한국지능정보시스템학회
02 춘계학술대회 논문집, pp. 305~314

사례기반 추론을 이용한 서적 추천시스템의 개발

이재식* · 정석훈**

Development of a Book Recommendation System using Case-based Reasoning

Jae Sik Lee · Suk Hoon Chung

Abstract

In order to adapt to today's rapidly changing environment and gain a competitive advantage, many companies are interested in CRM(Customer Relationship Management). Especially, the product recommendation system that can be implemented by personalizing the marketing strategy becomes the focus of CRM. In this research, we employed CBR(Case-Based Reasoning) technique that can overcome the limitation of CF(Collaborative Filtering) technique. Our system recommends the books that the customer is very likely to buy next time considering the factors such as 'Personal Features of Customer,' 'Similarity between Book Categories' and 'Sequence of Book Purchases.' Accuracy of predicting a book—not a particular book, but in the middle level of classification that contains about 190 categories—was about 57%.

Key words : Customer Relationship Management, Recommendation System, Case-based Reasoning.

1. 서 론

인터넷을 이용한 전자상거래의 발달은 기업 경영에 많은 변화를 가져다주게 되었는데, 특히, 웹페이지를 통하여 고객에 대한 데이터를 손쉽게 대량으로 얻을 수 있게 되었다. 초기의 온라인 상점들은 정량적인 성장에만 초점을 두었으며 고객에 대한 서비스 향상에는 노력을 기울이지 않았다. 하지만, 서비스 질의 저하가 고객이탈이라는 기업의 생존에 직결된 치명적인 부분으로 이어지자 각 기업에서는 이를 막기 위하여 새로운 경영 패러다임에 관심을 가지게 되었다. 즉, 웹페이지를 통하여 대량으로 수집한 고객에 대한 데이터를 어떻게 이용할 것인가에 대한 논의가 활발해 지게 되었다. 또한 컴퓨터 기술의 발달은 이러한 논의를 좀 더 구체적으로 진행할 수 있게 해주었으며 컴퓨터 기술을 이용한 데이터 마이닝(Data Mining)을 통하여 기업에 도움이 되는 고객에게 고품질의 서비스를 제공해야 한다는 경영 전략에 관심이 집중되었다.

이러한 환경에서 새롭게 관심의 초점이 되는 것이 고객관계관리(CRM: Customer Relationship Management)이다. 특히, 온라인 상황에서 수행되는 고객관계관리를 전자적 고객관계관리(eCRM: Electronic Customer Relationship Management)라고 하는데, 많은 온라인 기업들이 이에 관심을 갖게된 것이다. 고객관계관리의 개념 중, 수집된 다량의 고객 정보를 이용하여 고객에 대한지식을 추출하고 이런 지식을 바탕으로 각 고객에 대한 맞춤 서비스의 제공은 1:1 마케팅(One to One Marketing)의 수행으로 나타나고 있으며 이러한 1:1 마케팅의 일환으로 부각되고 있는 것이 상품 추천 시스템이다.

상품 추천 시스템에는 협력적 여과(Collaborative Filtering) 기법이 많이 사용되어져 왔으나, 이 기법은 단독으로 사용될 때에 몇 가지 문제점들을 가지고 있다. 본 연구에서는 이러한 문제점들을 사례기반 추론(CBR: Case-Based Reasoning)을 이용하여 해결 하고자 한다. 본 연구는 온라인 서점 A의 구매내역 데이터를 분석하여 고객이 다음 번에 구매할 책의 정보를 예측하는 것이다.

본 논문은 총 6절로 구성되어 있다. 제 2절에서는 본 추천시스템에 사용된 주 알고리즘인 사례기반 추론에 대하여 살펴본다. 제 3절에서는 고객관계관리와 그 일환인 추천시스템에 대하여 서술한다. 제 4절에서는 본 연구에서 개발된 사례기반 추론 시스템의 구조 및 방법에 대하여 설명하고, 제

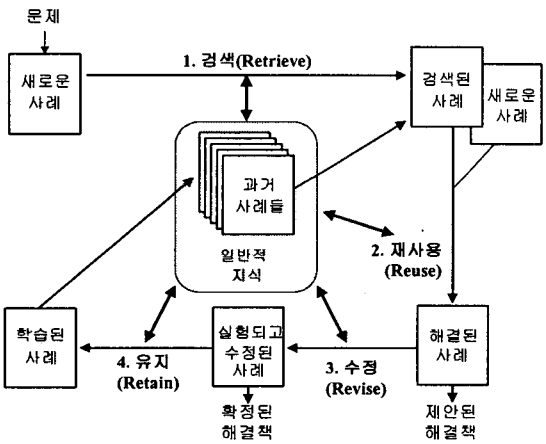
* 아주대학교 경영대학 교수
** 서강대학교 경영학과 박사과정

5절에서 실험 방법과 그 결과에 대하여 명기하였다. 끝으로 제 6절에서는 실험결과 얻게 된 결론과 미진했던 점 그리고 향후 연구방향에 대하여 논의한다.

2. 사례기반 추론 시스템

인간이 당면한 문제를 해결하기 위하여 사용하는 추론방법 중 자주 사용하는 방법은 과거에 이미 해결했던 문제들 중 현재 직면한 문제와 가장 유사한 문제를 기억해 내어 그 문제의 해답에 약간의 수정을 가한 후, 새로운 문제의 해답으로 사용하는 것이다[Riesbeck and Schank, 1989]. 사례기반 추론은 이러한 인간의 문제 해결방식을 모방한 기계학습(Machine Learning) 기법 중의 하나이다. 즉, 사례기반 추론은 과거에 한 번 발생한 문제는 또 다시 비슷한 형태의 문제로 발생할 가능성이 높으며 새로운 문제의 해답 역시 과거의 것과 유사할 것이다라는 가정에서 시작한다.

사례기반추론 시스템의 기본 순환구조는 <그림 2-1>과 같다[Aamodt and Plaza, 1994]. <그림 2-1>에서 보는 바와 같이 사례기반추론은 해를 구하고자 하는 새로운 사례가 입력되면 저장되어 있는 이전 사례들로부터 비슷한 사례를 검색하고, 적합한 형태의 응답으로 적응시켜 해를 도출한다. 그리고 도출된 해를 다시 저장함으로써 다음의 입력 사례에 대해 더욱 우수한 해를 제시해 줄 수 있도록 재사용된다.



<그림 2-2> 사례기반 추론의 순환 과정

3. 추천 시스템

3.1 고객관계관리

고객관계관리(CRM)란 가치 있는 고객으로부터 수익을 창출하고 고객과의 관계를 지속적으로 유지하기 위한 방법을 말한다. 이를 실현하기 위해선 고객관계관리가 일회성 행사로 끝나서는 안되며 기업 내부 및 외부의 고객과 관련된 데이터를 분석하고 통합하여 고객의 특성에 기반을 둔 마케팅을 펼치는 일련의 과정으로 이루어져야 한다. 컨설팅 회사인 언스트 앤드 영(Ernst and Young)의 고객관계관리에 관한 백서에서 언급하고 있는 고객관계관리의 정의는 아래와 같다[Ernst and Young, 2000].

> '기업의 수익성에 영향력이 큰 고객이나 고객그룹에 자원을 효과적으로 투입함으로써 고객으로부터 얻는 가치를 최대화하는 것'

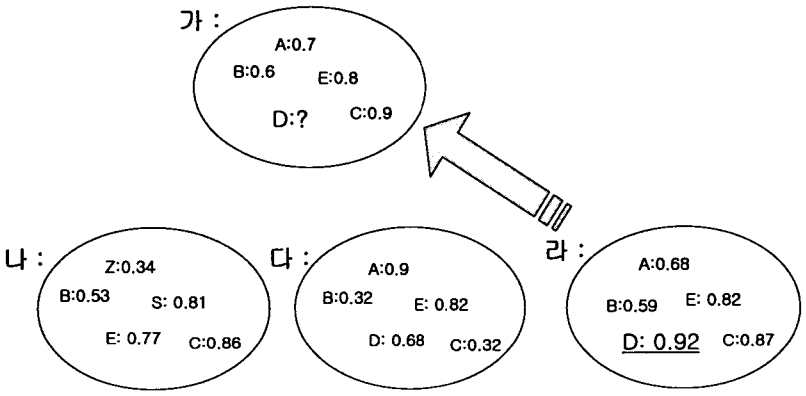
언스트 앤드 영의 정의에서도 '기업의 수익성에 영향력이 큰 고객이나 고객그룹'이라고 표현했듯이 고객관계관리는 모든 고객들을 대상으로 하는 것은 아니고, 기업에 실제로 수익을 가져다주는 고객들만을 대상으로 한다는 것이다. 이러한 개념이 나오게 된 이유는 두 가지로 들 수 있다. 첫 째는 실제 기업 수익의 대부분인 약 80%가 상위 20%의 고객들에 의해 얻어지며 나머지 고객들은 실제로 별 수익을 내주지 못하고 있다는 것을 기업이 자각하게 되었으며, 두 번째 이유는 기존 고객을 유지하는 비용보다 신규 고객을 유치하는 것이 훨씬 더 많은 비용을 요구하기 때문이다.

마케팅의 기법들은 정보·통신 기술의 발달과 더불어 <표 3-1>에서 보는 바와 같이 진화되어 왔다[Berson et al., 1999].

<표 3-1> 마케팅 기법의 진화

세대	마케팅 기법	관련 기술
암흑 세대	요행에 의한 마케팅	없음
르네상스 세대	개인적인 경험에 의한 마케팅	전화 인터뷰
산업혁명 세대	불특정 다수에 대한 마케팅	컴퓨터에 저장된 우송용 고객 명단
정보화 세대	데이터베이스를 이용한 마케팅	마케팅용으로 저장된 데이터 파일
최적화 세대	고객관계관리	데이터 웨어하우스, OLAP, 데이터 마이닝

최적화 세대의 마케팅 기법인 고객관계관리의 일환으로 대두되고 있는 1:1 마케팅 기법은 고객에 대한 정보를 손쉽게 대량으로 구할 수 있는 요즘 각광받고 있는 기법 중의 하나이다. 1:1 마케팅이 이루어지기 위해서는 기업이 고객의 성향, 구매특성 등 고객에 대한 많은 정보와 지식을 보유하고 있어야 하며, 이렇게 이미 잘 알고 있는 개인에 대하여 특화된 서비스가 준비되어 있어야 한다. 즉, 개인화된 서비스 제공이 가능해야 하는 것이다. 서론에서도 언급했던 것처럼 전자상거래의 발달과 컴퓨터 과학의 발달은 개인에 대한 데이터 수집을 용이하게 했으며 수집된 대량의 데이터를 신속하게 분석하여 가치 있는 정보나 지식 등으로 변화시킬 수 있었다. 이러한 기반들이 개인화된 서비스의 창



<그림 3-1> 협력적 여과 기법의 예

출과 보급을 가능하게 한 것이다.

3.2 추천 시스템에 대한 기존연구

추천 시스템은 고객으로부터 직접 입수한 정보나 웹에서의 고객 행동 패턴, 구매 내역 등을 분석하여 획득한 정보를 바탕으로 고객이 차기에 구매할 가능성이 높은 상품의 리스트를 작성하여 고객에게 추천해주는 시스템을 말한다[Maltz and Ehrlich, 1995].

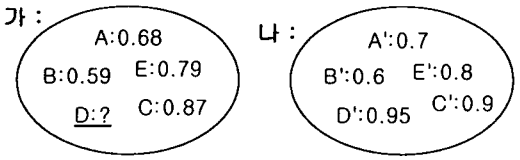
기존에는 추천 시스템에 협력적 여과(Collaborative Filtering) 기법이 자주 쓰였다. 협력적 여과 기법은 같은 문제영역(Domain) 내에 있는 아이템(Item)에 대한 사용자의 선호도 또는 우선순위 등과 같은 평가자료를 수집하여 동작하는데, 동일한 정보를 필요로 하는 사람 또는 동일한 성향을 가진 사람들을 연결 시켜주는 작업에 사용된다[황철현 외, 2000].

예를 들어, <그림 3-1>에서 고객 '가'의 아이템 D에 대한 선호도를 구하고자 할 때, 우선 고객 '가'가 구매한 아이템과 동일한 아이템을 구입했고, 또 각 아이템에 대하여 유사한 선호도를 갖는 고객들을 검색한다. 고객 '나'는 고객 '가'와 다른 아이템을 구입했으므로 제외되고, 고객 '다'와 고객 '라' 중에서 고객 '가'와 선호도가 가장 비슷한, 즉 선호도 차이가 가장 적은 고객의 아이템 D에 대한 선호도를 고객 '가'의 아이템 D에 대한 선호도로 사용하게 된다. <표 3-2>는 고객 '가'와 고객 '다', 그리고 고객 '가'와 고객 '라'간의 선호도 차이를 보여주고 있는데, 고객 '다'보다는 고객 '라'가 고객 '가'와의 선호도 차이가 적은 것을 알 수 있다. 그러므로, 고객 '가'의 아이템 D에 대한 선호도는 0.92가 된다.

<표 3-2> 아이템별 선호도 차이

	'가'와 '다'	'가'와 '라'
A	0.2(= | 0.7-0.9 |)	0.02(= | 0.7-0.68 |)
B	0.28(= | 0.6-0.32 |)	0.01(= | 0.6-0.59 |)
C	0.58(= | 0.9-0.32 |)	0.03(= | 0.9-0.87 |)
E	0.02(= | 0.8-0.82 |)	0.02(= | 0.8-0.82 |)
합 계	1.08	0.08

그러나 협력적 여과 기법을 추천 시스템에 사용하기에는 다음과 같이 몇 가지 단점이 있다. 첫 째는, 아이템 자체의 정보만을 사용한다는 것이다. 즉, 아이템에 관련된 여러 가지 속성들, 예를 들어 서적의 경우, 저자, 출판사, 주제별 분류 등의 정보가 무시되고 오직 서적 자체에 대한 정보만을 사용하고 있다. 특히 중요한 문제는 추천에 반드시 고려되어야 할 고객 정보를 전혀 반영하지 못하고 오직 구매 아이템에 대한 정보만을 사용하고 있는 것이다. 둘 째, 구매내역의 아이템들이 서로 완전 일치(Exact Matching)인 경우에만 사용 가능하기 때문에 아이템 A에 대해서 이와 상당히 유사한 아이템인 A′을 비교 대상으로 인식하지 못한다. 즉, <그림 3-2>에서 A와 A′, B와 B′, C와 C′, 그리고 E와 E′은 서로 상당히 유사한 아이템들이라고 가정했을 때, 협력적 여과 기법은 이 아이템들간의 유사한 정도를 반영하지 못하고 전혀 다른 아이템으로 인식하기 때문에, 이러한 상황에서는 사람 '나'에 대한 정보를 사람 '가'에 대하여 전혀 사용할 수 없게 된다. 넷 째로 협력적 여과 기법에서는 아이템의 구매 순서는 고려하지 않기 때문에 만약 순서에 의미가 있는 경우에는 적절치 못한 분석이 될 것이다.



<그림 3-2> 유사한 아이템간의 비교

협력적 여과 기법의 기존 연구를 보면 다음과 같다. Breese는 세 가지 서로 상이한 형태로 설계된 협력적 여과 기법을 서로 비교하는 연구를 수행했다[Breese *et al.*, 1998]. 첫 번째 형태는 상관계수에 기반한 기술을 사용한 것이며, 두 번째는 벡터(Vector) 기반 유사도 계산법을 사용한 것, 그리고 세 번째는 통계적 베이지안(Bayesian) 방법론을 사용한 것이다.

Burke는 추천시스템을 세 가지로 분류하였다. 첫 번째는 가장 많이 알려진 협력적 여과 기법, 두 번째는 내용-기반의 추천 시스템, 그리고 세 번째는 사례기반 추론과 같은 지식 기반 추천 시스템인데, 이들은 서로 보완적인 관계에 있으며 특히, 지식 기반 추천 시스템이 다른 방법들을 보완하는데 가장 좋은 방법임을 밝혀냈다[Burke, 1998].

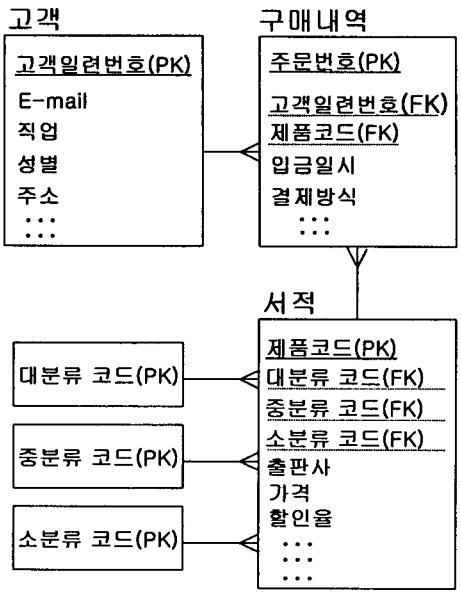
4. 서적 추천 시스템 CBRS의 개발

본 연구에서 개발한 시스템은 사례기반 추론을 이용하여 협력적 여과 알고리즘의 한계를 극복한 서적 추천 시스템이며 그 이름을 CBRS (Case-based Book Recommendation System)로 명명하였다. 본 연구에서는 추천할 대상 상품으로 서적이 사용되었지만 반드시 서적만 가능한 것은 아니며 본 시스템은 서적 이외의 다른 상품의 추천에도 사용할 수 있는 엔진의 형태로 되어 있다.

4.1 사용 데이터

본 연구에 사용된 데이터는 국내 온라인 서점 A의 5개월간의 구매내역 데이터 중 17,111건의 데이터이다. 본 데이터는 관계형 데이터베이스의 구조로 되어 있으며 고객, 구매내역, 그리고 서적 등의 3가지 테이블로 구성되어 있다. 데이터의 개체-관계도(ERD: Entity-Relationship Diagram)는 <그림 4-1>과 같다.

각 서적은 기본적인 정보 이외에 세 단계의 분류코드가 부여되어 있다. 대분류 코드는 전체 책을 주제별로 20가지로 분류한 후 각 분류마다 코드번호를 부여한 것이다. 대분류 코드의 예는 <표 4-1>과 같다.



<그림 4-1> 사용된 데이터의 개체 관계도

<표 4-1> 대분류 코드의 예

코드번호	주 제
1	가정과 가족
2	건강과 미용
3	문학
⋮	⋮
20	취미과 실용

중분류 코드는 대분류 코드를 다시 세분하여 코드값을 부여한 것이며, 소분류 코드는 중분류 코드를 다시 세분하여 코드값을 부여한 것이다. 이 분류 코드 값들은 해당 서적이 어떤 주제별 분류에 속하는지 알 수 있게 해준다.

<표 4-2> 구매 권수별 발생 건수 상황

총구매량(권)	발생 건수(명)
1	11,221
2	1,689
3	424
4	149
5	43
6	24
7	8
8	3
9	3
10	3
11이상	3544

고객 테이블에는 고객의 인구통계학적 정보 등이 저장되어 있으며, 구매내역 테이블의 경우 구매량이 한 권에서부터 수 백권에 이르기까지 다양한 분포를 갖지만 <표 4-2>와 같이 주로 구매내역이 1권에서 4권까지의 경우가 가장 많으며 나머지 경우는 매우 드물게 발생하고 있다. 따라서 본 연구에서는 구매내역이 4권까지인 내역만을 사용하고 나머지는 사용하지 않았다.

4.2 CBRS의 구조

원래의 데이터베이스는 구매내역이 <표 4-3>과 같은 형태로 되어 있다. 그림에서 볼 수 있듯이 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 'D'와 'E'를 구매했다. 데이터베이스에서 먼저 등장한 것일수록 구매한지 오래된 서적이며 나중에 등장하는 것이 가장 최근에 구매한 서적이다.

<표 4-3> 서적 구매 데이터베이스

고객	구매서적
가	A
가	B
가	C
나	B
나	A
다	D
라	E

이러한 구조로 된 원래의 데이터베이스를 <표 4-4>와 같은 형태로 변환시킨다.

<표 4-4> 변환된 데이터베이스

고객	1차 구매서적	2차 구매서적	3차 구매서적
가	A	B	C
나	B	A	null
다	D	null	null
라	E	null	null

본 연구에서 개발된 CBRS는 다수의 사례베이스로 구성되어 있다. 즉, <그림 4-2>에서 보는 바와 같이 원래의 데이터베이스를 구매내역의 건수에 따라 분할하여 사례베이스가 구성되어 있다.

제 1번 사례베이스

고객	1차 구매
다	D
라	E

제 2번 사례베이스

고객	1차 구매	2차 구매
나	B	A
마	D	F

제 3번 사례베이스

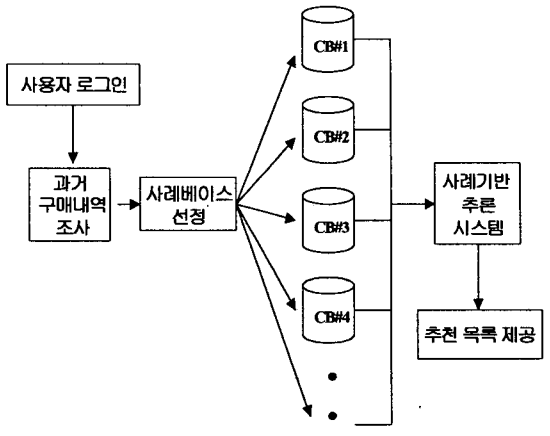
고객	1차 구매	2차 구매	3차 구매
가	A	B	C
사	Q	P	B

•
•

<그림 4-2> 분할된 사례베이스의 구조

그러므로, 추천을 하고자 하는 고객의 구매내역 건수에 따라 각각 다른 사례베이스가 적용된다. 즉, 구매내역이 한 권인 고객은 구매내역이 두 개로 구성되어 있는 제 2번 사례베이스가 적용되며, 구매내역이 두 권인 고객은 구매내역이 세 개로 구성되어 있는 제 3번 사례베이스가 적용된다. 이에 대한 자세한 설명은 제 4.4절에서 한다.

전체적인 CBRS의 구조는 <그림 4-3>과 같다. 사용자가 로그인(Log-in)을 하면 시스템은 먼저 그 사용자의 구매내역 건수를 조사하게 된다. 구매내역 건수를 조사한 후, 그 사용자에게 적용될 사례베이스가 선정된다. 사례베이스가 선정되면 사례기반 추론 시스템은 추론과정을 거쳐 고객에게 서적 추천 목록을 제공한다.



<그림 4-3> CBRS의 구조

4.3 입력속성과 목표값(Target Value)

CBRS의 입력속성은 크게 두 가지로 분류된다. 첫 째는 고객 즉, 사람에 대한 속성이며, 두 번째는 서적에 대한 속성이다. 고객에 대한 입력속성으로는 직업코드(job_code)와 성별(sex)이 사용되었으며, 서적에 대한 속성으로는 저자(writer), 출판사(publisher), 가격(price), 할인율(dc_rate), 대분류 코드(level_1), 중분류 코드(level_2)가 사용되었다. 그러므로 하나의 레코드는 <표 4-5>와 같이 기본적으로 고객 속성 두 개, 서적 속성 6개로 구성되어 있으며 구매 내역이 하나씩 늘어날 때마다 서적 속성이 6개씩 증가하게 된다. 구체적인 사례베이스의 모습은 다음절에서 살펴보기로 한다. CBRS 시스템의 목표값은 다음 번 구매할 것이라고 예측되는 서적의 중분류 코드(level_2)이다.

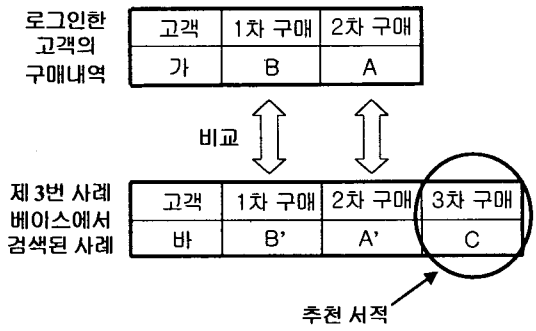
<표 4-5> 입력속성과 목표값

입력속성	형태	목표값	형태
직업코드	범주형	차기 구매서적의 중분류 코드	범주형
성별	이산형		
저자	문자형		
출판사	문자형		
가격	수치형		
할인율	수치형		
대분류 코드	범주형		
중분류 코드	범주형		

4.4 추천 방법 및 유사도 도출방법

4.4.1 추천방식

추천할 책을 선정하는 방법은 다음과 같다. 예를 들어 <그림 4-4>와 같이, 구매내역이 두 권인 고객이 로그인을 하면 구매내역이 세 권인 사례베이스(제 3번 사례베이스)가 사용될 사례베이스로 선정된다. 고객의 구매내역과 제 3번 사례베이스의 첫 번째와 두 번째 서적의 속성을 비교하여 유사도를 구한 후, 검색되어져 나온 사례들의 세 번째 서적을 추천 목록에 포함시킨다. 추천 서적의 개수는 총 10개로 한다.



<그림 4-4> 추천 목록 작성 방법

4.4.2 유사도 도출방법

본 시스템에서 사용한 속성 형태별 유사도 도출방법은 다음과 같다. 수치형 속성일 경우에는 (식 4.1)을 이용하여 유사도를 도출한다.

$$d = \left\{ \left| \frac{a-b}{\max} \right| \right\} \qquad \text{(식 4.1)}$$

d : 수치형 속성의 속성간 유사도
a : 사례베이스에 있는 사례의 속성값
b : 새로운 사례의 속성값
max : 사례베이스에 있는 속성 값 중 최대값

문자형 속성일 경우에는 완전히 일치하는 경우에만 점수를 부여하고 그렇지 않은 경우에는 점수를 부여하지 않는다. 새로운 사례 N에 대해서 사례베이스에 있는 사례 O의 총 유사도는 (식 4.2)와 같이 모든 속성별로 유사도를 구한 후, 각 속성별 유사도에 가중치를 곱한 후 이를 총합하여 구한다.

$$S(N,O) = \sum_{i=1}^{n} f(N_i, O_i) \times W_i \qquad \text{(식 4.2)}$$

N : 새로운 사례
O : 사례베이스에 있는 사례
$S(N,O)$: 새로운 사례 N과 사례베이스에 있는 사례 O간의 총 유사도
n : 사례의 속성 개수
N_i : 사례 N의 i 번째 속성
O_i : 사례의 O의 i 번째 속성
$f(N_i,O_i)$: 두 속성 N_i 와 O_i 사이의 유사도를 측정해 주는 함수
W_i : i 번째 속성의 가중치

사례베이스에 있는 모든 사례에 대하여 총 유사도를 구한 다음 총 유사도가 가장 큰 사례부터 내림차순으로 정렬하여 총 유사도 상위 10위까지의 사례를 가져온다.

4.4.3 분류코드의 유사도 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 4-6>과 같다.

<표 4-6> 분류수준별 코드의 개수

분류수준	개수
대분류	20개
중분류	190여개
소분류	300여개

분류코드 자체가 서적의 내용 및 주제 등의 유사성에 기반하여 부여된 것이라면 서적간의 유사도를 구하는데 매우 중요한 정보가 될 수 있다. 그러나 본 연구에서 수집한 온라인 서점 A의 구매내역 데이터의 서적 분류코드는 그와 같은 유사성에 근거하여 부여된 코드가 아니다. 그렇다고 해서 두

서적의 분류코드가 단지 서로 '같다', '다르다'만을 판정하는 이진적(Binary) 비교로는 사례기반 추론의 장점을 살리기가 힘들다. 따라서 본 연구에서는 대분류와 중분류의 코드 값들 상호간 유사성에 대하여 전문가의 의견을 참작하여 2차원 테이블의 형태로 유사도 점수를 부여하였다.

먼저 대분류 코드에 대하여 상호간 유사도 점수 테이블을 작성한 후, 각 대분류 코드로부터 다시 분류되는 중분류 코드들 상호간에 대하여 유사도 점수 테이블을 작성하였다. 유사도 점수는 최저 0점에서 최고 10점 사이의 정수로 부여하였는데, 유사도 테이블의 예는 <표 4-7>, <표 4-8>과 같다.

<표 4-7> 대분류 코드 유사도 테이블

코드		1	2	3	4	5	. . .	20
코드	의미	가정과 가족	건강과 미용	문학	인문	인물	. . .	취미와 실용
1	가정과 가족	10	1	1	1	1	. . .	2
2	건강과 미용	-	10	1	1	1	. . .	1
3	문학	-	-	10	7	1	. . .	1
4	인문	-	-	-	10	1	. . .	1
5	인물	-	-	-	-	10	. . .	1
⋮	⋮	-	-	-	-	-	10	⋮
20	취미와 실용	-	-	-	-	-	-	10

유사도 매트릭스는 대각선을 사이에 두고 대칭을 이루기 때문에 좌하단부의 수치는 표기하지 않았다. <표 4-7>을 보면 대부분의 유사도 점수가 1점인데, 이것은 대분류의 경우 분류간의 간격이 매우 넓기 때문에 상호간의 유사도가 떨어지기 때문이다. 각 대분류코드 안에서 다시 중분류 코드 상호간 유사도 점수를 부여하였다. <표 4-8>은 대분류 코드 1번(가정과 가족)에 속하는 중분류 코드 유사도 테이블이다.

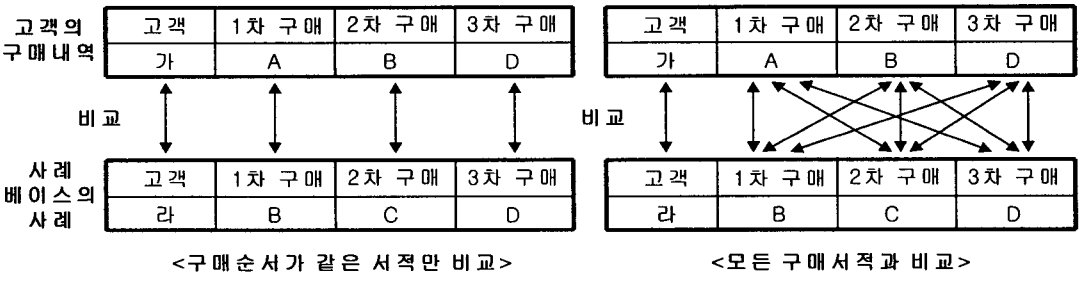
<표 4-8> 대분류 코드 1번의 중분류 코드 유사도 테이블

코드		23	24	26	39	42	. . .	54
코드	의미	결혼	임신 및 출산	태교	육아	가족 관계	. . .	홈 인테리어
23	결혼	10	9	8	7	5	. . .	4
24	임신 및 출산	-	10	9	9	6	. . .	1
26	태교	-	-	10	9	3	. . .	1
39	육아	-	-	-	10	3	. . .	1
42	가족 관계	-	-	-	-	10	. . .	1
⋮	⋮	-	-	-	-	-	10	⋮
54	홈 인테리어	-	-	-	-	-	-	10

4.4.4 사례베이스

현재 보유한 구매내역의 양으로는 4개까지의 사례베이스가 가능하다. 구매내역 다섯 건 이상의 사례들은 그 빈도가 너무 적어서 사례베이스로 역할을 하기에 무리가 따른다. 제 1번 사례베이스는 구매내역이 한 건인 사례들의 묶음이며 총 11,221개의 사례를 가지고 있다. 제 1번 사례베이스는 사례의 개수가 가장 많음에도 불구하고 본 연구에서는 사용하지 못한다. 왜냐하면 CBRS에서는 최소한 구매내역이 2권 이상이어야 사례베이스로 사용될 수 있기 때문이다.

본 연구에서는 제 2번 사례베이스부터 제 4번 사례베이스까지 사용되는데, 태스트와 평가를 위해서 <표 4-9>와 같이 사례를 나누어 사용하였다. 각 사례베이스 번호에 대해서 사례베이스의 사례 개수, 테스트용 사례 개수, 그리고 평가용 사례 개수의 비율은 8:1:1로 하였다.



<그림 5-1> 실험 1, 2번과 실험 3번의 차이

<표 4-9> 사례베이스용, 테스트 용 및 평가용 사례의 개수

사례 베이스 번호	사례베이스에 저장된 사례의 개수	테스트 사례 개수	평가 사례 개수	총 사례 개수
2	1,351	169	169	1689
3	340	42	42	424
4	119	15	15	149

5. 실험방법 및 결과

5.1 실험 방법

예측에 사용된 알고리즘은 사례기반추론이며 유사도 도출을 위한 속성(Features)간 비교 방법에 있어 다양한 비교방법을 수행한다. 또한 서적 분류 코드간 유사 정도를 점수에 반영하는 모델과 분류 코드가 완전히 같은 경우에만 점수를 부여하는 모델을 비교하여 최적의 서적 추천시스템을 개발한다. 본 연구에서 수행된 실험은 아래와 같이 네 가지로 나눌 수 있다.

[실험 1] 모든 속성들에 동일한 가중치 부여
[실험 2] 구매 서적별로 상이한 가중치 부여
[실험 3] 구매 순서를 고려하지 않은 유사도 측정
[실험 4] 서적의 분류코드만을 완전일치 방법으로 비교

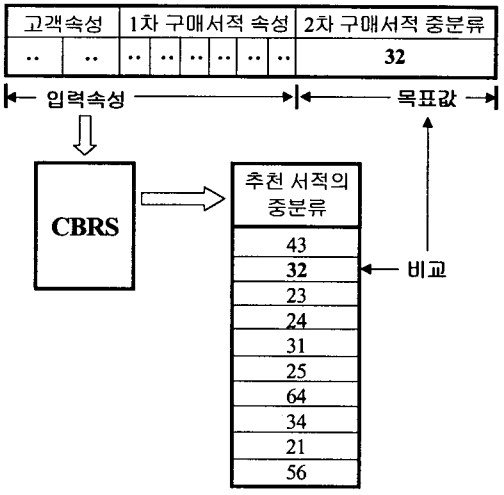
[실험 1]은 모든 속성에 가중치 '1'을 부여함으로써 실제로 가중치를 고려하지 않은 상태에서 실험을 수행한 것이며, [실험 2]는 첫 번째 구매서적, 두 번째 구매 서적, 세 번째 구매서적에 서로 다른 가중치를 부여함으로써 적중률의 변화를 관찰하는 실험이다. [실험 3]은 앞의 두 가지의 실험과는 다르게 구매순서를 고려하지 않은 실험이다. 즉, [실험 1]과 [실험 2]는 <그림 5-1>의 왼쪽 그림과 같이 새로운 사례의 첫 번째 구매 서적은 사례 베이스 사례의 첫 번째 구매 서적과만 비교하고, 두 번째 구매 서적은 두 번째 구매 서적과만, 그리고 세 번째 구매 서적은 세 번째 구매 서적과의 관계에서 유사도를 도출하지만 [실험 3]은 <그림 5-1>의 오른쪽 그림과 같이 새로운 사례 각각의 서적과 사례 베이스에 있는 사례의 모든 서적을 비교하여 유사도를 도출하는 것이다. 즉, 구매 순서를 고려하지 않고 모든 서적들끼리 비교하는 실험이다.

[실험 4]는 고객 속성은 삭제하고 서적 속성만 사용했으며, 속성비교에서 유사도를 사용하지 않고 속성끼리 완전일치(Exact Matching)인 경우에만 점수를 부여했다. 그리고 가중치 및 구매순서를 고려하지 않았다. 즉, [실험 4]는 사례기반 추론의 특징들을 모두 제거하고, 단순히 구매 내역만을 사용하여 데이터베이스 검색(Database Retrieval) 방식으로 실험한 것이다.

5.2 적중률 측정 방법

본 연구에서는 시점을 과거로 돌려 실험 대상이 되는 고객들이 최종에 구입한 서적을 아직 구입하지 않았다고 가정하고, 최종 구매 서적 바로 직전 서적까지만 입력하였을 때 최종서적에 대한 정보를 중분류 코드(level_2) 수준에서 얼마나 정확히 예측할 수 있는지 살펴보았다.

예를 들어, <그림 5-2>에서와 같이 실제 구매 내역이 두 권인 고객 '가'에 대한 적중률을 측정하고자 할 때, 먼저 고객 '가'를 구매 내역이 한 권인 사람으로 간주한다. 고객 '가'의 1차 구매 서적까지만 유사 사례 검색에 사용한 후, <그림 4-4>와 같은 방법으로 시스템이 추천해 주는 서적 목록에 고객 '가'의 실제 2차 구매 서적에 대한 정보가 포함되어 있는지를 검사한다. 일치하는 서적이 발견되면 추천이 적중한 것으로 판정한다.



<그림 5-2> 적중률 측정 방법

5.3 실험 결과

<표 5-1>은 본 연구에서 시도한 여러 가지 실험에서 보인 중분류에 대한 적중률의 비교표이다.

<표 5-1> 각 실험의 중분류 적중률 비교

실험 번호	제 2번 사례 베이스	제 3번 사례 베이스	제 4번 사례 베이스	평균
[실험 1]	56.80%	53.49%	53.34%	54.54%
[실험 2]	-	53.49%	60%	**56.75%**
[실험 3]	-	53.49%	51.16%	52.33%
[실험 4]	54.43%	48.84%	40%	47.76%

[실험 2]는 구매된 순서에 따라 서로 다른 가중치를 부여하고 실험하는 것이다. 즉, 구매 순서가 1차인 서적과 2차인 서적에 서로 다른 가중치를 부여함으로써 구매시기에 따른 중요도의 차이를 감안

하고자 하는 것이다. 제 2번 사례베이스에는 구매 내역이 한 권이어서 구매순서의 개념이 없기 때문에 [실험 2]에서는 사용되지 않았고, 제 3번과 제 4번 사례베이스가 실험에 사용되었다. 실험 결과, 각 사례베이스에서 조금 다른 최적 서적 가중치가 도출되었다. 제 3번 사례베이스에서는 모든 서적들에게 동일한 가중치 '1'을 부여하였을 때 가장 좋은 결과를 보였으나, 제 4번 사례베이스의 경우에는 구매 순서가 2차인 서적에는 가중치 '5'를, 그리고 3차인 서적에는 가중치 '10'을 부여하였을 때 가장 좋은 결과를 보였다.

[실험 3]은 구매순서를 고려하지 않고 모든 책들에 대해서 유사도를 도출하여 유사한 사례를 검색하는 실험이다. [실험 3] 역시 제 2번 사례베이스는 사용하지 못하고 제 3번과 제 4번 사례베이스만 사용하였다. 그 이유는 제 2번 사례베이스에서는 유사도 도출 시 적용되는 서적이 한 권뿐이어서 구매순서의 의미가 없기 때문이다. [실험 3]에서도 속성별 가중치들이 제 3번 사례베이스와 제 4번 사례베이스에서 조금 다르게 부여되었다.

<표 5-1>에서 볼 수 있듯이, 구매 순서를 고려한 [실험 2]의 경우가 가장 좋은 적중률을 보이고 있다. 한편, 사례기반 추론을 사용하지 않고, 단순한 데이터베이스 검색 방법으로 실험한 [실험 4]의 경우가 가장 낮은 적중률을 보이고 있다.

6. 결론 및 향후 연구과제

서론에서도 밝혔듯이 본 연구는 추천 시스템을 개발하는데 있어, 사례기반 추론이 고려할 수 있는 부분을 반영한 시스템을 개발하여 최적의 추천시스템 모델을 찾는 것이 그 목적이다. 최적의 모델을 찾기 위하여 다양한 모델을 개발하여 각 모델들을 비교하는 실험을 수행하였다. 실험 결과, 가장 좋은 적중률을 보인 모델은 [실험 2]에서 사용된 모델이다. [실험 2]에서 사용된 모델은 고객정보에 해당하는 속성은 제외하고, 구매 순서를 고려하였으며, 구매 순서에 따라 서적에 서로 다른 가중치를 부여한 것이다. [실험 2]의 모델에서 중분류 적중률은 제 3번 사례베이스에서 53.49%, 제 4번 사례베이스에서는 60%를 보여서 평균 57%에 달했다. 중분류 코드의 개수는 모두 190여개에 달한다. 이 중에서 10개를 선택했을 때에 우리가 원하는 서적의 중분류 코드가 포함될 확률은 5.26%(=10/190×100%)에 불과하다. 그러므로 57%의 적중률은 절대로 작다고 말할 수 없는 수치이다.

본 연구를 수행함에 있어 가장 큰 문제점은 사용된 데이터에서 거래내역에 관련된 데이터의 수량이 매우 부족했던 점이다. 앞에서도 살펴본 바와 같이 총 17,111건의 구매 내역 중 대부분을 차지하는 11,221건의 내역이 모두 구매내역이 한 권인 경우이다. 이렇게 거래내역 데이터의 수량이 적었던 이유는 연구에 사용될 데이터를 입수할 당시에 온라인 서점 A의 영업기간이 그리 오래되지 않아 거래내역의 발생건수가 적었기 때문이다. 이러한 문제점은 거래내역 건수가 늘어나면 자연스럽게 해결되리라 생각된다.

두 번째로는, 사례기반 추론에서 매우 중요한 부분을 차지하고 있는 속성 가중치 부여와 속성 선정에 관한 충분한 고려를 하지 못했다는 것이다. 본 연구에서는 속성을 사용함에 있어 파생변수는 전혀 사용하지 않았는데 결과에 좀 더 좋은 영향을 줄 수 있는 파생변수의 생성도 고려해야 할 것이다. 속성 가중치 부여에 있어서도 본 연구에서는 가중치를 임의로 부여하는 방법을 사용했다. 속성 가중치 부여에 대한 기존의 연구가 많이 진행되어 오고 있기 때문에 차후에는 이러한 가중치 부여 방법론들을 적용해서 연구 해볼 필요가 있을 것이다.

세 번째로는, 본 연구에서는 대분류와 중분류까지만 유사도 테이블을 작성하여 유사도를 비교하고 소분류는 고려하지 않았다. 유사도 테이블을 작성하여 사용한 이유는 실제 데이터의 분류 코드값이 일정한 체계에 의해 부여된 것이 아니기 때문에 코드값 자체로는 아무 의미도 찾을 수 없었기 때문이다. 유사도 테이블을 작성함에 있어서 그 개수가 300여 개에 이르는 소분류까지 유사도 테이블을 작성하는 것은 시간이 많이 요구되는 작업이다. 따라서 본 연구에서는 중분류 코드까지만 유사도 테이블을 작성했는데 이것은 목표값을 중분류 코드 값으로 할 수밖에 없는 연구의 한계점을 낳게 되었다. 만약 소분류 코드에까지 유사도 테이블을 작성하거나, 분류 코드를 코드값 자체만으로 유사도를 측정할 수 있도록 새롭게 부여하면 소분류까지 고려하여 유사도를 측정할 수 있을 것이고 적중률도 매우 증가시킬 수 있을 것이다.

참 고 문 헌

[1] 황철현, 박영길, 박용준, "협력적 여과에서 결측치(Missing Value) 예측에 관한 연구," 2000 한국지능정보시스템 학회 추계학술대회 논문집, (2000), 333~337.

[2] Aamodt, A. and E. Plaza, "Case-Based Reasoning: Foundational Issues, Methodological Variations, and system Approaches," *Artificial Intelligence Communications,* Vol. 7, No. 1, pp. 9-13, 1996.

[3] Berson, A., S. Smith and K. Thearling, *Building Data Mining Applications for CRM*, McGraw-Hill, 1999.

[4] Breese, J. S., D, Heckerman, and C, Kadie, "Empirical Analysis of Predictive Algorithms for Collaborative Filtering," *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*, 1998.

[5] Burke, R., "The Wasabi Personal Shopper: A Case-Based Recommender System," *Proceedings the 11th Annual conference on Innovative Applications of Artificial Intelligence*, (1999), 844~849.

[6] Dash, M. and H. Liu, "Feature Selection for Classification", *Intelligent Data Analysis*, (1997), 131~156.

[7] Ernst & Young, *CRM White Paper*, 2000.
[8] Kolodner, J., *Case-Based Reasoning*, Morgan Kaufmann publishers, Inc., 1993.
[9] Maltz, D. and K, Ehrlich, "Pointing the Way: Active Collaborative Filtering," *Proceedings of ACM CHI 1995 Conference on Human Factors in Computing Systems*, 1995, 202~209.
[10] Riesbeck, C. K. and R. L. Schank, *Inside Case-based Reasoning*, Lawrence Erlbaum Associates, 1989.
[11] Turban, E., *Expert Systems and Applied Artificial Intelligence*, Macmillan Publishing Company, Inc., 1992.
[12] Watson, I., *Applying Case-based Reasoning: Techniques for Enterprise Systems*, Morgan Kaufmann, 1997.