이용자 이용행위 및 콘텐츠 위치정보에 기반한 개인화 추천방법에 관한 연구

A Study on Personalized Recommendation Method Based on Contents Using Activity and Location Information

김 용(Yong Kim)*
김문석(Mun-Seok Kim)**
김윤범(Yoon-Beom Kim)***
박재홍(Jae-Hong Park)****

초 록

본 연구에서는 웹, IPTV 등의 콘텐츠 유통망에서의 개인화 추천서비스를 위하여 이용자의 콘텐츠 이용행위와 콘텐츠의 위치정보를 활용한 추천방법을 제안하고 있다. 추천방법의 성능향상을 위하여 이용자 및 콘텐츠 프로파일 생성방법과 함께, 이용자의 콘텐츠 이용행위를 암묵적 이용자 피드백으로서 학습과정에 적용하여 이용자 선호도를 분석하였다. 학습과정에서의 이용자 선호도 분석을 위하여 협업여과추천방법 및 내용기반추천 방법을 적용하였다. 또한 보다 정확한 추천을 위한 최종 콘텐츠 추천을 위하여 웹사이트 상의 콘텐츠에 대한 위치정보를 활용한 추천방법을 제안하고 있다. 이를 통하여 보다 효율적이고 정확한 추천 서비스의 제공이 가능할 수 있다.

ABSTRACT

In this paper, we propose user contents using behavior and location information on contents on various channels, such as web, IPTV, for contents distribution. With methods to build user and contents profiles, contents using behavior as an implicit user feedback was applied into machine learning procedure for updating user profiles and contents preference. In machine learning procedure, contents-based and collaborative filtering methods were used to analyze user's contents preference. This study proposes contents location information on web sites for final recommendation contents as well. Finally, we refer to a generalized recommender system for personalization. With those methods, more effective and accurate recommendation service can be possible.

키워드: 개인화, 추천, 협업여과추천, 내용기반추천
IPTV, personalization, recommendation, collaborative filtering, contents-based recommendation

* 전북대학교 문헌정보학과 조교수(yk9118@chonbuk.ac.kr) (제1저자)
** 전라북도 교육청 기록관리사(kk5077@naver.com) (공동저자)
*** 전북대학교 문헌정보학과 석사과정(sea_stars@naver.com) (공동저자)
**** (주) 유라클 대표이사(parkjh@uracle.co.kr) (공동저자)

■ 논문접수일자 : 2009년 2월 12일 ■ 최초심사일자 : 2009년 2월 18일 ■ 게재확정일자 : 2009년 2월 26일
■ 정보관리학회지, 26(1): 81-105, 2009. 〔DOI:10.3743/KOSIM.2009.26.1.081〕

1. 서 론

1.1 연구 배경 및 목적

인터넷과 정보기술의 폭발적인 발전과 성장은 기술적인 측면뿐만 아니라 사회문화적 측면에 정보의 생산 및 유통의 관점에서 우리에게 커다란 영향을 미치고 있다. 특히, '개방', '분산'과 '참여'라는 단어로서 대표되는 웹 2.0 개념의 출현 및 IPTV 등의 다양한 정보유통 채널의 등장은 전통적인 정보생산과 소비에 대한 패러다임을 변화시키고 있다.

초기 인터넷이 출현시 대부분의 도서관 및 정보센터를 포함한 정보제공자들은 인터넷을 통하여 정보를 제공함으로써 자관의 서비스 이용자 및 모든 정보이용자들의 요구를 충족시킬 수 있을 것이라는 기대로 전자도서관과 같은 정보저장소를 구축하였으나 정보생산을 위한 다양한 저작도구(authoring tool)의 발달과 인터넷의 확산 및 IPTV, 무선망 등의 발전에 따른 정보의 폭발적인 유통증가를 통한 정보과잉(information overflow)은 정보이용자에게 자신이 원하는 정보를 찾기 위하여 엄청난 시간과 노력을 요구 하는 결과를 초래 하였다. 이러한 문제를 해결하기 위한 다양한 방법들이 도서관 및 정보센터를 중심으로 시도되었으며 대표적인 방법으로서 정보여과시스템의 등장이라고 할 수 있다(정영미, 이용구 2002).

정보여과시스템은 전통적으로 도서관이나 정보센터에서 제공되던 선택적 정보배포(SDI: Selective Dissemination of Information)와 맞춤정보서비스가 등장 하게 되었다. 그러나 이러한 정보 서비스들은 이용자의 개인정보 등을 포함하는 이용자 프로파일 정보와 같은 단순 정보에 기반하여 해당 정보와 일치되는 모든 정보를 제공함으로써 여전히 정보과잉의 한계점을 극복할 수 없었다. 특히, 이용자의 정보요구행태 변화에 적절하게 대응 할 수 없는 제약점이 있다고 할 수 있다. 따라서 증가하는 정보의 효율적인 관리 및 이용자의 요구와 관심에 적합한 정보를 적시에 제공하여야 하는 도서관 및 정보센터로서는 전통적으로 정보를 제공하는 방법과는 다른 새로운 정보관리 및 서비스방법이 필요하게 되었다. 이러한 사회, 문화적인 요구에 따라 이용자요구에 적합한 정보의 추출과 제공을 위한 방법으로서 대량의 정보에서 개인별 맞춤화된 정보와 서비스를 제공할 수 있는 개인화 추천서비스에 대한 관심은 더욱더 높아지고 있다고 할 수 있다.

일반적인 관점에서 개인화라는 용어가 갖는 의미는 이용자의 요구에 적합한 정보를 추출하고 이를 신속하게 제공하는 일련의 결합된 방법을 개인화로서 정의할 수 있다(정경용 외 2004). 개인화 서비스를 제공하기 위해서는 가장 중요한 요소로서 논문정보, 콘텐츠 등의 이용자가 요구하는 정보에 대한 추천을 제공하는 추천 시스템이 요구된다. 현재 개인화서비스 제공을 위하여 최신의 정보여과 및 추출방법에 기반한 추천 시스템들이 제안되고 있으며 아마존(Amazon), 야후(Yahoo), 이베이(e-Bay) 등과 같은 대형 온라인 쇼핑몰과 포탈사이트에서 실제 적용되고 있다(Linden, Smith and York 2004). 이와 같은 사회적인 흐름에 비추어 이용자의 정보요구에 적절한 정보를 신속하고 정확하게 제공해야하는 도서관 및 정보센터의 입장에서 정보에 대한 개인화 서비스는 매우 중요

하고 필수적인 서비스방법이다. 특히 이용자를 세분화하여 이용자집단에 따른 적절한 정보서비스를 제공할 수 있으며, 대량의 정보를 효과적으로 처리하고 이를 적절하게 이용자에게 제공할 수 있다는 측면에서 폭발적으로 증가하는 전자정보의 처리가 요구되는 현재 도서관과 정보센터의 고민을 해결할 수 있을 것이다.

따라서 본고에서는 개인화서비스 시스템의 개발을 위한 알고리즘 및 적용 가능한 기술에 대하여 알아보고 개인화서비스 시스템의 전체적인 구성요소 및 전체시스템의 설계와 함께, 개인화서비스 시스템에서 가장 중요한 요소인 이용자의 성향을 분석하는 학습 시스템의 설계를 제안하고자 한다.

1.2 연구 범위 및 구성

본 연구에서는 멀티미디어자료를 포함하여 폭발적으로 증가하는 정보를 개인에게 맞춤화된 형식으로 추천서비스를 제공하기 위한 추천방법과 학습알고리즘의 제안 및 구현을 목표로 하고 있다. 특히, 개인화서비스의 대상을 텍스트 문서를 포함하여 최근 증가하고 있는 멀티미디어 콘텐츠로 확장하였으며 이를 위하여 기본적으로 요구되는 이용자의 개인정보와 정보를 탐색하는 과정에서 보여주는 행동 등에 대한 정보를 기준으로 생성되는 이용자 프로파일 및 콘텐츠에 대한 정보를 포함하고 있는 콘텐츠 프로파일의 생성 및 갱신을 위한 학습 알고리즘 및 콘텐츠 추천방법을 제안하고 있다. 이러한 요구사항의 만족을 위하여 본 연구에서 목표하고 있는 구체적인 연구 범위와 내용은 다음과 같이 요약될 수 있다.

첫째, 추천 시스템의 구성과 특징을 비교 분석한다.

둘째, 다양한 추천방법에 대한 기존 연구들을 분석하여, 본 연구에서 제안한 모형과 기반구조의 설계를 포함하는 기본 방향을 설정한다.

셋째, 대표적인 추천방법의 장단점을 분석하고 기존의 제한점을 극복하기 위한 학습알고리즘 및 추천 알고리즘을 적용한 추천 모형을 정의한다. 이를 위하여 이용자 프로파일, 콘텐츠 프로파일 및 이용자의 행동데이터에 대한 분석 등에 대한 분석방법론을 제안한다.

넷째, 이용자의 취향 변화를 반영키 위하여, 이용자 행동데이터에 대한 분석을 위한 이용자의 웹로그를 분석하고 이를 기반으로 이용자 프로파일 정보를 갱신하는 학습알고리즘을 제안한다.

다섯째, 항목의 분류 및 가중치부여를 통해 추천의 정확도 향상을 유도한다.

여섯째, 제안된 추천 모형을 동적으로 반영할 수 있는 응용 개발 환경을 설계한다.

이를 위하여 이용자 프로파일의 학습을 위한 이용자 프로파일과 콘텐츠 프로파일을 표현하는 방법을 제시하고, 웹에서의 이용자의 행동정보 및 웹사이트의 구조 정보, 콘텐츠 프로파일을 반영하여 이용자의 성향을 학습하는 알고리즘을 제시하고자 한다. 이를 위하여 제안된 추천 시스템은 추천서버와 학습서버 및 이용자 인터페이스로 구성되어 있는 처리 부분과 이용자 프로파일 및 콘텐츠 프로파일 정보를 담고 있는 데이터 저장 부분으로 구성되어 있다. 각 요소별 요구기능은 다음과 같이 정의 할 수 있다.

- 이용자 인터페이스(User interface) : 웹서버를 통하여 접속된 이용자와의 통신처리

를 위한 기능을 수행한다.

- 학습 시스템(Learning system): 이용자의 행위정보를 분석하여 이용자 프로파일 정보를 갱신하는 기능을 수행한다.
- 추천 시스템(Recommendation system): 이용자 프로파일정보와 일치하는 콘텐츠를 추천하는 기능을 수행한다.
- 데이터저장소(Data storage): 데이터저장소는 추천을 위한 각종정보 및 이용자 프로파일과 콘텐츠 프로파일을 저장하는 기능을 수행한다.

2. 개인화 추천 방법 및 시스템

2.1 개인화 추천 서비스를 위한 입력 데이터

개인화를 다른 관점에서 분류한다면 명시적 개인화(Explicit personalization)와 암묵적 개인화(Implicit personalization)로 구분할 수 있다. 명시적 개인화라는 것은 개인화의 대상이 되는 이용자로부터 정보를 직접 입력 받아서 그 정보를 이용하는 것이고 암묵적 개인화라는 것은 개인화의 대상이 되는 이용자의 구매행동이나 웹사이트를 이용하는 패턴 등 이용자의 행태를 근간으로 해서 개인화를 구현하는 것이다. 개인화를 위해서 사용하는 데이터는 개인화 방법에 따라 다양하지만 인구통계정보, 선호도, 이용자 입력 사항 등과 같이 이용자로부터 명시적인 입력 요청을 통해 확보해야 하는 데이터보다는 이용자의 구매이력, 장바구니이력, 클릭스트림 정보 등과 같이 이용자가 해당 웹사이트를 이용하는 과정 또는 정보를 탐색하는 과정에서 무의식적으로 행위하는 암묵적 데이터에 대한 분석 요구를 수용할 수 있도록 하는 것이 중요하다. 그러나 일반적으로 이용자가 의식적으로 제공하는 명시적 데이터가 있다면 해당 데이터를 보다 비중 있게 다루는 것이 좋다고 알려져 있다.

인터넷 웹사이트의 특징에 따라 차이가 있겠지만 개인화를 위해서는 아래와 같은 데이터들을 사용하게 된다.

- 이용자의 인구통계(Demographics) 정보: 나이, 성별, 결혼여부 등
- 콘텐츠 선호도: 콘텐츠군에 대한 선호도, 웹 페이지 분류군에 대한 선호도
- 이용자가 정보를 이용하였거나 구매한 이력
- 이용자의 웹 페이지를 클릭정보로서 구체적으로 웹 페이지 방문 횟수, 머문 시간, 웹 페이지 이동 경로(Navigation path)
- 이용자 입력 사항: 웹 페이지가 던지는 질문에 대해서 이용자가 답변한 내용으로 콘텐츠에 대한 만족도, 향후 이용 가능성 등

2.2 개인화 추천 방법 및 시스템

추천 문제가 학술적으로 처음 발표된 것은 90년대 중반부터 이다(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 추천 방법들이 제안되었으며 실제 추천 시스템의 구현에 적용되어 연구되어져 왔다. 대표적인 추천방법은 추천 과정에 따라 〈표 1〉과

같은 네 가지의 범주로 분류할 수 있다(Burke 2002). 이러한 추천방법들 중에서 최근에는 내용기반추천방법과 협업여과추천방법이 있다(Sarwar et al 2002; Billsus and Pazzini 2000).

내용기반 추천방법은 이용자의 과거 정보이용행태를 기반으로 하여 추천정보를 제공하는 방법으로서 사람에 의하여 추천메커니즘이 결정되고 추천을 위한 방법이 단순하고 통제가 용이 하다는 장점이 있다. 그러나 이용자가 과거의 자신이 경험한 것과 유사한 정보만을 취함으로써 데이터에 기반한 객관적인 이용자 행태 분석의 한계가 있다. 두 번째는 대부분의 내용기반 추천방법이 텍스트 자료에 한정된다는 것이다. 현재와 같은 멀티미디어 환경에서의 콘텐츠 추천에는 많은 한계점을 가지고 있다고 할 수 있다(김용 2006). 마지막으로 추천메커니즘이 사람에 의하여 결정됨으로써 풍부한 추천 노하우(Know-how)가 축적되지 않은 상태에서는 추천의 정확성에 한계가 있다. 한편, 협업여과 추천방법은 이용자가 제공한 정보를 사용하여 이용자를 비슷한 선호도를 가진 집단으로 나누어 유사한 선호도를 가진 집단의 선호도에 따라 추천하는 방식을 사용함으로써 정보이용자에 대한 초기 정보가 부족한 경우 가장 적절하게 이용될 수 있다. 즉, 이용자가 보여주는 정보탐색행동이나 정보탐색후의 피드백에 대한 정보가 충분치 않은 경우 해당 정보이용자와 비슷한 프로파일정보와 반응을 보여주는 이용자들의 선호도를 기준으로 정보를 제공하는 방법이다(Sarwar et al. 2001). 이러한 협업여과 추천방법은 기존의 추천방식에 비하여 많은 장점을 제공함으로써 현재까지 웹상에서 제공되고 있는 추천 방법에 있어서 가장 성공적인 추천방법이라고 할 수 있다. 그러나 협업여과 추천방법은 불완전하고, 적은 정보량을 토대로 이용자의 선호와 관심도에 대한 정보가 충분하지 않은 환경에서 적용하는 경우 전혀 적합하지 않은 정보를 제공할 가능성이 매우 높다고 할 수 있다(황성희 외 2001; 이기현, 고병진, 조근식 2002). 또한 추천을 위한 기계학습을 위해서 많은 시간을 요구한다.

최초의 추천시스템은 협업여과를 사용한 업무 메일링프로그램으로서 현재 콘텐츠 추천시스템은 일대일 마케팅을 비롯해 eCRM, 데이터마이닝, 콘텐츠관리, 그리고 검색엔진 등 거의 모든 인터넷 솔루션들이 '개인화'를 표방하고 있으며 Amazon, CD Now, Garden.com 등이 개인화 추천 서비스를 통한 대표적인 성공적 사이트로 평가 받고 있다(김용, 문성빈 2005). 이러한 개인화서비스를 통한 개인화 추천시스템에 대한 성능 평가에 관한 연구가 많은 부분에서 진행되고 있다. 현재 웹사이트에는 다양한

〈표 1〉 추천 방법의 분류

분 류	설 명
내용 기반 추천(Content-based)	이전에 이용자가 선호한 항목과 유사항목을 추천
인구 통계 기반 추천(Demographic-based)	인구 통계학적인 정보를 기반으로 유사한 이용자를 참조하여 항목을 추천
규칙 기반 추천(Rule-based)	자료를 통하여 규칙을 형성하고, 이 규칙에 따라 추천
협업 여과 추천(Collaborative filtering)	유사한 흥미나 선호도를 가진 이웃이 좋게 평가한 항목을 추천

〈표 2〉 평가 자료의 수정 주기별 추천 시스템의 구분

갱신주기	안정적	주기별	실시간
응용분야	영화, 음악 등	뉴스, 기사	웹페이지 등
추천 시스템 예	EachMovie Morse Firefly ⋮	Tapestry GroupLens Lotus Notes ⋮	Phoaks GAB Fab ⋮

방법을 적용한 여러 종류의 추천 시스템이 있다. 추천시스템은 응용목적과 자료의 구성 주기에 따라 〈표 2〉와 같이 크게 세 그룹으로 나눌 수 있다.

3. 제안된 개인화 추천방법

3.1 추천방법의 특징

내용기반 여과방법과 협업여과 추천방법은 서로 상호보완적인 특징을 가지고 있다. 따라서 두 방법의 장점을 수용하면서 이용자의 관심이 높은 콘텐츠를 추천하기 위하여 두 가지 방법을 결합시키는 연구들이 최근 늘어나고 있다(Fink 2002). 그러나 기존의 연구들은 실험 데이터의 불충분성과 접근 방법에 있어서 단순히 기존 추천방법을 동시에 고려한 모델들이라고 할 수 있다. 예를 들어 Fab 시스템에서는 이용자의 프로파일 생성을 위하여 이용자의 피드백을 고려하였으며 이를 위하여 이용자의 직접적 행위를 통하여 얻어진 명시적 피드백 정보를 이용하였다. 이러한 명시적 피드백은 이용자가 직접 콘텐츠에 대한 평가와 점수를 제공하는 것으로 명시적 피드백 정보만을 가지고 이용자 프로파일을 생성한다는 것은 매우 제한적인 데이터로 인하여 현실적으로 구현이 어렵다. 따라서 기존의 연구의 제한점을 해결하고 이용자에게 정확한 콘텐츠의 추천을 위하여 본 연구에서는 이용자가 웹상에서 보여주는 행위정보를 분석하여 획득한 암묵적 피드백에 기반한 개인화 추천방법을 제안하고 있으며 제안방법의 특징은 다음과 같다.

첫째, 콘텐츠의 추천을 위하여 콘텐츠 프로파일과 이용자 프로파일을 생성하였으며 두 가지의 프로파일을 구성하기 위하여 내용기반 추천과 협업기반 추천의 특징을 동시에 고려하면서, 보다 정확한 콘텐츠 추천을 위하여 이용자의 웹상의 위치정보를 이용하여 최종 추천 콘텐츠가 결정된다. 즉, 추천의 정확성을 높이기 위하여 이용자 프로파일과 콘텐츠 프로파일의 매칭과정을 통하여 1차 추천 콘텐츠가 결정이 되면 추가적인 웹사이트에서 이용자가 위치하고 있는 위치정보를 이용하여 2차 추천 콘텐츠를 최종적으로 제공함으로써 보다 적절한 이용자 요구사항을 반영한 콘텐츠를 추천할 수 있다. 여기에서 위치정보는 이용자가 정보 또는 콘텐츠를 탐색하는 과정 중에서 메인페이지, 1차 하위 페이지 또는 2차 하위페이지로의 이동 행위에 대한 정보를 의미한다.

둘째, 이용자 피드백정보의 정확한 분석을 위하여 이용자가 웹상에서 보여주는 행위정보를

분석시스템을 통하여 분석하고, 이를 유형별로 저장하여 학습 과정에서의 주요한 변수인 이용자 피드백에 대한 입력정보로서 반영하였다. 이와 같은 과정에서 행위정보는 이용자가 해당 페이지에 존재하는 정보 또는 콘텐츠에 대하여 구매, 다운로드, 단순클릭 등의 콘텐츠를 이용행위를 포함하고 있는 정보를 의미한다. 보다 세부적인 행위에 대한 구분은 〈표 8〉에서 보여주고 있다. 이와 같은 정보들은 실질적인 추천정보를 생성하기 위한 전처리 단계인 기계학습과정에서 이용자에 대한 피드백정보로서 활용함으로써 이전의 연구들에 비하여 이용자의 성향을 보다 정확히 반영함으로써 콘텐츠 추천에 있어서 정확도를 향상 시킬 수 있을 것이다.

셋째, 이용자 및 콘텐츠 프로파일의 생성 및 갱신을 위하여 이용자의 인구통계학적 정보와 함께, 이용자의 웹상의 행위정보에 대한 분석을 통한 이용자 피드백정보에 대한 분석 및 이용이 필수적이라고 할 수 있다(이수정, 이형동, 김형주 2004). 특히 본 연구에서는 이전의 Fab 시스템에서 적용한 방법과는 달리 이용자가 명시적으로 제공하는 피드백이 아닌 암묵적으로 이용자가 웹상에서 보여주는 행위를 분석하여 이를 수치화 하였다.

넷째는 기계학습 과정을 통한 프로파일의 갱신에 있어서 협업여과 추천방법을 통하여 콘텐츠가 속한 범주들에 대한 관계성을 분석하고 이를 수치화한 범주간의 가중치를 학습에 적용하였다.

마지막으로 본 연구에서 제안하고 있는 추천방법은 콘텐츠 추천을 위하여 이용자 프로파일을 생성하고 이를 지속적으로 갱신하는데 있어서 내용기반 여과방법의 장점과 함께, 이용자의 인구통계학적 정보 및 이용자 피드백 정보를 통하여 획득되는 협업적 요소로서 이용자의 특정 범주에 대한 선호도를 보여주는 범주 가중치(a)를 학습 과정에 적용하였다. 이용자에게 콘텐츠를 추천방법에 대한 하이브리드적인 특징은 함수(F)로 표현 될 수 있다. 각 구성요소에 대한 세부적인 내용은 아래에서 설명하고 있다.

$$HybridF(< user, user\,profile, feedbacks >) \rightarrow \{a, < Category, Keyword, Weight > set\} \quad (1)$$

3.2 이용자 모형 및 프로파일 구성

본 연구에서의 이용자 모형화 및 프로파일 구성의 범위는 이용자에 대한 정보를 포함하는 이용자 프로파일과 콘텐츠에 대한 정보를 포함하고 있는 콘텐츠 프로파일의 생성 및 갱신으로 정의할 수 있다. 세부적으로 이용자 프로파일은 이용자가 초기에 입력한 기본 정보를 바탕으로 구성되는 기본 프로파일, 이용자의 콘텐츠에 대한 선호 정보를 가지는 이용자 프로파일 및 관련 키워드 목록을 포함하고 있으며 콘텐츠 프로파일은 콘텐츠에 대한 키워드와 가중치를 가지는 콘텐츠 프로파일과 콘텐츠에 대한 키워드를 포함하고 있는 키워드 목록으로 구성되어져 있다. 각각의 프로파일에 대한 표현은 벡터공간모형(Vector space model)에 기반을 두고 있다. 즉, 각각의 프로파일들은 가중치가 부여된 키워드 벡터로 표현된다.

3.2.1 이용자 프로파일(User Profile)

이용자 프로파일은 이용자의 초기 입력 값과

전처리 과정을 통하여 생성된 통계정보에 의해 생성되는 기본 프로파일과 이용자의 콘텐츠에 대한 선호도에 대한 정보를 가지고 있는 이용자 프로파일로 구성된다.

〈표 3〉은 이용자 프로파일을 구성하는 기본 프로파일의 예를 보여주고 있다.

기본 프로파일(Long term profile)은 원칙적으로 가입자의 기본정보 입력에 따라서 정해진다. 이러한 기본 프로파일은 콘텐츠에 대한 선호도와 같은 일반적인 콘텐츠 평가방법, 내려받기, 구매, 단순클릭 등의 콘텐츠에 대한 이용자의 이용행위를 포함하는 이용자 피드백 등과 같은 명시적 방법(Explicit method)을 통하여 수정하고 추가할 수 있다. 또한 기본 프로파일은 추천과 관련하여 상대적으로 변화가 적은 항목들에 대한 정보를 유지하며 각각의 항목들은 기계학습과 추천에 반영되는 정보들이다. 기본 프로파일에 포함되어지는 정보는 이용자가 웹 서비스 시스템에 제공한 이용자에 대한 이름, 나이, 성별, 주소 등과 같은 인구통계학적 정보로서 〈표 3〉에서의 'a_1, a_2 ....a_i'의 항목들이 해당된다. 한편 범주별로 분류되어진 콘텐츠에 대한 메타 정보(장르, 작자, 위치 등)와 사이트의 구조가 포함된다. 그리고 추천을 위해 콘텐츠 분류에 따른 최상위 범주와 이용자가 회원 가입시에 선택 또는 입력한 콘텐츠에 대한 선호도에 대한 가중치가 포함되며 'c_1, c_2, c_3, ... c_i'가 해당 된다. 마지막으로 구매, 다운로드, 즐겨찾기 추가 등과 같은 이용자의 행위정보에 대한 분류와 이에 따른 가중치 값이 포함되며 'b_1, b_2, ..., b_i'로 표현된다. 각 항목에 대한 가중치는 실험을 위하여 초기에 임의로 입력되며 해당 값은 기계학습 과정에서 이용자의 콘텐츠 선호도 계산을 위한 입력정보가 된다.

가. 키워드 테이블

키워드 테이블은 〈표 4〉와 같이 간단히 표현될 수 있으며 각각의 키워드는 기본 프로파일의 각각의 속성 값에 대한 주제어로서 이용자의 입력 및 인구통계학 정보를 기준으로 키워드 테이블이 생성된다. 개별 키워드는 콘텐츠 추천을 위한 검색속도 증가를 위해 주기적으로 가중치 값이 높은 키워드를 접근점에 가까이 있도록 해 준다. 실제 구성에서, 영화의 장르 또는 선호음악과 같이 거의 변하지 않는 키워드들은 키워드 테이블에 포함 시킨다.

한편 〈표 4〉의 기본 프로파일에 대한 키워드 테이블과는 달리 콘텐츠의 매칭과 실제 학습에

〈표 3〉 개인 이용자(ID: JHP) 기본 프로파일 정보의 예

(1) 기본정보를 위한 이용자 프로파일

ID	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9	a_{10}	c_1	c_2	c_3	c_4	..
JHP	1	6	3	6	1	1	1	7	.	.	0.3	0.3	0.2	0.2	0

(2) 기타 접속 정보, 구매 기록, 인쇄 여부, 북마크 여부, 부정적 정보 등에 대한 이용자 프로파일의 추가 정보

ID	b_1	b_2	b_3	b_4	b_5	b_6	b_7	…….
JHP	5	.	.	.	.	.	.	...

〈표 4〉 기본정보와 관련된 키워드 테이블

#	성별	나이	직업	학력	결혼 여부	메일 수신	선호 콘텐츠	선호 음악	선호 영화	구매 여부	…	피드백 유형
속성 ID	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9	b_1		d_i
1	남자	5-9	대학생	초등	미혼	Yes	영화(c_1)	가요	액션	1~3	…	명시적
2	여자	11-14	대학원생	중재학	기혼	no	음악(c_2)	팝	코미디	4~7		클릭&보기
3		15-17	초중고생	고재학			오락(c_3)	영화음악	드라마	8~10		인쇄
4		18-20	컴퓨터, 인터넷	고졸업			생활(c_4)	락	멜로			즐겨찾기
5		21-24	금융, 회계	대재학				재즈	SF			장바구니
6		25-28	전자, 전기통신	대졸업				클래식	공포			구매
7		29-33	석유화학	대학원재				댄스	휴먼			검색
8		34-39	건설, 중공업	대학원졸				R&B	애니			
9		40-45	제조, 무역	기타				얼 터 너 티	미스터리			
10		46-50	운송업					브	어드밴처			
11		51-55	언론, 광고					뉴에잊	뮤지컬			
12		56-60	교육분야					가스펠				
13			보건, 의료					포크				
14			예술					국악				
15			스포츠									
16			군인, 공무원									
17			숙박, 요식업									
18			농, 임, 축, 수, 광업									
19			법률									
20			가사									
21			기타									

이용되는 이용자 프로파일을 생성하는데 있어서 이용되는 키워드를 가지고 있는 키워드 테이블을 생성한다. 〈표 5〉는 이러한 키워드 테이블의 예를 보여 주고 있다.

〈표 5〉에서 키워드 ID는 해당 키워드에 대한 문자열 값을 정수값으로 표현하기 위한 것으로서 실제 학습과정에서의 프로파일의 갱신 속도를 높이기 위한 것이다. 키워드는 초기 이용자의 입력 값과 함께 특정 콘텐츠에 대한 구매, 내려받기 등의 이용자 피드백 정보로부터 추출

〈표 5〉 이용자 프로파일을 위한 키워드 테이블

키워드 ID	영화(c_1)			음악(c_2)			그림(c_3)	
	장르(c_{11})	배우(c_{12})		장르(c_{21})	가수(c_{22})	.	.	.
1	액션	로버트드니로		댄스	HOT	.	.	.
2	코미디	발킬머		팝	조성모	.	.	.
3	멜로	기네스펠트로		영화음악	임창정	.	.	.
4	SF	존쿠삭		락	유승준	.	.	.
5	공포	이미연		째즈	박완규	.	.	.
6	휴먼	이정재		클래식	샵	.	.	.
7	애니	강수연		라틴	GOD	.	.	.
8	.	.		발라드	.	.	.	.
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
n	.	브루스윌리스	.	.	마돈나	.	.	.

한다. 한편, 다양한 콘텐츠 중에서 비슷한 종류의 콘텐츠는 같은 키워드를 갖게 되므로 이용자가 같은 키워드를 가지는 콘텐츠에 대하여 피드백 가중치가 높은 행위를 하는 경우 해당 키워드에 대한 가중치가 높아지게 된다.

나. 이용자 프로파일

이용자 프로파일은 위의 이용자에 대한 기본 정보를 포함하고 있는 기본 프로파일과 달리 실제 콘텐츠 프로파일과의 매칭 및 학습에 이용되는 프로파일이며 실질적인 추천에 있어서 중요한 기능을 수행한다.

〈그림 1〉에서도 볼 수 있듯이 이용자 프로파일은 크게 두 가지의 요소로 구성되어진다. 첫 번째 요소는 특정범주에 대한 이용자의 선호도를 보여주는 범주선호도로서 이는 해당 범주내의 콘텐츠에 대한 이용자의 선호도를 표현하는 정보를 갖고 있으며 〈범주, 가중치〉의 쌍으로 표현된다. 이를 수식으로 표현하면 수식(2)과 같이 표현 할 수 있으며 $W(C_i^p)$는 프로파일(P)에서 해당 범주(i)의 가중치 값을 표현한다.

$$P = \{(C_i^p, W(C_i^p))\} \quad (2)$$

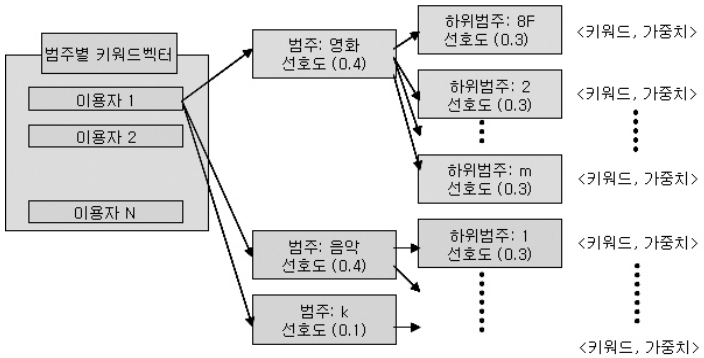
두 번째 요소는 각각의 범주에 대한 가중치가 부여된 키워드 벡터값이라 할 수 있다. 즉 위의 수식에서 C_i^p는 아래의 수식(3)으로 표현할 수 있다.

$$C_i^p = < w_{i,j}^p > \quad (3)$$

다음의 〈그림 1〉은 이용자 프로파일의 예를 보여 주고 있다.

이용자 프로파일은 이용자가 선호하는 콘텐츠를 찾기 위해서 필요한 이용자의 선호 정보를 가지고 있는데, 실제로 개인화 서비스를 위해서 가장 중요한 정보를 가지고 있으며 이러한 이용자 프로파일은 이용자의 초기 입력 정보와 이용자의 행동양식 정보를 바탕으로 구성된 이용자 각각의 프로파일을 바탕으로 개인의 콘텐츠에 대한 선호도를 갱신해 나간다.

추천할 콘텐츠를 결정하기 위해서는, 미리 선택된 추천 항목들을 일정 규칙에 따라서 추천하는 방법과 실시간에서 이용자의 행동 위치를 감시하여 이용자에게 적합한 추천 항목들을 추천하는 방법이 가능할 수 있다. 전자의 경우 추천이 매우 빨리 이루어 질 수 있으나 이용자

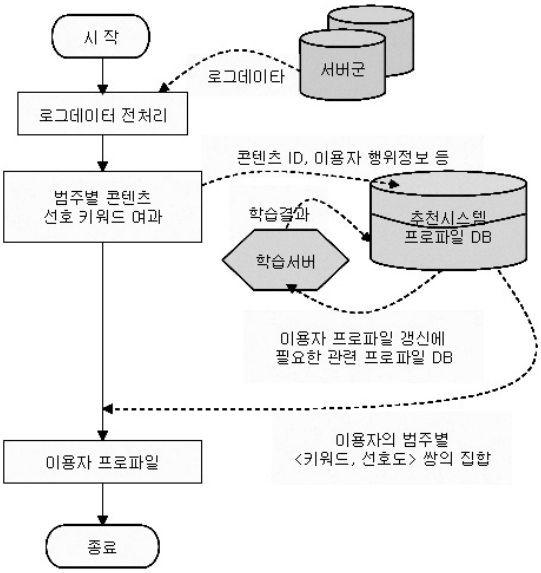


〈그림 1〉 이용자 프로파일 구조

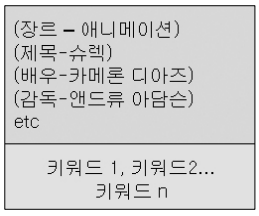
에게 보다 정확한 콘텐츠를 추천할 수 없다는 단점이 존재하고, 후자의 경우는 실시간 추천을 위한 모듈이 서버의 부하를 높일 수 있다. 〈그림 2〉는 전처리과정을 거쳐 이용자의 프로파일의 생성과 갱신 흐름을 표현한 것이다.

3.2.2 콘텐츠 프로파일(Contents profile)

콘텐츠들에 대한 프로파일은 콘텐츠의 특성을 나타내는 것으로, 콘텐츠 프로파일의 표현은 이용자 프로파일의 구성과 비슷하다고 할 수 있다. 콘텐츠 프로파일은 추천 시스템에서 정의된 필드들로 구성되어 있으며 각 필드들은 해당 콘텐츠가 속하는 장르(Genre), 제목(Title), 주인공(Actor), 감독(Director) 및 해당 콘텐츠에 대한 키워드로 구성되어져 있다. 〈그림 3〉은 영화 슈렉(Shrek)에 대한 가장 기본적인 콘텐츠 프로파일의 예이다. 각 필드에 따른 가중치 값은 중요도에 따라 임의로 결정한다. 예를 들어, 영화 콘텐츠에서 키워드로서 추출될 수 있는 주연배우와 조연배우에 대한 가중치는 다르다고 할



〈그림 2〉 이용자 프로파일 생성 및 갱신 흐름도



〈그림 3〉 콘텐츠 프로파일의 예(영화: 슈렉)

수 있다.

콘텐츠 프로파일에 포함되는 콘텐츠 키워드는 해당 키워드의 중요도에 따라 가중치가 부여된다. 따라서 각각의 콘텐츠는 키워드 벡터 값으로 아래와 같이 표현 할 수 있으며 $w_{i,j}$는 콘텐츠 C_i의 키워드 c_j에 대한 가중치 값을 표현한다.

$$C_i = < w_{i,j} > \quad (4)$$

〈그림 3〉에서와 같이 콘텐츠는 해당 콘텐츠의 장르, 감독, 배우 등과 같은 다양한 정보들과 함께 콘텐츠의 키워드 정보로 구성되어져 있다. 따라서 위에서 언급한 키워드 벡터는 정규화의 과정을 거쳐야 할 필요성이 있다.

콘텐츠 프로파일은 해당 콘텐츠에 대한 특징 또는 정보를 담고 있는 키워드를 가지고 있으며 해당 키워드들은 중요도에 따라 키워드 벡터로 표현되는 가중치가 부여된다.

$$K_i^c = < w_{i,j}^c > \quad (5)$$

수식(5)에서 '$w_{i,j}^c$'는 콘텐츠의 키워드에 대한 가중치를 의미하며 'i'는 키워드를 의미하고 'c'는 'K_i^c'가 이용자 프로파일과는 대비되는 콘텐츠 프로파일임을 의미한다. 또한 콘텐츠는 콘텐츠를 표현할 수 있는 대표적인 키워드들로 구성되며 따라서 콘텐츠는 아래와 같이 키워드 벡터의 집합으로 표현될 수 있다

$$C = \{K_i^c\} \quad (6)$$

다음 〈표 6〉의 콘텐츠 프로파일은 임의로 구성한 것이며, 다음절에서 기술할 프로파일 매칭과 관련한 일부분의 키워드 값만을 표시하였다. 콘텐츠 프로파일 테이블에서 CID는 각 콘텐츠의 등록번호를 나타내며, a_1 ~ a_k는 콘텐츠의 특성을 각각 나타내는 속성들이다. 예를 들어, a_3는 음악 범주의 장르 속성이 될 수 있

〈표 6〉 콘텐츠 프로파일의 예

	CID	SUB_C	a_1	a_2	a_3	a_4	a_5	a_6	...	a_k	b_1	...	콘텐츠 내용
c_1	6												
	18												
	36	c_{11}	7	33	4	44	38	121	5				
c_2	7	c_{21}	21	45	***1***	7	***4***	11	2	.	.	.	댄스 유승준 열정
	11	c_{21}	16	41	8	27	35	77	.	.	.	.	.
	17	c_{21}	76	39	8	9	3	3	.	.	.	.	발라드 임창정 연인
	8	c_{21}	.	.	.	.	.	.	.	.	.	.	
c_3	44	.	.	.	.	.	***4***	.	.	.	.	.	
	98	.	.	.	.	.	.	.	.	.	.	.	
	105	.	.	.	.	.	.	.	.	.	.	.	
c_4	85	.	.	.	.	.	.	.	.		.	.	

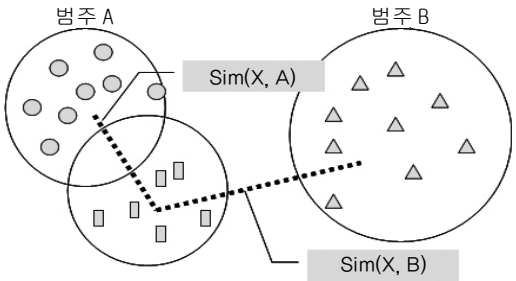
다. 콘텐츠 프로파일의 표현은 해당 서비스 사이트의 콘텐츠에 따라서 달라질 수 있으므로, 기존에 구축되어 있는 콘텐츠 구축정보를 참조하여 반영하도록 한다. 실제 개인화 서비스를 위한 시스템 구축을 위해서는 사이트에서 제공되는 콘텐츠의 특성에 따라 콘텐츠 프로파일의 필드들을 수정하여 적용 가능하다.

각 열(Row)의 속성(a_1, a_2, a_3, a_4, a_5, ..., a_k,..., b_1,..,)는 해당 콘텐츠의 특징을 기술하고 있는데, 다음 단계의 프로파일 매칭단계에서 보여지는 것처럼, 개인의 선호 프로파일과의 매칭 시에 속성값들이 사용되게 된다. 속성값 'b_1'은 프로파일 매칭에는 사용되지 않지만 콘텐츠를 표현하는데 꼭 필요한 정보를 담고 있다. 예를 들어 콘텐츠의 저장 위치나 유/무료 여부 등의 속성값이 될 수 있다.

3.2.3 범주 관계성(Category relation)

범주 관계성은 범주간의 연관관계를 표현하고 있다. 이러한 범주 관계성은 이용자가 이전에 경험하지 않았던 범주에 대한 가중치 값을 생성하기 위하여 적용되며 협업여과방법의 특징을 포함하고 있다. 즉, 특정 콘텐츠가 속한 범주가 c_1이라고 가정한다면 c_1에 대한 다른 범주인 c_2, c_3, ..., c_n에 대한 상대적 관계성이라고 할 수 있다. 본 연구에서는 범주간의 관계성을 획득하기 위한 가정으로서 하나의 콘텐츠 범주에 대해서(키워드, 가중치) 집합을 고려할 때 '서로 다른 두 범주 간에 있어서 동일한 키워드가 많을수록 두 범주간의 상호 관계성은 높다'고 가정한다. 이러한 범주 관계성을 얻기 위하여 임의의 값으로 초기값을 생성하고 이를 지속적으로 학습을 통하여 갱신한다.

범주간의 관계성은 〈그림 4〉와 같이 전체 콘텐츠 집합에서 기준이 되는 범주벡터의 중심점을 구한 후 이 거리로 측정할 수 있다. 이러한 범주간의 관계성을 수치화하여 표현한 범주 가중치는 상대적인 개념으로서 주체가 되는 측정 범주를 제외하고 해당 범주와 관계성이 있는 나머지 범주들의 가중치의 총합을 1로 가정함으로써 가중치의 과수치화에 따른 오류를 해결할 수 있다. 즉, 범주 가중치에 대한 정규화(Normalization)과정을 통하여 학습과정에서 발생할 수 있는 오류의 가능성을 사전에 제한함으로써 정확도를 향상시키고 있다. 예를 들어,(A, 0.3),(B, 0.4),(C, 0.5)가 범주 c_1에 속하고 c_2는(B, 0.4)(C, 0.6)(D, 0.1) 포함하고 있다고 가정하면, 콘텐츠 'B' 와 'C'에 대하여 c_1과



〈그림 4〉 범주간의 유사도를 통한 관계성

c_2의 범주 관계성에 있어서 C_1을 주체로 고려하였을 경우 콘텐츠 B에 대해서 c_1과 c_2는 같은 가중치를 가지고 있으며 c_2는 B에 대해서는 상호관계성이 1이 된다. 즉, 이를 비율로 계산을 하면 범주 c_1에 대해서 콘텐츠 B와 C는 각각 가중치가 (1)과 (1.2)로 계산할 수 있다. 또한 콘텐츠 B와 C에 대하여 c_3(1.4), c_4(0.8) 이라고 가정 한다면 각각 c_2(2.2), c_3(1.4), c_4(0.8) 이며 이를 정규화를 통하여 보정을 한다면 각각의 가중치는 c_2(0.5) c_3(0.32), c_4(0.18)이 될 수 있다. 즉, 범주 c_1에 대해서 c_2(0.5) c_3(0.32) c_4(0.18)의 관계성을 가진다.

이를 수식으로 표현 하면 다음과 같다.

$$\mathrm{sim}(\mathrm{x},\mathrm{y}) = \mathrm{C_x} \cdot \mathrm{C_y}$$
$$C_x = (c_{x1}c_{x2}...c_{xN}), C_y = (c_{y1}c_{y2}...c_{yN})$$
$$W_{\mathrm{x,y}} = \frac{1}{P_i}\sum_{k \in C^x} f_{ky} \qquad (7)$$

여기서 'k'는 기준 범주와 상대범주에 공통으로 포함되는 콘텐츠를 의미하며 'P_i'는 범주(i)에 속한 모든 콘텐츠들의 개수를 의미한다.

3.2.4 클러스터링(Clustering)

본 연구에서는 콘텐츠의 다중값 속성을 활용한 추천시스템 개발 방법론에 초점을 맞추고 있다. 이를 위하여 콘텐츠의 속성 정보중에서 범주 속성을 기반으로 K-means 클러스터링을 수행하였다. 보다 정확한 K-means 클러스터링을 위하여, 군집의 수를 각 각 4, 5, 6개로 변화시키며 실험을 수행하였으며 또한 계층적 클러스터링 방법을 통한 실험도 수행하였다. 군집의 수를 변화시킨 비계층적 방법과 계층적 방법을 통한 실험에서 얻어진 결과에 따라 군집의 수를 4개로 지정하여 K-means 클러스터링을 수행하였다. 한편, 다양한 속성 중에서 범주 속성별로 클러스터링을 수행하는 이유는 각각의 콘텐츠가 본래 하나의 범주 특성만을 가진 것이 아니라, 여러 범주의 특성을 가지고 있기 때문이다. 예를 들어, 영화 콘텐츠 중에서 '토이 스토리'는 어린이, 코미디, 애니메이션의 3가지 범주 특성을 가지고 있는 것이다. 이와 같이 대부분의 콘텐츠는 여러 범주의 특성을 지니고 있기 때문에, 하나의 범주로 단정하기에는 애매모호한 점이 많다. 하지만, 범주 속성별로 클러스터링을 수행하면 보다 포괄적인 범주로 재정의되어 어떤 종류의 영화인지 알 수 있는 것이다. 특히, 본 연구에서는 원본 데이터의 다양한 범주 중에서 범주속성을 대표하는 데이터만을 실험 데이터로 사용하였다. 이러한 이유는 다양한 콘텐츠가 속한 범주에서 극히 적은 데이터가 포함된 예외적인 범주를 대표속성으로 하는 경우 전체 콘텐츠에 대한 속성을 대표할 수 없으며 이러한 예외적 사례들은 모집단을 대표하지 못하고 따라서 실험을 왜곡시킬 수 있기 때문에(여운승, 2000; Hair et al. 1998), 전처리 과정에서 예외적인 속성값들은 제거하였다.

3.3 프로파일 매칭

이용자 프로파일과 콘텐츠 프로파일이 생성되면 학습 시스템은 이용자가 해당 사이트에 로그인하여 서비스를 이용할 때마다 이용자의 행위정보를 바탕으로 이용자 프로파일의 범주별 선호도를 갱신하게 된다. 그리고 이용자가 선호하는 콘텐츠를 찾기 위해서 이용자 프로파

일을 바탕으로 콘텐츠 프로파일과의 매칭을 수행하게 된다.

전통적인 벡터공간모형에서 질의어와 문서의 매칭은 질의어 벡터와 근접한 벡터값을 찾아서 관련 문서를 검색하는 방법으로서 일반적으로 두 벡터간의 유사도값을 찾는 방법으로는 상관관계 계수방법과 코사인 유사계수 방법이 사용된다. 본 연구에서는 콘텐츠 프로파일과 이용자 프로파일의 유사도를 계산하기 위하여 코사인유사계수 방법을 사용하였다. 이러한 유사계수는 두 프로파일의 스칼라곱(Scalar Product)을 통하여 계산할 수 있다.

$$Sim\,(V_i, V_j) = \sum_k w_{ik} \times w_{jk} \qquad (8)$$

이용자 프로파일과 콘텐츠 프로파일에 대한 유사도를 측정하기 위하여 먼저 일치하는 범주간의 유사도에 대한 계산이 이루어진다. 이때 범주별 유사도를 측정하기 위해서는 위의 수식(8)과 같은 방법을 적용할 수 있다. 한편, 범주별 유사도 값이 계산이 되면 콘텐츠 프로파일과 이용자 프로파일에서 일치하는 범주에 대한 유사도값이 수식(9)을 통하여 얻어지며 프로파일의 범주 가중치에 의해 계산된 범주별 유사도값의 총합이 콘텐츠 프로파일과 이용자 프로파일의 유사도값이 된다. 이러한 최종적인 두 프로파일의 유사도값은 수식(10)을 통하여 얻을 수 있다.

$$Sim\,(K_i^p, K_j^c) = \sum_k w_{ik} \times w_{jk} \qquad (9)$$

$$Sim\,(C, P) = \sum_i sim\,(K_i^c, K_i^p) \times (a) \qquad (10)$$

서로 다른 범주에 대한 유사도값은 해당 범주의 가중치에 의해 계산되어 수식(10)을 통하여 추가된다. 그러나 특정 범주 또는 키워드에 할당된 가중치의 범위가 1이상이 된다면 산출값에 대한 왜곡이 일어날 수 있으며 이러한 왜곡된 값의 반영은 추천의 정확성에 많은 문제점을 야기 할 수 있다. 따라서 가중치값에 대한 정규화 과정이 필수적이라고 할 수 있다. 정규화과정을 통하여 얻어지는 스칼라곱의 범위는 (-1, 1)로서 정규화과정을 통하여 범주에 대한 가중치 값을 제한 할 수 있다.

$$|C_i^p| = |C_i^c| = 1 \;\Rightarrow -1 \le Sim(C_i^p, C_i^c) \le 1 \;\forall i \qquad (11)$$

〈표 7〉은 이용자 프로파일과 콘텐츠 프로파일과의 프로파일 매칭을 위해서 이용자와 콘텐츠의 프로파일을 간단히 함께 나타낸 것으로 이용자 프로파일의 범주 c_2에 대하여 이용자가 선호하는 콘텐츠를 찾는 프로파일 매칭의 예를 보여주고 있다.

먼저 이용자 'JHP'의 최상위 범주 c_2가 음악 범주(Level 1)로서 하위 범주인 c_{21}이 음악의 가요장르(Level 2)에 대해서, 첫 번째 선호 키워드 번호가 '1'이고 선호도 값이 0.4이며 a_3 가 음악장르를 나타내고, a_5가 가수이름을 나타낸다고 가정한다. 즉, 가요음악 중에서 "댄스"에 대한 선호도가 0.4라는 것이고, 또 키워드 번호 '8', '3' 그리고 '4'도 역시 각각 선호하는 키워드를 나타내고 있다. 따라서 해당 키워드를 속성값으로 갖는 콘텐츠를 찾아보면 여러 개가 있을 수가 있고 이들 중에서 가장 유사도가 높은 콘텐츠 i+1을 추천을 위해서 우선적으로 선택할 수가 있다.

다음 예에서 매칭의 결과로 이용자 'JHP' 프로파일의 'c_{21}'범주와 관련이 있는 콘텐츠($i+1$)

〈표 7〉 이용자 프로파일과 콘텐츠 프로파일의 매칭

이용자 프로파일								
이용자 (JHP)	c_1	c_{11}	7	0.1	c_{12}	23	0.1	⋮
	c_2	c_{21}	1	0.4	c_{22}	21	0.1	⋮
	c_2	c_{21}	3	0.1	c_{22}	39	0.2	⋮
	c_2	c_{21}	8	0.2	c_{22}	27	0.3	⋮
	c_2	c_{21}	4	0.3	c_{22}	13	0.4	⋮
	c_3	c_{31}	65	0.6	c_{32}	79	0.7	⋮
	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	c_i	⋮	⋮	⋮	⋮	⋮	⋮	⋮

콘텐츠 프로파일									
	CID	Sub-C	a_1	a_2	a_3	a_4	a_5	a_6	⋮
콘텐츠(i)	m	c_{11}	7	33	4	44	38	91	⋮
콘텐츠 ($i+1$)	m_{i+1}	c_{21}	21	45	1	7	4	6	⋮
콘텐츠 $i+2$	m_{i+2}	c_{21}	16	41	8	27	35	77	⋮
콘텐츠 $i+3$	m_{i+3}	c_{21}	76	39	8	9	3	⋮	⋮
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
콘텐츠 $i+k$	m_{i+k}	C21	81	71	⋮	⋮	4	⋮	⋮

과 콘텐츠($i+3$) 등을 찾아낸다. 이렇게 선택된 콘텐츠들의 추천 가능성을 범주에 대한 유사도를 바탕으로 결정하여, 추천 후보 목록 테이블을 구성한다. 즉, 이용자의 선호도에 따른 콘텐츠 범주별 추천 후보 목록이 데이터베이스 테이블에 저장된다.

4. 콘텐츠 추천을 위한 학습방법

4.1 프로파일 학습과정

이용자 프로파일을 학습하는 과정은, 확보된 이용자 정보를 이용하여 이용자의 실제적인 관심과 선호도를 반영하기위한 과정으로 이용자 행위정보, 콘텐츠 가중치 및 이용자의 콘텐츠에 대한 범주 가중치를 이용하여 학습을 수행하고, 학습의 결과를 이용자 프로파일에 반영하는 과정이다. 일반적으로 추천 시스템에서는 추천 콘텐츠의 생성을 위하여 이용자 프로파일과 콘텐츠의 유사도를 계산하기 위하여 벡터 유사도(vector similarity)값을 이용하여 유사도 값을 계산하였다. 이는 이용자 프로파일의 키워드와 콘텐츠 프로파일의 콘텐츠 키워드의 가중치를 곱으로 계산한 것이다.

$$Sim(P, C) = \frac{P \times D}{\sqrt{P^2 C^2}} \quad (12)$$

위에서 P는 이용자 프로파일 벡터를 의미하며 C는 콘텐츠 프로파일 벡터를 의미한다.

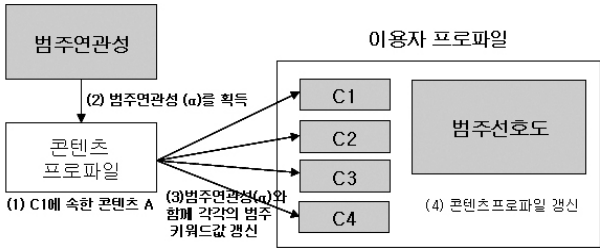
〈그림 5〉는 이용자의 웹상의 행위정보를 이용하여 이용자의 프로파일을 갱신하기 위한 학습 과정을 보여주는 것으로서 이를 보다 자세히 살펴보면 다음의 네 단계로 구분 할 수 있다. 첫 번째 단계는 범주 c_i에 속하는 콘텐츠 A에 이용자가 관심을 보인다고 가정하며 콘텐츠 A의 콘텐츠 프로파일정보를 입수한다. 두 번째 단계로서 범주 c_i와 다른 범주간의 범주 관계성을 구하며 이를 이용자 프로파일의 학습률로 적용한다. 세 번째 단계는 학습률(α)과 이용자의 행위정보(f_t)와 함께 이용자 프로파일을 갱신한다. 마지막 단계로서 이용자 프로파일의 범주선호도를 갱신함으로써 기계학습을 이용한 학습 과정이 마무리가 된다.

이용자 프로파일의 갱신을 위한 학습과정에서는 먼저 이용자 초기 프로파일을 구축한다. 이를 위하여 웹사이트의 구조정보를 이용하여 웹사이트를 다수의 범주로 세분화하고, 회원가입시 이용자가 입력한 기본정보, 이용자선호도 정보, 인구통계학적 정보 등을 이용하여 이용자 초기 프로파일과 웹사이트의 각 범주에 대한 가중치를 결정한다. 초기 이용자 프로파일이 구축되고 이용자의 콘텐츠에 대한 이용행위가 발생됨으로써 프로파일 갱신을 위한 학습과정이 수행된다. 학습과정에서는 구축된 이용자 프로파일과 본 연구에서 제안하고 있는 입력데이터와 함께 학습이 수행된다.

실질적인 학습과정에서는 이용자 피드백의 대상 콘텐츠가 속한 범주가 'c_1'이라 할 때 이용자 프로파일에서 'c_1'에 대한 범주의 선호도를 나타내는 범주 가중치(α)와 'c_1'에 속한 이용자 프로파일의 키워드와 가중치의 목록을 획득한다. 콘텐츠 프로파일에서는 콘텐츠의 키워드와 키워드의 중요도를 나타내는 가중치(d_i)를 추출하며, 이용자 행동데이터에서는 이용자 행동을 나타내는 피드백 값(f_t)을 추출한 후 이용자 프로파일의 변경값을 계산한다. 키워드 가중치 값의 갱신은 아래의 수식을 통하여 생성, 갱신된다.

$$W_i' = (W_i \times G_j) + (a \times F_t \times d_i) \quad (13)$$

(W_i): 이용자 프로파일의 키워드 가중치,(W_i') : 학습 후 이용자 프로파일의 키워드 가중치,(G_j): 그룹 선호도,(F_t): 피드백 값,(a): 범주 가중치,(d_i): 콘텐츠 프로파일의 키워드의 가중치



〈그림 5〉 이용자 프로파일 학습과정

예를 들어, 이용자 프로파일이 {(1, 0.1),(2, 0.2),(3, 0.3),(4 ,0.4)}과 같이 구성되고, 콘텐츠 프로파일이 {(1, 0.1),(2, 0.5),(4, 0.1),(5, 0.2)}와 같이 구성되고, 콘텐츠가 속한 범주 가중치(a)가 0.2, 그룹 선호도가(0.3), 이용자 피드백(f_t)값이 0.6이라고 한다면, 학습 후의 키워드 1에 대한 가중치는(0.1 * 0.3) +(0.2 * 0.6 * 0.1)로 계산 된다. 그러나 결과값이 일정 임계치 이상인 경우만 이용자 프로파일 및 범주 가중치를 변경시키고 임계치보다 작으면 변경하지 않는다. 이는 가중치 값의 변경이 적은 경우 추천을 위한 콘텐츠 선정과정에서 별 다른 영향이 없기 때문에다.

모든 키워드에 대하여 키워드의 가중치를 변경하고 이들 값을 다시 정규화(Normalization) 함으로써 이용자 프로파일의 변경이 종료되며 웹사이트의 모든 범주의 가중치를 변경하고 이들을 정규화 함으로써 이용자 프로파일의 학습이 끝난다.

4.2 학습을 위한 입력정보

기계학습을 위한 입력정보는 유사한 행위를 보여주는 이용자들을 그룹화하여 해당 이용자 그룹이 보여주는 콘텐츠에 대한 선호도와 함께, 각각의 이용자들이 콘텐츠를 이용하는 행위정보인 이용자 피드백정보가 기준이 된다. 이와 같은 이용자 피드백정보는 이용자의 피드백이 이용자 프로파일에 영향을 미치는 정도를 나타내는 학습률을 반영한다. 또한 추가적인 정보로서 콘텐츠 프로파일 가중치가 있으며 해당 정보는 이용자가 피드백의 대상이 되는 콘텐츠 프로파일에서의 해당 키워드에 대한 가중치로서 콘텐츠 프로파일의 속성들을 중요도에 적용한 것이다.

4.2.1 이용자 그룹 선호도

기존의 협업여과 방식이 콘텐츠의 특성을 반영하지 않아 이용자의 성향을 정확히 분석하지 못하는 문제점을 해결하기 위하여 본 연구에서는 콘텐츠의 특성 정보를 이용한 이용자의 콘텐츠 선호도를 반영하였다. 이용자의 콘텐츠에 대한 이용행위에 대한 피드백값은 2차원 배열로 표시되는데 이용자(i)가 콘텐츠(j)를 이용하였을 경우 배열의(i, j)값은 '1'이고 이용하지 않은 경우는 '0'으로 표시한다.

다음 수식(14)는 본 연구에서 적용하고 있는 협업여과방식을 통하여 이용자의 콘텐츠 선호도를 계산하기 위한 수식이다.

$$B(i,j)=\frac{1}{m(i)}\sum_{i=1}^{n}(CAT(j,j^{*})*A(i,j)) \qquad (14)$$

여기서, n은 전체 콘텐츠의 수이고, m(i)는 이용자(i)가 이용자의 콘텐츠에 대한 이용행위를 보여주는 콘텐츠의 수를 나타낸다. A(i, j)는 이용자의 콘텐츠에 대한 이용행위(구매, 내려 받기 등)내역으로서, 2차원 배열로서 이용자(i)가 콘텐츠(j)를 사용하였을 경우 배열의(i, j)원소가 1이 되고 사용하지 않았을 경우는 0이다. CAT(j, j^{*})는 콘텐츠(j)와 콘텐츠(j^{*})가 같은 범주에 속하면 1을, 그렇지 않으면 0을 전달하는 함수이다. B(i, j)는 콘텐츠 선호도로서, 2차원 배열이고 이용자(i)가 콘텐츠(j)를 얼마나 선호하는가를 나타낸다.

이용자의 콘텐츠 선호도를 구한 후, 비슷

한 콘텐츠 선호도를 가지는 이용자끼리 군집화(Clustering)를 수행한다. 동일한 군집내의 다른 이용자들의 콘텐츠 선호도를 이용하여 해당 이용자의 콘텐츠에 대한 이용행위에 대한 가능성을 구한다.

콘텐츠 이용 가능성 $P(a,j)$은 수식(15)에서 구해지는데, 이용자 "a", 콘텐츠 "j"에 대한 예상선호도 $P(a,j)$은 일반적인 협업여과방식에서와 같이 다음의 식을 이용하여 구할 수 있다.

$$P(a,j)=\mu(a)+k\sum_{i=C}(w(a,j)*(B(i,j)-\mu(a)) \quad (15)$$

수식(15)에서 "$\mu(a)$"는 이용자 "a"가 콘텐츠를 이용한 평균값이고, "k"는 정규화를 위한 파라미터이다. 그리고 $w(a,i)$는 이용자 "a"와 이용자 "i"의 콘텐츠 이용 내역에 대한 유사도를 나타내는데 보통 상관계수나 벡터유사도를 이용하며 본 연구에서는 벡터 유사도값을 적용하였다.

4.2.2 학습률

학습률은 이용자의 피드백이 이용자 프로파일에 영향을 미치는 정도를 나타내는 요소로서 학습률(a)을 결정하는 방법이 내용기반 추천시스템의 특징을 나타내며 대상이 되는 응용서비스 시스템의 종류에 따라 학습률을 결정하는 방법이 달라진다. 실제로 학습률(a)값을 결정하기 위한 방법은 여러 가지 고려가 필요하나 값은 $0 < a < 1$ 사이의 값이 된다. 이는 실험을 통해서 적정한 값을 결정하거나 다른 방법을 고려할 수 있다. 초기 학습률값이 설정이 되고 이용자의 피드백 유형을 통하여 지속적으로 이용자 프로파일에 미치는 영향을 분석하면서 학습률을 조절하여야 한다. 그러나 본 연구에서는 초기 학습률값을 동일하게 1로 설정하고 학습률의 효과를 거둘 수 있는 범주간의 관계성을 보여주고 있는 범주 가중치 값을 학습률로 적용하였다.

4.2.3 이용자 피드백

이용자 피드백은 피드백의 유형에 따라 그 중요도가 다르므로 이용자의 입력 값에 중요도를 반영한 결과값으로서 피드백의 유형에 따라 피드백값이 학습에 대한 기여도가 달라져야 한다. 따라서 고려 가능한 피드백의 종류를 기술하고 각 유형에 따른 가중치 값은 변화한다. 본 연구에서 고려하고 있는 이용자 피드백의 유형은 명시적 피드백과 암묵적 피드백으로 구분할 수 있으며 추가적인 피드백으로서 긍정적 피드백과 부정적 피드백으로 구분 할 수 있다. 본 연구에서의 이용자 피드백값은 부정적인 반응은 고려하지 않는다. 이용자 피드백값은 $0 < Ft < 1$의 범위로 설정하며 피드백 유형에 따른 가중치는 〈표 8〉과 같이 적용한다.

현재 위의 피드백유형의 각각에 대해서 보다 세분화된 가중치를 결정해야 하나 현재의 실험 데이터로서 이를 판단하기는 어려우며 따라서 본 연구에서는 위의 피드백 유형을 세 가지 타입으로 분류하였으며 향후 데이터가 충분히 확보된 후 가중치의 유형별 분류의 세분화 및 각 요소의 계층을 조정하고, 가중치를 조정할 필요가 있다.

추가적인 방법으로서 조회수, 사용시간, 순위화 등은 일정 가중치를 주는 것이 아닌 상태에 따라 가중치가 변화하는 규칙을 결정해야 한다. 이를 위하여 일정 임계치 값을 설정하여

〈표 8〉 피드백 유형별 가중치 중요도

	피드백 유형	가중치 중요도		피드백유형	가중치 중요도
1	내려받기	4	6	출력	4
2	실행(Play)	3	7	검색	3
3	구매	4	8	단순클릭	1
4	쇼핑카트담기	3	9	사용시간	--
5	조회/방문빈도수	--	10	기타(게시판사용)	1

해당 값 이상인 경우에 이를 이용자 피드백으로서 반영하였다.

4.2.4 콘텐츠 프로파일 가중치

콘텐츠 프로파일 가중치는 이용자가 피드백의 대상이 되는 콘텐츠 프로파일에서의 해당 키워드에 대한 가중치로서 콘텐츠 프로파일의 속성들을 중요도에 따라 가중치를 적용한다. 이러한 가중치 값은 콘텐츠의 종류에 따라 달라 질 수 있으며 본 연구에서는 대상이 되는 콘텐츠의 종류인 영화와 음악 등의 멀티미디어 콘텐츠로서 해당 가중치 값의 범위는 0 〈 Di 〈1 로서 전체 가중치의 합은 1이 되도록 정의한다. 예를 들어 영화 콘텐츠의 경우 주연배우가 조연배우보다 중요도가 높다고 할 수 있다. 〈표 9〉는 본 연구에서 정의하고 있는 콘텐츠 프로파일의 가중치 값을 보여주고 있다. 콘텐츠 프로파일의 가중치 역시 적용되는 사이트의 특성을 반영하여 변경된다.

유형 1의 경우 제목은 가장 높은 가중치 순위를 가지고 있으며, 대부분의 콘텐츠 이용자들의 선호도가 가장 높은 유형이며, 유형 2 ~ 5는 각각의 키워드에 대한 가중치별 순위를 콘텐츠 유형별로 정리하였으며 이러한 유형별 가중치 값은 초기에 임의로 결정을 하고서 필요시 변경이 가능하다. 또한 〈표 9〉의 가중치 유형별 가중치의 값은 정규화 과정을 통하여 전체 합이 1이 되도록 한다.

4.3 위치정보 기반의 콘텐츠 추천방법

학습을 통하여 이용자 프로파일의 생성 및 갱신이 이루어지면 이용자의 선호도에 따른 콘텐츠 추천목록이 생성된다. 이렇게 1차적으로 생성된 추천목록은 다시 현재 이용자의 해당 웹사이트에서의 위치하고 있는 정보를 기반으로 하여 다시 2차 추천목록을 생성하기 위한 여과과정을 거친다. 〈그림 6〉은 추천 시스템에서

〈표 9〉 콘텐츠 프로파일 항목별 가중치

가중치 유형 및 중요도	영화콘텐츠	음악콘텐츠
1	제목	제목
2	주연배우	가수
3	감독	작곡가, 작사가
4	장르	장르
5	기타	기타

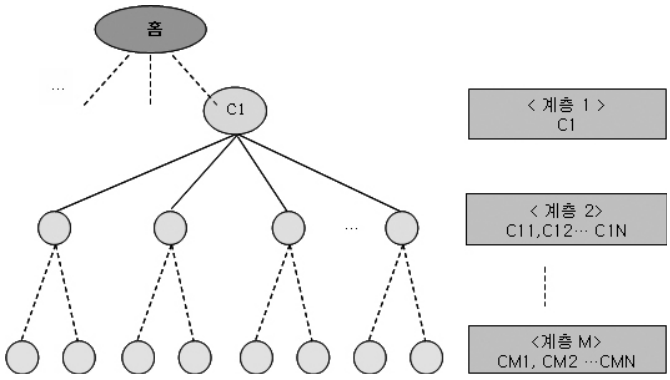
의 실시간 추천 과정을 설명하기 위한 전체 흐름도이다. 본 연구의 콘텐츠 추천 방법에서는 콘텐츠의 범주별 이용자의 선호 프로파일을 구성하고 그리고 실제 서비스에서 이용자가 해당 사이트에 접속하면 추천 시스템에게 이용자 선호 프로파일에 대한 정보를 미리 요구하게 된다. 이것은 본 연구에서 추천에 따른 실시간 추천 속도를 높이기 위한 것으로, 이를 위한 버퍼 관리자 등을 사용하여 추천 후보 목록의 색인 목록을 유지하는 방법을 사용하였다.

학습 과정을 통하여 추천 시스템으로 부터 추천 후보 목록을 받게 되면 후보 목록 중에서 이용자의 위치에 따른 추천을 제공하기 위한 추천 목록을 결정한다.

본 연구에서는 이러한 이용자 위치에 따른 추천 목록을 결정하기 위해서 웹사이트 계층을 〈그림 6〉과 같이 논리적 M 단계로 나누어 고려하였다.

웹사이트의 구조가 〈그림 6〉과 같이 표현될 수 있다는 것은 콘텐츠의 분류가 명확히 이루어져 있다는 것을 의미하는 것으로, 콘텐츠의 범주 분류에 따라서 그 구조와 계층이 정해질 수 있다. 그래서 해당 웹사이트의 계층, 다시 말해서 콘텐츠의 분류 구조와 이용자의 위치를 연관시켜서 이용자의 위치에 따른 적절한 콘텐츠의 추천을 가능하게 한다. 이용자의 프로파일이(키워드, 선호도)의 쌍으로 표현된다고 했을 때, 이를 반영한 이용자 프로파일의 구조를 〈그림 7〉과 같이 나타낼 수 있으며, 이용자는 각각의 범주 그룹에 대해서 계층 별(키워드, 선호도) 쌍의 집합을 가지는 것으로 생각할 수 있다.

〈그림 8〉은 추천 후보 목록을 웹사이트의 계

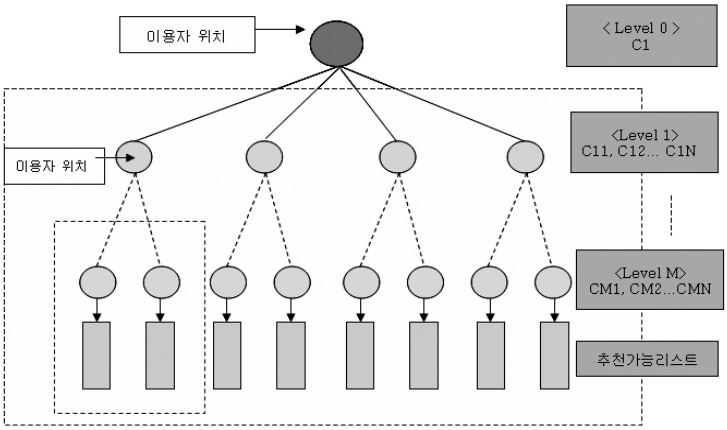


〈그림 6〉 이용자의 위치에 따른 콘텐츠를 추천 과정

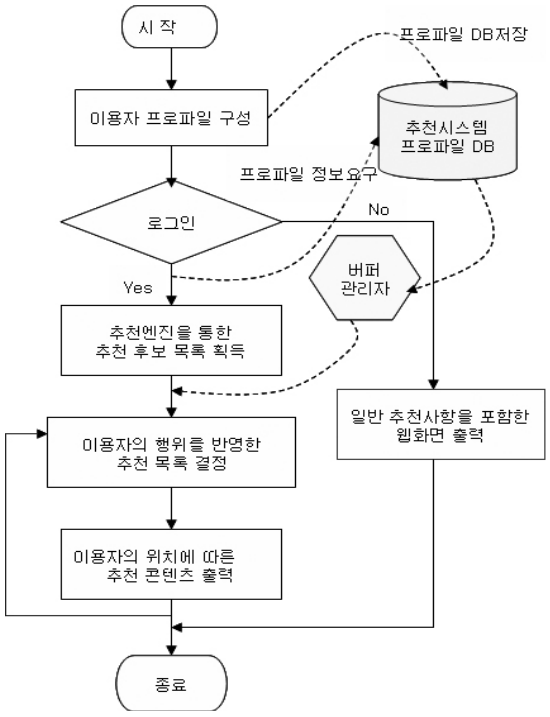
이용자 ID	C1 그룹	Lev1	…	LevM	계층별 <키워드, 선호도> 집합
	…	…	…	…	…
	Ck 그룹	…	…	…	계층별 <키워드, 선호도> 집합
	…	…	…	…	…

웹사이트계층

〈그림 7〉 계층 별 이용자 프로파일 구성도



〈그림 8〉 계층별 웹사이트 구조도



〈그림 9〉 위치정보에 따른 추천목록 결정방법

층에 따라서 표현한 것이다. 이제 분류된 추천 후보 콘텐츠들을 대상으로, 추천 정책에 따라서 지정된 상위 N개의 범주별 추천 콘텐츠 목록을 결정하여 이용자에게 출력한다.

본 연구에서의 추천 목록 결정과정에서는 〈그림 9〉와 같이 이용자의 현재 위치 이하의 범

주에 속한 추천 후보 목록들의 모든 콘텐츠들에 대하여 각각의 유사도를 비교하여 유사도가 높은 상위 N개의 콘텐츠를 현재 위치에서의 추천 콘텐츠로 선정하고 추천 콘텐츠와 유사도의 쌍으로 이루어진 추천 목록을 작성한다. 최종적으로 추천 콘텐츠의 출력과정은 위에서 결정된 최종 추천 목록을 화면에 출력하는 것으로 이용자 화면의 일부분이나 새로운 윈도우 창을 띄우는 등 이용자의 정의에 따라 다양한 방법으로 출력하는 과정을 말한다.

5. 결 론

급변하는 정보유통분야를 둘러싼 환경의 변화는 도서관 및 정보센터에게 있어서는 다양한 도전과 기회를 제공하고 있다. 변화하는 환경에 적절히 적응함으로써 이용자의 요구에 능동적으로 대응하기 위한 도서관과 정보센터의 일련의 변화의 결과물로서 선택적 정보배포 서비스와 맞춤정보서비스가 등장하게 되었다. 그러나 현재 제공되고 있는 맞춤정보서비스는 정보추천에 있어서 키워드에 기반하여 이용자가 초기에 입력하였던 선호도를 반영하는 단어와 콘텐츠가 포함하고 있는 단어의 매칭에 따라 이용자에게 정보를 제고한다. 따라서 이용자가 필요로 하는 정보를 추천하는 과정에서 매우 단순하고 낮은 정확도를 보여줌으로써 추천정보 또는 콘텐츠에 대한 이용자의 만족도는 매우 낮은 수준이라고 할 수 있다. 그러나 본 연구에서 제안하고 있는 방법은 이용자의 추천 만족도를 높이기 위하여 이용자가 암묵적 및 묵시적으로 보여주는 다양한 정보를 정밀하게 가공 및 적용하고 있다. 특히, 여전히 정보의 홍수 속에 살아가는 이용자에게 있어서 자신이 가장 필요로 하고 만족도 높은 정보를 자동적으로 제공해 주는 서비스는 도서관 및 정보센터에서 추구하고 있는 개인화된 정보서비스를 실현하기 위한 가장 빠른 방법이라고 할 수 있다.

그러나 이와 같은 보다 발전된 형태의 개인화 서비스를 제공하기 위해서는 해결해야 할 문제점이 많다. 우선 개인화를 위한 가장 일반적인 준비과정이 개인 정보의 수집이다. 즉, 이용자의 신상정보, 이용자의 선호도, 구매 행태 등에 대한 정보를 수집하여 이를 바탕으로 이용자에 가장 적절한 정보를 제공한다. 그러나 이러한 명분에도 불구하고 이용자의 개인정보를 수집하는 것에 대한 반대 의견도 만만치 않다. 즉, 고도의 개인화를 위한 지나친 개인정보 수집은 항상 사생활 침해의 우려를 안고 있다. 하지만 이러한 정보들을 이용하는 도서관 또는 정보센터의 입장에서 투명성과 약관을 통해 이용자들의 동의를 얻고, 얻어진 정보는 진정한 이용자만족을 위한 정보만으로 이용할 것을 보장함으로써 정보유통에 있어서 커다란 흐름인 개인화서비스를 한층 더 나은 단계로의 발전을 이루어 낼 수 있는 토대를 마련해야 할 것이다. 한편, 시스템의 개발에 있어서 기관의 특성에 적절한 학습알고리즘의 선택과 정보에이전트 개발과 함께, 이를 개인에게 직접 전송하거나 개인의 웹사이트에 전송할 수 있는 전송시스템의 개발이 요구된다. 따라서 본고에서는 개인화서비스 시스템의 개발을 위한 알고리즘 및 적용 가능한 기술에 대하여 알아보고 개인화서비스 시스템의 전체적인 구성요소 및 전체시스템의 설계와 함께, 개인화서비스 시스템에서

가장 중요한 요소인 이용자의 성향을 분석하는 학습 시스템의 설계를 제안하고 있다. 따라서 이러한 알고리즘 및 개인화기술을 적용하여 실제 서비스가 될 수 있는 실제 시스템을 구현하고 구현되어진 시스템을 통하여 이용자에 대한 시스템의 피드백을 통하여 보다 효율적인 개인화서비스 시스템의 지속적인 보완과정이 필요하다고 할 수 있다.

한편 개인화서비스를 제공하기 위해서는 기술적인 방법론 이외에 어떻게 인터넷에 적용하여 효과를 거둘 것인가에 대한 부분이 상대적으로 중요한 위치를 차지하게 된다. 특히 과거와는 달리 정보제공자와 이용자의 구분이 불명확해지면서 웹 2.0의 대표적인 특징으로서 정보흐름의 쌍방향성이 강조되면서 이용자의 정보이용행동에 기반한 개인화 추천서비스에 대한 연구는 도서관 및 정보센터의 궁극적인 목표인 이용자 정보요구에 적합한 정보를 정확하고 효율적으로 제공하기 위한 중요한 수단을 제공할 수 있다.

참 고 문 헌

김용. 2006.『멀티미디어 콘텐츠를 위한 이용빈도 기반 하이브리드 추천시스템에 관한 연구』. 연세대학교 문헌정보학과 박사학위논문.

김용, 문성빈. 2005. 학습알고리즘 기반의 하이브리드 개인화 추천시스템 개발에 관한 연구.『한국문헌정보학회지』, 39(3): 75-81.

이기현, 고병진, 조근식. 2002. 연관 규칙과 협력적 여과 방식을 이용한 추천 시스템.『한국지능정보시스템학회논문지』, 8(2): 91-103.

이수정, 이형동, 김형주. 2004. 사용자 경향에 기반한 동적 추천 기법: 영화 추천 시스템을 중심으로.『정보과학회논문지: 소프트웨어 및 응용』, 31(2): 153-163.

정경용, 김진현, 정헌만, 이정현. 2004. 개인화 추천 시스템에서 연관 관계 군집에 의한 아이템 기반의 협력적 필터링 기술.『정보과학회논문지: 소프트웨어 및 응용』, 31(4): 467-477.

정영미, 이용구. 2002. 필터링 기법을 이용한 도서 추천 시스템 구축.『정보관리연구』, 33(1): 1-17.

황성희, 김영지, 이미희, 우용태. 2001. 인구통계학적 특성에 따른 협동적여과 알고리즘의 추천 효율 분석.『2001 한국데이타베이스 학회 춘계 논문집』, 362-368.

Billsus, D. and M. Pazzaini. 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.

Deshpande, Mukund and George Kapyris. 2004. "Item-Based Top-N Recommendation

Algorithms." *ACM Transactions on Information Systems*, 22(1): 143-177.

Fink, J. and A. Kobsa. 2002. "User Modelling for Personalized City Tours." *Artificial Intelligence Review*, 18: 33-74.

Hair, J. F., R. E. Anderson, R. L. Tatham, and W. C. Black. 1998. Multivariate Data Analysis, 5th Edition, Prentice-Hall, Inc: New Jersey.

Hill, Will, Larry Stead, Mark Rosenstein and George 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.

Linden, G., Brent Smith, and Jeremy York. 2003. "Amazon.com recommendations: item-to-item collaborative filtering." *IEEE Internet Computing*, 7(1): 76-80.

Rensnick, P., N. Iacovou, M. Suchak, P. Nergstorm and J. Riedl. 1994. "GroupLens: An Open Architecture for Collaborative Filtering of Netnews," Proc. of the 1994 Computer Supported Cooperative Work conference, 175-186.

Sarwar, B., G. Karypis, J. Konstan and J. Riedl. 2002. "Getting to Know you: Learning New User Preferences in Recommender System for Groups of Users." Proc. of the 7th International conference on Intelligent user interfaces, 127-134, 2002.

Sarwar, B., G. Karypis, J. Konstan. and J. Riedl. 2001. "Item based collaborative filtering recommendation algorithms," Proc. of the 10th International World Wide Conference, 285-295, 2001.

Shardanand, U. and Maes, P. 2000. "Social information filtering: Algorithms for automating 'word of mouth'." Proc. of ACM CHI '95 Conference on Human Factors in Computing Systems, 210-217.