학습알고리즘 기반의 하이브리드 개인화 추천시스템 개발에 관한 연구

A Study on Development of Hybrid Personalization Recommendation System Based on Learning Algorithm

김 용(Yong Kim)*
문 성 빈(Sung-Been Moon)**

목 차

초 록

인터넷의 발전과 성장은 웹상에서의 정보의 량에 있어서 폭발적인 성장을 가져 왔다 .이러한 웹상에서의 정보량의 증가는 정보이용자의 요구와 필요에 맞는 정보 제공을 위한 서비스로서 웹기반의 개인화서비스에 대한 요구를 더욱 더 강조하게 되었다. 개인화서비스는 정보이용자의 요구와 필요에 의해 현실화 될 수 있으며 이러한 정보이용자의 관심사와 정보요구는 지속적으로 또한 급격하게 변화되고 있다. 웹상의 수많은 정보로부터 정보이용자의 요구와 필요를 만족시킬 수 있기 위하여 본 논고에서는 이용자의 관심과 요구를 표현하기 위하여 이용자 프로파일 정보를 이용하였으며 이러한 이용자의 프로파일 정보는 이용자의 요구와 흥미에 대한 변화를 반영하기 위하여 지속적으로 갱신하였다 .본고에서는 정보이용자의 정보요구와 흥미의 변화를 지속적으로 이용자프로파일에 반영하기 위한 방안으로서 학습알고리즘을 제안하였다. 정보이용자의 정보에 대한 피드백을 기반으로 이용자의 정보에 대한 흥미와 요구는 본 고에서 제안한 학습알고리즘을 통하여 지속적으로 갱신 되므로서 정보이용자에게 보다 정확한 정보를 제공할 수 있다고 할 수 있다 . 이러한 학습알고리즘은 보다 개선된 하이브리드 정보추천시스템에 적용하였다.

ABSTRACT

The popularization of the internet has produced an explosion in amount of the information. The importance of web personalization is being more and more increased. The personalization is realized by learning user's interest. User's interest is changing continuously and rapidly. We use user's profile to represent user's interest. User's profile is updated to reflect the change of user's interest. In this paper we present an adaptive learning algorithm that can be used to reflect user's interest that is changing with time. We propose the User's profile model. With this profile user's interest is learned based on user's feedback. This approach has applied to develop hybrid recommendation system.

키워드: 개인화서비스, 학습알고리즘, 하이브리드, 추천알고리즘, 추천시스템, SDI, 선택적 정보배포
Personalization, Recommendation, Hybrid, Learning Algorithm, SDI

* KT 마케팅연구소 스마트카드서비스개발실, 책임연구원(yongkim@kt.co.kr)
** 연세대학교 문헌정보학과 교수(sbmoon@yonsei.ac.kr)
논문접수일자 2005년 8월 15일
게재확정일자 2005년 9월 13일

1. 서 론

인터넷의 확산과 폭발적으로 증가하는 정보량에 따라 전통적으로 정보를 관리하고 이를 이용자에게 제공하는 도서관 및 정보제공기관에게 있어서 많은 어려움과 한계를 가져오고 있다. 따라서 도서관 및 정보제공기관에서는 이러한 어려움을 극복하기 위하여 보다 효율적으로 이용자의 정보요구에 부응하기 위한 다양한 방법들을 시도하고 있다. 특히, 이용자의 정보요구에 보다 적합한 정보를 적시에 제공하는 것은 현재 도서관과 정보제공기관에게 있어서 가장 중요한 관심사로 떠오르고 있다. 이러한 일련의 노력중의 하나로서 정보검색엔진의 등장이라고 할 수 있다. 그러나 검색엔진은 부분적으로 정보이용자의 욕구를 충족시켜줄 수는 있으나 근본적으로 이용자의 요구를 충족 시켜줄 수는 없었다. 한편, 도서관과 정보제공기관에서는 정보이용자의 정보욕구에 보다 적극적인 방안으로서 전통적으로 제공하던 선택적 정보배포 (SDI: Selective Dissemination of Information) 를 한 단계 발전시킨 형태로서 맞춤정보서비스를 제공하고 있다. 이러한 맞춤정보서비스는 정보이용자의 정보에 대한 선호도 및 프로파일에 기반한 방법으로서 현재 많은 대학도서관과 정보센터에서 제공되고 있다. 이러한 맞춤정보서비스는 보다 효율적인 정보서비스를 제공하기 위한 기초 자료를 제공하는 것으로서 최신정보제공서비스, 전자지정자료 관리, 도서관 정보서비스 관리 등의 서비스를 제공하고 있다. 현재 도서관 및 정보센터를 포함한 다양한 정보제공기관에서 제공되고 있는 맞춤정보서비스는 사이버공간상에서 이용자가 원하는 정보만을 효과적으로 검색하여 배달함으로써 이용자의 정보에 대한 맞춤화(Customization) 욕구를 충족시키고자 개발되었다. 이러한 일련의 노력으로서 다양한 맞춤정보서비스로서 웹상에서 개인별 프로파일정보에 따른 MyLibrary 서비스가 제공되고 있으며 그 효과는 약 75%이상의 이용자들이 해당 서비스에 만족을 표시하고 있다(김현희 2002). 비록 많은 이용자들이 현재 제공되고 있는 맞춤정보서비스에 대하여 많은 부분에서 만족하고 있다고는 하지만 실질적인 만족도에 대한 수준은 그리 높다고 할 수 없다. 그 이유는 대부분의 맞춤정보서비스를 통하여 제공되고 있는 정보는 신문기사 또는 단편적인 내용이기 때문이며 또한 정보를 추출하는 방법에 있어서 이용자의 입력에 따른 키워드매칭 방법이기 때문에 여전히 정보과잉에 대한 문제점은 존재한다. 특히 키워드매칭에 따른 정보의 적합성을 판별은 해당 키워드가 분야별로 다른 의미를 가지는 경우 이용자가 원하는 것과는 전혀 다른 문서가 검색될 수 있는 한계점을 가지고 있다(남궁 황 2003). 이러한 단순한 맞춤정보서비스에 대한 한계점을 극복하기 위한 방법으로서 e-CRM 분야의 핵심으로서 개인화서비스에 대한 관심이 매우 높아지고 있으며 현재 많은 기업뿐만 아니라 인터넷서비스 업체에서 서비스되고 있다. 이러한 사회, 경제적인 흐름과 함께, 정보서비스를 제공하는 도서관 및 정보센터에서도 MyLibrary 또는 MyPage라는 서비스로서 개인화서비스가 제공되고 있다. 특히 이러한 개인화서비스는 정보이용자의 요구에 적합한 정보를 제공하기 위하여 이용자의 성향을 파악하는 것이 중요한데, 이용자의 성향은 시간의 경과 및 상황에 따라 변화하므로 이를 반영

하기 위한 이용자의 성향을 반복적으로 학습하여 적용하는 학습과정이 필요하다. 이를 위하여 본고에서는 2장에서는 본 연구의 의의에 대해서 알아보고 있으며 3장에서 개인화 추천 시스템에서 간단히 알아보고 4장에서는 제안하고 있는 추천방법론에 대하여 기술하고 있으며 마지막으로 결론과 함께 향후 연구분야에 대하여 기술하고 있다.

2. 연구의 목적

정보기술과 웹의 발전은 폭발적인 정보의 생성과 유통을 초래하게 되었다. 따라서 증가하는 정보의 효율적인 관리 및 이용자의 요구와 관심에 적합한 정보를 적시에 제공하여야 하는 도서관 및 정보센터로서는 전통적으로 정보를 제공하는 방법과는 달리 새로운 정보관리 및 서비스 기법이 필요하게 되었다. 따라서 이러한 사회, 문화적인 요구에 따라 이용자요구에 적합한 정보의 추출과 제공을 위한 방법으로서 대량의 정보에서 개인별 맞춤화된 정보와 서비스를 제공할 수 있는 개인화서비스에 대한 관심은 더욱더 높아지고 있다고 할 수 있다. 이를 위하여 본 연구에서는 개인화서비스를 제공하기 위한 추천기법의 장단점에 대해서 알아보고 보다 효율적이면서 정확성을 높일 수 있는 추천기법을 적용한 개인화추천시스템을 구현하였다. 또한 전통적인 텍스트기반의 정보만이 아닌 다양한 동영상을 포함한 정보를 대상으로 하기 위하여 다양한 동영상 자료를 포함하고 있는 웹포탈사이트를 실험대상으로 하였다. 비록 실험사이트가 도서관이나 정보센터의 웹사이트가 아니지만 사이트의 특성이 정보제공이라는 측면에서 유사하므로 향후 도서관이나 정보센터에 적용하는 데에 있어서 적절하다고 할 수 있다.

2.1 개인화서비스 필요성

개인화서비스는 도서관 및 정보센터의 관점에서 정보이용자의 요구를 보다 정확하게 분석하고 이를 기반으로 이용자의 정보요구에 적합한 정보를 제공한다는 측면에서 매우 중요하고 필수적인 정보서비스로 고려되고 있다. 특히 이용자를 세분화하여 이용자집단에 적합한 맞춤서비스를 제공할 수 있으며, 대량의 정보를 보다 효율적으로 처리하고 이를 적절하게 이용자에게 제공할 수 있다는 측면에서 폭발적으로 증가하는 전자정보의 처리가 요구되는 현재의 도서관과 정보센터의 고민을 해결할 수 있을 것이다. 따라서 개인화서비스는 이러한 이용자의 요구를 충족하고 도서관이 처한 문제점을 보다 효과적으로 해결할 수 있다는 점에서 도서관 및 정보사서의 존재목적 자체를 충족시킬 수 있다고 할 수 있다.

2.2 연구 의의

폭발적으로 증가하는 정보와 과거와는 달리 멀티미디어정보를 포함하는 다양한 형식을 가지는 정보의 효율적인 처리와 관리와 함께, 이용자의 정보요구를 보다 적극적으로 충족시키기 위하여 본 연구에서는 웹에서 존재하는 정보를 대상으로 하였으며 다양한 형식의 정보를 대상으로 하였다. 전통적으로 도서관이나 정보센터에서 처리하는 자료의 형태는 대

부분이 텍스트자료였으며 일부 마이크로필름이나 디지털화된 자료를 포함하고 있다. 이러한 사회, 경제적인 변화는 도서관과 정보센터에 대하여 기존의 전통적인 텍스트자료와 함께 디지털화된 전자자료의 처리 및 관리에 대한 요구가 폭발적으로 증가하고 있다. 이러한 전자자료는 단순히 기존의 텍스트자료의 디지털화뿐만 아니라 동영상, 음성 등의 멀티미디어자료를 포함하고 있다. 따라서 증가하고 있는 멀티미디어자료의 개인화서비스를 위한 추천을 위해서는 단순한 추천기법만으로는 많은 한계점이 있다고 할 수 있다. 특히, 일반적으로 사용되는 추천 기법인 내용기반의 추천시스템은 주로 텍스트문서에 적용되어 왔으며 이용자의 선호도를 파악하는데 있어서 이용자의 적극적인 참여에 의존하므로서 보다 정확하고 개인화된 정보추천을 하는데 있어서 한계점이 있다고 할 수 있다. 이러한 문제점을 극복하기 위하여 본 연구에서는 내용기반 추천기법의 단점인 추천을 위한 규칙규정 및 고객행태분석의 한계점을 극복하기 위한 방법론으로서 이용자프로파일정보의 정확성 및 적합성을 높이기 위하여 추천대상이 되는 정보 범주 사이의 관계성을 분석하여 이를 적용하기 위한 방법론을 제시하고자 한다. 이를 위하여 이용자 프로파일의 학습을 위한 이용자프로파일과 컨텐츠프로파일을 표현하는 방법을 제시하고, 웹에서의 이용자의 행동정보 및 웹사이트의 구조 정보, 컨텐츠프로파일을 반영하여 이용자의 성향을 학습하는 알고리즘을 제시하고자 한다.

3. 개인화추천시스템

개인화추천시스템은 이용자가 제공한 정보나 로그데이터 또는 이용자 개인의 선호도와 숨어 있는 패턴을 발견하여 개인별로 적절한 정보를 추천해 주는 자동화된 정보필터링시스템을 말한다. 즉, 이용자에 대한 인구 통계학적 정보, 이용자의 선호도가 가장 높은 정보에 대한 탐색 패턴, 개인 선호도 등을 데이터마이닝(Data mining) 기법 등에 의해 분석을 하고 그 결과를 토대로 이용자가 구매하고 싶은 상품을 쉽게 찾을 수 있도록 고객만을 위한 맞춤화된 웹 페이지를 만들어 개인화 서비스를 고객에게 공급하는 시스템인 것이다(황성희 외 2001).

최초의 개인화시스템은 협업필터링을 사용한 업무 메일링프로그램으로서 이러한 개인화시스템은 현재 일대일 마케팅을 비롯해 eCRM, 데이터마이닝, 컨텐츠관리, 그리고 검색엔진 등 거의 모든 인터넷 솔루션들이 '개인화'를 표방하고 있으며 Amazon, CD Now, Garden.com 등이 개인화서비스를 통한 대표적인 성공적 사이트로 평가 받고 있다. 이러한 개인화서비스를 통한 개인화추천시스템에 대한 성능 평가에 관한 연구가 많은 부분에서 진행되고 있다. Manber et. al(2000)은 웹 상에서의 중개자(e-broker) 의 능률성 분석을 위해 실제적인 로그 데이터를 사용하였다. 중개자에 의해 제시된 결과가 능률적임을 논하고, 중개자의 행동을 분석하기 위해 클릭 수(click-through)와 반응시간과 언어, 관습과 같은 지역적인 요소들과 요인분석을 하여 클릭 수(click-through)와 지역적인 요소의 상관관계가 있음을 발견하였다. 또한 Schonberg et. al(2000)는 방문자의 로

그데이터를 통해 구매전환율인 'look-to-buy 매트릭스'을 제시하여 그 성과를 나타내었다. Yi-Hung et. al.(2001)의 연구에서는 이용자의 관심과 행동을 조합한 추천시스템인 OST를 기반으로 하는 3개의 추천 서비스를 제시하고 실제 13명의 이용자와 14개 분야로 나누어진 230개의 논문을 가지고 실험하였으며 해당 결과를 hit ratio, average ratio, miss ratio로 분류하여 추천 서비스의 채택율이 71.5%에서 94.5%임을 보여 주었다. 그러나 이 연구에서는 표본의 수가 너무 작아 타당성이 부족하였다.

3. 1 개인화추천 기법

개인화를 실행하기 위해서는 개인에 대한 정보를 수집한 다음 개인에 대한 분석을 하고 가정 적절한 서비스를 찾아서 고객에게 적절한 서비스를 실제로 제공할 수 있도록 하는 기술이 필요하다. 일반적으로 많이 쓰이는 기법들은 내용기반 필터링(Contents-based filtering), 협업 필터링(Collaborative filtering), 학습 에이전트(Learning agent) 등이 있다.

3. 1. 1 내용기반(Contents-based) 필터링

내용기반 필터링은 이용자들에게 몇 가지 질문들을 한 이후에 이 질문들의 답에 적합한 내용들을 전달하는 것이다. 내용기반 필터링에서 제공하는 질문은 이용자들을 구분하고 개개인을 구별하기 위한 목적으로 사용되며 매우 다양한 형태를 가질 수 있다. 가령 이용자의 우편번호를 물어보고 이용자의 거주지를 구분할 수 있고 인적사항 정보와 어떤 사항들에 대한 선호도 등을 물어보고 그 정보들에 따라 이용자들을 구분할 수 있다. 일반적으로 이용자의 인구통계학적(Demographic) 정보나 심지학적(Psychographic) 정보를 이용자 확인의 중요한 요소로 사용하게 되며 고차원의 개인화를 위해서는 이용자의 선호도 정보를 사용한다. 내용기반 추천기법은 이용자의 과거 정보이용행태를 기반으로 하여 관련정보를 제공하는 방법으로서 내용기반 추천기법은 사람에 의하여 추천메커니즘이 결정되고 추천을 위한 방법이 단순하고 통제가 용이 하다는 장점이 있다. 그러나 정보이용자가 과거의 자신이 경험한 것과 비슷한 정보만을 취하므로서 데이터에 기반한 객관적인 고객 행태 분석의 한계가 있다. 두번째는 대부분의 내용기반 추천기법은 텍스트 자료만에 한정된다는 것이다. 마지막으로 추천메커니즘이 사람에 의하여 결정되므로서 풍부한 추천 노하우(Know-how)가 축적되지 않은 상태에서 추천을 위한 규칙에 대한 규정의 어려움이 있다.

3. 1. 2 협업(Collaborative) 필터링

협업 필터링은 이용자가 자발적으로 제공한 정보를 사용하여 이용자를 비슷한 선호도를 가진 집단으로 나누어 그 집단 내에서 서로에게 추천하는 방식을 사용한다. 예를 들어, 영화에 대한 내용을 제공해야 할 경우 이용자들에게 먼저 주어진 영화들에 대해 평가를 해달라고 부탁한다. 가장 좋아하는 영화에는 5점, 보통 이상이라면 4점, 보통이면 3점, 2와 1은 각각 "좋지 않다"와 "매우 나쁘다"를 나타낸다. 컴퓨터는 이렇게 얻어진 평가 데이터를 패턴으로 만들고 패턴 인식 기술을 사용해 서로 비슷한 선호도를 가진 집단으로 나눈다. 나와 같은 그룹에 있는 선호도가 비슷한 사람이 내가 미처 보지 못한 어떤 영화를

보고 그 영화를 좋다고 생각했다면 컴퓨터는 그 영화를 나에게 추천 해주게 되며 그 사람과 선호성향이 비슷한 나는 그 영화를 좋다고 말할 가능성이 높다. 이러한 그룹 형성의 과정과 교차추천(Cross-recommendation)의 과정은 이용자가 처음 사용하는 경우라 하더라도 충분한 자료가 축적되어 있을 경우 이용자에게 즉시 서비스가 가능하고 또한 본질적으로 이용자의 개인정보를 공개하지 않아도 서비스 제공이 가능하기 때문에 최근의 개인정보 보호를 우선시하는 익명 개인화(Anonymous Personalization) 추세에도 잘 맞는 방법이다. 협업필터링 추천기법은 현재까지 웹상에서 제공되고 있는 추천시스템에서 가장 성공적인 추천기법이라고 할 수 있다. 이러한 기법은 정보이용자에 대한 초기 정보가 부족한 경우 가장 적절하게 이용될 수 있다. 즉, 정보이용자가 보여주는 정보탐색행동이나 정보탐색 후의 피드백에 대한 정보가 충분치 않은 경우 해당 정보이용자와 비슷한 프로파일정보와 반응을 보여주는 이용자들의 선호도를 기준으로 정보를 제공하여 주는 방법이다(George Karypis et. al. 2000)(Daniel Billsus and Michael J Pazzani 1998). 그러나 이러한 장점과 함께 협업필터링 추천기법은 불완전하고, 적은 정보량을 토대로 추천하므로서 정보이용자의 관심사에 대한 정보가 충분하지 않은 환경에서 적용하는 경우 전혀 적합하지 않은 정보를 제공할 가능성이 매우 높다고 할 수 있다(황성희 외 2001). 또한 학습시간에 있어서 많은 시간을 요구한다. 협업필터링 추천기법을 적용한 대표적인 사례로서는 GroupLens, Firefly와 MovieLens 등이 있다.

3. 1. 3 학습 에이전트(Learning Agent)

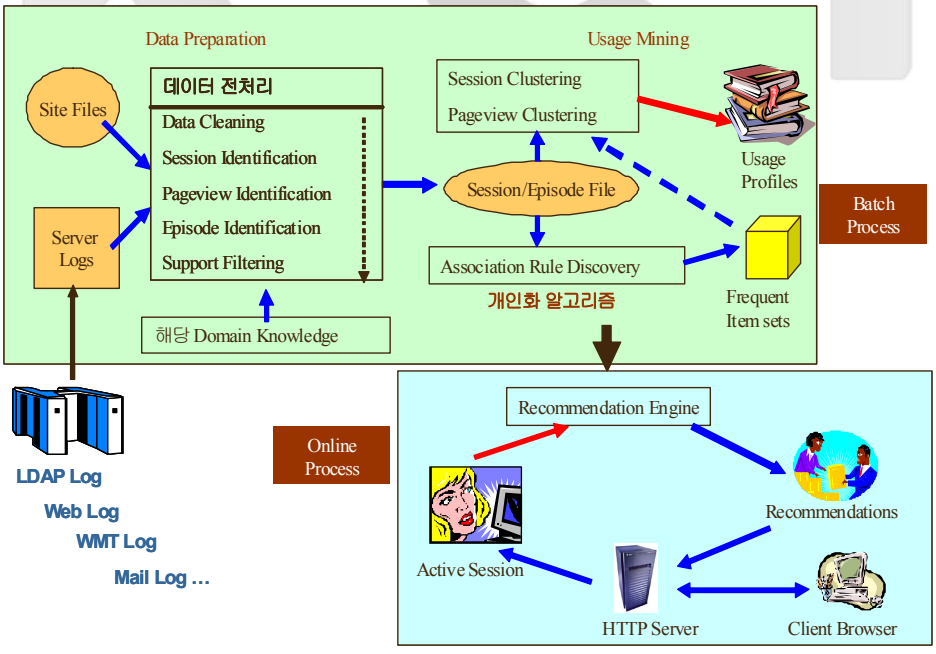
학습 에이전트(Learning Agent)는 이용자의 웹상에서의 활동을 관찰하고 이용자가 어떤 내용에 관심을 가지고 있는지 판단하여 이용자에게 알맞은 내용을 전달하도록 하는 것을 말한다. 이용자의 웹 내에서의 행동 중에서 중요하게 사용되는 것은 특정한 페이지를 보는 시간, 인쇄한 페이지, 전자상거래 사이트의 경우에는 구매한 상품과 쇼핑카트에 넣은 상품 등이다. 구축된 내용의 데이터베이스와 관찰된 이용자의 웹 사용 습관을 토대로 데이터 마이닝의 과정을 거쳐 이용자의 성향과 관심이 결정되고 이용자에게 알맞은 내용이 제공된다. 학습 에이전트는 다른 이용자 자료와의 비교를 필요로 하지 않기 때문에 이용자가 적은 경우에도 적절한 내용을 전달할 수 있다. 그러나 이용자의 웹에서의 행태를 일정 시간 이상 관찰한 이후에야 정보 제공이 가능하며 현재의 사이트의 내용을 학습 에이전트를 사용할 수 있는 환경으로 재조정해야 한다는데 어려움이 있다. 그러나 일단 이용자에게 적절한 내용을 제공할 수 있는 단계를 넘어서면 이용자 정보가 계속 축적되어서 더욱 효과적으로 이용자에게 알맞은 정보를 제공할 수 있다. 또한 이용자 확인만 되면 이용자의 신상정보를 공개할 필요가 없이 개인화 하는 것이 가능하게 되기 때문에 최근에 관심의 초점이 되고 있는 사생활 보호라는 측면에서도 바람직하다고 할 수 있다. 위의 기법중 어떤 한가지 기법이나 모델이 모든 상황에 가장 좋은 해답이 되는 것은 아니다. 웹 개인화의 목적에 따라 각각 다른 기법을 사용할 수 있다.

3. 2 개인화서비스 과정

개인화서비스의 과정은 먼저 개인화서비스를 위한 모델링 작업과 개인화서비스를 제공하는 과정으로 구분할 수 있다. 개인화모델링은 일괄처리모듈(Batch process module) 에서 수행되며 개인화 모델링을 위한 데이터를 구축하고 이를 웹데이터마이닝기법을 통하여 모델링 작업을 수행한다. 즉, 컨텐츠정보와 웹로그정보를 전처리과정을 통하여 수행하며, 이를 데이터웨어하우스에 구축하는 과정으로서 실제 개인화 구현을 위한 웹데이터마이닝작업으로서 여기에서 생성되는 결과물에는 이용자프로파일 및 콘텐트프로파일 들이라고 할 수 있다. 일괄처리모듈에서 개인화의 모델링 작업을 통하여 실제 개인화서비스가 제공되기 위하여 실시간처리모듈에서는 이용자의 세션 정보가 추천엔진에 입력되면 구축된 개인화 모델은 규칙에 따라서 서비스를 추천하는 작업이 수행된다. 개인화서비스의 처리과정은 〈그림 1〉과 같다.

4. 추천시스템의 구조 및 알고리즘

본 장에서는 제안하고 있은 추천시스템의 구조와 실험을 위한 웹사이트의 속성을 보여주고 있다. 또한 세부적으로 이용자의 선호도를 계량화하고 있는 이용자프로파일정보를 갱신하는 과정과 함께 이를 기반으로 추천정보를 생성하는 과정을 알아보고자 한다.



〈그림 1〉 개인화서비스 과정

4. 1 적용사이트의 속성

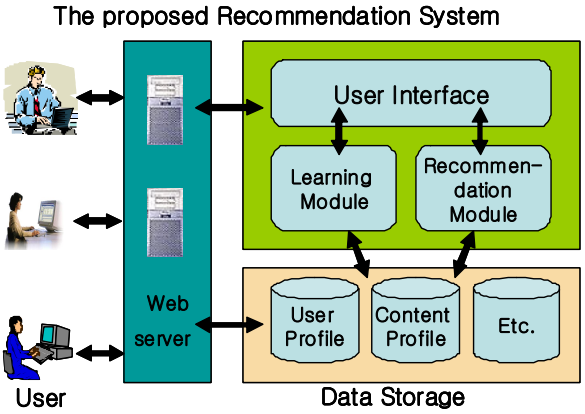
적용대상이 되는 사이트는 실시간 또는 사전 기록된 텍스트자료를 포함한 멀티미디어정보를 이용자에게 제공하는 웹사이트라고 할 수 있다. 비록 실험에 적용되는 사이트와 데이터가 도서관이나 정보센터와 정확히 일치 하지는 않은 한계점은 있으나 텍스트정보를 포함한 멀티미디어 정보를 이용자에게 제공한다는 측면에서 도서관이나 정보센터의 환경과 유사하다고 할 수 있으며 따라서 본 연구에서 제안하고 있는 알고리즘과 추천 시스템은 도서관이나 정보센터에 적용이 충분히 가능 하다고 할 수 있다. 현재 국내의 웹캐스팅사이트의 수는 지난 2000년을 기점으로 폭발적으로 증가하고 있은 추세에 있으며 현재까지 약 800여개의 사이트가 서비스를 제공하고 있다. 멀티미디어정보의 검색을 위하여 해당 메타데이터 정보를 기반으로 하고 있으며 제안된 추천시스템은 약 4백만의 이용자를 가지고 있는 파란(www.paran.com)에 구축하여 실험을 수행 하였다.

4. 2 제안된 추천시스템 구조

제안된 추천시스템은 추천모듈과 학습시스템 모듈 및 이용자 인터페이스 모듈로 구성되어 있는 처리 모듈과 이용자 프로파일 및 컨텐츠프로파일 정보를 담고 있는 데이터저장모듈로 구성되어 있다. 각 모듈과의 연관관계는 〈그림 2〉에서 보여주고 있다.

제안된 추천시스템은 〈그림 2〉에서와 같이 총 5개의 모듈로 구성되어 있으며 각 모듈의 기능은 아래와 같다.

- 이용자 인터페이스: 웹서버를 통하여 접속된 이용자와의 통신처리를 위한 기능을 수행한다.
- 학습모듈: 이용자의 정보이용행동을 분석하여 이용자프로파일정보를 갱신하는 기능을 수행한다.
- 추천 모듈: 이용자의 프로파일정보와 일치하는 정보를 추천하는 기능을 수행한다.
- 데이터저장소(Data storage): 데이터저장소는 추천을 위한 각종정보 및 이용자



〈그림 2〉 추천시스템의 구조

프로파일과 컨텐츠프로파일을 저장하고 있다.

현재까지 내용기반의 추천시스템은 주로 텍스트문서에 적용되어 왔으며 이용자의 선호도를 파악하는데 있어서 이용자의 적극적인 참여에 의존하여 왔다고 할 수 있다. 따라서 이러한 내용기반추천시스템의 한계점을 극복하기 위하여 본 연구에서는 이용자의 직접적인 선호도에 의한 가중치부여를 지양하고 이용자의 컨텐츠 이용행태와 컨텐츠범주사이의 관계성에 대한 연관관계를 파악하기 위하여 웹로그화일을 분석하여 그 연관성을 파악하였다.

4. 3 학습과정

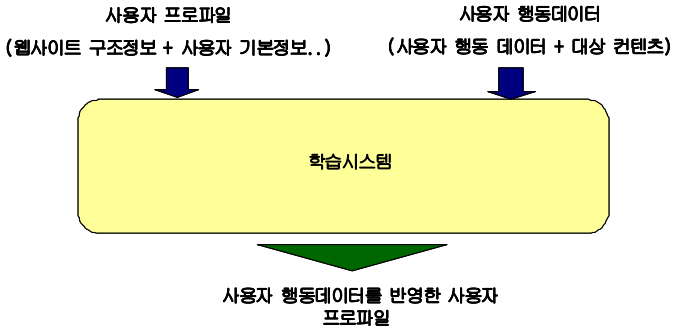
정보이용자의 주관적인 선호도는 지속적으로 또한 급격하게 변화한다. 따라서 추천시스템의 정확성을 향상하기 위해서는 이용자프로파일을 얼마나 빠르고 정확하게 생성하고 갱신하는 것이 중요하다고 할 수 있다. 아래 그림은 이용자의 행동데이터와 기존 이용자 프로파일정보를 기반으로 새로이 이용자프로파일정보를 갱신하는 과정을 보여주고 있다.

4. 3. 1 표현

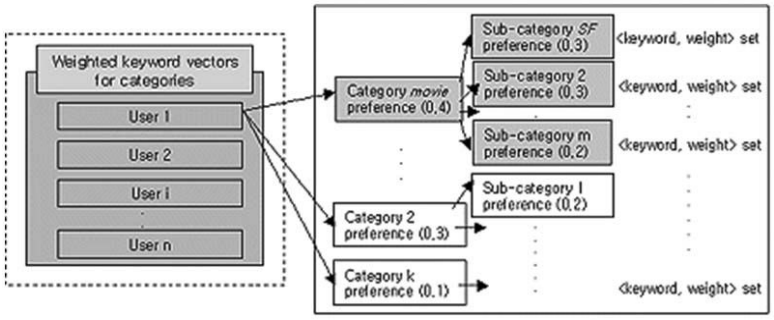
이용자프로파일과 컨텐츠프로파일의 표현은 벡터공간모델기법을 기반으로 가중치가 부여된 키워드벡터를 적용하였다.

4. 3. 2 이용자프로파일

이용자프로파일은 웹사이트의 특성을 반영하여 범주로 분류되어 있으며 각 범주별 이용자의 선호를 반영하기 위한 가중치를 가진다. 초기 이용자 프로파일의 범주 가중치는 인구통계학적 정보나 이용자 입력정보와 함께 웹사이트 구조정보 및 이용자그룹정보 등을 통하여 결정된다. 이용자프로파일은 두 개의 하위 분야로 구성되어 있다. 첫 번째는 이용자의 해당 정보범주에 대한 선호도를 표현하는 범주선호도로서 해당 범주의 컨텐츠 또는 정보에 대한 이용자의 선호도를 표현하기 위하여(범주, 가중치)의 쌍으로 표현된다. 두번째 분야는 개별 범주에 대한 가중치가 부여된 키워드 벡터로서(키워드, 가중치)의 쌍으로 표현되며 특정범주에 대한 이용자의 선호도 또는 관심도를 표현한다.



〈그림 3〉 개인화추천을 위한 학습과정



〈그림 4〉 이용자프로파일의 예

이러한 이용자프로파일은 이용자 행동데이터를 이용한 학습을 통하여 지속적으로 갱신되며 이러한 이용자프로파일은 이용자의 변화하는 성향을 반영하고 있다. 이러한 과정을 통하여 학습된 이용자프로파일은 추천시스템에서 이용자에게 유용한 정보 또는 컨텐츠를 추천하기 위해 사용된다.

4. 3. 3 이용자행동데이터

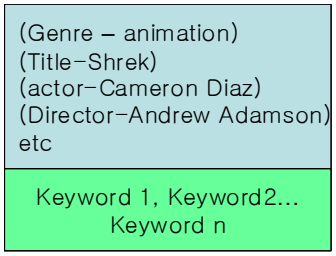
이용자의 행동데이터는 웹로그분석, 거래내역, 장바구니분석 등을 통해 이용자의 웹사이트에서의 행동을 추출하여 학습에 적합한 형태로 가공된 데이터로서 이용자 아이디, 대상 컨텐츠, 이용자피드백유형 등의 기본정보와 부가정보를 포함하고 있다. 이용자에게서 획득할 수 있는 피드백은 이용자에게 직접 정보를 입력받는 명시적 피드백과 이용자의 행위를 관찰함으로써 이용자 행위에 대한 정보를 획득하는 암시적 피드백으로 나눌 수 있다. 본 연구에서는 웹사이트에서의 이용자의 행동데이터를 크게 다섯 가지의 행동으로 구분하였으며 각각의 행동데이터에 가중치값을 부여하여 이를 이용자프로파일정보를 갱신하는데 적용하였다. 각각의 가중치값은 이용자피드백정보의 형태로 나타나며 보다 세부적인 획득방법은 4.4의 학습과정에서 살펴보기로 한다.

4. 3. 4 컨텐츠프로파일

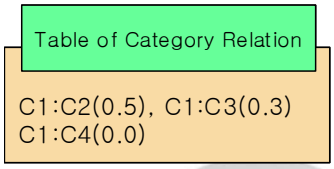
컨텐츠프로파일은 해당 컨텐츠에 대한 속성을 포함한다. 이러한 컨텐츠프로파일은 이용자프로파일과 비슷하게 표현되며 추천시스템에서 정의된 개별 필드로 구성된다. 현재 실험에 적용된 필드는 {genre, title, actor, director, a set of keyword}로 구성되었다. 〈그림 5〉은 영화 "슈렉"에 대한 컨텐츠프로파일을 보여주고 있다. 본 연구에서는 해당 필드의 특성에 따라 가중치 값을 부여 하였다.

4. 3. 5 범주관계성(Category Relation)

범주관계성(Category relation)은 개별 컨텐츠가 속한 범주사이의 연관성을 표현한다. 이러한 범주관계성은 이용자에게 제공되지 않았던 범주에 대한 가중치 값을 조정하기 위하여 이용되며, 이러한 관계성은 다양한 마이닝기법과 학습을 통하여 습득된다. 〈그림 6〉은 범주관계성의 사례를 보여주고 있다. 〈그림 6〉에서



〈그림 5〉 컨텐츠프로파일의 예(영화 슈렉)



〈그림 6〉 범주 관계성의 예

범주 C1과 C2의 관계성은 0.5로서 범주 C2에 속한 정보 또는 컨텐츠는 범주 C1에 속한 컨텐츠에 대하여 0.5의 관계성을 가지고 있으며 이는 범주 C1의 정보 또는 컨텐츠를 이용자가 선호하는 경우 C2에 속한 컨텐츠 또한 선호할 수 있다는 것을 의미한다.

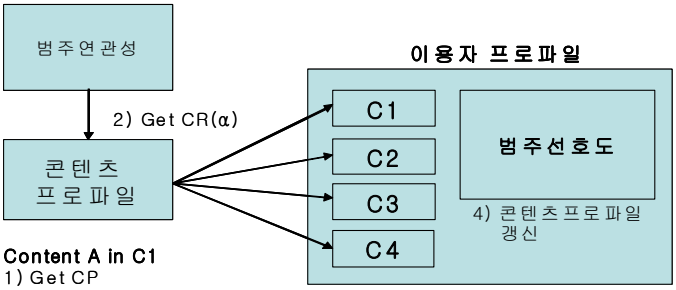
4. 4 이용자프로파일 학습과정

이용자프로파일을 학습하는 과정은, 확보된 이용자 데이터를 이용하여 이용자의 실제적인 관심을 반영하기 위한 과정으로 이용자 범주의 가중치를 이용하여 학습을 수행하고, 학습의 결과를 이용자 프로파일에 반영하는 과정이다.

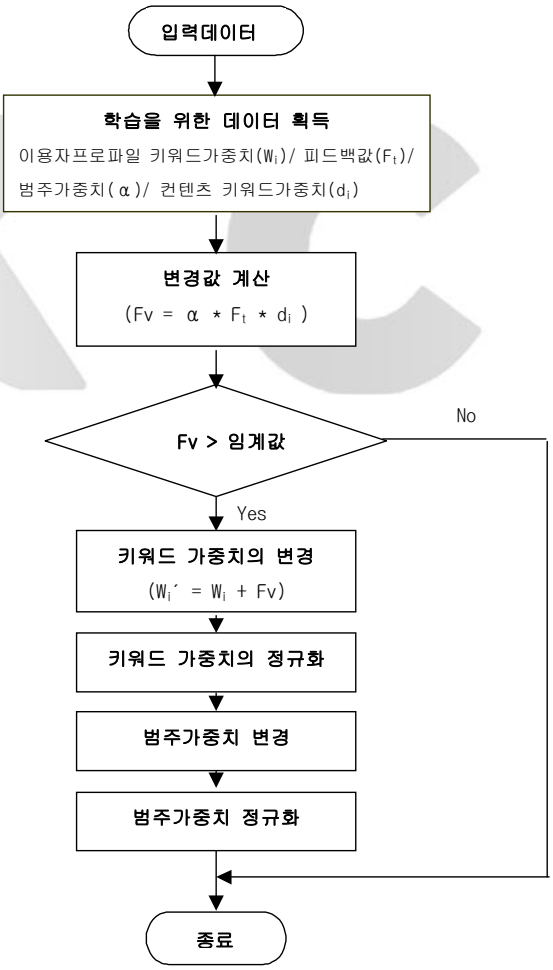
본 연구에서 적용된 연구방법에 있어서 초기 프로파일은 이용자에 의하여 제공된 선호도를 통하여 생성되었으며 이용자프로파일을 갱신하기 위하여 이용자의 웹 이용행태를 분석하여 이용자의 선호도를 갱신하였다. 〈그림 7〉은 이용자의 선호도 갱신을 위한 학습과정을 보여 주고 있으며 이러한 일련의 과정은 아래와 같다.

1. 범주 Ci에 속한 컨텐츠 A에 이용자가 반응한다고 가정하여 A에 대한 컨텐츠 프로파일을 구한다.
2. 범주 Ci와 다른 범주에 대한 범주관계성을 구한다.
3. 범주가중치(α)와 이용자의 피드백으로서 이용자행동변수(F_t)를 통하여 이용자프로파일을 갱신한다. 이용자프로파일 갱신을 위한 수식은 아래와 같다.
4. 이용자프로파일에 대한 범주선호도를 갱신한다.

이를 보다 세분화하여 살펴보면 다음의 〈그림 8〉과 같다.



〈그림 7〉 이용자프로파일 학습과정



〈그림 8〉 이용자프로파일 학습과정 흐름도

이용자프로파일정보의 갱신과정은 전통적인 벡터조정모델과 비슷하다고 할 수 있으며 학습기법은 NewsT에서 사용하는 방법과 같은 방법을 적용하고 있다.

본 연구에 적용된 이용자프로파일 갱신을 위한 수식은 다음과 같다.

$$W_i' = W_i + (\alpha * F_t * d_i) \quad \text{----------(1)}$$

* W_i : 이용자 프로파일의 키워드가중치
* W_i' : 학습 후 이용자 프로파일의 키워드가중치
* F_t : 피드백 값,
* α : 범주 가중치
* d_i : 컨텐츠프로파일 키워드의 가중치

각 값에 대한 자세한 설명은 다음과 같다.

(1) 피드백 값(F_t) : 이용자의 피드백 유형에 따른 피드백 값으로 웹사이트의 특성과 이용자 행위에 따라 유형을 분류하고 값을 지정한다. 예를 들어, 전자상거래 사이트의 경우 동영상을 보거나 다운 받는 행위보다는 구매행위로 간주 될 수 있다.

〈표 1〉 이용자 피드백 유형

피드백 유형	Action
1	이용자의 명시적 피드백
2	다운로드, 구매 등
3	Play, 검색어, 쇼핑카트 담기 등
4	단순클릭
5	기타

유형 1의 경우 명시적 피드백은, 이용자가 추천에 대해 직접 평가하거나 특정분야에 대한 선호도변경 등 명시적으로 피드백을 주는 행위에 대한 것으로 정확도와 기여도가 높으나, 이용자로부터 획득하기 어렵기 때문에 유형 2 ~ 4 등의 암시적 피드백을 통한 학습이 주로 이루어진다. 피드백에는 부정적 피드백도 있으나 이용자의 의도를 알아내는데 어려움이 있으므로 이용자의 의도가 확실한 경우가 아니면 학습에 이용하지 않는다.

일반적인 경우 피드백 값은 $0 < F_t < 1$을 가지며, 피드백의 중요도에 따라 값이 결정된다 유형 5는 조회수/사용시간 등이 이에 속하며, 이들은 일정 가중치를 주는 것이 아니라 이용자의 반응에 따라 값에 변화를 주어야 한다

(2) 범주가중치(α): 이용자의 피드백이 이용자프로파일에 영향을 미치는 정도를 나타내는 것으로, $0 < \alpha < 1$ 사이의 값이다.

(3) 이용자프로파일의 가중치(W): 컨텐츠와 같은 범주에 속하는 이용자프로파일의 키워드의 가중치이다.

(4) 컨텐츠프로파일 가중치(D_i): 이용자가 피드백의 대상이 되는 컨텐츠/페이지의 프로파일의 키워드가중치로 컨텐츠프로파일 어트리뷰트들을 중요도에 따라 가중치를 준다. 즉 영화 컨텐츠의 경우 주연배우가 조연배우보다 중요도가 높기 때문에 가중치 값은 높아진다. 컨텐츠프로파일의 가중치 역시 적용되는 사이트의 특성을 반영하여 변경된다. 예를 들어, 대학도서관시스템의 경우 논문자료, 책자형자료, 뉴스레터, 신문기사 등과 같이 해당 기관의 성격에 따른 정보자료의 선호도에 대한 가중치를 차별화 함으로서 정보자료에 대한 중요도를 반영할 수 있다. 즉, 범주 관계성이 〈그림 6〉과 같고 이용자가 범주 C1에 속하는 컨텐츠 A 를 내려 받

기를 한다고 가정한다면, 범주 C1은 각각의 범주 C2, C3과 C4에 대해서 C2(0.5), C3(0.3) and C4(0.0)의 범주관계성을 가지고 있다고 할 수 있다. 따라서 본 연구에서 제안된 학습알고리즘은 범주 C2, C3 그리고 C4에 대한 키워드벡터를 갱신하며 또한 피드백은 각각 0.5, 0.3과 0.0이 된다. 학습이 이루어지는 과정은 이용자 피드백의 대상이 되는 컨텐츠의 각 키워드에 대해서 키워드의 가중치와 범주 가중치, 피드백 유형에 따른 피드백 값을 이용하여 이용자프로파일의 해당 키워드의 값을 식(1)을 적용하여 변경시켜주게 된다. 즉, 이용자프로파일 = {{1, 0.1i}, {2, 0.2 }, {3, 0.3 }, {4 ,0.4}} 이와 같이 구성되어 있고, 컨텐츠프로파일 = {{1, 0.1}, {2, 0.5 }, {4, 0.1 }, {5, 0.2}} 와 같이 구성되어 있으며, 컨텐츠가 속한 범주의 가중치는 0.2, 피드백(F_t)값이 0.6 이라고 가정한다면, 학습 후의 키워드 1의 가중치 는 0.1 +(0.2 * 0.6 * 0.1)로 계산된다. 계산된 변경 값이 임계치 이상인 경우만 이용자 프로파일 및 범주 가중치를 변경시키고 임계치보다 적으면 변경하지 않는다. 모든 키워드에 대하여 키워드의 가중치를 변경하고 이들 값을 다시 정규화(normalization)함으로써 이용자프로파일의 변경이 종료된다. 이를 통하여 최종적으로 웹사이트의 모든 범주의 가중치를 변경하고 이들을 정규화함으로써 이용자프로파일에 대한 학습과정이 종료된다.

4. 5 추천리스트 생성과정

추천리스트생성을 위하여 본 연구에서는 추천기준으로서 이용자프로파일과 컨텐츠프로파일을 사용하였다. 특히, 본 연구에서는 위에서 기술한 이용자프로파일생성과 함께 하이브리드 추천방법을 적용하였다. 이러한 접근방법은 내용기반추천기법과 협업추천기법의 특징을 혼합하여 적용하였다는 점에서 특징적이라고 할 수 있으며 추천의 정확성을 높이는데 기여 할 것이라고 생각한다. 3장에서 언급한바와 같이 내용기반의 추천기법은 이용자의 관심 및 요구를 파악하기 위하여 컨텐츠의 내용 자체를 사용한다. 이전의 비슷한 경험을 가진 이용자군이 보여주는 특징과 유사한 특징을 지닌 콘텐츠가 추천된다고 할 수 있다. 따라서 내용기반추천방법은 매우 간단하기는 하지만 이용자의 다양한 경험의 결여에 따른 다양한 콘텐츠를 추천하는데 있어서 제약점이 있다. 반면에 협업추천기법은 콘텐츠의 추천을 위하여 다른 이용자의 적극적이고 숫자화된 평가에 기반하여 추천콘텐츠를 위한 이용자선호도를 예상한다. 이러한 협업추천기법은 내용기반의 추천기법을 보완할 수 있으며 현재 많은 개인화서비스에 적용되고 있으며 가장 좋은 성능을 보여주고 있는 기법이라고 할 수 있다. 그러나 협업추천기법이 좋은 성능을 보여주고 있으나 여전히 확장성 및 계위성에 문제점을 보여주고 있다. 따라서 이러한 문제점을 해결하기 위하여 본 연구에서는 위의 추천기법들의 제한점을 극복하고 장점만을 취한 새로운 하이브리드 추천기법을 제안하고 이를 구현하였다. 제안하고 있는 하이브리드 추천기법은 이용자의 인구통계학적 정보와 이용자의 피드백정보의 정도에 의해 결정되는 자질요소는 이용자가 웹상에서의 수행하는 이용자행동데이터를 포함하고 있으며 이러한 행동데이터는 적합성피드백 과정을 통하여 얻는다. 이용자프로파일과 콘텐츠프로파일에 대한 유사도값을 얻기 위하여 코

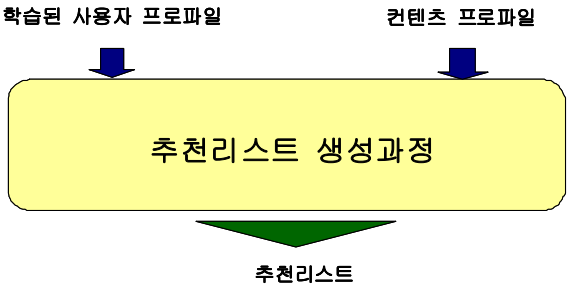
사인유사도값(Cosine Similarity)을 적용하여 추천리스트를 생성하였다. 한편 학습과정에서는 자질요소추출을 위하여 협업추천기법을 적용하여 최종추천리스트를 구할 수가 있다. 최종추천리스트는 계층적으로 범주화되고 보완되었다. 학습과정에서의 적용된 하이브리드추천기법의 특징은 아래의 함수 f로 표현 될 수 있다.

Hybrid f(〈user profile〉, 〈user feedbacks〉,〈group profile〉)
≈ Hybrid f(〈user preference〉, 〈user feedbacks〉,〈group similarity〉)
—〉 {〈*category, keyword, weight*〉 set}

개인화된 추천서비스는 크게 두개의 영역으로 구분 할 수 있다. 먼저 이용자가 웹상에서 보여주는 행동을 통하여 이용자추천정보를 추출하고 학습과정을 통하여 생성되는 추천리스트에 해당 사이트나 기관의 비즈네스규칙을 적용하는 단계와 함께 다음단계로서 이용자가 웹사이트를 이동하므로서 이용자의 포지션에 따라 이용자가 선호하는 콘텐츠를 추천하는 단계이다. 또한 웹사이트의 구조정보를 적용하므로서 계층화된 이용자프로파일과 추천리스트를 생성하였으므로 웹사이트별 위치적용추천이 가능하다고 할 수 있다.

5. 결 론

본 연구에서는 웹에서 개인화서비스를 제공하기 위한 내용기반 추천기법에 기반하여 이용자 행위정보를 이용한 학습방법을 제시하고 있다. 즉 기존 추천기법의 제한점을 극복하고 장점만을 적용한 새로운 하이브리드방식의 추천기번을 제안하고 있다. 이를 통하여 내용기반의 추천시스템의 한계점을 극복 하면서 내용기반의 장점을 수용할 수 있은 고유한 하이브리드방식의 추천시스템에 적용 할 수 있다. 특히, 변화하는 이용자의 성향을 반영하기 위해 웹사이트의 구조와 이용자의 행동데이터를 반영한 이용자 프로파일과, 웹사이트의 각 범주에 대한 가중치를 이용하여 학습을 수행하는 것을 특징으로 한다. 이를 통해 이용자에게 효과적인 개인화서비스를 제공함으로써, 서비스의 차별화를 통한 고객의 충성도와 함께 보다 적절한 컨텐츠를 개인에게 제공하므로서 서비스 이용률을 높일 수 있다.



〈그림 9〉 추천 리스트 생성과정

그러나 본 연구에서의 실험대상이 인터넷상의 웹캐스팅사이트로서 도서관이나 정보센터의 개인화서비스에 즉시적으로 적용은 어렵지만 본 연구에서 제안하고 있는 알고리즘을 기반으로 도서관 및 정보센터의 웹사이트의 특성을 추가하여 이를 적용한 추천 시스템을 구축한다면 기존 도서관이나 정보센터에서 제공되고 있는 개인화서비스에서 제공되는 서비스의 질과 함께, 제공되는 정보의 적합성을 향상 할 수 있을 것이라 생각되며 이를 위하여 향후 도서관 및 정보센터의 사이트의 특성과 함께 정보컨텐츠에 대한 이용자 선호도에 조사가 필요로 하리라 생각된다.

참 고 문 헌

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

남궁 황. 2003. 학습시스템에 기반한 개인화 정보 서비스에 관한 연구. 『정보관리학회지』, 20(4): 112-134.

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

Balabanovic, Marko, and Yoav Shoham. 1997. "Content-Based Collaborative Recommendation." *Communications of the ACM*, 40(3): 66-72.

Billsus, Daniel, and Michael J Pazzani. 1998. *Learning Collaborative Information Filters*. Berkeley: University of California Press.

Dahlen, B.J. and Konstan, J.A., Herlocker, J.L. 1998. Jump-starting movielens: User benefits of starting a collaborative filtering system with 'dead data'. University of Minnesota TR 98-017.

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

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

Manber, Udi, Ash Patel, and John Robison. 2000. "Experience with personalization on yahoo!" *Communication of the ACM*, 43(8): 35-39.

Resnick, Paul, Neophytos Iacovou, Mitesh Suchak, et al. 1994. "GroupLens: An Open Architecture for Collaborative Filtering of Netnews." Proceedings of

the Conference on Computer supported Cooperative Work.

Schonberg, Edith, Thomas Cofino, Robert Hoch, Mark Podlaser, and Susan L. Spraragen. 2000. "Measuring Success." *Communications of the ACM*. 43(8): 53-57.

Shardanand, Upendra, and Patti Maes. 1995. "Social information filtering: Algorithms for automating "Word of Mouth". Proceedings of the ACM CHI '95 Conference on Human Factors in Computing Systems, 210-217.

Sheth, Beerud Dilip. 1994. *The learning approach to Personalized information Filtering*. Boston: MIT Press.

Wu, Yi-hung, Yong-chuan Chen, and Arbee l. P. Chen. 2001. "Enabling Personalized recommendation on the Web based on User Interests and Behaviors." Proceedings of Eleventh International Workshop on Data Engineering.