또한 순위를 기준으로 판단하기 위해서는 각 콘텐츠 집합별로 고려할 콘텐츠 개수는 동일하여야 한다. 이렇게 개별 후보 추천 콘텐츠에 공통으로 적용되는 제한수를 "M" 이라고 정의한다. 개별 추천방법에 적용되는 가중치에 대한 정의는 다음과 같다.

- 최상위 이용빈도를 가지는 콘텐츠 집합에 대한 가중치(wp) : 다이용 콘텐츠 목록(POP)에 대한 가중치로서, 다이용 콘텐츠를 최종 추천 콘텐츠로 선정하는데 있어서 반영하는 정도를 나타낸다.
- 협업여과 추천방법 기반 추출 콘텐츠집합에 대한 가중치(wc) : 유사이용자를 기준으로 추출된 후보 추천 콘텐츠(COL)에 대한 가중치로서, 최종 추천 콘텐츠를 선정하는데 있어서의 반영도를 나타낸다.
- 연관규칙 기반 추출 콘텐츠집합에 대한 가중치(wr) : 연관규칙을 기준으로 추출된 후보 추천 콘텐츠(REL)에 대한 가중치로서 콘텐츠 간의 연관성을 기준으로 최종 추천 콘텐츠를 선정하는데 있어서 반영하는 정도를 나타낸다.

```
Algorithm COMBINATION
        IN : POP, REL(u), COL(u), 최종 콘텐츠 반환개수 N
        OUT : RECOM, ordered contents list
begin
        RECOM_CON = POP, REL, COL 내에 포함된 콘텐츠의 합집합
        (위에 기술한 대로, 각각 M 개만 고려된다)
        RECOM_WGT = {} // 가중치를 포함한 집합, 중간 계산에만 사용
        For all RECOM_CON내의 모든 콘텐츠(c_n)에 대해서 다음을 계산한다.
        begin
                POP에서 c_n의 순위를 o_pop, REL 에서 c_n의 순위를 o_rel,
                COL에서 c_n의 순위를 o_col 라고하면 c_n의 결합순위는 다음과 같다.
                        o_recom = (wp)o_pop + (wr)o_rel + (wc)o_col
                        단, 이 때 해당 집합에서 해당 콘텐츠가 없다면, 순위는
                        M+1 이 된다.
                RECOM_WGT에 (c_n, o_recom)을 추가한다.
        end
        o_recom이 적은 N을 선택하여 해당 콘텐츠를 RECOM에 순서대로 추가 한다.
end
```

〈그림 6〉 최종 추천 콘텐츠 선정을 위한 결합추천방법

4. 콘텐츠 추천 실험 및 평가

4.1 실험환경 및 데이터

본 연구에서는 해당 사이트에서 14일간의 데이터베이스에 기록된 이용자의 음악 콘텐츠에 대한 총 72,074,534건의 이용로그를 수집하였으며 이를 기반으로 이용자의 음악 콘텐츠를 이용하는 행위 즉, 콘텐츠에 대한 이용빈도를 추출하였다. 여기에서 전반 7일간의 36,007,553건의 이용내역은 콘텐츠 추천을 위한 실험에 적용하고 후반 7일간의 36,066,981건의 이용내역은 제안된 하이브리드 추천방법을 기반으로 구축된 추천시스템에 대한 검증에 적용하였다. 한편, 본 연구에서 정의하고 있는 이용자의 콘텐츠 이용행태에는 음악 콘텐츠를 사이트에서 직접 감상하거나 이용자의 컴퓨터에 내려받기(download)를 수행하는 두 가지 이용행위를 의미한다.

실험을 위한 환경에 있어서 제안된 하이브리드 추천방법을 통하여 후보 콘텐츠 추출 및 결합추천을 수행하는 추천시스템의 구현 및 평가

는 Windows XP Professional운영 체제에서 MS SQL Server 2000의 데이터 처리용 DBMS를 가지고 Visual Studio .NET, C++/STL의 개발환경에서 Pentium 3.0GHz/800MHz/2MB 환경에서 수행 되었다.

4.1.1 평가 방법

본 연구에서는 기존의 협업여과 추천방법과 연관규칙 기반 추천방법 및 통계적으로 이용빈도에 기반한 다이용 콘텐츠의 추천을 수행하는 추천방법을 혼합한 하이브리드 추천을 위한 기반구조와 방법론을 제안하였으며 제안된 추천방법을 기반으로 추천시스템을 구현하였다. 구축된 하이브리드 추천시스템에 대한 성능을 평가하고 검증을 수행하기 위하여 정확률 및 재현율과 함께, 본 연구에서는 성공률과 NC 포함률을 평가척도로서 활용하였다.

성공률(success ratio)은 검증을 위하여 이용자에게 추천된 콘텐츠 수에 관계없이 추천된 콘텐츠 중에서 적어도 한 개 이상 이용자에게 이용된 경우에 이를 추천에 대한 성공을 의미한다. 따라서 이용자에 대한 추천에 있어서 성공률은 1과 0의 값만 갖는다.

NC 포함률(normal contents ratio)은 추천 콘텐츠에서 다이용 콘텐츠가 아닌 일반 이용빈도를 가지는 일반 콘텐츠가 포함된 비율을 의미하는 척도로서 제안된 하이브리드 추천방법의 성능평가에 있어서 중요한 의미를 담고 있다.

재현율은 검증데이터가 되는 후반부 7일 간의 이용내역으로부터 추출한 추천대상 이용자가 이용한 콘텐츠에 대하여 실험데이터가 되는 전반부 7일간의 이용내역에서 추출된 추천 콘텐츠 중에서 이용자에 의해 이용된 콘텐츠에 대한 비율을 의미하는 것으로서 수식 4와 같이 표현 할 수 있다.

$$\text{재현율}(recall) = \frac{\text{추천 콘텐츠 중에서 이용자가 이용한 콘텐츠 수}}{\text{검증데이터에서 이용자가 이용한 콘텐츠 수}} \quad (4)$$

정확률은 실험데이터를 통하여 추출된 추천 콘텐츠에 대하여 이용자가 이용한 콘텐츠에 대한 비율로써 수식 5와 같이 표현된다.

$$\text{정확률}(Precision) = \frac{\text{추천 콘텐츠 중에서 이용자가 이용한 콘텐츠 수}}{\text{이용자에게 추천된 콘텐츠 수}} \quad (5)$$

이러한 재현율과 정확률 값은 서로 반비례의 관계에 있으므로 적절한 조정과정이 필요하다. Lewis 등(1994)은 재현율과 정확률의 문제점을 보완하기 위하여 재현율과 정확률을 결합한 F-Measure 개념을 제안하였다.

$$F_\beta = \frac{(\beta^2+1)*presion*recall}{\beta^2*(presion+recall)} = \frac{2(presion*recall)}{presion+recall} \quad (6)$$

β는 정확률과 재현율 값의 중요도에 따른 가중치를 나타낸다. 본 연구에서는 제안된 하이브리드 추천방법의 정확성을 검증하기 위하여 $\beta = 1$ 즉, 정확률과 재현율 값에 동일한 가중치를 적용하여 추천의 성능을 평가하였다.

〈표 2〉 다이용 콘텐츠 비율(HC ratio) 변경 실험결과

HC ratio	HU 수	HC 수	최대 유사 이용자수	전체 누락 유사이용자	전체 유사 이용자 수	유사 이용자가 없는 이용자수	연관규칙 대상 콘텐츠 수	연산 시간
0.01	344	1,891	18	633,542	5,564,038	20,508	3,444,298	167.52
0.02	344	3,783	18	622,007	5,575,573	17,638	2,449,089	174.06
0.05	344	9,458	18	615,904	5,581,676	15,868	1,375,694	171.81
0.1	344	18,916	18	614,550	5,583,030	15,461	823,069	174.50
0.2	344	37,833	18	614,142	5,583,438	15,333	439,689	174.12

4.2 분할 임계값 및 후보 추천 콘텐츠 추출

4.2.1 이용자 및 콘텐츠 분할 임계값 추출

1) 콘텐츠 분할 임계값 변경 실험

이용빈도를 기준으로 콘텐츠를 분할하는 목적은 크게 두 가지로 구분할 수 있다. 첫 번째는 콘텐츠에 대한 이용자의 선호도를 반영하고 있는 이용빈도에 따른 후보 추천 콘텐츠를 추출하기 위한 것이다. 두 번째는 협업여과 추천방법과 연관규칙 기반의 추천방법에 적용하기 위한 콘텐츠 집합을 분할하기 위한 것이다. 본 실험에서는 전체 실험대상이 되는 콘텐츠에서 다이용 콘텐츠(HC) 비율을 콘텐츠 이용분포에서 급격한 변화를 보여주는 구간 내에서 1%, 2%, 5%, 10%, 20%로 선정하여 실험을 수행하였다. 한편, 고정값으로서 다이용 이용자 비율은 0.1%, 유사이용자의 비율은 다이용 이용자에 대하여 5%의 비율로 설정하였다.

〈표 2〉에서 볼 수 있듯이 다이용 콘텐츠의 비율이 높아지면서 연관규칙의 대상이 되는 콘텐츠의 수가 300만 이상에서 100만 이하로 완만하게 줄어든다. 한편, 유사이용자 수에 있어서 다이용 콘텐츠의 비율이 증가하면, 유사이용자를 추출하기 위한 이용벡터의 길이가 늘어나기 때문에 추천대상 이용자가 이용한 콘텐츠와 유사이용자와의 공통 콘텐츠가 누락되는 경우가 적어지게 된다. 따라서 이용자 개인당 유사이용자 수는 미세하나마 증가되는 효과가 있다. 한편, 연산시간에 있어서 다이용 콘텐츠 비율(HC ratio)이 변화하더라고 연산시간은 급격한 변동이 없다.

2) 이용자 분할 임계값 변경 실험

이용자 분할은 협업여과 추천방법에서 유사이용자 추출과 밀접한 관련이 있다. 따라서 이용자 분할에서 적절한 분할 임계값의 추출은 최종 추천 콘텐츠를 선정하기 위한 전단계로서 협업여과 추천방법을 적용하여 후보 콘텐츠를 추출하는데 있어서 직접적인 영향을 주게 되며 결과적으로는 추천의 정확성 및 연산시간에서의 밀접한 영향을 줄 수 있다. 이를 검증하기 위하여 본 연구에서는 다이용 콘텐츠 추출실험

〈표 3〉 다이용 이용자 비율(HU ratio) 실험결과

HU 비율	HU 수	HC 수	최대 유사 이용자	전체 누락 유사이용자	전체 유사이용자	유사 이용자가 없는 이용자	평균 유사 이용자 수	연산 시간
0.0001	34	189	4	163,705	1,213,535	48,046	3.52	77.92
0.0005	172	189	18	853,779	5,343,801	41,120	15.52	118.58
0.001	344	189	34	1,665,089	10,041,451	41,038	29.16	176.17
0.002	688	189	68	3,315,644	20,097,436	41,038	58.37	311.27
0.005	1721	189	172	8,410,287	50,811,033	41,038	147.57	1179.44

과 같은 환경 하에서 다이용 콘텐츠의 비율을 0.1%와 유사이용자의 비율을 다이용 이용자에 대하여 10%의 비율로 고정하고, 다이용 이용자 비율을 각각 0.01%, 0.05%, 0.1%, 0.2%, 0.5%로 조절하여 실험을 수행하였다.

〈표 3〉에서는 다이용 이용자 비율 조절에 따른 유사이용자 수와 누락된 유사이용자 수에 대한 변화를 보여주고 있다. 특히, 주목할 수 있는 부분은 다이용 이용자 비율(HU ratio) 조절에 따라 유사이용자의 수에 있어서 급격한 변화가 발생된다는 것을 알 수 있다. 이러한 이유는 다이용 이용자의 수가 적을 때는 이용자는 전체 다이용 이용자와 공통으로 이용한 콘텐츠가 적기 때문에 유사이용자가 적게 추출되지만 다이용 이용자가 많아지면 연관성 있는 유사이용자의 증가가 단순히 산술적으로 증가하는 것이 아닌 배수의 형태로 증가하기 때문이라고 판단할 수 있을 것이다.

3) 유사이용자 비율 조절 실험

유사이용자 추출에 있어서의 비율조절은 추천 대상 이용자에 대한 최대 유사이용자 수를 조절하는 것으로서 본 실험에서 의미하는 유사이용자 비율은 다이용 이용자 중에서 유사이용

〈표 4〉 유사이용자 비율(RelUser ratio) 실험결과

유사이용자 비율	HU 수	HC 수	최대 유사이용자	전체 누락된 유사이용자	전체 유사이용자	유사이용자가 없는 이용자	연산시간
0.01	344	189	3	123,200	909,730	41,038	139.67
0.02	344	189	6	247,127	1,818,733	41,038	134.75
0.05	344	189	17	750,311	5,102,959	41,038	159.72
0.1	344	189	34	1,665,089	10.041.451	41,038	176.17
0.2	344	189	68	3,821,982	19,591,098	41,038	281.31
0.5	344	189	172	14,814,591	44,406,729	41,038	560.13

자를 추출하는 것을 의미한다. 따라서 전체 이용자 344,410에서 다이용 이용자 비율(HU ratio)이 0.001이고 유사이용자 비율(RelUser ratio)이 0.1이라고 가정하면 다이용 이용자는 344명이 되며 여기서 최종 유사이용자의 수는 34명이 된다는 것을 의미한다. 그러나 전체 이용자를 대상으로 유사이용자를 선정하기 때문에 위의 가정에서 산술적으로 추출할 수 있는 유사이용자의 총 수는 (344,410 * 34 = 11,706,540)이 된다. 본 실험에서는 다이용 이용자 비율을 0.1%와 다이용 콘텐츠의 비율을 0.1%로 설정하고 유사이용자의 비율을 1%, 2%, 5%, 10%, 20%, 50%로 변경하여 실험을 수행하였다. 해당 조건에서 수행한 실험결과를 〈표 4〉에서 보여주고 있다.

4.2.2 후보 추천 콘텐츠 추출을 위한 대상 목록 추출

후보 추천 콘텐츠 추출은 최종의 추천 콘텐츠를 추출하기 위한 전단계로서 본 연구에서 적용된 각각의 추천방법별로 후보 추천 콘텐츠를 추출하기 위한 추천방법에 따른 목록을 선정하는 과정이다. 본 실험을 통하여 추출되는 결과물은 각각의 추천방법에 따라 다이용 콘텐츠 목록(PopList), 연관규칙 목록(ARList), 및 유사이용자 목록(RelUserList)이다. 추출된 목록을 기반으로 후보 추천 콘텐츠가 선정된다.

1) 다이용 콘텐츠 목록 추출

다이용 콘텐츠는 실험의 대상이 되는 전체 이용자 집합의 선호도를 반영한다고 할 수 있다. 〈표 5〉는 HC(0.1)을 기준으로 추출된 결과로서 전체 콘텐츠 189,169에서 총 189개의 콘텐츠가 추출되며 콘텐츠 번호가 140496인 콘텐츠가 최상위 이용빈도를 가지기 때문에 해당 콘텐츠의 이용빈도를 1로 산정하고 나머지 콘텐츠에 대한 이용빈도를 상대적 비율로 계산된 결과 테이블이다.

〈표 5〉 다이용 콘텐츠 목록 추출결과

콘텐츠 번호	선호도(popularity)
⋮	⋮
135850	0.502802
140243	0.310997
140496	1
140569	0.0723915
143187	0.279007
⋮	⋮

2) 유사이용자 목록 추출

유사이용자를 다이용 이용자(HU)에서 추출하기 때문에 유사이용자를 추출하기 위해서 먼저 다이용 이용자(HU)를 추출하여야 한다. 추출된 다이용 이용자 목록으로부터 유사이용자를 추출하기 위하여 최대 유사이용자 수를 결정하여야 하며 이를 위하여 최대 유사이용자 수를 3, 8, 18, 34, 68, 172로 한정하여 실험을 수행하였다. 유사이용자 추출을 위하여 다이용 이용자 비율(HU ratio), 다이용 콘텐츠 비율(HC ratio) 및 다이용 이용자 중에서 유사이용자의 추출 비율(RelUser ratio)에 대한 임계값을 설정함으로써 추천대상 이용자

〈표 6〉 추출된 유사이용자 목록

이용자 ID	유사이용자 ID	유사도(similarity)
608870	910412	0.404916
608870	802545	0.344673
608870	695681	0.344552
608870	617136	0.309837
608870	987326	0.303697
608870	609301	0.273162
608870	676407	0.271430
608870	869104	0.261425
608870	799714	0.255469
608870	823675	0.235441
608871	975079	0.214397
.	.	.
.	.	.
.	.	.

와 유사이용자에 대한 유사도가 계산된다.

한편, 〈표 6〉은 다이용 이용자 비율(HU(0.03)), 다이용 콘텐츠 비율(HC(2))과 다이용 이용자에서 0.1의 비율로 유사이용자를 추출한 결과테이블로서 이용자 간의 유사도 값을 포함하고 있다. 위의 설정값을 기준으로 각각의 추천대상 이용자별로 추출할 수 있는 최대 유사이용자 수는 10명이 된다.

3) 연관규칙 목록 추출

연관규칙을 통하여 후보 추천 콘텐츠를 추출하기 위하여 먼저 콘텐츠 분할 임계값을 기준으로 일반 이용빈도를 가지는 콘텐츠(NC)를 추출하고 해당 콘텐츠에 대한 빈발항목집합을 생성한다. 본 연구에서는 빈발항목집합을 기준으로 연관규칙을 생성함으로써 추천 후보 콘텐츠를 생성하게 된다. 입력데이터에는 규칙추출을 위한 최소 지지도, 최소 신뢰도 및 최대 빈발항목집합에 대한 개수를 포함하는 설정값과 실험데이터를 포함하는 원본파일이 포함되며 출력데이터는 실험에 대한 결과내역을 포함하는 결과파일, 빈발항목집합파일, 연관규칙을 추출할 수 있는 선행부와 결과부에 대한 내용을 포함하고 있는 규칙파일이 포함된다.

〈표 7〉에서는 실험데이터 집합에서 연관규칙 기반의 추천방법을 통한 후보 추천 콘텐츠 추출의 결과를 포함하고 있다. 실험결과에서 볼 수 있듯이 빈발항목집합을 통한 후보 추천 콘텐츠를 추출하기 위한 연관규칙의 척도를 변경하면서 결과값을 조정할 수 있으며 이를 통하여 가장 적합한 빈발항목집합을 기준으로 연관규칙을 추출할 수 있다.

〈표 7〉 빈발항목집합을 기준으로 생성된 규칙파일

선행부ID	결과부ID	지지도	신뢰도	향상도	선행부빈도	결과부빈도
48529	53999	0.00589875	0.86684	2.24307	2343	133059
48529	140496	0.00559380	0.82202	1.74055	2343	162610
48529	135604	0.00537597	0.79001	3.83441	2343	70939
48529	48900	0.00535854	0.78745	4.29156	2343	63177
55707	140496	0.00117046	0.76616	1.62226	526	162610
55707	53980	0.00114141	0.74715	2.44289	526	105306
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.

4.3 제안된 하이브리드 추천방법에 대한 성능평가

제안된 하이브리드 추천방법에 대한 성능평가를 위하여 수집된 14일간의 콘텐츠 이용내역 중에서 후반 7일간의 36,066,981건의 이용내역 데이터를 추출하고 이를 검증을 위한 전처리 과정을 통하여 28,847,452건의 데이터를 추출하였다. 그러나 해당 데이터에는 동일한 이용자에게 동일한 콘텐츠가 중복되어 이용된 경우가 있으므로 검증실험의 신뢰도를 위하여 중복되어 이용된 내역을 1회로 산정하여 추출된 이용빈도는 11,832,631건으로 계산 되었다. 해당 이용내역에서 이용된 콘텐츠는 193,617 종류이며 1회 이상의 콘텐츠를 이용한 이용자 수는 338,150명이다. 이러한 검증데이터와 함께, 적절한 이용자 및 콘텐츠 분할을 위한 임계값을 추출하기 위하여 다이용 콘텐츠 및 다이용 이용자 비율에 대한 변경을 통한 추천의 정확성에 대한 평가를 수행하였다. 이렇게 추출된 분할 임계값을 적용하여 최종적으로 제안된 하이브리드 추천방법과 기존 추천방법에 대한 비교평가를 위한 검증데이터는 전체 이용자에서 0.5%(1,668명)의 비율로 추출하였으며 추천대상 이용자에게 추천 콘텐츠의 수를 10, 20, 40, 100 개로 변경하면서 정확률, 재현율, F-Measure 및 성공률의 네 가지 평가척도를 가지고 검증을 수행하였다.

4.3.1 추천방법별 추천성능 비교 및 가중치 조절 실험

추천성능에 대한 평가실험을 위한 입력 데이터와 매개변수의 값은 〈표 8〉과 같다. 특히, 추천성능에 대한 평가에 있어서 개별 추천방법에 대한 이용자 및 콘텐츠 분할 임계값과 매개변수(parameter)의 조절에 따라 추천성능이 달라지므로 개별 추천방법의 성능을 최적화할 수 있는 분할 임계값 및 매개변수의 값을 추출하기 위한 실험이 선행되었다.

1) 개별 추천방법 추천성능 비교

본 실험에서는 제안된 하이브리드 추천방법

〈표 8〉 실험 입력 데이터

입력 데이터 종류	항목
추천방법별 추출 목록	유사이용자 목록(RelUserList)
	연관규칙목록(ARList)
	다이용 콘텐츠 목록 (PoPList)
변경 매개변수 값	다이용 콘텐츠 목록 가중치(wp)
	연관규칙 목록 가중치(wr)
	유사이용자 목록 가중치(wc)
검증데이터	추천성능 검증데이터

〈표 9〉 개별 추천방법별 추천 실험결과

항목	성공률	정확률	재현율	연산시간(sec)	NC 포함률
다이용 콘텐츠	68.66%	0.298621	0.343982	9.218000	0.00%
연관규칙	29.06%	0.142347	0.132227	14.359000	67.16%
협업여과추천	22.94%	0.131239	0.111086	17.796000	76.74%

에서 고려하고 있는 세 가지의 추천방법을 조합하지 않고 각각의 추천방법만으로 독립적인 추천에 대한 성능을 평가하였다. 즉, 본 연구에서 고려하고 있는 세 가지의 추천방법에서 하나의 추천방법에 대한 가중치를 1로하고 나머지의 추천방법에는 가중치를 0으로 하여 실험을 수행하였다. 먼저, 실험을 위하여 이용자 분할 임계값으로서 이전의 실험을 통하여 추출된 다이용 이용자 비율(HU ratio)은 0.05%, 콘텐츠 분할 임계값으로서 다이용 콘텐츠 비율(HC Ratio)은 0.1%로 설정하였다. 또한, 추천방법별 매개변수 선정에 있어서 연관규칙 기반의 추천방법에 있어서는 최소 지지도 변경에 따른 추천성능에 대한 선행실험에서 추천의 정확성 및 연산시간에서 가장 좋은 결과를 보여주는 최소 지지도인 0.0005를 설정하였으며, 협업여과 추천방법에 있어서는 다이용 이용자 비율 조절에 따른 추천성능에 대한 선행실험에서 추천의 정확성 측면에서 가장 좋은 성능을 보여주었던 0.1%의 다이용 이용자 비율을 적용하였으며, 연산시간에 대한 효율성을 고려하여 유사이용자 비율은 다이용 이용자에 대하여 10%로 설정하였다.

〈표 9〉는 실험결과를 보여주고 있으며, 특이한 점은 통계적 이용빈도에 기반하여 추출된 다이용 콘텐츠를 추천한 결과의 성공률, 정확률, 재현율이 월등히 높은 성능을 보여주고 있다는 것이다. 특히 협업여과 추천방법의 유사이용자 기반의 추천은 대용량의 이용자와 콘텐츠 환경에서는 해당 추천방법의 대표적인 단점

인 희소성의 문제로 인하여 매우 낮은 평가결과를 내는 것을 알 수 있다. 한편, 연관규칙 기반의 추천방법에서는 성공률의 비율이 30%에 근접하기 때문에 협업여과 추천방법에 비하여 상대적으로 나은 효과를 보여주고 있으나 다이용 콘텐츠 추천방법과 비교하면 역시 비효율적임을 알 수 있다. 그러나 단순히 다이용 콘텐츠에 대한 추천성능이 높게 나온다는 것만으로 해당 추천방법의 성능이 월등이 좋다고는 볼 수 없다. 다이용 콘텐츠의 특성 자체가 해당 데이터 집합에서 특정 기간 동안에 일시적으로 가장 많이 이용된 콘텐츠이기 때문에 추천평가에서 좋은 결과를 보이는 것은 지극히 당연한 결과라고 할 수 있다. 그러나 다이용 콘텐츠의 특성에 비추어보면 대부분의 다이용 콘텐츠는 일정기간이 경과함에 따라 이용자에게 거의 이용되지 않는다. 특히, 다이용 콘텐츠 추천방법의 결과에서 주목해야 하는 부분은 NC 포함률이라고 할 수 있다. 다이용 콘텐츠 추천방법을 통하여 선정된 추천 콘텐츠에는 일반 이용빈도를 가지는 콘텐츠의 포함여부를 보여주는 NC 포함률에 대한 결과값이 0%이다. 따라서 다이용 콘텐츠 추천방법을 통하여 추출된 콘텐츠는 위에서 언급한 바와 같이 이용자의 개인성향을 반영한 개인화 추천서비스가 아닌 단순히 특정 기간 동안에 이용자에게 일시적으로 높은 이용빈도를 가지는 콘텐츠를 추출한다는 의미라고 할 수 있다.

2) 다이용 콘텐츠 비율 가중치 조절 실험

본 연구에서는 개별 추천방법에 대한 가중치 조절을 통하여 제안된 하이브리드 추천방법의 성능을 최적화할 수 있는 추천방법별 가중치 추출을 위한 실험을 수행하였다. 이를 위하여 다이용 콘텐츠 추천방법을 기준으로 연관규칙과 협업여과 추천방법에 대한 각각의 가중치에 대한 비율을 8:7로 부여하여 실험을 수행하였다. 예를 들어 다이용 콘텐츠 가중치가 0.4 일 경우에 나머지 0.6의 가중치를 협업여과 추천방법에 대한 가중치는 0.28(0.6 * (7/15))로 부여하고 연관규칙에 대한 가중치는 0.32(0.6 * (8/15))의 비율로 나누어 부여하였다. 특히, 협업여과 추천방법에 비하여 연관규칙에 대한 가중치 비율을 보다 많이 부여한 것은 이전의 실험을 통하여 연관규칙이 보다 나은 추천결과를 보여주었기 때문이며 이러한 방법을 통하여 다이용 콘텐츠 목록(POP)에 대한 비율을 조절함으로써 추천방법별 가중치가 자동적으로 조절될 수 있다. 한편, 적용된 설정값은 개별 추천방법별 추천실험과 같이 다이용 이용자 비율(HU ratio)은 0.05%, 다이용 콘텐츠 비율(HC ratio)은 0.1%를 적용하였으며 연관규칙에서 가장 좋은 결과를 보여주는 최소 지지도로서 0.0005, 협업여과 추천방법에서 가장 좋은 결과를 보여주는 유사이용자 비율로서 다이용 이용자에 대하여 10%의 비율을 적용하여 실험을 수행하였다.

〈표 10〉에서는 다이용 콘텐츠의 추가에 따른 실험결과에 대한 일부 왜곡을 방지하기 위한 평가척도로서 고려되고 있는 NC 포함률에 대한 실험결과를 포함하고 있다. NC 포함률은 추천된 콘텐츠에서 다이용 콘텐츠가 아닌 일반 이용빈도를 포함하고 있는 비율을 의미하는 것으로서 이용자에게 단순히 인기 콘텐츠만을 제공하지 않는 비율만큼 이용자에게 새로운 콘텐츠를 추출하는 것이다. 따라서 정확률과 재현

〈표 10〉 다이용 콘텐츠 가중치 조절에 따른 성능평가

wp	wc	wr	성공률	정확률	재현율	F-Measure	연산시간	NC ratio
0.00	0.35	0.40	0.1817	0.0203	0.0199	0.0201	20.7180	1.0000
0.12	0.35	0.40	0.6751	0.1211	0.2563	0.1645	19.8900	0.4580
0.21	0.35	0.40	0.7326	0.1327	0.2717	0.1783	19.9060	0.4138
0.40	0.35	0.40	0.7800	0.1676	0.3128	0.2182	19.9680	0.2558
0.57	0.35	0.40	0.7884	0.1832	0.3302	0.2356	22.2810	0.1533
0.73	0.35	0.40	0.7890	0.1910	0.3373	0.2439	20.5000	0.0879
0.87	0.35	0.40	0.7896	0.1957	0.3415	0.2488	20.7960	0.0344
0.93	0.35	0.40	0.7908	0.1976	0.3435	0.2509	22.1400	0.0147
1.00	0.00	0.00	0.7866	0.1986	0.3440	0.2518	9.2180	0.0000

율에 대한 결과를 낮추지 않고 NC 포함률이 상대적으로 높은 성능을 가지는 수치값을 다이용 콘텐츠에 대한 가중치값으로 선정할 때 본 연구에서 제안하고 있는 하이브리드 추천방법을 통하여 최대의 추천성능을 얻을 수 있다. 이러한 기준에 근거한 실험결과를 바탕으로 다이용 콘텐츠 가중치를 0.4 ~ 0.5의 구간으로 설정할 때 제안된 하이브리드 추천방법을 통해 최대의 추천성능을 얻을 수 있다고 할 수 있다.

4.3.2 추천성능에 대한 비교평가

제안된 하이브리드 추천방법에 대한 성능평가를 위하여 본 실험에서는 연관규칙 및 협업여과 추천방법에 대한 추천의 정확성에 있어서 상대적인 비교평가를 통하여 추천성능을 검증한다. 선행실험을 통하여 후보 추천 콘텐츠의 추출과정이 이미 일괄처리(batch processing)되기 때문에 본 실험에서는 실시간의 추천에 따른 추천의 정확성에 대한 평가를 수행한다. 먼저 선행실험을 통하여 각각의 추천방법에서

〈표 11〉 추천방법별 성능평가를 위한 비교실험 설정값

추천방법	항목	설정값
연관규칙	다이용 콘텐츠 비율	0.001
	최소 지지도	0.0005
협업여과 추천방법	다이용 이용자 비율	0.0005
	다이용 콘텐츠 비율	0.001
	유사이용자 비율	0.1
제안된 추천방법	추천방법별 가중치	POP(3.5):COL(3):REL(3.5)

〈표 12〉 연관규칙 기반의 추천실험 결과

추천 개수	성공률	정확률	재현율	F-Measure	연산시간
10	0.213909	0.172437	0.130274	0.148419	14.796000
20	0.251079	0.158241	0.135461	0.145968	17.593000
40	0.290647	0.144956	0.122227	0.132625	20.937000
100	0.335612	0.132884	0.130030	0.131441	20.203000

〈표 13〉 협업여과 추천방법 기반의 추천실험 결과

추천 개수	성공률	정확률	재현율	F-Measure	연산시간
10	0.122782	0.114854	0.114421	0.114637	21.296000
20	0.114388	0.114111	0.100449	0.106845	12.750000
40	0.129376	0.125239	0.110109	0.117187	17.796000
100	0.135372	0.125597	0.111493	0.118125	12.265000

〈표 14〉 제안된 하이브리드 추천방법 기반의 추천실험 결과

추천 개수	성공률	정확률	재현율	F-Measure	연산시간
10	0.675659	0.255216	0.163744	0.199494	28.890000
20	0.731415	0.208993	0.228426	0.218278	19.781000
40	0.780576	0.187386	0.313011	0.234429	18.359000
100	0.838729	0.176829	0.433165	0.251137	19.218000

최상의 추천성능을 보여주는 분할 임계값 및 매개변수 값을 추출하여 개별 추천방법에 적용함으로써 평가를 수행한다. 실험에 대한 평가를 위하여 추천 대상 이용자는 검증데이터에서 0.5%를 추출한 1,668명에게 추천 콘텐츠의 수를 10, 20, 40, 100으로 조절하여 성공률, 재현율, 정확률, F-Measure를 기준으로 실험을 수행하였다. 〈표 11〉은 본 실험에서의 요구되는 분할 임계값과 매개변수를 보여주고 있으며 각각의 수치값들은 선행실험들을 통하여 추천방법별로 최적의 성능을 보여주고 있는 수치값을 추출한 것이다.

위의 설정값을 기준으로 네 가지의 평가척도를 측정한 결과가 〈표 12〉, 〈표 13〉 및 〈표 14〉와 같다.

위의 결과를 각각의 평가척도를 기준으로 보다 구체적으로 분석하면 다음과 같다.

1) 성공률 평가

성공률은 본 연구에서 제안하고 있는 평가척도의 하나로서 추천된 콘텐츠 수에 관계없이 이용자가 추천된 콘텐츠를 이용하는 경우를 추천의 성공으로 고려하는 팡가척도로서 추천시스템을 통하여 이루어지는 추천행위의 성공여부에 대한 기준이라고 할 수 있다.

〈표 12〉, 〈표 13〉과 〈표 14〉에서 볼 수 있듯이 다른 추천방법에 비하여 상대적으로 제안된 하이브리드 추천방법의 추천성능이 월등히 높은 결과를 보여주고 있다. 특히, 추천되는 콘텐츠 수가 증가함에 따라 성공률은 지속적으로 증가하고 있는 모습을 볼 수 있다. 이는 하이브리드 추천방법이 다른 추천방법과는 달리 일부 다이용 콘텐츠를 포함하고 있으며 각 추천방법에서 가장 높은 이용자 선호도를 갖는 콘텐츠가 포함되어 있기 때문에, 추천대상 이용자가 추천된 콘텐츠 중에서 이용자가 이용할 확률이 매우 높기 때문이라고 할 수 있다. 또한 추천시스템을 통하여 이루어지는 각각의 추천행위에 대하여 이용자가 선호하는 콘텐츠를 포함하고 있는 확률이 높다는 것을 의미한다. 한편, 이전에 수행된 개별 추천방법에 대한 추천성능의 평가에서와 같이 개별 추천방법에 최적화된 매개변수와 분할 임계값을 적용한 경우에도 대용량의 이용자 및 콘텐츠 환경에서는 협업여과 추천방법에 비하여 연관규칙 기반의 추천방법이 보다 나은 성능을 보여주고 있음을 알 수 있다.

2) 정확률 평가

본 연구의 특징은 대용량의 환경에서 멀티미디어 콘텐츠를 추천하는 것이라고 할 수 있다. 따라서 대규모의 콘텐츠에서 이용자의 선호에 맞는 콘텐츠를 추천하는 것은 상대적으로 매우 어려우면서도 중요한 요소라고 할 수 있다. 따라서 정확률에 대한 성능은 다른 중소규모의 환경에서 추천서비스를 제공하는 것에 비하여 상대적으로 낮은 정확률을 보일 것이라고 예측할 수 있다. 이러한 예측은 제안된 하이브리드 추천방법 뿐만 아니라 기존의 추천방법들에게도 동일하게 적용될 수 있을 것이다. 이전의 개별 추천방법별 성능평가의 정확률 측정에 있어서 각각의 추천방법들은 매우 낮은 성능을 보여주었다. 그러나 본 실험에서는 개별 추천방법이 가장 좋은 성능을 나타낼 수 있는 매개변수와 분할 임계값을 적용함으로써 정확률에 있어서 이전의 실험과의 수치값에 있어서 보다 나은 성능을 발견할 수 있다. 이러한 정확률에 있어서의 성능의 차이는 본 연구에서 제안하고 있는 분할 임계값을 기준으로 추천방법별 적용대상을 달리함으로써 추천성능을 높일 수 있음을 증명할 수 있을 것이다.

〈표 12〉, 〈표 13〉과 〈표 14〉에서 볼 수 있듯이 제안된 하이브리드 추천방법의 경우에 정확률은 다른 추천방법에 비하여 상대적으로 높은 결과를 보여주고 있으나 추천 콘텐츠의 수가 증가하면서 정확률은 떨어지는 경향을 보여주고 있다. 이러한 원인은 제안된 하이브리드 추천방법의 대상이 되는 대규모의 이용자와 콘텐츠가 많아짐에 따라 자연스럽게 정확률이 떨어졌다고 생각할 수 있다. 또한 하이브리드 추천방법에 비하여 연관규칙도 일정 부분 추천 콘텐츠의 수가 증가함에 따라 정확률이 떨어지는 경향을 보여주고 있으나 하이브리드 추천방법에

비하여 상대적으로 기울기는 낮다고 할 수 있다. 한편, 협업여과 추천방법은 다른 추천방법에 비하여 상대적으로 정확률에 있어서 가장 낮은 결과를 보여주고 있다. 그러나 특이하게 추천 콘텐츠의 수가 증가함에도 불구하고 정확률의 결과는 변함이 없이 미미하게 증가하는 경향을 보여주고 있다. 이러한 결과는 협업여과 추천방법의 경우에 상대적으로 정확률이 낮음에도 불구하고 일부 이용자에게는 가장 적합한 추천 콘텐츠를 제공한다는 예상이 가능하다.

3) 재현율 평가

재현율은 전체 검증데이터에서 추출된 추천대상 이용자가 이용했던 콘텐츠와 일치하는 콘텐츠를 추천시스템이 실험데이터로부터 정확하게 추출하였다는 것을 의미하는 것으로 본 실험에서 추천 콘텐츠의 수가 증가하면 이용자의 선호도에 일치하는 콘텐츠를 포함할 확률이 높아진다는 것을 의미한다고 할 수 있다. 일반적으로 재현율은 정확률과는 서로 반비례의 관계를 보여준다. 본 실험의 결과에서도 이러한 현상을 정확하게 보여주고 있다. 이전의 정확률 평가실험의 결과 그래프와 비교하면 하이브리드 추천방법은 추천 콘텐츠의 수가 많아지면 정확률은 감소하였지만 재현율에 있어서는 〈표 12〉, 〈표 13〉과 〈표 14〉에서 볼 수 있듯이 추천 콘텐츠의 수가 가장 적은 10개의 경우에 있어서 나머지 두 개의 추천방법과 비슷한 결과를 보이고 있다.

한편, 추천 콘텐츠의 수가 증가함에 따라 연관규칙 추천방법과 협업여과 추천방법에는 변화가 거의 없으나 제안된 하이브리드 추천방법의 경우에는 재현율이 가파르게 상승하는 경향을 보여주고 있다. 이는 제안된 하이브리드 추천방법을 통하여 추출된 추천 콘텐츠가 일정 규모의 다이용 콘텐츠를 포함하고 있기 때문에 추천 콘텐츠의 수가 증가하면서 보다 많은 다이용 콘텐츠가 포함되었기 때문이라고 추측할 수 있다. 특히, 다이용 콘텐츠의 특성상 많은 이용자에게 이용된 콘텐츠로서 검증데이터에 포함되는 추천대상이 되는 이용자가 이용할 가능성이 높기 때문이라고 예상할 수 있다. 한편 정확률에 대한 평가결과에 있어서 추천 콘텐츠의 수가 적은 경우에는 연관규칙 기반의 추천방법이 협업여과 추천방법에 비하여 매우 높은 성능을 보여주고 있으나, 재현율에 대한 측정결과는 추천 콘텐츠의 수가 증가함에도 불구하고 별다른 영향이 없음을 알 수 있다.

4) F-Measure 평가

F-Measure 측정방법은 정확률과 재현율에 대한 보정을 위한 평가방법으로서 재현율과 정확률에 대한 중요도에 따라 의 조정을 통하여 평가가 이루어진다. 그러나 본 실험에서는 재현율과 정확률에 대한 가중치가 같도록 $\beta = 1$로 설정하여 실험을 수행하였다.

〈표 12〉, 〈표 13〉과 〈표 14〉에서는 F-Measure를 기준으로 평가한 결과값을 보여주고 있는데 해당 수치값은 재현율에 대한 실험결과 비슷한 값을 보여주고 있다. 이러한 이유는 재현율과 정확률에 대한 가중치를 같게 하였기 때문에 나타나는 현상이라고 할 수 있다. 그러나 정확률과 재현율에 대한 가중치를 같게 하여 측정을 하였음에도 불구하고 제안된 하이브리드 추천방법의 경우에는 추천 콘텐츠 수가 증가함에 따라 F-Measure의 값은 정

확률 측정의 결과와는 달리 완만하게 상승하는 모양을 보여주고 있다. 이러한 이유는 제안된 하이브리드 추천방법이 재현율에 있어서 월등히 높은 성능을 보여주고 있기 때문에 정확률 성능이 떨어지는 비율을 상쇄하는 효과를 가지고 있기 때문이라고 생각할 수 있다.

5. 결 론

본 연구에서는 추천서비스의 기준이 되는 이용자의 콘텐츠 선호도 분석을 위하여 이용자의 묵시적 피드백으로서 콘텐츠 이용빈도를 기반으로 분석을 수행하였다. 또한 하이브리드 추천방법을 위하여 협업여과 추천방법, 연관규칙 및 통계적으로 높은 이용빈도를 가지는 다이용 콘텐츠 추천방법이 이용되었다. 본 연구에서 제안하고 있는 하이브리드 추천방법에서는 '이용자-항목' 매트릭스의 누락을 최소화하면서 희소성 문제를 해결하고 대용량의 이용자 및 콘텐츠 환경에서 효과적인 개인화 추천서비스를 제공하기 위한 방법을 제안하였다. 이러한 하이브리드 추천방법을 통하여 본 논문에서는 개별 추천방법의 장점은 수용하면서 기존의 협업여과 추천방법의 대표적인 문제점인 희소성 문제와 연관규칙에서의 규칙추출에 따른 계산의 복잡성을 해결하면서 동시에 초기 이용자 및 초기 평가문제를 해결하고자 하였다. 본 연구의 중요한 기여는 다음과 같다. 본 연구를 통하여 대용량 콘텐츠 및 이용자 환경에서 개인화된 콘텐츠 추천을 위한 방법(framework)을 위하여 콘텐츠 이용빈도에 있어서 상위 이용자 및 콘텐츠를 적절히 분리하고 활용하는 새로운 형태의 추천방법과 대용량 추천시스템에 적합한 연관규칙, 협업여과 추천방법 및 통계적 이용빈도를 통하여 추출된 후보 추천 콘텐츠에 대한 적절한 결합방법을 제안하였으며 제안된 하이브리드 추천방법에 대한 성능평가를 위하여 기존 추천방법과의 비교실험을 수행하였다. 실험을 위하여 음악 콘텐츠 서비스를 제공하는 사이트에서 14일간의 이용내역을 추출하여 전반 7일간의 이용내역은 추천 콘텐츠 선정을 위한 실험에 적용하고 후반 7일간의 이용내역은 추천방법의 성능평가를 위한 검증에 적용하였다. 실험의 결과로서 제안된 하이브리드 추천방법은 네 가지 평가척도에서 가장 좋은 성능을 보여주고 있다. 제안된 하이브리드 추천방법이 상대적으로 높은 추천 정확성을 보여주는 반면에 대용량의 이용자 및 콘텐츠 환경이라는 조건 때문에 연관규칙 및 협업여과 추천방법은 예상보다 낮은 결과를 보여주고 있으며 특히, 협업여과 추천방법의 경우에는 연관규칙에 비하여 상대적으로 더 낮은 성능을 보여주고 있다. 이것은 협업여과 추천방법의 대표적인 문제점인 희소성 문제가 대용량의 이용자 및 콘텐츠 환경에서는 더 많이 발생하게 되기 때문이다. 이러한 문제점은 추천 대상 이용자에 대한 유사이용자의 수를 증가시켜서 일정정도는 해결할 수 있으나 유사이용자의 수가 증가하면 유사도 계산을 위한 연산시간이 급격하게 증가하기 때문에 근본적인 해결방법은 아니다. 따라서 제안된 하이브리드 추천방법은 많은 추천서비스를 제공하는 사이트에서 이용되는 연관규칙 및 협업여과 추천방법과 함께, 일정기간 동안에 많은 이용자에게 공통적으로 이용빈도가 높은 다이용 콘텐츠를 동시에 적용

하기 위하여 각각의 추천방법을 위한 대상을 콘텐츠 이용빈도를 기준으로 분할하여 추천방법별로 적용하면서, 한편으로는 중요도에 따라 추천방법별 가중치를 적용함으로써 가장 좋은 추천성능을 얻을 수 있다는 것을 증명하였다.

결론적으로 제안된 하이브리드 추천방법을 통하여 대용량의 콘텐츠 및 이용자 환경에서의 추천의 정확성 및 추천 콘텐츠 추출에 따른 연산시간에 있어서의 높은 효율성을 보여주고 있으며, 이와 같은 추천서비스의 정확성과 효율성을 위하여 고려해야 하는 요구사항을 발견할 수 있었다. 이를 통하여 다음과 같은 결론을 얻을 수 있었다.

- 대용량 이용자 및 콘텐츠 환경에서 개인화 추천서비스의 제공을 위해서는 추천의 정확성과 추천 콘텐츠 선정을 위한 연산시간에 대한 고려가 동시에 이루어져야 한다.
- 기존의 추천방법들은 추천의 대상이 되는 이용자와 콘텐츠의 규모에 영향을 받는다. 따라서 콘텐츠와 이용자의 적절한 분할과 추천대상별로 적절한 추천방법의 적용을 통하여 추천의 정확성을 높일 수 있다.
- 추천 콘텐츠를 선정하는 과정에서 통계적으로 이용빈도가 높은 다이용 콘텐츠에 대한 적절한 비율과 가중치에 대한 고려가 필요하다.
- 콘텐츠 및 이용자의 분할을 통한 추천방법별 적용은 추천의 정확성 및 연산시간에 따른 문제점을 동시에 해결할 수 있다.
- 추천방법별로 요구되는 매개변수는 추천의 정확성에 있어서 매우 중요하다고 할 수 있다. 따라서 최종 결합추천의 선행과정으로서 개별 추천방법에서 요구되는 매개변수의 최적값을 추출하고 이를 적용함으로써 추천의 정확성 및 연산시간의 단축과 같은 추천의 효율성을 얻을 수 있다.

한편 추천서비스를 제공하기 위해서는 기술적인 방법론 이외에 어떻게 인터넷에 적용하여 효과를 거둘 것인가에 대한 부분이 상대적으로 중요한 위치를 차지하게 된다. 인터넷이라는 매체가 가지는 특성은 양방향 커뮤니케이션이 가능한 상호 작용적 매체라는 점이다. 특히, 근래에 웹을 대상으로 가장 많은 관심을 받고 있는 주제인 웹 2.0에서의 핵심은 정보를 제공하는 쪽이나 정보를 제공 받는 쪽만의 시각으로 문제를 해결할 수 없음을 시사하고 있다. 따라서 웹을 기반으로 정보이용자와 정보제공자들의 상호작용을 통하여 제공되는 서비스의 활성화를 위한 출발점으로서 이용자의 관심에 가장 적합한 정보를 제공하고 이에 대한 피드백을 통한 개선된 추천서비스는 매우 중요한 문헌정보학의 연구분야로서 지속적인 연구가 수행되어야 할 것이다.

참 고 문 헌

김병만 외. 2004. 추천시스템을 위한 내용기반 필터링과 협력필터링의 새로운 결합기법. 『한국정보과학회논문지 : 소프트웨어 및 응용』, 31(3) : 332-342.

김현희, 구내영. 2002. 춤정보서비스를 위한 MyCyberLibrary 모형설계와 평가에 관한 연구. 『한국정보관리학회지』, 19(2) : 132-157.

박우창 외. 2003. 『데이터 마이닝 개념 및 기법』. 서울 : 자유아카데미(주).

하단심, 황부현, 2000. 데이터의 상대 지지도를 이용한 다단계 연관규칙탐사기법. 『한국정보과학회 추계 학술발표논문집』, 27(2) : 195-197.

Balabanovic, M. and Y. Shoham. 1997. colla "Fab : Content-based, borative recommendation." *Communications of the ACM*, 40(3) : 66-72.

Basu, C., H. Hirsh and W. Cohen. 1998. "Recommendation as classification using social and content-based information in recommendation." *Proc. of the Fifteenth International Conference on Artificial Intelligence* (AAAI-98), 714-720.

Billsus, D. and M. Pazzani. 1998. " Learning collaborative information filters." *Proc. of the International conference on Machine Learning*, 46-54.

Billsus, D. and M. Pazzani. 2000. "User modeling for adaptive news access." *User Modeling and User Adaptive Interaction*, 10(2-3) : 147-180.

Burke, R. 2002. "Hybrid Recommender Systems : Survey and Experiments." *User Modeling and User Adapted Interaction*, 12(4) : 331-370.

Claypool, M. Gokhale, Miranda. T. Murnikov, P. Netes.D, and M. Sartin. 1999. "Combining content-based and collaborative filters in an online newspaper." *Proc. of ACM SIGIR Workshop on Recommender Systems.*

Goldberg, D., D. Nichols, B. M. Oki, and D. Terry 1992. "TAPESTRY : using collaborative filtering to weave an information." *Communications of the ACM*, 35(12) : 61-70.

Good, N., J. Schafer, J. Konstan, A. Borchers, B. Sarwar, J.

Herlocker, AND J. Riedl 1999. "Combining collaborative filtering with personal agent for better recommendation." *Proc. of the AAAI Conference*, 439-446.

Gupta, D., M. Digiovanni, H. Narita, and K. Goldberg 1999. "Jester 2.0 : A new linear-time colla borative filtering algorithm applied to jokes." *Proc. of ACM SIGIR Workshop on Recommender Systems : Algorithms and Evaluation.*

Hill, W., L. Stead, M. Rosenstein, and G. Furnas 1995. "Recommending and evaluating choices in a virtual community of use." *Proc. of CHI'95 Conference on Human Factors in Computing Systems*, 194-201.

Lang, K. 1995. "Newsweeder : Learning to filter netnews." Proc. of the 12th International Conference on Machine Learning, 331-339.

Lewis, D and Gale W. A.. 1994. "A sequential algorithm for training text classifiers." *Proc. of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 3-12.

Popescul, A. et al. 2001. "Probabilistic models for unified collaborative and content-based recommen dation in Sparse-Data Environments." *Proc.of the 17th Conference on Uncertainty in Artificial Intelligence*, 437-444.

Rensnick, P. et al. 1994. "GroupLens : An open architecture for collaborative filtering of netnews." *Proc. of the 1994 ComputerSupported Cooperative Work conference*, 175-186.

Shardanand, U. and P. Maes. 1995. "Social information filtering : Algorithms for automating 'word of mouth'." *Proc. of ACM CHI' 95 Conference on Human Factors in Computing Systems*, 210-217.

Wasfi, A. M.. 1999. "Collecting user access patterns for building user profiles and collaborative filtering." *Proc. of International Conference on Intelligent User Interfaces*, 57-64.