Apriori 알고리즘 기반의 개인화 정보 추천시스템 설계 및 구현에 관한 연구*

A Study on Design and Implementation of Personalized Information Recommendation System based on Apriori Algorithm

김 용 (Yong Kim)**

초 록

정보기술과 인터넷의 발전에 따른 정보의 폭발적인 증가와 함께, 이용자에게 있어서 적합한 정보의 획득을 위한 방법이 절실하게 요구되고 있다. 이를 위하여 정보검색 및 여과시스템이 개발 및 발전되어 왔다. 또한 보다 적극적인 서비스를 제공하기 위한 방법으로써 개인화 정보추천서비스에 대한 요구가 높아지고 있다. 본 연구에서는 도서관에서 적극적인 정보서비스를 위한 방법으로 이용자의 관심과 선호도에 적합한 정보를 제공하기 위한 연관규칙 기반의 개인화 정보추천시스템을 설계 및 구현하였다. 이를 위하여 기존의 추천방법에 대한 장단점을 분석하고 기존 추천방법에 대한 문제점을 해결하기 위한 방법으로써 대용량 콘텐츠 및 이용자 환경에서 이용자의 묵시적 정보이용행위에 관한 정보를 포함하고 있는 로그파일을 통하여 연관규칙 생성을 위해 요구되는 항목을 추출 및 변환하여 연관규칙 생성프로그램을 통하여 연관규칙의 생성 및 정보추천을 위한 방법을 제안하였다.

ABSTRACT

With explosive growth of information by recent advancements in information technology and the Internet, users need a method to acquire appropriate information. To solve this problem, an information retrieval and filtering system was developed as an important tool for users. Also, users and service providers are growing more and more interested in personalized information recommendation. This study designed and implemented personalized information recommendation system based on AR as a method to provide positive information service for information users as a method to provide positive information service. To achieve the goal, the proposed method overcomes the weaknesses of existing systems, by providing a personalized recommendation method for contents that works in a large-scaled data and user environment. This study based on the proposed method to extract rules from log files showing users' behavior provides an effective framework to extract Association Rule.

키워드: 개인화, 추천, 연관규칙, 선호도, 정보여과, Apriori 알고리즘
Personalization, Recommendation, Association Rule, User Preference, Information Filtering, Apriori Algorithm

---

* 이 논문은 2012년도 전북대학교 연구기반 조성비 지원에 의하여 연구되었음.
** 전북대학교 문헌정보학과 부교수, 인문영상연구소 연구원(yk9118@jbnu.ac.kr)
논문접수일자 : 2012년 11월 22일 논문심사일자 : 2012년 11월 29일 게재확정일자 : 2012년 12월 17일

1. 서 론

1.1 연구 배경 및 목적

웹의 폭발적인 성장과 함께 정보유통 환경의 변화와 저작도구(authoring tool)의 발달은 다양한 유형의 정보생산의 증가에 따른 정보과잉(information overflow)을 가져왔고, 이는 이용자 측면에서 다양한 정보에 대한 접근성을 가능하게 해준 반면에 대용량의 정보환경에서 자신이 원하는 정보를 찾기 위하여 많은 시간과 노력이 필요하게 되는 결과로 이어졌다. 이러한 문제점에 대한 대안으로써 검색엔진이 등장하게 되었다. 그러나 검색엔진은 궁극적으로 이용자가 원하는 정보를 얻기 위해서는 직접 검색을 수행하고 검색결과에 대한 판단을 스스로 결정을 수행하여야 한다. 웹상의 수많은 정보에서 필요로 하는 정보를 찾는 과정에서 이용자는 여러 부분에서 원하는 정보를 최종적으로 확보하는데 있어서 많은 어려움에 직면할 수 있다. 따라서 전통적으로 도서관 분야에서는 보다 적극적인 정보서비스 제공의 관점에서 이용자가 원하는 정보를 미리 선별하여 제공할 수 있는 선택적 정보배포(Selective Dissemination of Information, 이하 SDI) 서비스를 제공하였다. 그러나 SDI 서비스의 개념은 이용자에 의하여 입력된 키워드와 같은 명시적 정보(explicit information), 즉 이용자에 의하여 입력된 단어와 일치되는 단어를 포함하는 모든 정보를 제공함으로써 정보제공에 있어서의 정확도가 매우 낮은 결과를 초래하였다. 기존의 대부분의 개인화 추천방법에서 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. 그럼에도 불구하고 기존의 추천분야에 대한 연구에서는 이러한 이용자의 명시적 피드백을 추천의 중요한 고려요소로 적용하고 있으며 따라서 이러한 방법론을 추천시스템에 적용하여 구축한다는 것은 현실적으로 매우 부적절한 방법이라고 할 수 있다.

한편, 웹 2.0 환경에서 웹 상에서 생산되는 정보를 이용자에게 자동적으로 제공하는 서비스의 방법으로써 RSS(Really Simple Syndication, 이하 RSS) 기반 서비스가 있다. 이와 같은 단순한 정보기술에 기반을 둔 SDI 및 RSS 서비스는 세 가지 측면에서 한계를 가진다. 첫째, 이용자가 자신의 관심 키워드를 등록한 경우에 대해서만 서비스가 이루어진다. 둘째, 이용자의 변화하는 관심 분야를 시스템이 능동적으로 파악할 수가 없다는 점이다. 특히 최근과 같은 정형, 비정형 정보를 포함한 대용량의 정보가 생산되는 빅 데이터[1]시대에서 이용자의 요구에 적합한 정보를 찾는다는 것은 매우 어려운 문제라고 할 수 있다. 시장조사기업인 IDC(2012)는 2012년 전 세계적으로 생성되는 디지털 정

보량이 2.7Zb[2]에 이를 것으로 추정하고 있는 상황에서 디지털정보의 폭발적 증가에 따른 빅데이터(Big Data) 시대의 도래는 이용자의 정보요구를 충족시켜야 하는 도서관의 입장에서 매우 어려운 문제점을 제시하고 있다. 셋째, 단순한 키워드 매칭에 의한 정보추출로 인하여 이용자에게 제공되는 정보의 정확도가 매우 떨어진다. 현재와 같은 정보환경에서 정보서비스의 가장 중요한 목적은 이용자의 관심 및 선호도에 적합한 정확도 높은 정보를 제공하는 것임에도 불구하고 단순한 키워드 매칭을 통한 정보추출은 정확도에 있어서 매우 낮을 수 있다. 따라서 이용자의 개인적인 선호도와 관심을 실시간으로 반영하면서 정확도 높은 정보추천을 위한 방법이 요구되고 있다. 이러한 요구에 따라 도서관을 포함한 정보서비스 기관 및 온라인 쇼핑몰, 온라인 서점 등의 기업에서의 개인화서비스에 대한 관심이 더욱 높아지고 있다.

개인화서비스는 웹 서비스의 새로운 패러다임으로써 정보개방, 공유, 참여를 주요 원리로 하고 있는 웹 2.0 및 웹 3.0에서 궁극적으로 추구하는 최종의 목표가 되는 개념이다. 따라서 도서관 관점에서 이용자의 정보요구를 보다 정확하게 분석하고 이를 기반으로 이용자의 정보요구에 적합한 정보를 제공한다는 측면에서 매우 중요하고 필수적인 정보서비스로 고려되고 있다. 일반적으로 개인화(personalization)라는 용어는 이용자의 정보요구에 부합되는 정보를 제공한다는 의미로 광범위하게 사용되고 있다(Shahabi and Chen 2003). 개인화는 소극적 개인화와 적극적 개인화로 구분할 수 있다. 소극적 개인화는 이용자가 특정한 정보 또는 상품 검색과정에서 이용자의 프로파일을 기반으로 이용자에 적합한 정보 및 상품을 제공하는 것으로 주로 정보검색 분야에서 용어의 중의성 해소 및 질의어 확장 등의 분야에서 연구되어지고 있다. 즉, 이용자의 과거 검색행태를 분석하여 이용자의 검색질의어를 확장하여 이용자에게 적합한 검색결과를 제공하는 분야라고 할 수 있다(윤흥준 외 2010). 적극적 개인화는 서비스 제공자가 새로운 정보 또는 상품이 입수되면 이용자의 선호도 및 관심을 반영하고 있는 이용자의 프로파일을 기반으로 이용자의 선호도에 적합한 정보를 이용자의 요청이 있기 전에 제공하는 서비스로써 개인화 추천 서비스라고 할 수 있으며 주로 데이터마이닝 분야에서 활발히 연구되고 있다. 일반적으로 개인화 추천서비스는 뉴스, 영화, 도서, 음악, 모바일앱 등의 추천에 널리 응용되고 있다. 특히, 개인화 추천서비스는 대용량의 정보환경에서 이용자의 정보요구를 만족시키기 위하여 도서관이 제공할 수 있는 효과적이면서 적극적인 정보서비스라고 할 수 있다.

본 연구에서는 도서관 관점에서 개인화된 정보추천서비스를 위한 방법으로써 연관규칙을 추출하기 위한 Apriori 알고리즘 기반의 추천시스템의 설계 및 구현을 목표로 한다. 이를 위하여

---

1) 기존 데이터베이스 관리도구의 데이터 수집·저장·관리·분석의 역량을 넘어서는 대량의 정형 또는 비정형 데이터 세트 및 이러한 데이터로부터 가치를 추출하고 결과를 분석하는 기술을 의미한다. [cited 2012.5.26]. 〈http://en.wikipedia.org/wiki/Big_data〉.

2) 1Zb(제타바이트) = 10_{21}Byte, 고화질 영화 1편은 2Gb로 1Zb는 고화질 영화 5천억 편에 해당
Byte 〈 Kilo(10_3) 〈 Mega(10_6) 〈 Giga(10_9) 〈 Tera(10_{12}) 〈 Peta(10_{15}) 〈 Exa(10_{18}) 〈 Zeta(10_{21})

최근의 개인화 추천서비스와 관련된 최근의 연구동향과 추천서비스 제공을 위한 추천방법의 장단점에 대해서 알아보고, 개인화 추천을 위한 연관규칙 추출프로그램의 설계 및 구현을 통한 실질적인 개인화 추천을 위한 방법을 제안하였다.

1.2 연구 범위 및 구성

본 연구에서는 대용량의 정보환경에서 이용자의 정보요구에 적합한 정보를 개인에게 제공하기 위한 연관규칙 기반의 추천방법과 서비스 시스템의 설계 및 구현을 목표로 하고 있다. 특히, 개인화 추천서비스의 대상을 정형, 비정형의 정보를 포함하면서 이용자의 정보요구에 적합한 정보를 추천하기 위하여 요구되는 이용자의 행위정보를 기반으로 추출된 원시데이터에 대한 전처리 작업 및 전처리과정을 통하여 추출된 정보의 저장 및 활용을 통하여 실질적인 연관규칙의 생성 및 추천방법을 제안하고 있다. 이를 위하여 첫째, 개인화 추천서비스에 대한 최근의 연구동향 및 추천방법에 대한 장단점 및 특징을 비교 분석한다. 둘째, 다양한 추천방법에 대한 기존 연구들을 분석하여, 본 연구에서 제안한 모형과 기반구조의 설계를 포함한 기본 방향을 설정한다. 셋째, 연관규칙을 추출하기 위해서 입력데이터 생성을 통한 전처리작업 및 결과 데이터의 저장방법을 제안한다. 마지막으로 연관규칙의 생성과 함께, 연관규칙 기반의 정보추천방법을 제안한다.

2. 개인화 정보추천서비스

개인화라는 용어가 갖는 의미는 이용자의 요구에 적합한 정보를 추출하고 이를 신속하게 제공하는 일련의 결합된 방법을 개인화로서 정의할 수 있다(정경용 외 2004). 최근에는 웹의 활성화에 따라 웹상에서의 개인화에 대한 연구가 활발히 진행되고 있다. 웹상에서의 개인화는 웹사이트에 접속하는 이용자의 성향과 행태별로 세분화하여, 이용자가 선호할 수 있는 적절한 정보 또는 상품을 제공함으로써 보다 적극적인 서비스를 제공하는 것을 의미한다. 이와 같은 서비스 전략은 이용자의 요구를 만족시킴으로써 해당 웹사이트에 대한 이용자의 충성도를 높여줄 뿐 아니라 타깃 마케팅(target marketing)과 일대일 마케팅(one-to-one marketing)을 가능하게 해준다는 점에서 의의가 크다(김용, 문성빈 2006). 특히 최근의 u-Commerce 환경에서 개인화는 핵심요소로서 새로운 혁신으로 인식되고 있다(Sheng et al. 2005; 한현수, 임동수 2008).

2.1 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. 전통적인 추

천 시스템은 사용자의 선호정보를 가지고 이와 부합하는 콘텐츠를 추천하고자 했던 규칙기반(rule-based) 방법이 주를 이루었으나 1990년대 이후의 연구 방향은 내용기반(contents-based) 방법 및 협업여과(collaborative filtering) 방법으로 구분되고 있다(Adomavicius and Tuzhilin 2005).

대표적인 추천방법은 적용 방법에 따라 〈표 1〉과 같이 분류할 수 있다(김용 외 2009).

2.1.1 협업여과 추천방법

협업여과라는 용어는 1990년 초에 제록스(Xerox)의 PARC 연구소에서 개발한 Tapestry에서 처음 사용되었다(Goldberg et al. 1992). 협업여과 추천방법은 특정 이용자의 성향과 비슷한 다른 이용자의 성향을 통합/분석하여 새로운 정보 또는 상품과 같은 항목에 대한 선호도를 예측하는 방법으로서 추천 대상 사용자와 취향이 유사한 사용자가 선호했던 아이템의 선호도를 예측하거나(Ahn 2008), 추천 대상 사용자가 선호했던 아이템을 선호하는 다른 사용자가 선호한 다른 아이템에 대한 선호도를 예측하는 방법이다(Sarwar et al. 2001). 협업여과 추천방법은 자동으로 분석이 어려웠던 영상, 음향, 아이디어, 감정 등의 속성이 이용자의 평점으로 평가되기 때문에 기존의 내용기반 추천방법에서 다루지 못했던 정보 또는 상품들에 대한 추천이 가능하다. 특히, 다른 이용자의 경험을 바탕으로 하기 때문에 이용자가 기존에 선호해 왔던 정보와는 다르지만, 다른 이용자가 높이 평가할 수 있는 정보에 대한 추천도 가능하다. 그러나 이용자의 행위 데이터의 수가 적으면 공통된 정보에 대해 평점을 내린 이용자 집단이 작기 때문에 최근접 이웃을 찾기 어렵기 때문에 시스템의 성능이 저하된다(Linden et al. 2003). 따라서 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.

〈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 기법을 주로 이용	- 추천 시간이 짧음 - 확장성 및 희소성문제 일부 해결	- 대용량의 콘텐츠에 체계적 적용이 어려움 - 개인성향 반영 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 반영가능 - 초기평가와 희소성 문제를 부분적으로 해결	- 멀티미디어 자료에 적용이 어려움 - 계산의 복잡성 - 신규이용자 문제 - 용어의 중의성

2.1.2 내용기반 추천방법

내용기반 추천은 정보검색 분야에서 그 기원을 두고 있는 것으로 사용자가 과거에 구매를 하였거나 관심을 보인상품의 프로필과 유사한 상품간의 비교를 통하여 추천이 이루어진다(Wu and Chen 2001). 즉, 내용기반 추천방법은 이용자의 항목에 대한 평가 정보 혹은 구매내역을 바탕으로, 미리 정의된 항목에 대한 특징들(features)에 의해 이용자 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. 주로 텍스트 기반의 뉴스나 인터넷 기사, 도서, 모바일 앱, 영화, 음악 등에 대한 추천 시스템에서 주로 사용된다. 그러나 내용기반 추천방법은 다룰 수 있는 정보의 유형이 대부분 텍스트로서 그 범위가 좁으며 추천 정보가 하나의 분야나 경향에 집중되기 쉽다. 또한 이용자에 대한 추천은 반드시 프로파일이 생성되어야만 추천할 수 있다.

2.1.3 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 기반으로 통계적 방법에 의해 연관성이 있는 항목들 사이의 규칙성을 추출하는 연관규칙이 많이 사용된다.

2.2 연관규칙 기반 개인화 정보추천

2.2.1 정의

연관규칙은 데이터마이닝의 여러 기법들 중 하나로써 개인화 추천서비스에 널리 활용되고 있으며 장바구니 분석(market basket analysis)이라고도 불리는 것으로써 항목집합(itemset)들로 이루어진 데이터베이스에서 항목들의 동시출현성향에 대한 관계성을 표현한다(김미성 외 2012). 즉, 조건부 확률로써 “사건 A가 일어났을 때, 사건 B가 일어나는 것”을 나타내며, 추천 시스템에서는 “임의의 고객이 조건 A를 만족할 경우, 조건 B를 만족한다.”를 의미한다. 여기서 조건 A는 고객의 나이, 성별 등의 특징이 될 수 있으며 또는 이용자가 선택한 항목들을 나타낼 수 있다. 그리고 조건 B는 이용자에게 추천할 항목으로 정의된다. 연관규칙 추출의 대상인 항목집합(itemset)은 트랜잭션(transaction)이라고 정의된다. 예를 들어 슈퍼마켓에서

한 번에 여러 개의 물건을 샀을 때, 해당 물건들의 리스트 자체가 하나의 항목집합이 되며 충분히 빈번하게 나타나는 항목집합을 빈발항목집합(large itemset)이라고 한다. 두 개의 빈발항목집합 X와 Y에 대해 연관규칙은 X → Y의 형태로 표현되며 의미적으로 빈발항목집합 X를 포함하는 트랜잭션은 또 다른 빈발항목집합 Y도 함께 포함하는 경향이 있음을 의미한다. 예를 들어, (Bread, Butter) → (Milk)라는 연관규칙이 있다고 하면, 이는 빵과 버터를 함께 구매하는 고객은 우유도 함께 구매하는 경향이 있음을 의미한다. 연관규칙은 항목집합들 간의 규칙은 아니며, 트랜잭션에 함께 출현하는 항목집합들 간의 관련성을 규칙의 형태로 표현한 것이다.

연관규칙과 관련하여 지지도(support)와 신뢰도(confidence) 및 향상도(lift)라는 세 가지 값이 존재한다. 지지도는 연관규칙을 구성하는 항목집합을 포함하는 트랜잭션이 전체 트랜잭션에서의 비율을 의미하는 것으로써 위의 예에서는 전체 매장의 구매 고객 수 중에서 빵과 버터와 우유 세 가지 모두를 구매한 고객수의 비율이라고 할 수 있다.

신뢰도는 항목간의 관련성의 성립 정도로서 빵과 버터를 사는 고객이 우유까지 함께 구매하는지에 대한 비율을 위미한다. 신뢰도가 규칙의 강도를 나타낸다면 지지도는 해당 규칙이 전체 데이터베이스에서 가지는 통계적인 중요성을 표현한다.

향상도는 고객이 빵을 사는 경우 해당 트랜잭션이 버터를 동시에 구매하는 경우와 버터가 임의로 구매되는 경우의 비율을 나타내주며 연관관계를 파악할 수 있다.

항목집합으로 이루어진 데이터베이스에서 연관규칙을 추출하는 작업은 최소지지도와 최소신뢰도가 주어졌을 때, 주어진 최소지지도 값보다 높은 지지도를 가지면서, 주어진 최소신뢰도보다 큰 신뢰도 값을 갖는 모든 연관규칙을 추출하는 것으로서 최소지지도 이상을 가지는 항목집합이 빈발항목집합이 된다.

2.2.2 특징

생성된 연관규칙을 해석하고 활용하는 방법은 통계적 방법을 이용한 가설의 검증과는 다르다. 예를 들어, 인간의 유전자와 그에 의해 발현되는 형질에 대한 연구에서

(유전자 a) → (형질 g): 23 %
(유전자 b) → (형질 g): 19 %
(유전자 a, 유전자 b) → (형질 g): 85 %

와 같은 형태의 연관규칙이 발견되었다고 가정하면 통계적 방법을 이용하여 유전자 a가 나타

$$\text{지지도}(Support) = P(A \cap B) = \frac{A\text{와 }B\text{를 동시에 포함하는 거래의 수}}{\text{전체 거래수}} \quad \langle \text{수식 1} \rangle$$

$$\text{신뢰도}(Confidence) = P(B \mid A) = \frac{P(A \cap B)}{P(A)} = \frac{A\text{와 }B\text{를 포함하는 거래수}}{A\text{를 포함하는 거래수}} \quad \langle \text{수식 2} \rangle$$

$$\text{향상도}(lift) = \frac{P(B|A)}{P(B)} = \frac{P(A \cap B)}{P(A)P(B)} = \frac{A\text{와 }B\text{를 동시에 포함하는 거래수}}{A\text{를 포함하는 거래수} * B\text{를 포함하는 거래수}} \quad \langle \text{수식 3} \rangle$$

날 때 형질 g가 나타나는 경우와 유전자 b가 나타날 때 형질 g가 나타나는 경우, 그리고 유전자 a와 b가 모두 나타날 때 형질 g가 나타나는 경우 등 모든 가짓수를 나열하고 통계치를 계산한다면, 연관규칙과 같은 형태의 지식을 얻을 수 있다. 그러나 유전자가 수백, 수천가지라고 한다면 가능한 모든 조합의 수는 지수적(exponential)으로 증가하기 때문에 통계적 방법으로는 처리하기에는 너무나 많은 연산량이 요구된다(윤종찬, 윤성대 2007). 따라서 연관규칙의 생성은 대용량의 데이터베이스에서 항목들 간의 관계성을 분석할 수 있는 효과적인 방법론으로써 과거에 알지 못했거나 규칙으로 표현될 가능성이 있는 조합이 너무 많아서 통계적으로 처리 하지 못했던 연관성들을 찾아 주는 유용한 방법이라고 할 수 있으며 연관규칙을 사용하면 이용자의 거래정보 없이 연관규칙을 추출하는 것만으로 추천시스템을 구동할 수 있다는 장점이 있다(오재영, 전종훈 2004).

본 연구에서는 변형된 Apriori 알고리즘 기반의 연관규칙추출 및 정보추천방법의 특징은 다음과 같이 정리할 수 있다. 첫째, 연관규칙 추출에 있어서 〈선행부 → 결과부〉로 표현되는 연관성 분석 결과를 해석하는데 있어서 이해하기 쉽고 이를 즉각적으로 실제에 적용하기 용이하다. 둘째, 서술적 모델(descriptive model)이라고도 불리는 비지도 학습(unsupervised learning) 분석기법으로서 대부분의 데이터마이닝 기법들이 예측적 모델(predictive model)로써 지도 학습적(supervised learning) 특성을 가지고 있기 때문에 목적변수가 뚜렷하지 않은 경우에는 적용이 어려운데 비하여 연관성 분석은 이러한 상황에서 적절한 해결 방법이 될 수 있다. 셋째, 사용이 편리한 분석 데이터의 형태라는 점이다. 이는 트랜잭션 내용에 대한 데이터를 복잡한 변환 없이 이용할 수 있는 간단한 자료구조를 갖는 분석방법이다. 넷째, 연관규칙 추출에 따른 연산의 용이성에 있다. 연관규칙 추출에 있어서 대용량 데이터일 경우에 계산의 수가 크게 증가하지만 분석을 위한 계산은 상당히 간단하며, 단순한 결과만을 위해서는 기초적인 워크시트의 사용법만을 알고 있는 분석자도 분석이 가능하다. 다섯째, 이용자의 이용행위를 유형별로 구분하고 중요도에 따른 가중치를 통하여 추천에 따른 정확도를 높일 수 있도록 설계되었다. 마지막으로 대용량의 데이터를 처리할 수 있는 시스템 구조를 기반으로 효과적인 연관규칙을 추출할 수 있도록 설계되었다.

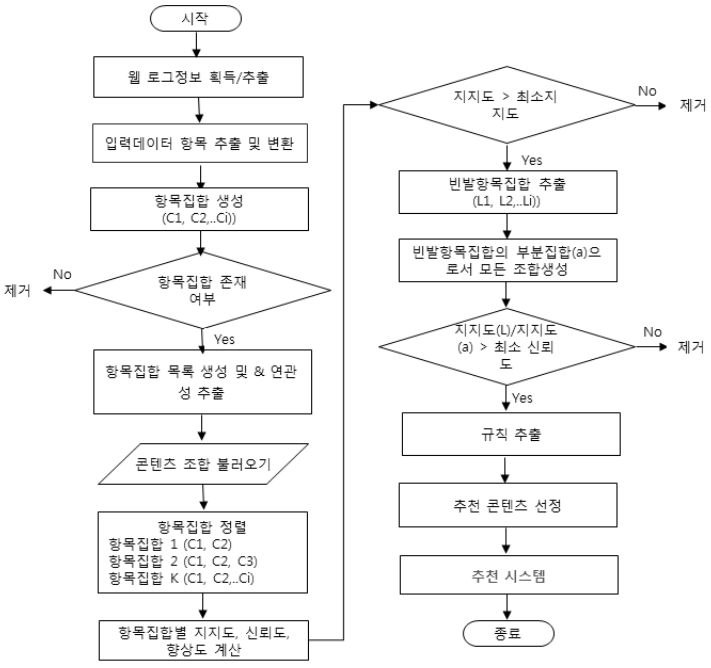
3. 연관규칙 추출 및 정보추천 방법

본 연구에서는 대용량의 정보환경에서 효율적인 정보추천 시스템 구현을 위하여 연관규칙 기반의 정보추천 방법을 제안하였다. 연관규칙 추출 알고리즘으로는 Apriori, OCD, SETM, DHP알고리즘 등이 있으며(Wang et al. 2005) 본 연구에서는 구현이 용이하고 이해하기 쉬우면서 적절한 결과를 생성할 수 있으며 이진연관규칙에 대한 빈발항목집합을 찾아내는데 유용한 알고리즘으로써 Apriori 알고리즘을 활용하였다. Apriori라는 명칭은 알고리즘이 빈발항목집합 특성인 사전지식(prior knowledge)을 활용한다는 점에서 비롯되었다(박우창 외 2003).

연관규칙 활용에 가장 큰 문제점은 수행시간

이 오래 걸릴 수 있다는 점과 최소신뢰도, 최소지지도와 같은 적절한 수행 파라미터 설정이 어렵다는 점이다. 항목 4개만을 담고 있는 하나의 트랜잭션에서 발생할 수 있는 연관규칙의 조합수만 하더라도 50개나 되며, 장바구니 분석과 같은 일반적인 응용분야에서 10~20개의 항목을 담고 있는 수 십 만개 이상의 트랜잭션으로 구성된 데이터베이스를 대상으로 한다. 그리고 연관규칙 탐사과정은 많은 조합 중에서 지정된 지지도와 신뢰도를 만족시키는 규칙들을 추출하는 것이기 때문에 여기에는 많은 주기억장치와 계산량이 소요된다. 따라서 본 연구에서는 이와 같은 문제점을 해결하기 위하여 대용량의 데이터를 다룰 수 있으면서 상대적으로 컴퓨터 처리비용이 낮은 시스템에서 안정적으로 동작할 수 있는 시스템 구조 및 서비스 환경을 제안하였다.

본 연구에서 제안하고 있는 연관규칙 기반의 정보추천방법의 단계별 과정은 다음과 같다. 먼저 이용자의 웹상의 행위 정보를 담고 있는 웹로그를 전처리 과정을 통하여 분석하여 이용자 번호(user number), 항목번호(contents number), 이용자 행위유형에 대한 정보(user behavior information) 등의 세부항목정보를 추출하는 단계, 두 번째, 연관규칙 추출을 위한 데이터 형식 변환 및 데이터베이스에 저장하는 단계, 세 번째, 연관규칙을 적용하여 추천항목을 추출하는 과정으로서 임계값 이하 범위에 포함되는 항목을 통하여 빈발항목집합을 생성하고 이를 통하여 연관규칙을 추출하는 단계, 마지막으로 추출된 연관규칙을 기반으로 최종의 추천항목을 선정하는 단계로 구분할 수 있다. 〈그림 1〉은



〈그림 1〉 제안된 연관규칙 기반의 정보추천 흐름도

본 연구에서 제안하고 있는 방법을 통하여 정보 추천을 수행하는 과정을 도식화하여 보여주고 있다.

3.1 연관규칙 생성

제안된 방법을 통하여 연관규칙을 추출하고 이를 기반으로 추천항목을 선정하는 방법은 다음과 같다. 먼저, 이용자의 이용행위를 담고 있는 로그파일에서 필요로 하는 항목을 추출한다. 둘째, 해당 항목에서 Apriori 알고리즘을 적용하여 각각의 항목집합에서 최소지지도를 만족하는 빈발항목집합을 추출한다. 셋째, 추출된 빈발항목집합을 기준으로 연관규칙을 추출하고, 추출된 연관규칙은 데이터베이스에 저장한다. 마지막으로 추출된 연관규칙을 통하여 해당 항목(c)에 대한 관계성이 높은 추천항목(R(c))을 추출한다. 연관항목(R(c))은 다음과 같이 정의된다.

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

여기서 "c_k"는 항목에 부여된 식별번호(identification number)를 의미하며 "m"은 해당 항목과의 연관성을 의미하는 값으로서 일반적으로 항목에 대한 지지도를 의미한다.

연관규칙 추출을 위하여 본 연구에서는 변형된 Apriori 알고리즘 활용하였다. 이전의 연구에서는 콘텐츠간의 연관성 추출을 위하여 데이터마이닝의 연관성 추출기법을 그대로 활용했으나, 본 연구에서는 연산처리의 개선 및 효과적인 추천을 위하여 다음과 같이 변형된 연관성 추출기법을 사용하였다. 첫째, 선행부와 결과부에는 각각 하나의 항목집합만을 가지는 규칙이 생성됨으로써 결과부 또는 선행부에 두 개 이상의 항목을 가지는 규칙은 허용되지 않는다. 둘째, 규칙을 추출하는데 있어서 향상도가 1 이상인 것만 추출한다. 이를 통하여 통계적으로 의미가 있는 규칙만을 추출할 수 있다. 셋째, 추출된 연관규칙 중에서 하나의 동일한 선행부에 대응되는 결과부의 경우의 수를 제한한다. 여기서 결과부에 나타날 수 있는 항목에 대한 경우의 수는 동일한 선행부에 대하여 동일한 수의 규칙만이 추출된다. 한편, 연관규칙의 추출을 위하여 먼저 빈발항목집합을 추출하여야 하며 해당 빈발항목집합을 추출하는데 있어서는 지지도를 적용하였다.

3.1.1 Apriori 알고리즘

Apriori 알고리즘은 구현이 간단하고 성능 또한 만족할 만한 수준을 보여주는 알고리즘으로써 패턴분석에 적합한 알고리즘으로써 해당 데이터베이스에 있는 특정한 길이(항목의 수)를 가지는 항목집합을 선택하는 반복적인 알고리즘이다. 빈발항목집합의 생성을 위하여 특정 데이터베이스에서 발생하는 모든 트랜잭션을 분석하고 항목에 대한 지지도를 이용하여 동시에 자주 나타나는 항목들을 정제하고 빈발항목집합에서 생성된 규칙들을 신뢰도를 이용하여 정제하는 방식으로 후보항목집합(candidate itemset)에서 각각의 지지도를 계산한 후 이용자가 정의한 지지도보다 크거나 같은 조건을 만족하는 데이터로 빈발항목집합을 구성한다(하단심, 황부현 2000).

Apriori 알고리즘은 트랜잭션과 각 트랜잭션에 대한 항목집합이 주어졌을 때, 항목집합과 다

른 항목집합들 사이에 연관관계에 대한 규칙을 찾는 것을 목표로 한다. 예를 들어 90%의 연구자들은 '스마트폰' 자료를 내려받기를 할 때, '모바일앱' 자료를 동시에 같이 내려받기를 한다는 가정 하에서 '스마트폰'과 '모바일앱' 항목에 대한 연관관계에 대한 규칙을 발견하고자 한다면, 연관관계에 대한 계산은 항목들의 출현빈도를 이용하여 계산하게 된다. 여기서는 두 개의 항목이 동시에 출현하게 될 경우의 확률에 대해서 지지도 s라고 정의하고, 항목 X가 출현할 때 항목 Y가 동시에 출현하게 될 조건부 확률을 신뢰도 c라고 가정한다. 〈표 2〉에서 주어진 데이터를 기준으로 최소 지지도를 50%, 최소 신뢰도를 50%라고 가정하여 연관규칙 생성과정을 알아보면 다음과 같다.

〈표 2〉 예제 데이터

Transaction ID	내려받기 항목
1	모바일앱, 네트워크, 스마트폰
2	모바일앱, 스마트폰
3	모바일앱, 갤럭시
4	네트워크, 옵티머스, 아이폰

우선 모바일앱과 스마트폰의 연관관계를 살펴보면 '모바일앱' 자료를 내려받기를 할 때 '스마트폰' 자료를 동시에 내려받기를 하는 조건부 확률은 66.6% 정도가 될 수 있으며, 전체 항목에서 '모바일앱'과 '스마트폰'이 동시에 출현할 확률은 50%가 된다. 따라서 지지도는 50%, 신뢰도는 66.6%가 됨을 알 수 있다. 이와 반대의 경우로 '스마트폰' 자료를 내려받기를 할 때 '모바일앱' 자료를 동시에 내려받기를 할 확률은 지지도 50%와 신뢰도 100%가 됨을 알 수 있다. 이를 연관성관계로 표현하면 다음과 같다.

모바일앱 => 스마트폰 (50%, 66.6%)
스마트폰 => 모바일앱 (50%, 100%)

연관규칙 생성을 위해서는 최소지지도 이상의 지지도를 가지는 항목집합들의 모든 집합들을 추출하여야 하며 여기서 추출된 항목집합을 빈발항목집합이라고 한다. 〈표 2〉의 데이터를 기준으로 계산하면 최소 지지도를 50%로 주었기 때문에 50% 이하의 지지도를 가지는 아이템을 모두 제거하면 〈표 3〉과 같은 항목집합이 나온다.

〈표 3〉 항목들에 대한 지지도

항목집합	지지도(Support)
{모바일앱}	75%
{네트워크}	50%
{스마트폰}	50%
{모바일앱, 스마트폰}	50%

이를 이용해 후보집합을 추출하면 {모바일앱, 네트워크}, {모바일앱, 스마트폰}, {네트워크, 스마트폰}의 세 가지가 됨을 알 수 있다. 이를 다시 이용해 빈발항목집합을 〈표 2〉의 데이터를 기반으로 다시 계산하면 〈표 4〉와 같은 결과가 나오게 된다.

〈표 4〉 최종 빈발항목집합

항목집합	지지도(Support)
{모바일앱, 스마트폰}	50%

하나의 빈발항목집합이 생성되었고, 더 이상 후보집합을 구할 수 없으므로 {모바일앱, 스마트폰}이 최종 빈발항목집합이 된다. 이를 이용해 연관 규칙을 추출하면 〈표 5〉와 같다. 여기서 신뢰도가 50% 이상인 규칙을 가져오면 원하는 항목이 추출된다.

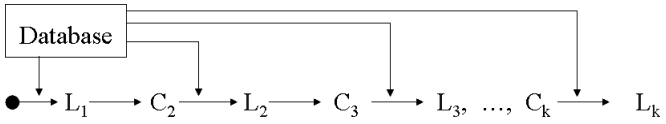
〈표 5〉 생성된 연관규칙

항목집합	지지도(Support)
모바일앱 =〉 스마트폰	2/3 = 66.6%
스마트폰 =〉 모바일앱	2/2 = 100%

3.1.2 빈발항목집합(Large Itemset) 탐색

빈발항목집합의 추출을 위한 탐색과정은 〈그림 2〉와 같다. 데이터베이스로 부터 후보항목집합을 생성하고, 이를 데이터베이스 트랜잭션과 비교하여 빈발항목집합을 찾아내는 과정을 반복하며 최종적으로 빈발항목집합에 공집합이 될 때까지 반복하는 과정을 거친 후 최종 빈발항목집합을 생성하게 된다.

〈그림 3〉에서는 Apriori 알고리즘을 기반으로 연관규칙 생성을 위하여 요구되는 후보항목집합과 빈발항목집합을 생성하는 알고리즘을 보여주고 있다.



〈그림 2〉 빈발항목집합 탐색과정

Input :
C_k // 후보항목집합(candidate itemset)
$s_{\min}$ // 최소지지도
Output :
L_k // 빈발항목집합(large itemsets)
단계 0. 최소 지지도($s_{\min}$)를 정한다.
k=1
$C_1 = [\{i_1\}, \{i_2\}, \ldots . \{i_m\}]$
$L_1 = \{c \in C_1 | p(c) \geq s_{\min}\}$
단계 1. k = k+1
L_{k-1}로부터 C_k형성 (Apriori-gen 함수)
단계 1-1. (join) L_{k-1}의 집합들을 접합하여 k-항목집합군을 형성함.
$C = L_{k-1} * L_{k-1}$
단계 1-2. (prune) C의 (k-1)-항목부분집합이 L_{k-1}에 속하지 않을 때 이를 모두 제거한 후 C_k를 형성한다.
$C_k = \Phi$이면 Stop.
단계 2. C_k의 집합 중 지지도가 최소지지도($s_{\min}$) 이상인 것을 모아 L_k 생성함
$L_k = \{c \in C_k | p(c) \geq s_{\min}\}$
단계 1 반복.

〈그림 3〉 빈발항목집합 생성 알고리즘

〈그림 3〉에서 C_k는 새로운 후보항목의 집합이고, L_k는 자주 발생하는 길이가 k-항목들의 집합인 빈발항목집합이 된다. 첫 번째에서 항목의 발생빈도수를 계산하기 위하여 트랜잭션 데이터베이스를 읽는다. 다음 단계로써 최소지지도를 설정하여 후보1-항목집합들을 추출하고 후보1-항목집합으로부터 최소지지도를 만족하는 빈발1-항목집합들(L_1)을 추출할 수 있다. 빈발2-항목집합들을 탐색하기 위하여 모든 부분집합도 역시 최소지지도를 가져야 하므로 Apriori 알고리즘은 후보항목집합(C_2)를 생성하기 위하여 $L_1 * L_1$을 이용한다. 여기서 ($*$)은 접합(join) 연산자이다. 다음 단계로써 다시 트랜잭션 데이터베이스를 읽고 C_2에 속한 후보항목집합의 지지도가 계산된다. 빈발2-항목집합들(L_2)은 C_2에 속한 후보2-항목집합의 지지도에 따라 결정된다. 후보항목집합(C_3)은 빈발2-항목집합들(L_2)에서와 같은 과정으로 생성된다. 빈발3-항목집합들(L_3)에서 더 이상 다른 후보4-항목집합을 구할 수 없다면 Apriori 알고리즘은 빈발3-항목집합(L_3)를 구성한 후 빈발항목집합을 탐사하는 과정을 마친다.

3.2 정보추천 알고리즘

연관규칙의 생성은 빈발항목집합들을 이용하게 되는데 모든 빈발항목집합(L)에 대하여 'L'의 모든 공집합이 아닌 부분집합들을 찾는다. 각각의 부분집합 'A'에 대하여, 만약 Support(A)에 대한 Support(L)의 비율이 적어도 최소 신뢰도 이상이면 A → (L-A)의 형태의 규칙을 출력한다. 연관규칙에 대한 계산은 〈수식 1〉과 〈수식 2〉에 나타난 지지도와 신뢰도 계산을 통해 특정 임계값을 설명하는 최소신뢰도 이상인 규칙들을 추출한다. 연관규칙을 통하여 콘텐츠 간의 상호 연관성을 통하여 추천 콘텐츠를 선정하는 과정으로써 추천대상 이용자가 과거에 정보이용행위에 대한 내역을 기반으로 해당 이용자의 정보에 대한 선호도를 항목 간의 연관성을 통하여 해당 이용자에게 추천하는 방법을 이용한다. 항목 간의 연관성에 대한 기준은 연관규칙의 지지도가 된다. 이를 위하여 콘텐츠 간의 상호 연관성을 분석하여 추천하는 연관규칙 기반의 추천정보(RI)는 〈그림 4〉와 같이 정의된다.

$RI(u) = \{(c_1, s), (c_2, s), (c_3, s), \ldots, (c_n, s)\}$
c_k = 연관규칙 기반의 추천항목
s = 해당 항목의 선호도

〈그림 4〉 추천항목에 대한 정의

추천항목을 선정하기 위하여 해당 이용자가 이용한 항목들에 대한 목록과 추출된 연관규칙을 통하여 생성된 연관규칙 목록을 기반으로 〈그림 5〉의 과정을 반복하면서 추천항목(RI)에 대한 선호도를 생성한다.

```
Algorithm Calculate_RI
    RI = { }
    이용자 행위정보 목록에 포함된 각 항목(c_k) 및
    연관항목 내의 (c_n, m) 각 항목에 대해 다음을
    반영한다.
        RI내에 c_j가 없으면, (c_n, m)을 추가한다.
        RI내에 c_j가 포함되어 있으면, 기존의 값과 비교
        하여 최대값 (max)을 취한 m으로 변경한다.
    Return RI
```

〈그림 5〉 제안된 Apriori 알고리즘 기반 정보추천방법

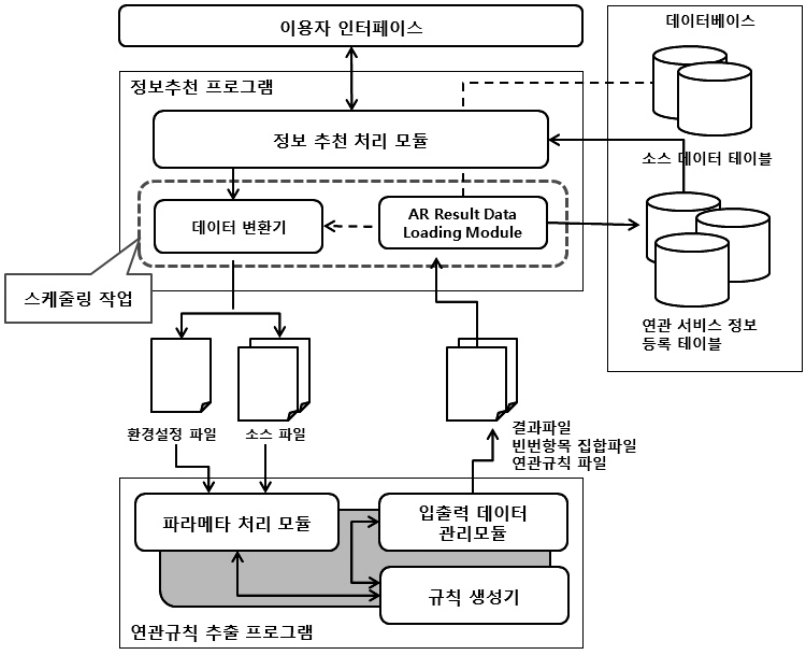
4. 시스템 설계 및 구현

4.1 구현환경

본 연구에서 제안하고 있는 연관규칙 기반의 정보추천 방법의 구현을 위하여 이용자의 이용행위를 포함하고 있는 소스파일에 대한 분석을 통하여 빈발항목집합을 추출하고 이를 통하여 연관규칙을 생성한다. 생성된 연관규칙을 통하여 최종적으로 정보추천이 이루어진다. 이를 위하여 본 연구에서 정의하고 있는 이용자의 이용행위는 웹상에서 이용자가 해당 항목에 대하여 내려받기, 구매, 단순클릭 등의 직접적 행위를 의미하며 각각의 이용자 행위는 분석과정에서 중요도에 따라 가중치가 설정된다. Apriori 알고리즘 기반의 정보추천 방법의 구현 및 평가는 Windows 7 64bit OS 운영체제에서 MS SQL Server 2003의 데이터 처리용 DBMS를 가지고 Visual Studio .NET, C++/STL의 개발환경에서 구현되었다.

4.2 시스템 구조

본 연구에서 제안하고 연관규칙 기반의 정보추천방법은 데이터마이닝 분야에서 대표적인 비지도 학습과정을 통하여 유사항목 간의 관계성을 추출하여 이를 이용자에게 직접 추천하는 방법으로서 온라인 환경에서 이용자의 정보탐색 및 이용행위를 암묵적으로 보여주는 이용자 로그파일을 통하여 추출된 이용자의 행위정보를 기반으로 연관규칙을 추출하기 위한 것이다. 이를 위하여 로그파일을 통하여 추출된 분석 데이터를 시스템을 통하여 처리할 수 있는 입력 형식에 맞도록 변형하는 전처리 과정을 통하여 추출된 데이터를 실제 연관규칙 추출프로그램에서 사용할 수 있는 형태로 변형하는 후처리 과정이 요구된다. 이러한 과정은 자동화된 방식으로 처리가 되며 설정된 특정 시간에 정해진 로그파일로부터 추출하여 연관규칙 추출과정을 수행하고 해당 결과를 정해진 형식으로 전달하게 된다. 이러한 과정을 통하여 생성된 연관규칙을 기반으로 이용자에 대한 정보추천이 이루어진다. 〈그림 6〉에서 볼 수 있듯이 입력데이터 파일은 연관규칙 추출을 위하여 시스템에 설정되어져 있는 지지도, 신뢰도 및 향상도 등의 파라미터 값을 기준으로 연관규칙을 생성하게 된다. 이러한 과정에서 최소지지도의 설정에 따라 결과의 개수에 많은 영향이 있을 수 있다. 값이 너무 큰 경우에는 결과값이 하나도 없을 경우가 있으며 값이 너무 작은 경우에는 결과값이 기하급수적으로 늘어날 수 있다. 특히 최소지지도가 너무 작은 경우 연관규칙 추출프로그램이 너무 오랫동안 수행하는 상황이 발생할 수 있다. 따라서 입력되는 파라메타의 값은 신중히 결정하여야 한다. 한편, 생성된 연관규칙에 대한 결과정보를 실질적인 정보추천에 적용하기 위해서는 데이터베이스에 저장되어야 한다. 이러한 과정은 정형화된 스케줄링 과정을 통하여 수행되며 이용자의 행위정보를 기록하고 있는 로그파일인 소스파일에서 입력데이터의 추출, 연관규칙 생성 및 정보추천의 과정이 순차적으로 수행된다. 제안된 추천 시스템은 정보추천 프로그램 추천서버와 학습서버 및 이용자 인터페이스로 구성되어 있는 처리 부분과 이용자 프로파일 및 콘텐츠 프로파일 정보를 담고 있는 데이터 저장 부분으로 구성되어 있다. 각 요소별 기능은 다음과 같다.



〈그림 6〉 연관규칙 추출 및 정보추천 프로그램 구조

- 이용자 인터페이스(User Interface): 웹서버를 통하여 접속된 이용자와의 접점으로서 이용자의 행위정보를 수집하고 이용자의 요청을 해당 서버에 전달 및 추천정보를 표현하는 기능을 수행한다.
- 연관규칙 추출프로그램(Association Rule Extraction Program): 이용자의 행위정보를 분석하여 이용자의 이용자 프로파일 정보를 갱신하는 기능을 수행한다.
- 정보추천 프로그램(Information Recommendation Program): 이용자 프로파일 정보와 일치하는 콘텐츠를 추천하는 기능을 수행한다.
- 데이터베이스(Database): 데이터베이스는 추천을 위한 연관규칙, 각종정보 및 이용자 프로파일과 콘텐츠 프로파일을 저장하는 기능을 수행한다.

4.3 연관규칙 생성 및 저장

4.3.1 전처리 과정

전처리과정은 로그파일로 부터 연관규칙 추출을 위하여 요구되는 항목을 추출하고 이를 연관규칙 추출프로그램에서 요구되는 형식으로 변환하는 과정으로서 소스파일로 부터 연관규칙 추출프로그램에서 사용할 수 있는 형태로 변환하는 과정이다. 실질적으로 연관규칙을 추출하기 위해서 분석에 사용할 데이터는 전처리 과정을 통해 연관규칙추출엔진에 전달하게 될 정수형(integer)의 항목으로 구성된 텍스트 혹은 이진 파일로 변형된다. 예를 들어 웹 로그를 이용한 사용자들의 사이트 이용패턴을 분석하고

자 할 때 스케줄을 이용해 일별 웹 로그 파일을 읽어오게 된다. 웹 로그파일이 〈표 6〉과 같다고 가정하면

〈표 6〉 입력 데이터로 전달받은 원시 소스파일의 예

61.33.35.AAA	/game/run.jsp?id=P132
61.33.35.AAA	/game/run.jsp?id=S367
61.33.35.AAB	/game/run.jsp?id=A937
211.105.83.AAA	/game/run.jsp?id=P132

소스화일에서 필요한 항목의 추출과 추출된 항목에 대한 설명화일을 참고하여 최종적으로 정보추천에 필요로 하는 형식으로 변환된 변환 파일이 본 과정에서 최종적으로 생성된다.

4.3.2 연관규칙 생성

이진형태(8 bytes binary)의 데이터를 입력으로 받아 연관규칙을 추출하는 작업을 수행하기 위해서는 최소신뢰도, 최소지지도 등의 파라메타에 대한 설정값 등과 같은 설정이 필요한 입력사항은 미리 생성한 환경파일을 이용하거나 명령라인의 파라미터를 이용하여 입력된다. 이러한 환경설정을 통하여 수행된 결과물로는 실행상태를 파악할 수 있는 실행결과 파일, 빈발항목집합 파일, 연관규칙 파일, 데이터베이스에 저장하기 유용한 형태의 연관규칙 패턴 파일이 생성된다. 빈발항목집합 파일과 연관규칙 파일은 이진파일로 생성된다.

1) 빈발항목집합 결과파일(LARGE FILE)

빈발항목집합과 그에 대한 지지도를 이진 포맷과 텍스트 포맷 파일로 제공한다. 지지도는 실수 값으로 표시된다. 〈그림 7〉은 빈발항목집합에 대한 이진형식으로의 표현방법에 대한 실례를 보여주고 있으며 전체 항목이 여섯 개의 항목으로 구성되어 있고 항목의 개수는 3개가 되는 (23, 225, 4)인 빈발항목집합이 지지도 20.5%를 유지한다는 것을 표현하고 있다.

```
표현방법 : total_size #of_items item_list support ending_zero
표현 예 : 6 3 23 225 4 0.205 0
```

〈그림 7〉 빈발항목집합의 이진형식표현 예

2) 연관규칙 결과파일(RULE FILE)

연관규칙은 이진포맷 형식으로 생성되며 생성된 연관규칙 파일에서는 연관규칙과 지지도, 신뢰도, 향상도에 대한 정보를 포함하고 있다. 〈그림 8〉에서는 결과물로써 연관규칙추출 과정을 통하여 생성된 결과물에 대한 이진포맷으로 표현된 예를 보여주고 있으며 전체는 아홉 개의 항목으로 구성되어져 있다. 첫 번째 규칙은 2개의 항목으로 두 번째 규칙은 1개의 항목으로 구성되었으며, 각각의 값은 (23, 445) → (4) 라는 것을 알 수 있으며 연관규칙의 지지도는 20.5%, 신뢰도는 36.5%, 향상도는 2.1라고 알 수 있다.

```
표현방법: total_size antecedent_size consequent_size items support confidence lift ending-zero
표현예 : 9 2 1 23 445 4 0.205 0.365 2.1 0
```

〈그림 8〉 연관규칙결과 이진형식표현 예

3) 연관규칙 패턴화일(PATTERN FILE)

생성된 연관규칙 결과를 데이터베이스에 저장하기 위한 형태로 생성된다. 각각의 연관규칙을 'RuleID'로 표현하여 지지도, 신뢰도, 향상도,

연관규칙 항목 정보를 텍스트 형태로 나타낸다. 〈그림 9〉는 생성 및 변환된 연관규칙의 패턴에 대한 예를 보여주고 있으며 구체적으로 Ruleid가 10007번인 규칙의 경우에 있어서 지지도 2%, 신뢰도 50%, 향상도가 5.56인 연관규칙으로 규칙을 구성하는 항목은 2개이고 head 바스켓에 한 개, tail 바스켓에 각각 한 개의 항목이 존재한다는 것을 의미한다. 추가적으로 연관규칙 패턴 파일의 연관규칙 항목 정보를 자세히 설명한 파일도 동시에 생성된다.

```
표현방법 : rule_id support confidence lift Head_size
Tail_size Item_size
표현예 : 10007 0.02 0.5 5.56 1 1 2
```

〈그림 9〉 연관규칙 패턴화일 표현 예

4.4 시스템 구현 및 정보추천

4.4.1 입력데이터 추출 및 변환

이용자의 행위정보를 기록하고 있는 기존의 로그파일로 부터 연관규칙을 추출하기 위해서는 연관규칙 생성을 위하여 요구되는 필수항목을 추출하여 연관규칙 추출프로그램에서 정의하고 있는 형식으로 변환과정을 통하여 입력데이터를 생성하여야 한다. 예를 들어 웹 사이트에서 사용자들의 소비 성향을 알고 싶은 경우 기간을 설정하여 주기적으로 이용자 구매이력 등과 같은 이용자 행위정보를 해당 테이블에서 데이터를 추출하는 단계라고 할 수 있다.

이러한 전처리 과정에서는 소스파일에서 추출된 항목이 무엇을 의미하는지, 어떤 형태로 구성된 것인지를 알려주는 원시데이터에 대한 설명파일을 같이 전달하며 〈표 7〉과 같이 전달된 설명파일을 해석하여 〈표 8〉의 형태로 변환파일을 생성한다.

〈표 7〉 원시데이터에 대한 설명 파일의 예

1	사용자 IP	string
2	방문 URL 주소	string

〈표 8〉 변환 파일의 예

#사용자IP	치환된 값
61.33.35.AAA	1
61.33.35.AAB	2
211.105.83.AAA	3
#방문 URL 주소	치환된 값
/game/run.jsp?id=P132	1
/game/run.jsp?id=S367	2
/game/run.jsp?id=A937	3

생성된 변환파일은 연관규칙에 대한 분석 결과로 나온 파일들을 실질적인 정보의 추천이 이루어지는 정보추천 프로그램에서 사용할 수 있는 형태로 재변환하는 과정에서 이진형태로 저장된 연관규칙을 해석하는 과정에서 활용된다. 원시데이터는 변환파일에 표시된 형태대로 값을 치환하여 최종적으로 연관규칙 추출프로그램에 전달할 형식으로 변환 및 생성된다.

한편, 전처리 과정에서는 로그파일에서 연관규칙 및 정보추천을 위하여 필요로 하는 항목의 추출과 함께, 이용자의 이용행위를 세분화하여 행위별 가중치를 적용한다. 즉, 이용자의 이용행위에 있어서 단순읽기, 내려받기, 실행 등으로 세분화하여 중요도에 따른 가중치를 부여하여 추천에 따른 정확도를 높이도록 하였다. 이용자의 행위정보를 보여주는 이용자 프로파일을 구성하는 항목 중에서 구매내역이 〈그림 10〉

과 같다면 연관규칙 추출프로그램에 제공할 소스데이터를 이용자의 구매이력을 그룹으로 정렬하여 파일로 추출하고 소스데이터에 문자열 타입의 데이터가 포함된 경우는 정수형 타입으로 변환할 수 있는 변환 테이블을 이용하여 정수형 타입으로 코드화된 형태의 데이터로 변환이 〈그림 11〉과 같이 이루어진다.

4.4.2 규칙생성

〈그림 12〉는 연관규칙 추출과정에 대한 개념도를 보여주고 있다. 먼저 입력파일로서 환경파일(configuration file)과 소스파일(source file)

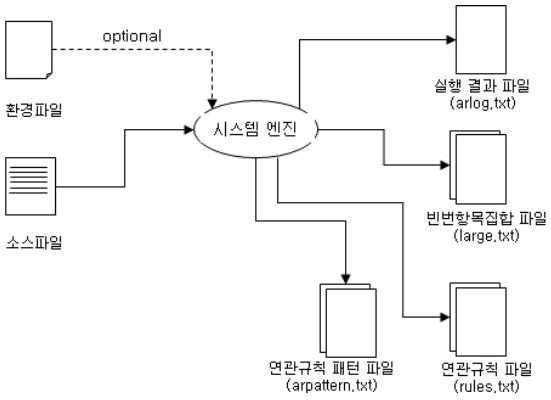
사용자 ID	사용자 이름	구입도서 ID	구입도서명
37007	금다래	C03825	자동차
37007	금다래	P10083	컴퓨터
37007	금다래	H58641	유머
37009	산머루	F25075	금융
37013	오들희	S01938	스포츠
37017	윤이나	L01748	주택

〈그림 10〉 구매이력 데이터 예

사용자 ID	구입도서명
37007	자동차
37007	컴퓨터
37007	유머
37009	금융
37013	스포츠
37017	주택

```
37007  1
37007  2
37007  3
37009  4
37013  5
37017  6
```

〈그림 11〉 추출한 항목 및 변환된 데이터 예



〈그림 12〉 연관규칙 생성 흐름도

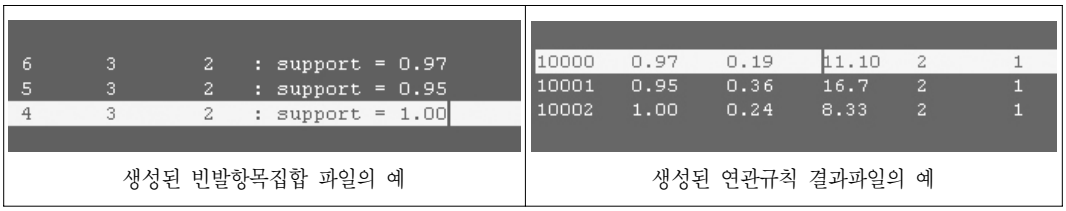
이 존재하며 환경파일은 입력데이터 파일이름, 입력데이터 형식, 최소지지도, 최소신뢰도, 빈발항목집합 등에 대한 설정값을 포함하는 파일로써 연관규칙의 추출에 매우 중요한 역할을 수행하는 파일이다. 소스파일은 이용자의 행위정보를 로그파일을 통하여 추출된 파일로써 텍스트와 이진파일 형태로 구성된다. 입력파일을 통하여 연관규칙의 추출과정을 통하여 최종적으로 생성된 출력파일은 기본적으로 수행 결과에 대한 상태파일(result file)과 이진파일 형태의 빈발항목집합파일(large file), 연관규칙파일(rule file) 및 그리고 DB에 저장하기 위한 결과 연관규칙 패턴파일들(pattern file)이 생성된다. 연관규칙 파일과 텍스트 포맷 파일들은 환경파일의 설정값에 따라 생성 여부를 결정할 수 있다.

소스파일을 분석한 결과 〈표 9〉와 같은 특정 구매패턴을 발견하였다고 가정하면 먼저 1차적으로 항목간의 관계성을 보여주는 빈발항목집합을 추출한 파일이 생성되며 생성된 빈발항목집합파일을 기준으로 연관규칙이 생성된다. 생성된 결과를 기준으로 분석모듈에서의 분석을 통하여 연관규칙 패턴파일과 연관규칙 상세파일로부터 RuleID에 해당하는 연관성 정보와 각 항목정보를 테이블에 저장함으로써 최종적인 연관규칙 생성이 이루어진다. 특히, 연관규칙 상세파일로 추출된 항목의 경우는 코드 변환을 통해 정수형 타입으로 변형된 데이터였으므로 코드변환 후처리를 통해 프로그램에서 사용하는 데이터 형태로 변형하여 테이블에 데이터를 저장하게 된다. 이러한 과정을 통하여 정보서비스 기관 또는 쇼핑몰에서는 연관규칙 테이블로부터 필요한 데이터를 조회하여 이용자의 이용행위를 통하여 분석된 이용자의 선호도를 기준으로 정보 또는 상품에 대한 추천이 이루어질 수 있다.

〈표 9〉 소스파일 분석 결과에 따른 연관성 결과

구매패턴	지지도	신뢰도
(주택, 컴퓨터) → (자동차)	97%	19%
(스포츠, 컴퓨터) → (자동차)	95%	36%
(유머, 컴퓨터) → (자동차)	100%	24%

예를 들어 소스파일에 대한 분석을 통하여 지지도와 신뢰도에 대한 계산이 〈표 9〉와 같이 연관성 결과가 추출되었다고 가정한다면 연관규칙 추출프로그램에서 설정된 파라미터를 기준으로 항목간의 연관성을 확인할 수 있는 빈발항목집합파일과 연관규칙 결과파일이 생성된다. 아래의 〈그림 13〉은 〈표 9〉의 예를 기준으로 연관규칙 추출 프로그램을 통하여 생성된 연관규



〈그림 13〉 생성된 빈발항목집합 파일 및 연관규칙 결과파일

칙생성 결과들을 대한 실례를 보여주고 있다.

4.4.3 정보추천

정상적으로 추출된 연관규칙은 〈그림 13〉에서 볼 수 있듯이 이진파일 형태로 생성된다. 결과파일들은 정보추천 시스템에서 사용할 수 있도록 후처리 과정을 통하여 〈그림 14〉의 형식을 다시 변형되어 저장된다.

〈그림 15〉에서는 본 연구에서 제안하고 있는 방법을 도서구매 사이트에 적용한 실례를 보여주고 있다. 즉, 이용자들의 웹 사이트에서의 행위정보를 추출하고 이를 연관규칙 추출프로그램을 통하여 구매패턴의 연관성을 분석한 결과를 보여주고 있다. 이와 같은 결과를 기준으로 향후 해당 이용자에 대한 도서추천을 위한 세부 주제분야 및 도서명에 대한 추천이 이루어진다. 〈그림 15〉에서 보여주는 상품간의 연관성을 분석하면 {주택} → {컴퓨터}의 지지도 0.7, 신뢰도 1.6 향상도는 2.3으로써 특정 기간 동안 이용고객의 0.7%는 {주택, 컴퓨터}를 동시에 구매하였으며 {주택}을 구매한 고객의 1.6%는 동시에 컴퓨터를 구매하였다. 또한 {주택} 구매와 {컴퓨터} 구매 사이에는 2.3배의 연관성이 있다는 것을 의미한다. 이러한 결과는 〈수식 3〉의 향상도를 추출하는 방법을 통하여 구체적으로 산출되었으며 이를 통하여 해당 이용자에 대한 정보추

RULEID	SUPPORT	CONFIDENCE	LIFT	HEADSIZE	TAILSIZE	TOTALSIZE
10000	0.97	0.19	11.1	2	1	3
10001	0.95	0.36	16.7	2	1	3
10002	1	0.24	11.1	2	1	3

〈그림 14〉 데이터베이스에 저장된 연관규칙 예

Antecedent	Consequent	Support (%)	Confidence (%)	Type	Lift	설명
컴퓨터	유머	0.5568	100	-	13.081	
자동차	컴퓨터	0.7965	14.759	-	3.4784	
유머	여행	0.7965	18.774	-	3.4784	
컴퓨터	스포츠	0.67466	14.561	-	3.3737	
스포츠	운동	0.67466	15.631	-	3.3787	
국악	바이올린	0.60557	11.22	-	1.5168	
바이올린	국악	0.5568	14.652	-	1.3058	
웹 프로그램	SNS	0.5568	77.841	-	1.2184	
여행	유머	2.3044	53.39	-	0.83571	
주택	컴퓨터	0.7300	1.62	-	2.3645	

〈그림 15〉 제안된 방법을 통한 추천 예

천은 {컴퓨터}에 대한 추천이 이루어진다. 실질적인 추천과정에 있어서의 추천 후보 항목 중에서 최종 추천을 위한 항목선정을 위한 기준이 되는 임계값은 서비스제공에 있어서 추천 항목의 수, 추천주기, 추천유형 등의 환경적 요소 등을 고려하여 설정한다.

4.5 평가

현재 추천서비스를 제공하고 있는 대부분의 사이트에서는 이용자에 의하여 입력된 명시적 정보와 인구통계학적 정보를 기반으로 구축된 추천시스템을 운영하고 있다. 그러나 정보기술의 발전을 통한 정보환경과 이용자 요구의 변화에 따라 기존의 추천방법을 추천시스템에 직접적으로 적용하기에는 몇 가지 문제점이 지적되고 있다. 첫째, 이용자가 직접 입력하는 이용자 정보의 부정확성 문제이다. 이전의 연구에서 이용자에 의해 초기에 입력되는 개인정보는 추천을 위한 중요한 요소로 고려되고 있다. 그러나 이용자 이름, 주민번호, 주소, 성별 등의 개인정보에 대한 사회적인 민감성과 함께, 이용자 성향에 비추어볼 때 대부분의 서비스 이용자는 자신의 개인정보를 정확히 제공하지 않고 있다. 따라서 현실적으로 이러한 부정확한 정보를 추천에 이용하는 것은 오히려 추천의 정확성을 낮추는 결과를 초래할 수 있다. 이와 관련하여 김진수 등(2004)은 이용자에 의해 직접 입력되는 정보를 추천을 위한 요소로 고려하는 것에 대한 문제점을 지적하고 있다. 두 번째, 콘텐츠의 내용정보를 이용하여 추천을 제공하기 위해서 콘텐츠에 대한 속성 값으로 키워드를 이용하는 경우 용어의 중의성 문제와 함께, 대용량의 콘텐츠 환경에서의 콘텐츠와 이용자 간의 관계성 및 가중치를 추출하는 과정에 많은 계산량이 요구된다. 따라서 이전의 중소규모의 이용자 및 콘텐츠 환경에서 좋은 성능을 보여주었던 추천방법이 변화된 정보환경에서는 많은 한계점을 보여주고 있다. 셋째, 대표적인 추천방법으로써 알려진 협업여과 추천방법의 경우에 있어서 유사 이용자에 대한 성향분석을 통해 추천이 이루어진다. 그러나 협업여과 추천방법의 문제점으로 지적되는 희소성의 문제가 빈번하게 발생할 수 있다.

이러한 문제점에 대한 인식과 함께, 추천서비스를 둘러싸고 있는 새로운 환경적인 변화에 따른 문제점을 해결하고 보다 높은 추천의 정확성과 대규모의 이용자 및 콘텐츠의 추천을 위하여 본 연구에서는 첫째, 이용자의 개인정보 및 콘텐츠 속성을 활용하지 않고 이용자의 콘텐츠에 대한 이용행위를 주요한 고려요소로서 설정하고 있다. 둘째, 연관규칙 추출을 위하여 요구되는 형식으로의 변환을 위하여 변환 및 저장에 최적화 할 수 있는 방법을 제시하였다. 예를 들어 기존의 로그화일에서 항목추출을 위해서는 필요한 항목을 텍스트 파일로 추출하여 저장하였으나 본 연구에서는 필요항목을 추출하고 이를 이진파일 형태 저장된다. 이와 같이 생성된 변환파일은 연관규칙에 대한 분석 결과로 나온 파일들을 실질적인 정보의 추천이 이루어지는 정보추천 프로그램에서 사용할 수 있는 형태로 재변환하는 과정에서 이진형태로 저장된 연관규칙을 해석하는 과정에서 활용된다. 셋째, 이용자의 이용행위를 세분화하여 행위별 가중치를 적용함으로써 실질적인 정보추천단계에서 이용자의 묵시적 행위정보에 따라 추천정보가

결정된다. 즉, 이용자의 이용행위에 있어서 단순읽기, 내려받기, 실행 등으로 세분화하여 중요도에 따른 가중치를 부여하여 추천에 따른 정확도를 높이도록 하였다. 넷째, 연관규칙추출에 있어서 문제점으로 지적되고 있는 연산의 복잡성에 대한 해결을 위하여 임계값 이하의 범위에 포함되는 콘텐츠만을 대상으로 콘텐츠 간의 연관성을 분석하고 분석결과를 기준으로 규칙을 추출함으로써 보다 정확한 규칙추출 및 연관성 계산에서의 복잡성을 제거할 수 있었다. 이용도가 높은 상위 콘텐츠를 제외한 콘텐츠를 연관규칙을 위한 대상 콘텐츠로 적용하는 방법은 정보검색 분야에서 문헌을 대표하는 색인어 추출을 위한 방법론으로서 문헌집단에서 출현빈도가 낮을수록 색인어의 중요도가 높다는 가정에 근거한 역문헌빈도와 유사한 개념이라고 할 수 있다. 즉, 대상이 되는 콘텐츠 간의 높은 연관성으로 인하여 너무 많은 빈발항목집합이 생성되고 이러한 빈발항목집합을 통하여 생성된 규칙은 대부분의 콘텐츠들이 상호간을 서로 추천하는 결과를 초래하게 됨으로써 추천 콘텐츠를 선정하는 과정에서 콘텐츠 간의 변별력이 낮아지므로 규칙의 유효성이 매우 낮아진다. 따라서 이용자의 이용행위에 있어서 공통적으로 이용행위가 자주 일어나는 콘텐츠를 빈발항목집합에서 제외함으로써 계산량을 줄이고 변별력을 높일 수 있는 방법을 제시하고 있다. 이를 보다 자세히 알아보면 이전의 연구에서는 콘텐츠간의 연관성 추출을 위하여 데이터마이닝의 연관성 추출방법을 그대로 활용했으나, 본 연구에서는 효과적인 규칙추출 및 정보추천을 위하여 다음과 같이 변형된 Apriori 알고리즘을 적용하였다. 예를 들어, 항목 A와 B가 있다고 가정하고 A와 B에 대한 연관규칙에 대해, 왼쪽 항목집합 A를 선행부라 하고, B를 결과부라 가정하고 추천 처리를 단순화하기 위하여 다음과 같은 제한을 둔다. 첫째, 선행부와 결과부에는 각각 하나의 항목집합만을 가지는 규칙이 생성된다. 예를 들어 (A) → (Z), (B) → (Y)와 같은 규칙이 추출되며 (A) → (Z, Y), (B, A) → (Z)와 같은 규칙은 허용되지 않는다. 둘째, 규칙을 추출하는데 있어서 향상도가 '1' 이상인 것만 추출한다. 이를 통하여 통계적으로 의미가 있는 규칙만을 추출할 수 있다. 셋째, 추출된 연관규칙 중에서 하나의 동일한 선행부에 대응되는 결과부의 경우의 수를 제한한다. 여기서 결과부에 나타날 수 있는 항목에 대한 경우의 수를 제한하는데 예를 들어 제한수가 '3'인 경우에 있어서 동일한 선행부에 대하여 '4'개 이상의 규칙이 추출될 수 없다. 이렇게 결과부에 나타나는 규칙의 수를 제한하는 이유는 연관규칙이 상위 지지도를 가지는 콘텐츠들 사이에서만 집중되는 경향이 있으므로 특정 콘텐츠에 규칙이 집중되는 문제점을 방지함으로써 규칙추출에 따른 유효성을 확보하기 위한 것이다.

이와 같이 본 연구에서 제안하고 있는 방법이 기존의 추천방법과는 차별화 될 수 있는 특징을 정리하면 다음과 같다. 첫째, 대용량의 멀티미디어 콘텐츠 및 이용자를 대상으로 추천서비스를 제공한다. 둘째, 이용자 정보 및 콘텐츠 속성을 활용하지 않고 이용자의 콘텐츠에 대한 이용행위를 고려요소로 설정하였다. 셋째, 연관규칙 추출을 위하여 콘텐츠의 이용에 따른 가중치를 적용하여 특정 임계값 이상에 포함되는 일부 콘텐츠를 제거하고 나머지 콘텐츠만을 대상으로 활용하였다. 즉, Apriori 알고리즘의 변형된 형

태로써 빈발항목집합의 추출과정에서 많은 이용자에 의하여 공통적으로 이용되는 콘텐츠를 제거하여 빈발항목집합을 구성함으로써 연관규칙에 있어서 문제점으로 지적되고 있는 계산량을 줄이면서 동시에 추천의 정확성을 확보할 수 있다.

결론적으로 기존의 추천방법들이 단지 이용자 개인의 선호도 정보 또는 유사이용자의 선호도 정보를 기반으로 추천에 따른 희소성의 문제를 간과한 것과는 달리 이용자의 묵시적 행위정보를 기반으로 변형된 Apriori 알고리즘을 기반으로 연관규칙을 추출하면서 연관규칙의 대표적 단점으로 지적되고 있는 계산의 복잡성을 해결함으로써 대용량의 정보환경에서 보다 정확한 추천서비스를 제공할 수 있다.

5. 결 론

본 연구에서는 도서관 등의 정보서비스 제공기관에서의 정보추천을 위하여 대용량의 정보환경에서 적합한 연관규칙을 기반으로 하는 연관규칙 생성 및 정보추천 시스템의 설계 및 구현을 수행하였다. 이를 위하여 이용자에 의하여 입력되는 명시적 정보가 아닌 이용자의 정보이용행위를 실질적으로 보여주는 묵시적 정보를 포함하고 있는 이용자의 로그파일을 통하여 이용자의 묵시적 정보를 기준으로 연관규칙의 생성 및 정보추천 시스템을 구현하였다. 또한, 연관규칙 생성과정에서 항목 수의 증가에 따른 연산량을 최소화하면서 정확도 높은 연관규칙의 생성을 통하여 대용량의 정보환경에서 정보추천의 효과를 극대화 할 수 있는 방안을 적용하였다. 현재 도서관 등에서 제공되고 있는 SDI 또는 맞춤정보서비스는 이용자에 의하여 입력된 키워드에 대한 매칭을 통하여 정보추천이 이루어지기 때문에 매우 단순하고 낮은 정확도를 보여주고 있다. 그러나 본 연구에서는 이용자의 만족도를 높이기 위하여 이용자에 대한 묵시적 행위정보를 기반으로 이용자의 정보이용행위와 관련된 다양한 정보를 정밀하게 가공 및 적용하고 있다.

본 연구에서는 개인화된 추천서비스제공을 위한 시스템 개발을 위한 알고리즘 및 적용 가능한 기술에 대하여 알아보고 개인화 추천서비스 시스템의 구성요소 및 시스템의 설계와 함께, 개인화 추천을 위한 이용자의 성향을 분석하여 규칙을 생성하는 연관규칙 생성시스템의 설계 및 구현을 수행하였다. 결론적으로 제안된 추천방법을 통하여 대용량의 콘텐츠 및 이용자 환경에서의 추천의 정확성 및 추천 콘텐츠 추출에 따른 연산시간에 있어서의 높은 효율성을 확보할 수 있으며, 최근의 웹 환경의 변화에 따른 추천서비스의 정확성과 효율성을 위하여 고려해야 하는 요구사항을 발견할 수 있었다. 이를 통하여 다음과 같은 결론을 얻을 수 있었다. 첫째, 대용량 이용자 및 콘텐츠 환경에서 개인화 추천서비스의 제공을 위해서는 추천의 정확성과 추천 콘텐츠 선정을 위한 연산시간에 대한 고려가 동시에 이루어져야 한다. 둘째, 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. 셋째, 도서관의 정보추천서비스 제공은 웹 2.0 시대를 넘어서 웹 3.0

시대에 있어서 개인화된 정보서비스 제공을 위한 중요한 방법으로써 지속적인 연구가 요구되는 분야라고 할 수 있다.

향후 연구과제로서 본 연구에서 제안하고 있는 알고리즘 및 연관규칙 추출프로그램을 디지털도서관 시스템 등에 실질적인 적용을 통하여 이용자 행위정보의 추출 및 정보서비스의 최종의 목표인 개인화서비스 제공을 위한 시스템의 지속적인 보완과정이 필요하다고 할 수 있다. 특히, 본 연구에서는 제안된 방법을 기반으로 추천시스템을 구현하였으나 다른 추천시스템과의 실질적인 비교평가를 수행하지 못하였다. 따라서 다른 추천시스템과의 상대적인 비교평가를 통하여 제안하고 있는 방법에 대한 차별성을 증명하여야 할 것이다. 정책적인 측면에서의 추가 연구로써는 실질적인 추천서비스를 제공하기 위해서는 기술적인 방법론 이외에 웹 서비스에 대한 적용방안에 대한 정책적인 고려가 요구된다. 웹 2.0 및 3.0의 도래와 함께, 쌍방향적 정보흐름은 향후 정보서비스의 목표와 방법에 대한 중요한 시사점이 될 수 있다. 정보추천서비스는 정보이용자와 정보제공자 간의 상호작용을 통하여 효과가 극대화 될 수 있으며 궁극적으로 개인화된 정보서비스의 중요한 분야가 될 수 있다. 따라서 지속적인 연구와 적용을 통하여 이용자의 정보요구에 대한 적극적 서비스로써 개인화 정보추천서비스를 지속적으로 발전시켜야 할 것이다.

참 고 문 헌

김미성, 김남규, 안재현. 2012. 연관규칙 마이닝에서의 동시성 기준 확장에 대한 연구. 『지능정보연구』, 18(3): 23-38.

김용, 김문석, 김윤범, 박재홍. 2009. 이용자 이용행위 및 콘텐츠 위치정보에 기반한 개인화 추천방법에 관한 연구. 『정보관리학회지』, 26(1): 81-105.

김용, 문성빈. 2006. 멀티미디어 콘텐츠를 위한 이용빈도에 기반한 하이브리드 추천시스템에 관한 연구. 『정보관리학회지』, 23(3): 91-127.

김진수 외. 2004. 이용자 로그분석과 클러스터 내의 문서 유사도를 이용한 동적 추천시스템. 『한국정보과학회논문지: 소프트웨어 및 응용』, 31(5): 586-594.

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

오재영, 전종훈. 2004. 전자상거래에서 연관규칙을 이용한 추천 시스템의 설계 및 구현. 『한국정보과학회 2004년도 봄 학술발표논문집』, 31(1): 121-123.

윤종찬, 윤성대. 2007. 스쿼스 연관규칙을 이용한 개인화 웹 마이닝 설계. 『멀티미디어학회 논문지』, 10(9): 1106-1116.

윤흥준, 노준호, 김재광, 이병정, 강수용, 장재영. 2010. 개념 네트워크 기반 개인화 검색 시스템의 설계 및 구현. 『한국인터넷 정보학회 학술발표대회 논문집』, 147-152.

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

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

한현수, 임동수. 2008. U-commerce에서 개인화가 미치는 연향에 대한 연구.『지식정보산업연합회 창립기념 학술대회 논문집』, 183-192.

Adomavicius, G. and A. Tuzhilin. 2005. "Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-art and Possible Extensions." *IEEE Transactions on Knowledge and Data Engineering*, 17, no. 6(2005): 734-749.

Ahn, H. J. 2008. "A New Similarity Measure for Collaborative Filtering to Alleviate the New User Cold-starting Problem." *Information Sciences*, 178, no. 1(2008): 37-51.

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

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.

Hill, W., L. Stead, M. Rosenstein, and G. Furnas. 1995. "Recommending and Evaluating Choices in a Virtual Community of Use." *Proceedings of CHI '95 Conference on Human Factors in Computing Systems*, 194-201.

IDC. 2012. The 2011 Digital Universe Study: Extracting Value from Chaos. Hopkinton, Mass: EMC [online]. [cited 2012.9.20]. 〈http://www.emc.com/leadership/programs/digital-universe.htm〉.

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

Linden, G, B. Smith and J. York. 2003. "Amazon.com Recommendations: Item-to-item Collaborative Filtering." *IEEE Internet Computing*, 7(1): 76-80.

Rensnick, P. et al. 1994. "GroupLens: An Open Architecture for Collaborative Filtering of Netnews." *Proceedings of the 1994 Computer Supported Cooperative Work conference*, 175-186.

Sarwar, B., G. Karypis, J. Konstan and J. Riedl. 2001. "Item-based Collaborative Filtering Recommendation Algorithm." *Proceedings of WWW 10*, 285-295.

Shahabi, Cyrus and Yi-shin Chen. 2003. "Web information personalization: Challenges and Approaches." *Proceedings of 3rd Workshop on Databases in Networked*

Information System, 5-15.

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

Sheng et al. 2006. "An experiemental study on U-commerce Adoption: Impact of Personalizationand Privacy Concern." *Proceedings of the fifth Annual Workshop on HCI Research in MIS*, Milwaukee, WI.

Wang, Dexing et al. 2005. "Association Rules Mining on concept Lattice using Domain Knolwedge." *Proceedings of the First International Conference on Machine Learning and Cybermetrics*, 2152-2154.

Wu, Y. H. and A. Chen. 2000. "Index Structures of User Profiles for Efficient Web Page Filtering Services." *Proceedings of IEEE Conference on Distributed Computing Systems*, 644-651.