사례기반 추론을 이용한 인터넷 서점의 서적 추천시스템 개발

Development of a Book Recommender System for Internet Bookstore using Case-based Reasoning

이재식(Jae Sik Lee)*, 명훈식(Hun Sik Myoung)**

초 록

오늘날 인터넷의 전반적인 보급 및 전자상거래의 확산으로 인하여 정보의 홍수를 이루게 되었고, 고객들은 자신이 원하는 제품이나 서비스를 선택하기 위해서 정보를 탐색하는 작업이 더욱 어려워지게 되었다. 이러한 고객들에게 좀 더 편리하게 자신이 원하는 제품이나 서비스를 선택하도록 도와주는 것이 추천 시스템으로서, 고객 관계 관리의 중요한 부분으로 자리 잡게 되었다. 본 연구에서는, 인터넷 서점을 이용하는 고객에게 그가 관심을 가질만한 서적을 추천하여 줌으로써 구입할 서적의 선택을 도와주는 서적 추천 시스템을 개발하였다. 기존의 서적 추천 시스템 개발에 협업 필터링 기법이 주로 활용되어 왔다. 하지만 협업 필터링 기법을 적용하기 위해서는 각 서적에 대한 구매자들의 평가치가 수집되어야 하는데, 이러한 평가치들은 시스템 개발 이전에 오랜 기간에 걸쳐 정교한 계획 하에서 수집되어야 한다. 더욱이 구매자들이 평가치 제공에 협조하지 않을 경우에는 추천 시스템 자체의 작동이 불가능하게 된다. 그러므로 본 연구에서는 고객들의 구매기록만으로 서적 추천을 수행할 수 있도록 사례기반추론 기법을 활용하여 시스템을 개발하였는데, 서적의 소분류 코드를 예측하는 상황에서 약 40% 수준의 적중률을 보였다.

ABSTRACT

As volumes of electronic commerce increase rapidly, customers are faced with information overload, and it becomes difficult for them to find necessary information and select what they need. In this situation, recommender systems can help the customers search and select the products and services they need more conveniently. These days, the recommender systems play important roles in customer relationship management. In this research, we develop a recommender system that recommends the books to the customers of Internet bookstore. In previous researches on recommender systems, collaborative filtering technique has been often employed. For the collaborative filtering technique to be used, the rating scores on books given by previous purchasers have to be collected. However, the collection of rating scores is not an easy task in reality. Therefore, in this research, we employed case-based reasoning technique that can work only with the book purchase history of customers. The accuracy of recommendation of the resulting book recommender system was about 40% on the level 3 classification code.

키워드 : 서적 추천시스템, 사례기반 추론, 고객 관계 관리
Book Recommender Systems, Case-based Reasoning, Customer Relationship Management

* 교신저자, 아주대학교 경영대학 e-비즈니스학부 교수
** (주) 애버커스 솔루션 사업본부 기술기획팀
2008년 10월 15일 접수, 2008년 10월 31일 심사완료 후 2008년 11월 07일 게재확정.

1. 서 론

인터넷 이용의 보편화는 전자상거래의 급속한 발전을 가져왔고, 이는 기업과 고객 모두에게 행태의 변화를 요구하고 있다. 기업은 실제 수익을 주는 고객과 충성도가 높은 고객을 식별하여 다른 고객과는 차별화된 서비스를 제공해야 한다. 한편, 고객의 입장에서는 기업에 대한 풍부한 정보를 바탕으로 좀 더 합리적인 구매 의사결정을 내릴 수 있게 되었지만, 반면에 엄청난 양의 정보의 홍수 속에서 중요한 정보를 식별하기 어렵게 되는 딜레마에 빠지게 되었다.

추천시스템은 이러한 환경적인 변화에 대응하기 위한 하나의 도구이다. 고객에 대해서 수집된 다량의 정보를 분석하여 고객에 대한 지식을 추출하고, 이런 지식을 바탕으로 고객이 필요로 하는 정보만을 제공하거나, 고객이 좋아할 만한 아이템만을 추천함으로써 의사결정의 짐을 덜어 줄 수 있는 시스템이다. 여기서 아이템이란 구매 대상이 되는 제품이나 서비스를 지칭한다. 추천 시스템은 개인화 서비스의 일종으로 간주할 수 있다. 개인화 서비스란 고객들이 어떠한 제품이나 서비스를 필요로 하는지를 명시적으로 묻지 않고 관련 정보를 제공하는 것을 뜻하는데[13], 인터넷 서비스 제공자들의 중요한 성공요인으로 인식되고 있다[3]. 근래에 와서 이러한 추천 시스템은 고객관계관리(CRM : Customer Relationship Management)와 맞물려 운영되어 진다. 즉 각 고객에 대한 맞춤 서비스의 차원에서 고객과의 관계 강화를 추구하며 1 : 1 마케팅의 수단으로 활용되고 있다.

추천시스템은 이미 Amazon이나 CD Now 등 인터넷 쇼핑몰에서 많이 사용되고 있다. 추천시스템은 다양한 기법을 통해 구현될 수 있는데 최근 전자상거래 분야에서 쓰이는 기법 중에서 대표적인 것이 협업 필터링(Collaborative Filtering)이다. 협업 필터링은 고객들이 아이템들을 경험하고 나서 부여한 평가치들을 기반으로 목표고객이 평가하지 않은 아이템들 중에서 높게 평가할 것이라고 예상되는 아이템을 추천하는 기법이다[11, 14]. 즉, 대상이 되는 수많은 아이템들에 대해서 기존의 많은 고객들이 경험한 후에 부여한 평가치들이 존재하여야만 작동이 되는 기법이다. 하지만, 현실적으로 고객들로부터 평가치를 수집하는 것이 쉽지 않을 뿐만 아니라, 수집된 평가치들의 진정성도 확신할 수 없다. 그러므로 협업 필터링에 대해서는 희박성(Sparsity), 확장성(Scalability) 그리고 투명성(Transparency) 등의 한계점이 지적되고 있다[2, 12, 15, 18].

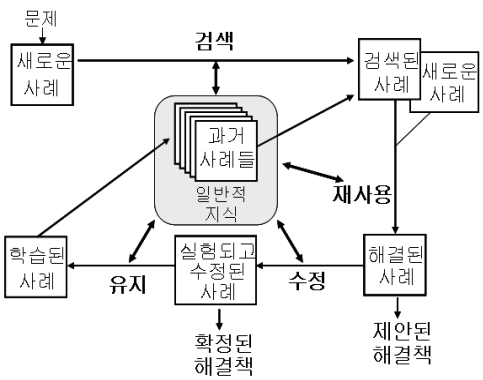
본 연구에서는 이러한 한계점들로부터 자유로운 기법인 사례기반추론을 이용하여 서적 추천시스템을 개발하고자 한다. 본 연구에서 개발하는 추천 시스템에서는 아이템의 평가치가 필요하지 않으며, 단지 고객별 구매기록만 있으면 작동이 가능한 시스템이다. 우리는 이 시스템을 CbBR(Case-based Book Recommender) 시스템으로 명명하였다. 본 논문은 다음과 같이 구성되었다. 제 2장에서는 본 연구에서 사용한 기법인 사례기반추론에 대해서 간략하게 기술한다. 제 3장에서는 본 연구에서 사용한 데이터인 서적 구매 데이터에 대한 설명 및 준비 과정을 설명한다. 제 4장에서는 CbBR 시스템의 핵심 기능인

입력 사례의 생성 및 유사 사례의 검색 방법에 대해서 기술한다. 제 5장에서는 CbBR의 파라메터들을 변화시키면서 성능을 측정하고, 제 6장에서 한계점의 기술과 함께 결론을 맺는다.

2. 사례기반추론

사례기반추론(CBR : Case-Based Reasoning)은 분류 및 예측 문제 모두에 효과적으로 적용 가능한 기계학습(Machine Learning) 기법이다. CBR은 Exemplar-based Reasoning, Instance-based Reasoning, Memory-based Reasoning, Analogy-based Reasoning등 다양한 용어로 사용되지만, 그 기본 개념은 유사하다[4]. CBR은 두 개의 기본 사상에 기반 하는데 하나는 유사한 문제는 유사한 해법을 가진다는 것이고, 다른 하나는 한번 발생한 문제는 자주 발생할 수 있다는 것이다. 따라서 과거에 현재의 문제와 유사한 문제가 존재하였고 그것이 어떻게 해결됐는지를 안다면, 과거의 경험을 바탕으로 현재 문제의 해결책을 추론할 수 있다는 것이다. 새로운 문제 해결을 위해 과거 사례의 해결책을 재사용 한다는 이러한 특성은 CBR이 다른 기계학습 기법들과 구별되는 접근 방식이라고 할 수 있다. CBR의 문제 해결 방식은 인간의 문제 해결 방식과 유사하기 때문에 그 결과를 이해하기 쉽고, 새로운 사례를 단순히 저장하는 것만으로도 추가적인 작업 없이 학습이 진행된다는 장점을 가진다. CBR은 다양한 현실 문제 해결에 적용되고 있으며, 고장 진단[6, 16, 17], 헬프 데스크[5, 7], 전략 수립[4] 등은 성공적으로 CBR이 적용되었던 응용 영역이다. 최근에 CBR은 유비쿼터스 컴퓨팅 시스템의 상황인식 기능[9, 10] 및 개인화 서비스의 구현[8]에도 활용되고 있다.

Aamodt와 Plaza[1]는 CBR의 문제 해결 과정을 <그림 1>과 같이 크게 검색, 재사용, 수정, 유지의 4 단계로 구분하였다.



〈그림 1〉 사례 기반 추론의 과정

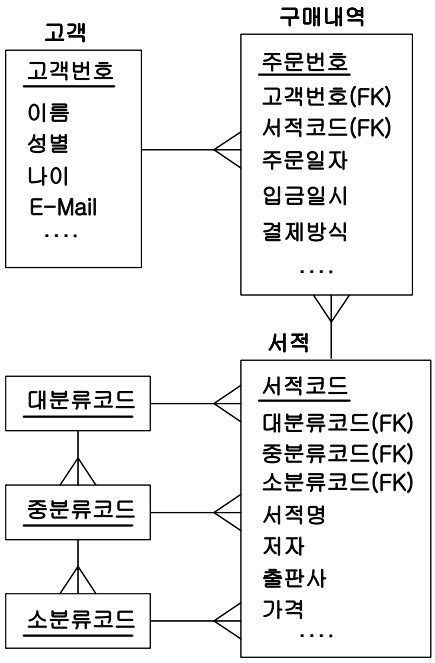
1) 검색(Retrieve) : 검색은 현재 문제와 가장 유사한 과거 사례들을 사례 베이스로부터 찾아내는 것이다.
2) 재사용(Reuse) : 재사용은 검색을 통해 찾아진 유사 사례들의 해법을 현재 문제 해결을 위해 사용하는 것이다.
3) 수정(Revise) : 수정은 현재 문제의 해결을 위해 검색된 유사 사례들의 해법을 현재 문제에 적합한 형태로 조정 하는 것이다.
4) 유지(Retain) : 유지는 새롭게 해결된 문제와 해법을 새로운 문제 해결을 위한 목적으로 사례베이스에 저장하는 것이다.

사례베이스로부터 유사 사례를 찾기 위한 대표적인 검색 방법으로는 귀납적 검색(Inductive Retrieval)과 최근접 이웃 검색(Nearest Neighbor Retrieval)이 있다. 귀납적 검색은 사례를 가장 잘 구분시켜주는 속성들을 찾아서 이 속성들을 사용하여 유사 사례를 검색 하는 방법이다. 귀납적 검색은 사례의 검색 및 구성을 위해 의사결정나무 형태의 구조를 사용한다. 최근접 이웃 검색은 현재 문제의 유사 사례 검색을 위해 현재 문제를 표현하는 사례와 사례 베이스에 있는 모든 사례와의 유사도를 측정한 후에, 유사도가 높은 일정 개수의 사례들을 검색하여, 유사 사례를 찾는 방법이다. 최근접 이웃 검색에서 일반적으로 가장 많이 사용되는 방법은 현재 사례와 가장 유사한 k개의 사례를 검색해 주는 k-NN(k Nearest Neighbors) 방법이다. 본 연구에서는 k-NN 방법을 사용하여 유사 사례를 검색하였는데, 자세한 방법은 제 4.4절에서 설명한다.

3. 사례 베이스의 구축

3.1 사용 데이터

본 연구에 사용된 데이터는 국내 인터넷 서점인 A사의 5개월 간의 구매내역 데이터이다. 본 데이터는 관계형 데이터베이스 구조로 되어 있으며 크게 고객, 구매내역, 그리고 서적의 세 개의 테이블로 구성되어 있고 각각의 서적을 분류한 대·중·소분류 테이블을 가지고 있다. 데이터의 개체-관계도(ERD : Entity-Relationship Diagram)는 <그림 2>와 같다.



〈그림 2〉 사용 데이터의 개체 관계도

<그림 2>와 같은 구조의 최초 원시 데이터는 고객 테이블 224,355건, 서적 테이블 257,196건, 그리고 구매내역 테이블 119,952건의 레코드로 구성되어 있었으나 본 연구에서는 이를 정제하여 48,193건의 데이터를 사용하였다. 정제과정에서 구매회수가 한 번인 경우의 데이터는 제거 되었다. 그 이유는 본 연구에서 추천 시스템의 적중률을 파악하기 위해 시점을 과거로 돌려 측정하기 때문이다. 예를 들어 두 번의 구매가 있는 경우에는 한 번의 구매가 있는 것으로 간주하여 첫 번째 구매에 관한 정보를 사용하여 두 번째 구매할 것이라고 예상되는 책을 추천하게 되고 이를 실제 고객이 구매한 서적과 비교하여 적중률을 계산하게 된다. 따라서 구매회수가

한 번인 고객의 데이터는 추천의 결과가 맞는지를 확인하는데 필요한 두 번째 서적에 관한 정보가 없으므로 제거되었다.

고객 테이블은 고객의 이름, 성별, 나이, E-Mail, 주소 등의 인구통계학적 정보를 저장하고 있다. 구매내역 테이블은 주문일자, 입금일시, 결제방식과 같은 서적 구매에 대한 일반적인 정보를 저장하고 있는데, 고객별로 한 번에서부터 수십 번에 이르기까지 다양한 구매회수를 가지고 있다. 각 서적은 기본적인 정보 이외에 대・중・소 세 가지 분류코드를 가지고 있다. 대분류 코드는 전체 책을 주제별로 20가지로 분류하여 각 분류마다 코드번호를 부여한 것이다. 중분류 코드는 대분류 코드를 다시 193개로 세분한 것이며, 소분류 코드는 중분류 코드를 다시 337개로 세분한 것이다. 이 분류 코드는 해당 서적이 어떤 주제에 속하는지를 알게 해준다.

3.2 사례 베이스의 구축

<표 1>은 구매회수별 사례 발생 건수를 표시한 것이다.

<표 1>에서 보는 바와 같이 구매회수가 2회에서 4회까지의 경우가 전체의 76.8%로서 전체 구매의 약 3/4을 차지한다. 각 구매회수별 사례 발생건수 비율을 보면 두 번, 세 번, 네 번 구매한 경우가 각각 39.3%, 23.1%, 14.4%로 높은 반면 5회 이상 구매한 경우는 그 비율이 매우 낮다.

구매실적의 발생은 고객-구매서적의 쌍으로 발생하는데, 이러한 데이터들을 고객별로 취합하여야 한다. 예를 들어 <표 2>에서 보듯이, 고객 '가'는 A, B, C 세 가지 서적을 구매했고, 고객 '나'는 B, A 서적을 구매했으며, 그리고 고객 '다'와 '라'는 각각 서적 D와 E를 구매했다고 하자. 구매순서는 데이터베이스에서 먼저 등장할수록 최근의 구매한 서적이며 나중에 등장할수록 구매한지 오래된 서적이 된다.

이러한 구조로 된 원래의 데이터베이스를 <표 3>과 같은 형태로 변환시킨다. 실제로 고객 '가'는 서적 A, B, C를 동시에 구매했을 수도 있다. 하지만, 데이터베이스에는 한

〈표 1〉 구매회수별 사례 발생 건수

구매 회수	발생건수	비 율(%)	누적 발생건수	누적 비율(%)
2회	18,952	39.3	18,952	39.3
3회	11,111	23.1	30,063	62.4
4회	6,941	14.4	37,004	76.8
5회	3,713	7.7	40,717	84.5
6회	2,497	5.2	43,214	89.7
7회	1,530	3.2	44,744	92.8
8회	886	1.8	45,630	94.7
9회	619	1.3	46,249	96.0
10회	479	1.0	46,728	97.0
10회 이상	1,465	3.0	48,193	100.0

〈표 2〉 서적 구매 데이터

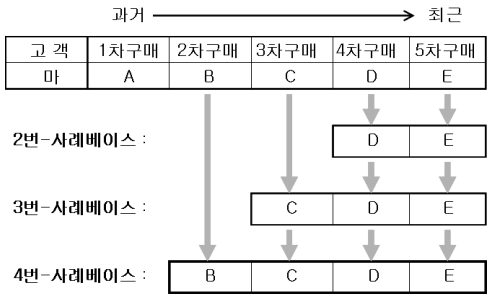
고 객	구매서적
가	C
가	B
가	A
나	A
나	B
다	D
라	E

〈표 3〉 변형된 서적 구매 데이터

고 객	1차 구매서적	2차 구매서적	3차 구매서적
가	A	B	C
나	B	A	Null
다	D	Null	Null
라	E	Null	Null

명의 고객에 대해서 하나의 서적씩 기록되기 때문에 <표 2>와 같이 기록되어 있는 것이다. 기록되는 순서는 인터넷 서점의 구매화면에서, 고객이 서적을 장바구니에 담는 순서이다. 이 기록을 우리는 <표 3>과 같이 변형함으로써, 고객 '가'가 서적 A, B, C를 마치 시간 차이를 두고 구매한 것처럼 취급하고 있다. 하지만, 서적 A, B, C가 동시에 구매되었다는 사실이 본 연구의 본질을 흐리지는 않는다. 고객 '가'가 서적 A, B를 구매하였다면, 그가 동시이건 시간 차이를 두던 간에 서적 C를 구매할 것이라고 추천하는 것이 본 연구의 목적이기 때문이다.

<표 1>의 설명에서 언급하였듯이, 2회, 3회, 4회의 구매실적이 가장 많이 발생하였다. 그러므로 본 연구에서는 5회 이상의 구매실적도 2회, 3회, 4회 구매한 형태로 사례를 변형하여 각각 2번-사례베이스, 3번-사례베이스, 4번-사례베이스를 만들었다. <그림 3>은 사례베이스를 구성하는 예를 보여준다.



〈그림 3〉 사례의 구성 방법

<그림 3>에서 보는 바와 같이 총 다섯 번의 구매가 있는 고객이 있다고 가정을 하면 가장 최근 구매한 것과 그 바로 이전 구매를 결합하여 이것을 저장하는 2번-사례베이스를 만들기 위한 사례를 만들고, 다시 여기에 바로 전 구매 하나를 더 추가하여 최근 시점으로부터 세 번의 구매를 저장하는 3번-사례베이스를 만들기 위한 사례를 만들고, 마지막으로 3번-사례베이스에 바로 전 구매 하나를 더 추가하여 최근 시점으로부터 네 번의 구매를 저장하는 4번-사례베이스를 만들기 위한 사례를 만든다. 이러한 과정을 거쳐서, 2번-사례베이스용으로 11,920개의 사례, 3번-사례베이스용으로 2,448개의 사례, 4번-사례베이스용으로 913개의 사례를 만들었다. 이 사례들 중에서 2번-사례베이스용과 3번-사례베이스용으로는 1,500개의 사례를 무작위로 추출하여 사용하였고, 4번-사례베이스용으로는 900개의 사례를 무작위로 추출하여 사용하였다. 생성된 사례를 전부 사용하지 않고, 일부만을 추출하여 사용한 이유는, 사

례기반 추론 기법이 협업필터링과는 달리 단지 고객별 구매기록만 있으면 작동이 가능한 시스템임을 보이는데 본 연구의 목적이 있기 때문이다. 생성된 사례를 전부 사용하여도 문제는 없으나, 일부 사례만을 사용하여도 본 연구의 목적을 충분히 달성할 수 있기 때문에 적은 개수의 사례로 시스템을 구축한 것이다. <표 4>는 각각의 사례베이스용으로 생성된 사례의 총 개수와, 실제로 시스템 구축에 사용된 사례의 총 개수를 나타낸다.

〈표 4〉 생성된 사례 개수와 시스템 구축에 사용된 사례 개수

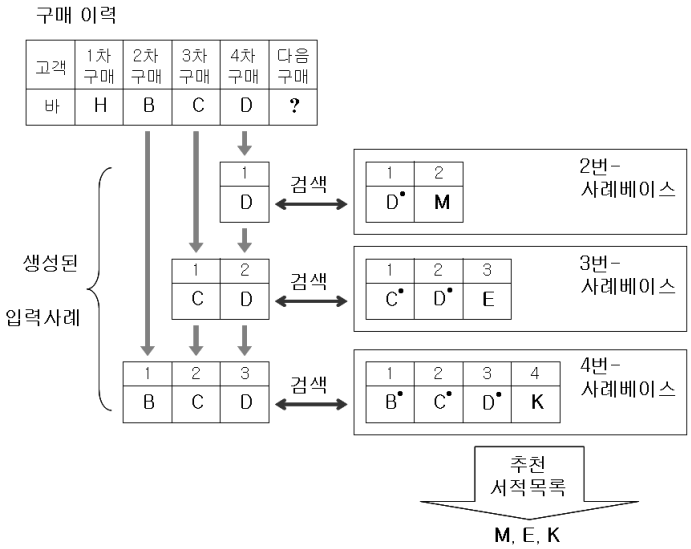
사례베이스 번호	생성된 사례의 개수	시스템 구축에 사용된 사례의 개수
2번	11,920	1,500
3번	2,448	1,500
4번	913	900

사례베이스를 하나로 만들지 않고, 이렇게 두 번, 세 번, 네 번 구매한 경우로 구분하여 만든 이유는, 구매회수별로 독립적으로 사례베이스를 가짐으로써 시스템의 속도 향상을 기대할 수 있기 때문이다. 즉, 구매회수가 두 번인 경우로 변형된 경우에는 2번 사례베이스만을 검색하고, 다른 사례베이스(3번-사례베이스, 4번-사례베이스)를 검색할 필요가 없기 때문에 모든 사례들을 하나의 사례베이스에 저장하여 사용하는 것보다 검색해야 할 사례의 개수가 줄어든다.

4. 입력 사례의 생성 및 유사 사례의 검색

4.1 입력 사례의 준비

고객의 차기 구매 서적 추천을 위해 고객의 구매 이력 데이터는 준비과정을 통해 사



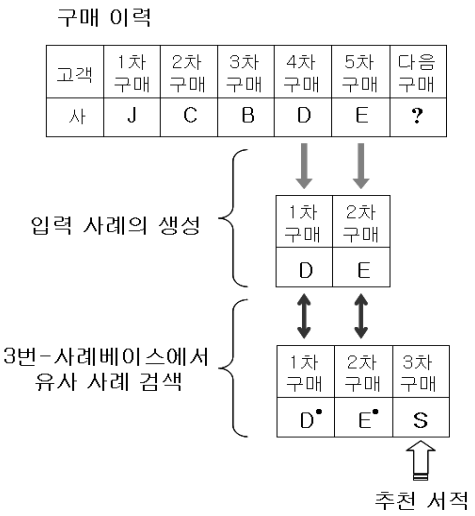
〈그림 4〉 입력 사례의 준비

례베이스와 비교할 수 있는 형태로 변형되어야 한다. <그림 4>는 네 번의 구매 경험이 있는 고객의 차기 서적 추천을 위한 준비 과정을 보여주고 있다.

<그림 4>에서 보는 바와 같이 차기 구매 서적을 추천 받기 원하는 고객의 구매 이력 데이터도, 사례베이스와 마찬가지로, 최근 구매를 시점으로 재구성된다. 즉, 2번-사례베이스와의 비교를 위해 고객의 4차 구매 서적 'D'를 가지고 와서 입력 사례를 생성하고, 3번-사례베이스와의 비교를 위해 고객의 3차 구매 서적 'C'와 4차 구매 서적 'D'를 가지고 와서 입력 사례를 생성한다. 그리고 마지막으로 고객의 2차 구매 서적 'B', 3차 구매 서적 'C', 4차 구매 서적 'D'를 가지고 와서 4번-사례베이스와의 비교를 위한 입력 사례를 생성한다. <그림 4>에서는, 각 사례베이스별로 2번-사례베이스로부터는 서적 'M'을, 3번-사례베이스로부터는 서적 'E'를, 그리고 4번-사례베이스로부터는 서적 'K'를 추천하였다.

4.2 추천 방식

CbBR 시스템은 고객에게 총 10권의 서적이 포함된 추천 목록을 전달한다. 추천목록은 추천을 받고자 하는 고객의 사례와 유사한 사례들로부터 검색된 서적으로 구성되는데, 추천목록에는 동일한 서적들이 포함될 수 있다. 예를 들어, 다섯 번의 구매 경험이 있는 고객에게 3번-사례베이스를 이용하여 여섯 번째로 구매할 서적을 추천한다고 하자. 이 경우에, 입력 사례의 생성과 사례베이스에서 유사 사례의 검색은 <그림 5>와 같다.

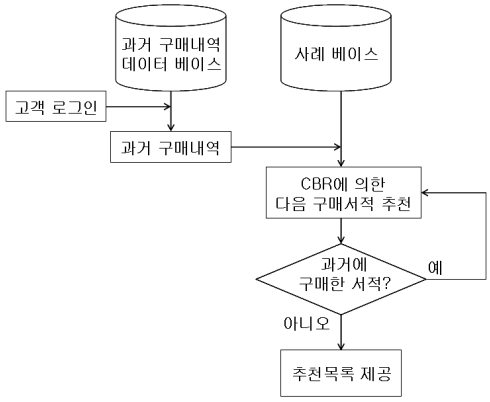


〈그림 5〉 3번-사례베이스를 이용한 추천

<그림 5>에서 보듯이, 그 동안의 구매내역 중에서 가장 최근에 이루어진 4차 구매의 서적 정보와 5차 구매의 서적 정보를 가져와 사례를 구성한다. 이것을 3번-사례베이스의 첫 번째 서적과 두 번째 서적의 속성과 비교하여 유사도를 구한 후, 유사도가 높은 순으로 사례베이스의 사례를 정렬하여 상위 10개 사례의 세 번째 서적을 초기 추천 목록에 포함시킨다. 고객이 이미 구매한 서적을 추천 목록에 포함시키면 안되므로 초기 추천 목록에 포함된 서적에 대해서는 고객의 구매 이력과 비교하여 과거에 구매한 서적인지를 검사한다. 만약 초기 추천 목록에 이전에 구매한 서적이 있다면 그 서적은 추천 목록에서 제거되고, 유사도가 차상위인 사례의 세 번째 서적이 추천 목록에 포함된다. 이러한 과정은 추천 목록에 포함된 서적들 중에 고객이 과거에 구매한 서적이 없을 때까지 반복적으로 수행된다. 추천 목록이 완성되면 총 10권의 서적이 포함된 최종 추천 목록을 고

객의 여섯 번째 구매를 위해 추천하게 된다.

CbBR의 서적 추천 과정을 정리하면 <그림 6>과 같다. <그림 6>에서 보는 바와 같이, 차기 구매 서적을 추천 받기 원하는 사용자가 로그인을 하게 되면 시스템은 사용자의 그 동안의 구매내역을 검색한다. 구매내역을 검색한 후, CbBR은 제 4.1절에서 언급한 바와 같이 사례베이스와 비교하기 위한 입력사례를 구성하고, 본 절에서 언급한 바와 같은 방식으로 유사도 점수가 높은 10개의 사례를 검색하여 사용자에게 10권의 서적을 추천한다.



〈그림 6〉 CbBR의 서적 추천 과정

4.3 입력 속성과 목표값

CbBR 시스템의 입력속성은 <표 5>와 같이 고객에 대한 속성과 서적에 대한 속성으로 구분되는데, 고객에 대한 입력속성으로는 직업코드(job)와 성별코드(sex)가 사용되었으며, 서적에 대한 입력속성으로는 저자(writer), 출판사(publisher), 가격(price), 할인율(discount), 대분류 코드(lev_1), 중분류 코드(lev_2), 소분류 코드(lev_3)가 사용되었다.

<표 5>와 같이 기본적으로 고객 속성 두 개, 서적 속성 7개로 구성되어 있으며 구매 내역이 하나씩 늘어날 때마다 서적 속성은 7개씩 증가하게 된다. 즉 2번-사례베이스는 총 16(= 2 + 7 + 7)개의 속성으로 구성되고, 3번-사례베이스는 총 23(= 2 + 7 + 7 + 7)개의 속성으로 구성 되며, 4번-사례베이스는 총 30(= 2 + 7 + 7 + 7 + 7)개의 속성으로 구성되어 있다. 출력 속성은 <표 5>에서 보는 바와 같이 다음번에 구매할 것이라고 예측되는 서적의 소분류 코드(lev_3)이다.

〈표 5〉 사례의 입력속성과 출력속성

	입력 속성	유 형	개 수	출력 속성	유 형	개 수
고 객	직업코드(job)	범주형	2	차기 구매서적의 소분류코드 (lev_3)	범주형	1
	성별코드(sex)	이산형				
서 적	저자(writer)	문자형	7			
	출판사(publisher)	문자형				
	가격(price)	수치형				
	할인율(discount)	수치형				
	대분류코드(lev_1)	범주형				
	중분류코드(lev_2)	범주형				
	소분류코드(lev_3)	범주형				

4.4 유사도 계산 방법

새로운 입력 사례 N과 사례베이스에 있는 사례 O간의 총 유사도 $S(N, O)$는 식 (1)과 같이 속성 i 별로 유사도 점수인 $f(N_i, O_i)$를 구하고, 각 속성의 가중치를 곱한 후 이를 총합하여 계산한다.

$$S(N, O) = \frac{\sum_{i=1}^{n} f(N_i, O_i) \times W_i}{\sum_{i=1}^{n} W_i} \qquad (1)$$

여기서,

n : 속성의 개수

$f(N_i, O_i)$: 사례 N과 O의 i 속성 간의 유사도 점수

W_i : i 속성의 가중치

사례간의 유사도는 식 (1)에 의하여 0에서 1사이의 실수 값으로 표현되는데, 0에 가까울수록 두 사례의 유사성이 낮다는 것을 의미하고, 1에 가까울수록 유사성이 높다는 것을 의미한다. 속성 간의 유사도 점수인 $f(N_i, O_i)$는 속성의 유형이 수치형이냐 범주형이냐에 따라, 또는 개발자의 전문지식 활용에 따라 달라진다. 가격(price), 할인율(discount)과 같은 수치형 속성일 경우에는 식 (2)를 이용하여 유사도 점수를 계산한다.

$$f(N_i, O_i) = 1 - \frac{a_i - b_i}{\max_i} \qquad (2)$$

여기서,

a_i : N_i의 값

b_i : O_i의 값

$\max_i$: 사례베이스에 있는 i번째 속성의 값들 중 최대값

문자형, 범주형, 이산형 속성의 경우에는 두 속성의 값이 완전히 일치하는 경우에만 유사도 점수 1점을 부여하고 그렇지 않은 경우에는 0점을 부여한다.

제 4.2절에서 언급하였듯이 사례베이스에 있는 모든 사례에 대하여 총 유사도를 구한 다음 총 유사도가 가장 큰 사례부터 내림차순으로 정렬하여 상위 10위 까지의 사례에서 추천하는 서적을 일단 가져온다. 이렇게 만들어진 목록 중에 고객이 이전에 구매한 서적을 제거하고 제거한 만큼의 서적을 총 유사도가 큰 사례에서 추천하는 서적을 통해 다시 보충하는데, 이러한 과정은 총 10권의 서적이 모아질 때까지 반복한다.

4.5 분류코드와 유사도 점수 매트릭스

각 서적에 부여된 세 개의 분류코드(대분류, 중분류, 소분류)는 서적의 특징을 파악하는데 매우 중요한 정보로 사용될 수 있다. 각 분류수준별 코드의 개수는 <표 6>과 같다.

〈표 6〉 각 분류수준별 코드의 개수

분류수준	개 수
대분류	20
중분류	193
소분류	337

분류코드 자체가 서적의 내용 및 주제를 고려한 유사성에 기반 하여 부여된 것이라면 서적간의 유사도를 구하는데 매우 중요한 속

성이 된다. 그러나 본 연구에서 수집한 온라인 서점 A의 구매내역 데이터의 서적 분류 코드는 그와 같은 유사성에 근거하여 부여된 코드가 아니다. 그렇다고 해서 서적의 특징을 주제별로 다양하게 표현하고 있는 범주형의 분류코드를 문자형 변수와 같이 정확하게 일치한 경우에만 점수를 부여한다면 사례기반추론을 적용하는 의미가 없어진다. 따라서 본 연구에서는 대·중·소분류의 코드 값들 상호간 유사성에 대하여 전문가의 의견을 참조하여 2차원 테이블의 형태로 유사도 점수 매트릭스를 작성하였다.

유사도 점수는 최저 0점에서 최고 10점 사이의 정수로 부여하였는데, 분류 코드별 유사도 점수 매트릭스의 예는 <표 7>, <표 8>, <표 9>와 같다.

<표 7>에서 대부분의 유사도 점수가 1점인 것은 대분류의 경우 분류간의 간격이 매

〈표 7〉 대분류 코드 유사도 점수 매트릭스

	코 드	1	2	3	4	5	6	…	20
코드	코드명 / 코드명	가정과 가족	건강과 미용	문 학	인 문	인 물	종 교	…	취미와 실용
1	가정과가족	10	1	1	1	1	1	…	2
2	건강과미용		10	1	1	1	1	…	1
3	문 학			10	7	1	1	…	1
4	인 문				10	1	2	…	1
5	인 물					10	1	…	1
6	종 교						10	…	1
…	…							…	…
20	취미와실용								10

〈표 8〉 중분류 코드 유사도 점수 매트릭스

	코 드	23	24	26	39	42	48	…	54
코드	코드명 / 코드명	결 혼	임신 및 출산	태 교	육 아	가족관계	가정의례	…	홈인테리어
23	결 혼	10	9	8	7	6	5	…	4
24	임신 및 출산		10	9	9	6	3	…	1
26	태 교			10	9	3	2	…	1
39	육 아				10	3	2	…	1
42	가족관계					10	2	…	1
48	가정의례						10	…	1
…	…							…	…
54	홈인테리어								10

〈표 9〉 소분류 코드 유사도 점수 매트릭스

	코 드	136	137	139	140	141	143
코드	코드명 / 코드명	운영체제일반	윈도우	유닉스	리눅스	Mac OS	운영체제기타
136	운영체제일반	10	9	8	7	6	5
137	윈도우		10	8	8	6	3
139	유닉스			10	9	3	2
140	리눅스				10	3	2
141	Mac OS					10	2
143	운영체제기타						10

우 넓기 때문에 상호간 유사도가 떨어지기 때문이다. 따라서 광범위하게 묶여진 대분류를 통해 유사도를 고려하는 것은 어렵다고 할 수 있다. 이에 각 대분류 코드 안에서 다시 중분류 코드 상호간 유사도 점수를 부여하였다. 한 개의 대분류 코드에 한 개의 중분류 코드 유사도 점수 매트릭스가 생성되기 때문에 총 20개의 중분류 코드 유사도 점수 매트릭스가 만들어진다. <표 8>은 대분류 코드 1번(가정과 가족)에 속하는 중분류 코드 유사도 점수 매트릭스이다. <표 9>는 운영체제라는 중분류에 포함된 소분류 코드간의 유사도 점수를 보여주고 있다. 각각의 중분류 코드에 대해 다시 소분류 유사도 점수 매트릭스가 만들어지며, 소분류를 가지고 있지 않은 중분류에 대해서는 중분류 코드를 그대로 상속받아 소분류 코드로 사용하게 되며, 다른 소분류 코드와의 유사도 점수는 0점이 부여된다.

5. CbBR의 성능 측정

5.1 측정 방법

CbBR의 성능을 측정하기 위해서 먼저, 준비된 사례들을 8 : 1 : 1의 비율로 학습용 사례, 테스트용 사례, 그리고 평가용 사례로 구분하였다. <표 10>은 각 번호별 사례베이스의 사례 개수와 측정에 사용된 학습용, 테스트용, 평가용 사례 개수를 나타내고 있다.

학습용은 CbBR의 사례베이스 자체로 사용하였고, 테스트용은 속성 가중치 등 모델의 파라메터들을 조정하기 위한 용도로 사용하였고, CbBR의 최종 성능은 평가용을 사용하여 측정하였다. 제 4.3절에서 언급한 바와 같이 각 사례베이스는 기본적으로 2개의 고객 속성과 한 구매 건수 당 7개씩의 서적 속성으로 구성되어 있으며 구매내역이 늘어남

〈표 10〉 각 번호별 사례베이스의 성능 측정용 구분

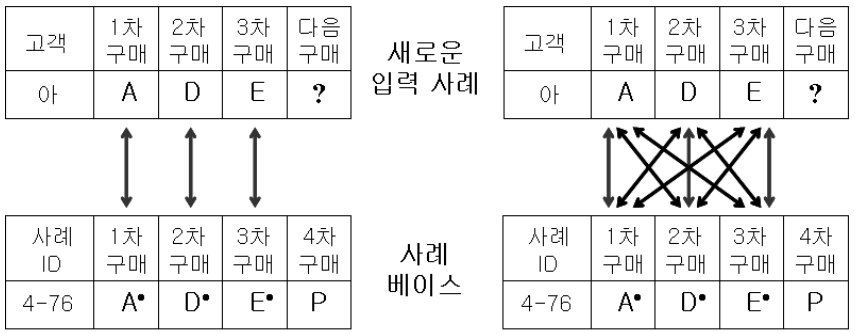
사례베이스 번호	학습용	테스트용	평가용	총 사례 개수
2번	1,200	150	150	1,500
3번	1,200	150	150	1,500
4번	720	90	90	900

에 따라 7개의 서적 속성이 더불어 증가하게 된다.

본 연구에서는 <표 11>과 같이 네 가지 모델을 구축하여 성능을 측정하였다. [모델 1]과 [모델 2]는 서로 비교하기 위한 모델들인데, [모델 2]에서는 유사도를 반영하지 않고 단순한 데이터베이스 검색을 하는 것으로서 속성끼리 완전 매칭이 되는 경우에만 10점의 점수를 받고 그렇지 않은 경우에는 0점을 받는다. 이러한 점수 부여 방식은 수치형, 문자형, 범주형, 이산형 등 모든 속성에 적용된다. [모델 1]의 성능을 [모델 2]의 성능과 비교해 봄으로써, 수치형의 유사도 계산 및 유사도 점수 매트릭스의 효과가 과연 있는지를 측정해 볼 수 있다. [모델 3]과 [모델 4]는 속성들에게 상이한 가중치를 줄 때에 추천 적중률의 향상이 있는 지를 알아보는 것이다. 본 연구에서는 속성 가중치를 0 또는 1로 부여함으로써 속성 선정의 목적으로 사용하였다. 즉, 추천의 적중률을 향상시키는 최적의 가중치를 찾기보다는 사례에 포함된 속성 중에서 적중률에 부정적인 영향을 주는 속성을 제거하기 위함이었다. 만일 어떤 속성에 가중치를 0으로 부여한 경우에 시스템의 추천 적중률이 증가하거나 변화가 없다면, 이 속성은 추천 적중률 향상에 부정적인 영향을 미친다고 판단하고 이를 제거하였다. 이러한 이진(Binary) 가중치들은 한번에 하나의 속성의 가중치만을 변화시키면서 추천 적중률의 변화를 관찰하여 획득하였다. 이를 통해 어떤 속성이 추천 적중률에 긍정적 효과 또는 부정적인 효과를 주는지도 확인하게

〈표 11〉 각 모델의 특성

	속성 가중치	유사도	구매 순서
[모델 1]	동일한 가중치 '1' 부여함	반영함	고려함
[모델 2]	동일한 가중치 '1' 부여함	반영하지 않음	고려함
[모델 3]	상이한 가중치 부여함	반영함	고려함
[모델 4]	상이한 가중치 부여함	반영함	고려하지 않음



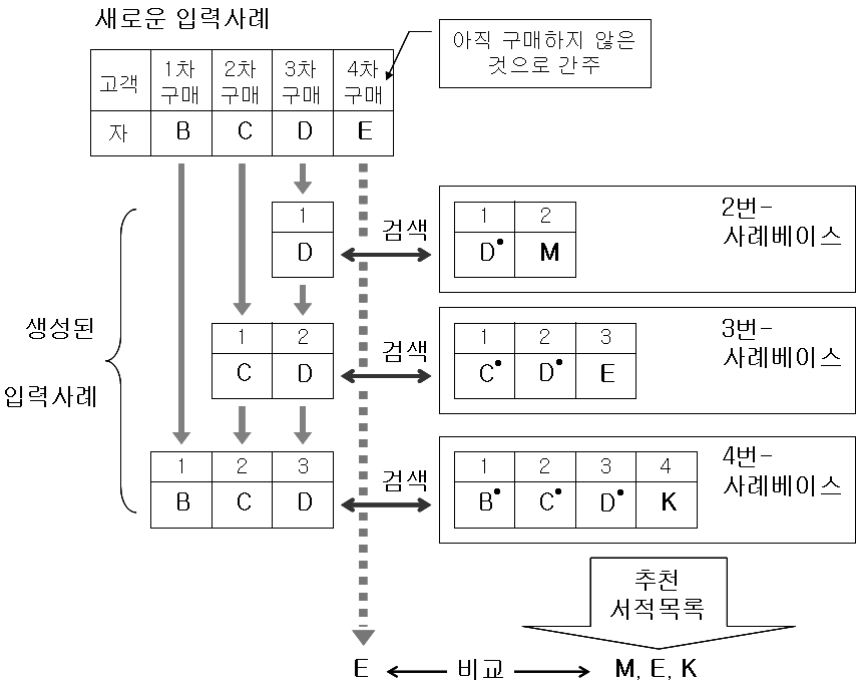
〈그림 7〉 구매 순서를 고려하는 모델과 고려하지 않는 모델

된다.

[모델 1], [모델 2], [모델 3]에서는 구매순서를 고려하였다. [모델 4]는 구매의 순서를 고려하지 않고 성능을 측정 하여 [모델 1], [모델 2], [모델 3]의 성능과 차이가 있는 지를 관찰하는 것이다. <그림 7>은 구매순서를 고려한 경우와 고려하지 않은 경우의 차이를 그림으로 보여주고 있다.

<그림 7>에서 보는 바와 같이 구매순서를 고려하는 경우에는 새로운 서적과 사례베이스 내의 서적을 비교할 때에 구매 순서에 따라 비교하지만, 구매순서를 고려하지 않는 경우에는 하나의 구매 서적과 사례베이스 내에 모든 서적과 교차 비교를 수행한다. 예를 들어 새로운 사례의 1차 구매 서적은 사례베이스의 첫 번째 서적, 두 번째 서적, 그리고 세 번째 서적 모두와 비교가 이루어진다. 마찬가지로 2차 구매서적과 3차 구매서적도 사례베이스의 모든 서적과 비교된다.

일반적으로 추천시스템의 실제 적중률은 추천된 것들 중에서 실제로 구매가 발생했는지 또는 추천시스템으로 인해 매출의 향상이 있었는지에 대해 사후적으로 데이터를 분석한 후에나 측정이 가능하다. 따라서 본 연구에서는 시점을 과거로 돌려 연구 대상이 되는 고객들이 최종에 구입한 서적을 아직 구입하지 않았다고 가정하고, 최종 구매 서적 바로 직전 서적까지만 새로운 사례로 입력하였을때 최종서적에 대한 정보(소분류 코드 : lev_3)를 얼마나 정확히 예측할 수 있는지를 살펴보았다. <그림 8>은 입력 사례의 생성과 추천 방식을 보여 주고 있다.



〈그림 8〉 입력 사례의 생성과 추천방식

추천 적중률은 각각의 사례베이스를 통한 추천이 실제 추천과 일치하는 지를 비교하여 계산하는데, <그림 8>에서 2번-사례베이스는 고객의 4차 구매 서적에 대해 'M'을 추천하였고, 3번-사례베이스는 'E'를 추천하였고, 4번-사례베이스는 'K'를 추천하였다. 이 중에 'E'가 실제 구매와 일치하여 추천이 적중하였다. 적중률의 계산은, 추천이 적중한 경우를 모두 세고 이것을 입력 사례의 총 개수로 나누어 계산한다. 예를 들어, 2번-사례베이스의 평가용 사례를 입력사례로 사용했을 때에 모두 60개의 사례에 대해 소분류 코드가 적중하였다면, 평가용 사례의 총 개수가 150개 이므로 추천 적중률은 40%(= 60/150)가 된다.

5.2 측정 결과

본 연구의 목표는 차기 구매 서적의 소분류 코드를 예측하는 것이지만 대분류 코드 및 중분류 코드의 적중률 측정도 수행하였으며, 평가용 사례에 대한 네 가지 모델의 결과는 <표 12>와 같다.

〈표 12〉 평가용 사례에 대한 적중률(%)

모델 번호	대분류 코드	중분류 코드	소분류 코드
[모델 1]	79.6	56.4	38.9
[모델 2]	79.8	56.6	37.5
[모델 3]	79.2	57.9	39.6
[모델 4]	74.5	52.8	37.0

<표 12>에서 보듯이, 모델의 결과 간에는 미세한 차이만을 보이고 있지만 [모델 3]이 소분류 코드 예측에서 가장 높은 적중률을 보이고 있다. 이는, 속성마다 상이한 가중치를 부여하고, 유사도를 반영하며 구매순서를 고려하여 서적을 추천할 때에 가장 높은 적중률을 보인다는 것을 말해주고 있다.

소분류 코드의 개수는 337개로서, 무작위로 10개를 추천했을 때에 그 중에 실제 구매 서적의 소분류 코드가 포함되어 있을 확률은 10/337로 약 3.0% 정도가 된다. [모델 3]의 소분류 코드 적중률은 <표 12>에서 보듯이 39.6%이다. 즉, CbBR 시스템은 무작위 추천의 적중률보다 무려 13배에 달하는 적중률 증가를 보이고 있는 것이다.

6. 결론 및 향후 과제

본 논문에서는 온라인 서점의 서적 추천 시스템의 개발 기법으로 사례기반 추론을 선정한 후에, 다양한 모델을 개발하여 우수한 적중률을 내는 모델을 찾아내는 연구를 수행하였다. 본 연구에서는 먼저, 한 고객의 구매 이력 레코드로부터 구매 회수별로 복수개의 사례를 생성함으로써 사례 베이스를 풍부하게 확보하였다. 이러한 사례 베이스를 대상으로 하여, 사례기반 추론 기법 적용시 고려하여야 할 사항에 변화를 줌으로써 4개의 모델을 구축하였고, 그들의 적중률을 비교하였다. 이 모델들 중에 가장 우수한 성능을 보인 것은 [모델 3]으로서 소분류 코드 적중률에서 39.6%를 보였다. [모델 3]은 속성마다 상이한 가중치를 부여하고, 유사도를 반영하며 구매순서를 고려하여 서적을 추천하는 모델이다.

본 연구의 한계점은 다음과 같다. 첫째, 모델 구축에 사용될 속성의 선정에 관해 별도의 연구가 없었다는 것이다. 다시 말해서, 원시 데이터 내에 있는 수많은 속성들 중에서 어느 것들을 모델구축에 사용할 것인가의 결정을 할 때에, 통계적 또는 실험적 분석을 수행하지 않고, 본 저자의 판단에 따라 사용할 속성을 선정하였다. 본 연구에서 구현한 일부의 모델에서 속성의 가중치를 0 또는 1로 부여하며 속성 선정 과정을 부분적으로 수행하기는 하였지만, 제한적이었다. 분석적 기법을 도입하여 속성 선정을 수행한다면, 적중률의 향상을 기대할 수도 있을 것이다. 둘째, 성능 측정을 통해 여러 가지 모델간의 추천 적중률의 차이를 기술 하였으나, 적중률 간의 차이가 통계학적으로 의미가 있는지에 관한 검증을 하지 못하였다. 즉, 10번 이상의 무작위 추출을 통해 성능 측정에 사용되는 사례 베이스를 여러 개 만들어서, 사례 베이스의 구성의 차이로 인하여 적중률의 차이가 생기는지를 확인하는 연구를 수행하지 못하였다. 그러므로 본 연구의 의의는 어떤 모델이 다른 모델보다 우수함을 보였다는 관점보다는, 사례기반 추론 기법이 협업 필터링 기법을 적용할 수 없는 상황에서도 무리 없이 적용될 수 있으며, 무작위 추천 보다 매우 우수한 성능을 보인다는 면에 있다고 하겠다. 셋째, 분류 코드 간 유사도 점수 매트릭스는 변할 수 있다는 것이다. 본 연구에서 사용한 분류 코드 간 유사도 점수 매트릭스는 본 저자와 온라인 서점의 담당 실무자의 논의를 통하여 작성한 것이다. 하지만, 출판 업계의 다른 실무자 또는 다른 연구자는 본 연구와는 상이한 분류 코드 간 유사도 점수 매트릭스를 작성할 수도 있고, 그에 따라 적중률도 달라질 수가 있다. 하지만, 이것은 유사도를 사용하는 사례기반 추론 기법의 특성으로 이해해야 할 것이다. 향후에는 전문가집단에 대한 설문조사와 논의 등의 방법을 통하여 좀 더 일반화되고 검증된 분류 코드 간 유사도 점수 매트릭스를 마련하는 것이 바람직하다고 하겠다.

참 고 문 헌

[1] Aamodt, A. and E. Plaza, “Case-Based Reasoning : Foundational Issues, Methodological Variations, and System Approaches,” Artificial Intelligence Communications, Vol. 4, No. 3, 1996, pp. 39-59.

[2] Adomavicius, G. and A. Tuzhihin, “Toward the Next Generation of Recommenders Systems : A Survey of the State-of-the-Art and Possible Extensions,” IEEE Transactions on Knowledge and Data Engineering, Vol. 17, No. 6, 2005, pp. 734-749.

[3] Ansari, A., S. Essegaier, and R. Kohli, “Internet Recommender Systems,” Journal of Marketing Research, Vol. 37, No. 3, 2000, pp. 363-375.

[4] Chanchien, S. W. and M. Lin, “Design and Implementation of a Case-based Reasoning System for Mar-

keting Plans," Expert Systems with Applications, Vol. 28, 2005, pp. 43-53.

[5] Goker, M. H. and T. Roth-Berghofer, "The Development and Utilization of the Case-Based Help-Desk Support System HOMER," Engineering Applications of Artificial Intelligence, Vol. 12, 1999, pp. 664-680.

[6] Kuo, R. J., Y. P. Kuo, and K. Y. Chen, "Developing a Diagnostic System through Integration of Fuzzy Case-Based Reasoning and Fuzzy Ant Colony System," Expert Systems with Applications, Vol. 28, 2005, pp. 783-797.

[7] Law, Y. F. D., S. B. Foong and S. E. J. Kwan, "An Integrated Case-Based Reasoning Approach for Intelligent Help Desk Fault Management," Expert Systems with Applications, Vol. 13, 1997, pp. 265-274.

[8] Leake, D., A. Maguitman, and T. Reichherzer, "Cases, Context, and Comfort : Opportunities for Case-Based Reasoning in Smart Homes," Lecture Notes in Artificial Intelligence, Vol. 4008, 2006, pp. 109-131.

[9] Lee, J. S. and J. C. Lee, "Music for My Mood : A Music Recommendation System based on Context Reasoning," Lecture Notes in Computer Science, Vol. 4272, 2006, pp. 190-203.

[10] Lee, J. S. and J. C. Lee, "Context Awareness by Case-based Reasoning in a Music Recommendation System," Lecture Notes in Computer Science, Vol. 4836, 2007, pp. 45-58.

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

[12] Li, P. and S. Yamada, "A Movie Recommender System Based on Inductive Learning," IEEE Conf. on Cybernetics and Intelligent Systems, 2004, pp. 318-323.

[13] Mulvenna, M., S. S. Anand, and A. G. Büchner, "Personalization on the Net using Web Mining : Introduction," Communications of the ACM, Vol. 43, No. 8, 2000, pp. 122-125.

[14] Resnick, P., N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl, "Grouplens : An Open Architecture for Collaborative Filtering of Netnews," Proceedings of the ACM Conf. on Computer Supported Cooperative Work, 1994, pp. 175-186.

[15] Sarwar, B., Sparsity, Scalability, and Distribution in Recommender Systems, Ph. D. Diss., Dept. of Computer and Information Sciences, Univ. of Minnesota, 2001.

[16] Varma, A. and N. Roddy, "ICARUS : Design and Development of a Case-Based Reasoning System for Loco-

motive Diagnostics," Engineering Applications of Artificial Intelligence, Vol. 12, 1999, pp. 681-690.

[17] Wang, H. C. and H. S. Wang, "A Hybrid Expert System for Equipment Failure Analysis," Expert Systems with Applications, Vol. 28, 2005, pp. 615-622.

[18] Yang, W., Z. Weng, and M. You, "An Improved Collaborative Filtering Method for Recommendations' Generation," IEEE Int'l Conf. on Systems, Man and Cybernetics, 2004, pp. 4135-4139.

저 자 소 개

이재식	(E-mail : leejsk@ajou.ac.kr)
1977년	서울대학교 경영학과 (경영학사)
1979년	KAIST 산업공학과 (공학석사)
1989년	University of Pennsylvania, Wharton School 경영정보시스템 (경영학박사)
현재	아주대학교 경영대학 e-비즈니스학부 교수
관심분야	Data Mining, Ubiquitous Computing, Recommender Systems, Intelligent Information Systems, AI Application to Business Problem Solving

명훈식	(E-mail : acell@iabacus.co.kr)
2000년	아주대학교 (경영학사)
2003년	아주대학교 대학원 경영정보학과 (경영학석사)
현재	㈜ 애버커스 솔루션 사업본부
관심분야	Data Mining, Recommender Systems, 통신 회선 품질 Mining, 통신 분야 Collaborative CRM