

데이터 마이닝을 이용한 인터넷 쇼핑몰 상품추천시스템

김경재
동국대학교 경영대학 정보관리학과
(*kjkim@dongguk.edu*)

김병국
동국대학교 경영대학 정보관리학과
(*ubf@dongguk.edu*)

전자상거래의 확산에 따라 인터넷 쇼핑몰에서의 구매활동은 일반적인 현상이 되었다. 그 결과, 유사한 업종이나 형태의 인터넷 쇼핑몰이 범람하게 되었고 업체들 간의 경쟁도 심화되어 차별화된 서비스를 제공하지 않는 업체는 도태되기 쉬운 상황이다. 본 연구에서는 치열한 경쟁환경 하에서 인터넷 쇼핑몰들의 차별화된 마케팅 서비스의 수단으로써 이용되고 있는 상품추천시스템의 개선된 모형을 제시하고자 한다. 본 연구에서 제안하는 모형은 전역 최적화 기법 중의 하나인 유전자 알고리즘을 데이터 마이닝의 도구로 활용한 인터넷 쇼핑몰에서의 개인화된 상품추천시스템 모형이다. 유전자 알고리즘은 추출하기가 어려운 소비자의 성향을 데이터를 통해 추출하고 이에 맞는 상품군을 선택할 수 있도록 해 주는 최적화 기법으로 상품추천시스템의 추천엔진으로써 유용할 것으로 기대된다. 본 연구에서는 제안하는 유전자 알고리즘에 기반한 추천규칙들이 장착된 웹 기반의 개인화된 상품추천시스템의 프로토타입을 개발하고 이에 대한 실제 사용자들의 이용 만족도를 확인함으로써 본 연구에서 제안한 방법론의 유용성을 확인하고자 한다.

논문접수일 : 2005년 5월

게재확정일 : 2005년 6월

교신저자 : 김경재

1. 서론

e-비즈니스의 활성화로 인하여 수많은 인터넷 쇼핑몰 업체가 등장하였으며 극심한 경쟁에서 생존하기 위한 고객관계관리의 노력도 계속되고 있다. 이러한 노력은 판매자의 영리추구를 위해서 진행되고 있지만, 이 중 일부는 고객들에게 상품에 대한 유용한 정보를 제공함으로써 최적의 구매의 사결정을 할 수 있도록 도움을 주기도 한다. 상품에 대한 정보가 과다하게 제공되는 전자상거래의 현실을 고려하면 고객에게 반드시 필요하고 도움이 되는 정보를 선택적으로 제공하는 것만이 판매자와 소비자 양자의 효익을 최대화할 수 있는 수단이 될 수 있다.

이에 따라 상품정보의 과부하 현상을 겪고 있는 고객의 어려움을 완화할 수 있는 방안으로써 일대일 마케팅 전략이 주목을 받고 있으며, 일대일 마케팅을 위한 정보기술의 하나로써 많은 인터넷 쇼핑몰들이 상품추천시스템(product recommender systems)을 사용하고 있다. 상품추천시스템은 인터넷 상거래 업체가 고객에게 개인화 서비스를 제공할 수 있도록 하여 고객의 충성도를 높일 수 있는 경쟁력으로 작용하며, 고객의 측면에서는 정보 과부하 현상을 완화함으로써 상품 선택의 탐색 비용을 감소시켜, 판매자와 소비자 양측 모두에게 유용하다. 그러나 고객의 요구를 잘못 판단하여 잘못 되거나 과다한 추천을 하게 되면 오히려 고객의 불만원인이 될 수 있다. 따라서 상품추천시스템의

효익만을 향유하기 위해서는 정교하고 절제된 추천을 해 줄 수 있는 추천시스템의 개발이 무엇보다 중요하다.

선행 연구에서 제안된 추천시스템 구축을 위한 기법들은 다양하다. 협업필터링(collaborative filtering) 기법(Resnick et al., 1994)은 지금까지 가장 성공적인 추천 기술로 알려져 있으며 내용기반(content-based) 기법과 함께 추천시스템 연구의 두 축을 이룬다. 그러나 이 두 가지 기법은 기본적으로 고객간 또는 상품간의 연관성을 기반으로 추천을 함으로써 희박성(sparsity)의 문제 즉, 구매를 상대적으로 적게 하는 고객이나 구매가 상대적으로 적게 이루어지는 상품에 대해서는 정교한 추천을 하지 못할 가능성이 있다. 또한 협업필터링의 경우에는 확장성(scalability)의 문제 즉, 거래 데이터가 증가함에 따라 유사성 계산에 필요한 연산량이 기하급수적으로 증가하는 문제가 있다. 따라서 협업필터링 등의 기법의 경우에는 모든 상품에 대한 거래가 빈번하게 일어 나지 않는 중소 규모의 인터넷 쇼핑몰이나 사업 초기 단계의 인터넷 쇼핑몰의 경우에는 추천의 성과가 낮을 가능성이 있고, 거래가 과다하게 일어나는 대형 인터넷 쇼핑몰에서는 신속한 추천을 하지 못할 가능성이 있으며, 이러한 점들은 중요한 한계점으로 지적될 수 있다.

본 연구에서는 기존의 상품추천시스템들의 이러한 한계점을 보완할 수 있는 유전자 알고리즘(genetic algorithms) 기반의 추천기법을 제안한다. 본 연구에서 제안하는 기법은 기존 방법들의 한계점을 보완할 수 있는 규칙기반 추천기법의 일종이지만, 규칙기반 추천기법의 한계점이라고 할 수 있는 규칙 추출의 어려움을 보완한 기법이다. 규칙 추출을 위해서는 일반적으로 의사결정나무, 연관성 규칙 등이 활용 되고 있으나, 유전자 알고

리즘은 주어진 문제 영역에서 전역 탐색(global search)을 통해 정교하고 절제된 최적해를 찾아 가는 데이터 마이닝 기법으로서, 많은 고객 데이터를 활용하여 추천규칙을 생성하는데 매우 효과적으로 사용될 수 있다. 제안하는 방법은 추천규칙을 미리 추출하고 각 규칙에 대응되는 사건이 발생할 때 추천을 하게 되므로 협업필터링 등의 단점인 데이터 증가에 따른 확장성의 문제가 발생하지 않는다. 또 각 상품에 대한 선호도를 기반으로 추천을 하는 협업필터링과 달리, 본 연구에서 제안하는 방법은 고객의 인구통계학적인 데이터를 기반으로 추천규칙을 추출하게 되므로 선호도 입력의 어려움에 따른 희박성 문제의 가능성도 줄일 수 있다.

본 연구에서는 유전자 알고리즘에 기반한 상품추천시스템을 제안하고 이의 유용성을 확인하기 위해 가상의 인터넷 쇼핑몰을 구축하고 사용자들의 이용 만족도를 확인하는 방식으로 제안된 기법의 유용성을 확인하고자 한다.

본 연구는 다음과 같이 구성되어 있다. 먼저, 선행연구에서 사용된 상품추천기법들을 살펴 보고, 그 장단점을 확인한다. 확인된 장단점을 기반으로 단점을 보완할 수 있는 기법을 제안하고 제안 기법의 주요구성요소인 유전자 알고리즘에 대해 살펴 본다. 3장에서는 제안추천기법을 활용하여 추천시스템을 설계하고 이의 유용성을 확인할 수 있는 가상의 인터넷 쇼핑몰을 설계한다. 4장에서는 추천시스템 구축과 유용성 확인에 활용된 데이터에 대해 살펴 보고 이를 활용한 실험과정을 기술한다. 5장에서는 가상 인터넷 쇼핑몰에서의 이용자 만족도를 확인하고, 6장에서는 연구의 결론과 한계점을 제시한다.

2. 관련연구

2.1 상품추천시스템 주요 추천기법

상품추천시스템에서 사용되는 추천기법은 크게 협업필터링, 내용기반추천, 규칙기반추천 등의 세 가지 종류로 분류할 수 있다.

협업필터링(collaborative filtering) 추천기법은 가장 성공적인 추천 방식 가운데 하나이며, 사용자 사이의 연관성을 기반으로, 선호도 또는 구매 패턴이 유사한 고객군을 분류하고, 유사한 고객군에 속하는 다른 사람들이 선호하는 상품을 추천하는 방법이다 (Funakoshi & Ohguro, 2000). 협업필터링에 대한 초기 연구는 Tapestry, GroupLens (Resnick et al., 1994) 등의 사례가 대표적이며, Ringo와 Video Recommender 등과 같은 이메일과 웹 기반의 협업필터링 기법에 의한 추천시스템 등이 있다 (Sarwar et al., 2000). 최근에도 협업필터링을 이용한 추천시스템 연구는 계속 되고 있다 (김재경 등, 2002; 김종우 등, 2004; 조운호 등, 2004; 김재경 등, 2005 참고). 협업필터링은 일반적으로 고객들이 동질적인 평가결과를 보이는 상품군에 대해 상대적으로 높은 예측력을 보이며, 데이터가 충분한 경우에는 다른 기법들에 비해 상대적으로 높은 예측력을 보이는 장점을 가지고 있다 (Konstan et al, 1997; Pazzani, 1999). 반면, 고객의 프로필 정보를 활용하지 못하는 점, 충분한 거래 데이터가 확보되어야만 좋은 추천성적을 보일 수 있는 점, 고객들이 이질적인 평가 결과를 보이는 상품군에 대해서는 예측력이 현저하게 떨어질 가능성이 있는 점, 고객군의 크기가 커질수록 많은 연산량이 요구되는 점 등은 협업필터링 기법의 단점이라 할 수 있다 (황성희 등, 2001; 안현철과 한인구, 2002).

한편, 내용기반(content-based) 추천기법은 상품이 지닌 속성을 파악하고 고객의 구매 이력을 바탕으로 고객이 구매한 상품의 속성들과 유사한 상품을 추천하는 방법으로서 제품 정보를 중심으로 추천하는 방법이다 (Balabanovic, 1997; Schafer et al., 1999; Lawrence et al., 2001; Sarwar et al., 2000).

상기의 두 가지 기법은 여러 장점을 가지고 있지만 많은 한계점도 가지고 있다. 내용기반 추천기법의 경우에는 유사한 아이템을 찾아내기 위해서 각각의 개별 아이템의 특성을 명확하게 추출하여야 하는데 이것이 용이하지 않은 경우가 많다. 또 추천의 결과가 주로 사용자가 이미 구매를 했거나 평가를 내린 아이템과 관련된 것으로 결정되는 문제도 존재한다.

한편, 협업필터링의 경우에는 우선 거래 데이터가 분석에 충분할 만큼 확보되어야 한다는 점이다. 또 다른 단점으로는 아무리 충분한 고객 데이터를 확보하였다 하더라도 그 고객이 구매하거나 평가를 내린 아이템이 적은 경우나 아이템의 종류가 너무 많아서 제대로 된 추천결과를 내지 못할 수 있는 가능성 등이다 (Balabanovic & Shoham, 1997). 충분한 거래 데이터가 확보되지 않아서 구매 이력이 없거나 상품선호도가 입력되지 않은 상품들에 대해서는 정확한 추천을 하지 못할 가능성이 있으며, 이런 현상을 희박성(sparsity)의 문제라고 한다(김재경 등, 2002; 조운호 등, 2004; 김재경 등, 2005). 김종우 등(2004)은 희박성의 효과를 직접 분석하여 희박성 정도와 추천의 성과가 역의 관계에 있음을 실증하였다. 희박성의 문제는 중소 규모의 인터넷 쇼핑몰이나 사업 초기단계의 인터넷 쇼핑몰에서는 치명적인 한계점이 될 수 있다. 또 협업필터링은 고객 데이터의 규모가 커질수록 유사성을 계산하기 위한 연산의 양이 기하급수적

으로 증가하는 확장성(scalability)의 문제를 발생시킬 수 있다(김재경 등, 2002; 조운호 등, 2004; 김재경 등, 2005). 확장성의 문제는 해결해야 하는 문제가 제시된 이후에야 추론을 시작하는 ‘게으른 학습방법(lazy learning technique)’의 일반적인 특징으로, 신속한 응답을 원하는 인터넷 사용자의 특성을 감안할 때 고객의 이탈을 유도할 수 있는 치명적인 한계점이다.

규칙기반(rule-based) 추천기법은 사용자의 프로필 데이터, 구매 데이터, 웹로그 데이터 등에 근거하여 조건문 형식의 규칙을 이용한 개인화된 추천을 제공하는 방법이다. 관련 연구로는 지식기반의 전문가 시스템에 기반한 추천시스템을 제안한 Chun et al.(2001)이 있으며, 연관성분석을 활용하여 추출한 규칙을 이용하는 추천시스템을 제안한 김종우와 이경미(2000), Agrawal et al.(1993) 등이 있다. 이 방법은 규칙과의 대응여부를 기반으로 추천을 하게 되므로 추천에 소요되는 시간이 짧고 추천규칙을 직관적으로 쉽게 이해할 수 있다는 장점이 있다. 또한 일단 생성된 규칙은 새로운 규칙이 추가될 때까지 변동이 없으므로 데이터의 증가에 따른 확장성의 문제가 거의 없으며, 규칙을 추출할 수 있을 정도의 데이터가 확보되거나 미리 추출된 규칙을 이용할 경우에는 희박성의 문제가 전혀 발생하지 않는다고 할 수 있다. 반면, 규칙 추출은 주로 전문가의 직관이나 지식에 의존하여 수행해야 하는데, 다양한 상품군이 존재하거나 분석을 위한 데이터가 많고 복잡할 때에는 체계적으로 규칙을 추출하기 어려운 단점이 있다.

2.2 유전자 알고리즘

유전자 알고리즘은 데이터 마이닝의 대표적인 기법 중의 하나라고 할 수 있다. 유전자 알고리즘

은 자연계에 존재하는 개체집단이 적자생존과 자연도태를 통해 다음 세대의 집단을 형성해 가는 과정을 수학적인 알고리즘 형태로 구성한 것이다. 즉, 유전자 알고리즘은 선택, 교배, 돌연변이 등의 과정을 통해 생성된 가능한 탐색공간을 설정하고 전역 탐색을 통해 최적 또는 최적에 가까운 적임자를 찾아 가는 과정을 구현한 알고리즘이라고 할 수 있다.

실제 유전자 알고리즘의 탐색과정은 매우 복잡하다(Goldberg(1989); Adeli & Hung(1995) 참고). 따라서 본 연구에서는 탐색과정을 이해하기 쉽도록 간단하게 요약하여 설명하도록 한다. 유전자 알고리즘은 우선 임의의 값을 가진 초기화된 개체집단을 생성한다. 즉 전체 탐색공간 내에서 임의의 n 개의 개체들을 선택하여 개체집단(population)을 형성하게 된다. 이렇게 생성된 개체집단이 해결해야 하는 문제에 얼마나 적합한지를 적합도 함수(fitness function)를 이용하여 평가한다. 다음 세대의 개체 생성은 평가된 개체집단을 확률적으로 선택(selection)하거나, 교배(crossover)와 돌연변이(mutation) 등의 과정을 통해 이전 세대와 다른 새로운 개체를 생성한다. 다시 적합도 함수에 의해 평가를 하여 종결조건에 도달할 때까지 즉, 목표로 한 적합도 수준에 도달하거나 미리 정해진 최대 진화 수를 넘을 때까지 반복적으로 다음 세대의 개체집단을 생성한다(조희연과 김영민, 2003). 선택된 모든 세대에 대한 평가가 끝나면 모든 세대의 적합도 정도를 파악하고 가장 최적의 적합 정도를 나타낸 세대를 선택하게 된다.

유전자 알고리즘은 전역 탐색을 통해 문제에 대한 최적 또는 최적에 가까운 해를 찾아 주므로 복잡한 경영학적인 문제를 해결하는데 많이 이용되고 있다. 유전자 알고리즘은 적합도 함수를 적절하게 정의하면 특정한 경영학적 현상에 대한 상황이

나 해를 규칙 형태로 추출할 수 있도록 사용할 수 있다. 경영학 분야에서 유전자 알고리즘이 활용된 세부분야로는 주식시장에서의 거래 규칙 추출 (Shin et al., 1999), 기업부도예측(Shin & Lee, 2002), 마케팅 응용(Hurley et al., 1995) 등이다.

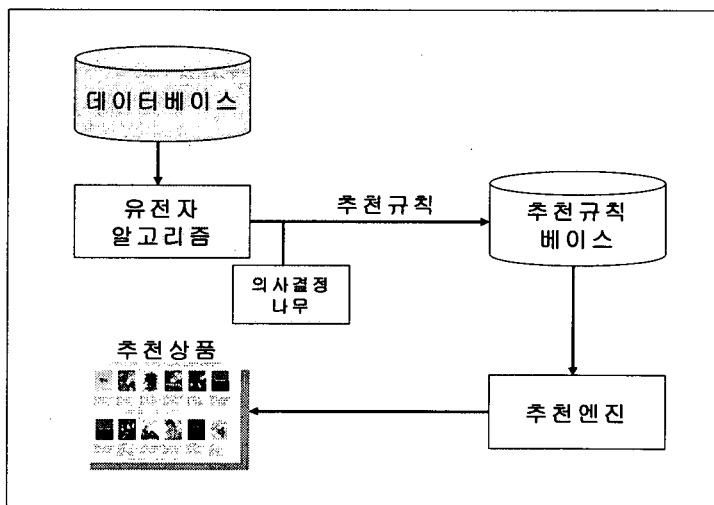
3. 제안 추천시스템

선행연구에서 지적된 협업필터링 등의 한계점을 감안하여 볼 때, 규칙 추출의 어려움을 보완할 수만 있다면 규칙기반의 추천기법이 더 유용할 것으로 기대된다. 따라서 본 연구에서는 유전자 알고리즘을 활용하여 규칙을 자동적으로 생성하고 이를 활용하여 상품을 추천하는 추천기법을 제안한다. 본 연구에서 제안하는 추천기법은 확장성과 희박성의 문제가 한계적인 기존의 추천기법의 한계점을 보완할 수 있는 규칙기반 추천기법의 일종이지만, 데이터로부터의 규칙 생성과정을 자동화한 기법으로 규칙기반 추천기법의 단점을 보완한 방

법이라고 할 수 있다.

한편, 유전자 알고리즘은 최적화 알고리즘으로서 탐색공간을 탐색하여 최적 또는 최적에 매우 가까운 해만을 제공해 준다. 즉, 높은 구매가능성을 가진 고객의 프로필과 그 고객이 구매할 가능성이 매우 높은 상품들을 추천해 주는데, 이 추천의 정도가 매우 확실한 경우에만 규칙을 생성하게 된다. 이는 유전자 알고리즘이 최적화 알고리즘으로서 최적에 가깝지 않은 불확실한 후보 규칙에 대해서는 규칙을 생성하지 않는 특징 때문이다. 그러나 일반적으로 인터넷 쇼핑몰에서는 구매가능성이 높지 않은 고객에 대해서도 상품추천을 해주어야 하므로 본 연구에서 제안하는 모형을 보완하기 위해 의사결정나무를 활용한 추천모듈을 보조적으로 사용하게 된다.

따라서 본 연구에서 제안하는 상품추천시스템은 두 개의 추천 모듈을 가지게 된다. 첫 번째 모듈에서는 유전자 알고리즘을 활용하여 생성된 마케팅 규칙들이 장착되고, 두 번째 모듈은 의사결정나무를 활용하여 생성된 마케팅 규칙들이 장착되



[그림 1] 제안 추천시스템 모형

는데, 후자의 규칙들은 전자의 유전자 알고리즘에 의해 생성된 규칙을 보조하는 목적으로만 이용된다. 즉, 규칙이 충돌하는 경우에는 유전자 알고리즘이 생성한 규칙을 우선적으로 적용하는 것이다.

규칙을 생성하기 위해서 각각의 모듈은 고객 데이터와 거래 데이터를 활용한다. 유전자 알고리즘과 의사결정나무 분석을 통해 도출된 상품 추천 규칙은 모델 기반의 추천 엔진에 통합되어 저장된다. 사용자 자료와 각 모듈 내의 규칙은 순서대로 대응되어 개인화된 맞춤 상품을 추천하게 된다. [그림 1]은 본 연구에서 제안하는 추천시스템의 모형을 보여준다.

4. 데이터 분석

4.1 분석용 데이터

본 연구에서 사용된 데이터는 인터넷 설문조사를 통해 수집되었으며 연구 모형을 통해 상품추천 규칙을 생성하기 위한 데이터로 사용된다. 인터넷 설문조사는 2004년 6월 10일부터 15일까지 총 6일간 대학생 및 일반인들을 대상으로 실시되었으며, 설문 참여자는 개인 신상정보, 맞춤 서비스를 위한 부가 정보, 인터넷 쇼핑몰 이용 현황에 관한 정보를 입력한 후, 8개 장르의 영화 DVD 상품 36개 가

<표 1> 초기 변수군

기본 개인정보 입력내용		
변수명	내용	범주의 수
Gender	성별	2
Age	연령	7
Phone	이동전화	7
Mail	이메일	9
Add	주소	28
개인화 서비스를 위한 맞춤정보 입력내용		
변수명	내용	범주의 수
Edu	학력	8
Occ	직업	20
Income	월소득	7
House 1	주거형태	10
House 2	주거소 유무	9
Mar	결혼 유무	2
Car	자동차 보유 유무	2
Pref	선호 장르	8
쇼핑몰 이용현황정보 입력내용		
변수명	내용	범주의 수
ltime	구매 빈도	5
Item	주 구매/검색 품목	7
Pay	주 결제수단	3

운데 구매를 원하는 상품을 최대 5개까지 선택하게 하였다. 이를 통해 총 134명 고객의 신상에 관한 데이터 및 615건의 거래 데이터를 수집하였다. 인터넷 설문조사를 통해 수집하는 변수들은 안현철과 한인구(2002) 등의 선행연구에서 고객의 성향을 파악하는 데에 유용한 것으로 조사된 것들을 참고하였으며, 수집된 데이터의 변수들은 <표 1>과 같다.

4.2 실험설계

유전자 알고리즘을 수행하기 위해서는 탐색공간 내의 다양한 변수 집합을 염색체(chromosome)라고 불리는 선형 스트링(string)에 사상(mapping)하는 과정이 필요하다. 사상의 방법은 여러 가지가 있는데 해당 사례 내에서 유전자 알고리즘이 효율적인 탐색을 하도록 하기 위해서는 탐색 목적에 적합한 효과적인 사상방법을 찾아야만 한다. 본 연구에서 유전자 알고리즘을 수행하는 목적은 고객 데이터를 활용하여 특정 상품을 구매하는 고객을 예측할 수 있는 추천규칙을 찾는 것이다. 즉, <표 1>에 열거된 16개의 변수와 그 가능한 속성범주의 조합으로 모집단 염색체를 구성하고 유전자 알고리즘 고유의 연산과정인 선택, 교배, 돌연변이 연산과 이에 대한 적합도 평가를 하게 된다. 유전자 알고리즘의 탐색과정을 통해 찾고자 하는 것은 추천규칙이므로 특정 상품을 구매할 고객의 속성과 속성값, 그리고 여러 개의 속성을 연결할 수 있는 "AND" 또는 "OR" 연산자의 가능한 조합들을 탐색하게 된다. 유전자 알고리즘은 탐색공간 내에서 적합도 평가를 통해 적합도가 높은 추천규칙을 생성하게 된다. 유전자 알고리즘 분석을 위해 사용된 분석 도구는 Palisade사(社)의 Evolver 4.0 (<http://www.palisade.com>)이며 선행연구들의 유

전자 알고리즘 실험환경 설정수준을 참조하여 모집단의 크기는 50, 교차율은 0.5, 돌연변이율은 0.06으로 설정하였다. 일반적으로 염색체 내의 유전자(gene)들은 후보해의 모수를 표현한다. 즉, 유전자는 연구자가 최적화하고자 하는 요소의 모수들을 표현하고 이 모수들의 집합이 염색체이며, 가능한 염색체들의 집합이 모집단이 된다. 본 연구에서 염색체는 총 147개의 유전자로 구성된다. 이에 16개 변수들의 가능한 속성범주값 134개를 표현하기 위한 유전자가 131개이고, 16개의 각 변수를 규칙 내에 포함시킬 것인가를 결정하기 위한 유전자가 16개이다. 즉 후자인 16개의 유전자를 통해 각 변수는 선택 또는 제거된다. 그림 2는 염색체 구성을 위한 변수의 사상방법의 예를 나타낸 것이다.

[그림 2]를 보면, 각 변수(성별, 학력, 주 결제수단)에는 다양한 속성범주가 존재하며 각 속성범주는 0과 1의 이진(binary)값 스트링으로 표현된다. 성별, 결혼 여부, 자동차 보유 여부와 같은 변수처럼 변수 내 속성범주가 두 개인 경우에는 하나의 유전자로 구성된다. 즉, 성별이 '1'이면 '남', '0'이면 '여'를 표현하게 되고, 결혼 유무는 '1'이면 '기혼', '0'이면 '미혼'을, 자동차 보유 여부는 '1'이면 '보유', '0'이면 '미보유'를 의미한다. 변수 내 속성이 k 개 ($k \geq 3$, 정수)를 가진다면 그 변수는 k 개의 이진값 스트링으로 표현된다. [그림 3]은 하나의 염색체 구조와 속성값이 입력된 예를 표현한 그림이다.

[그림 3]에서 변수 선택 유무를 표현하는 스트링은 16개의 유전자를 가지고 있는데, 이는 16개의 변수 가운데 어느 변수들을 추천규칙 내에 포함시킬 것인지를 나타낸다. '0'은 해당 변수를 선택하지 않음을 의미하고, '1'은 해당 변수를 선택한다는 것을 의미한다. 각 유전자는 [그림 3]에 열거된 변수의 순서대로 첫 번째 유전자는 '성별'을, 마지막

'OR(internal disjunction)연산'으로 구성된다. 한편, 규칙 조건부에 포함된 각 변수들은 'AND(conjunction)연산'을 수행하게 된다. 따라서 [그림 3]의 염색체는 특정 DVD 상품군을 구매할 확률이 높은 고객의 속성과 속성범주값을 표현하고 있으며, 이는 곧 상품을 추천할 수 있는 규칙을 표현하는 것이다. [그림 3]에서 제시된 염색체 구조의 예를 추천규칙으로 표현하면 아래와 같다.

```
IF    ('성별' = '여성') AND
      (('연령' = '20대') OR ('연령' = '30대'))
      AND (('학력' = '대학교 재학') OR ('학력' = '대학교 졸업') OR ('학력' = '기타'))
      AND (('주 결제수단' = '신용카드 결제'))
THEN "XXX DVD 상품군 추천"
```

4.3 적합도 평가

유전자 탐색 과정에서 필요한 적합도(fitness) 평가의 목적은 특정 DVD 상품군에 대한 구매 또는 비구매에 대한 오분류율을 최소화함과 동시에 특정 상품을 구매할 수 있는 고객들의 속성을 파악하기 위함이다. 본 연구에서 활용된 적합도 함수와 그 평가과정은 아래와 같다.

먼저, 염색체 변이과정을 통해 생성되어 염색체에 표현된 일련의 규칙 r 과 고객데이터 중 하나인 사례 u 의 조건부가 일치하는지를 판단한다. 다음으로 특정 목표상품군 p 에 대하여 고객데이터 중 하나인 사례 u 의 결론부가 일치하는지를 판단한다. 마지막으로 이전 두 단계의 결과가 일치하는 사례들을 추출한다. 즉, 목표상품군 p 를 위한 추천 규칙 r 모두가 고객신상데이터와 거래데이터에 일치하면 'true', 일치하지 않으면 'false'로 판단한다. 적합도함수는 이전 단계에서 'true'인 사례들과

'false'인 사례들 간의 차이를 최대화하는 것이며, 이 함수값을 가장 크게 하는 규칙들을 선택하게 된다.

4.4 의사결정나무 분석

전술한 바와 같이 유전자 알고리즘은 최적화 기법 가운데 하나로서 주어진 문제 영역 내에서 전역 탐색을 통해 최적의 해를 찾는 방법이다. 유전자 알고리즘은 적합도 함수를 최대화하는 소수의 추천규칙만을 생성하게 된다. 이는 유전자 알고리즘이 최적화의 과정을 통해 추천의 확률이 매우 높은 추천규칙들만 생성해 주기 때문이다. 그런데 비록 유전자 알고리즘을 통해 생성된 추천규칙에 적용되지 않는 고객일지라도 해당 상품을 구매할 가능성은 있다. 즉, 상품추천시스템의 경우에는 구매의 가능성이 낮은 고객에 대해서도 추천을 해야만 하는 특성이 있다. 따라서 구매의 확률이 낮지만 구매할 가능성이 있는 경우에는 이에 대해서도 상품을 추천할 수 있는 추천기법을 활용하여 유전자 알고리즘을 보조하게 하는 것이 현실적인 방법이 될 수 있다.

본 연구에서는 유전자 알고리즘 기반의 추천 모듈을 보조하기 위해, 의사결정나무 기반의 추천 모듈을 사용하여 생성된 규칙을 보조적으로 사용하게 된다. 본 연구에서 사용하는 의사결정나무 분석기법은 Gini index 최소화를 분기기준으로 이용하는 CART(classification and regression tree; Breiman et al., 1984) 알고리즘이며, 분석에 사용된 툴은 SAS Enterprise Miner 4.0이다. 분석결과와 과대적합 가능성을 피하기 위해 하나의 의사결정노드로부터 분기되는 최대 가지수는 2개로, 전체 가지의 최대 깊이는 6개 가지까지로 하였다.

장착되어 있다. [그림 4]는 프로토타입의 사용자 신상자료 입력화면이고, [그림 5]는 추천된 상품의 예와 그에 대한 사용자의 만족도 평가화면이다.

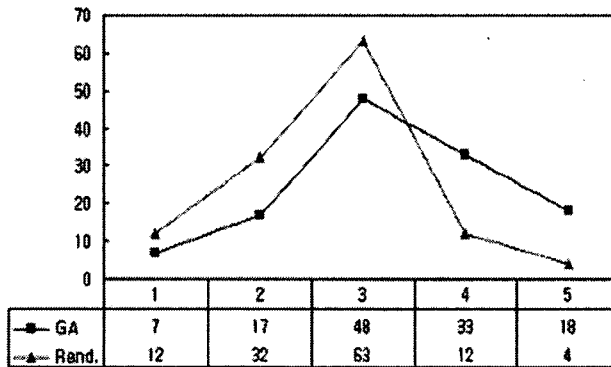
프로토타입 시스템의 개발 환경은 운영체제로 Microsoft Windows XP Professional를 사용하였고, 개발언어는 Microsoft Active Server Pages와 Javascript를, Database Management System은 Microsoft Access 2003을 사용하였다.

5.2 유용성 평가 및 결과

제안 추천시스템에 대한 사용자들의 만족도 평가를 위한 인터넷 설문조사는 2004년 8월 18일부터 22일까지 총 5일간 진행되었으며 123명의 사용

자가 참여하였다. 설문 참여자는 상품 추천을 위해 필요한 자료를 입력한 후, 본 연구에서 제안한 유전자 알고리즘에 기반한 추천규칙에 의해 선정된 상품과 무작위로 선정되어 추천되는 상품에 대하여 각각의 유용성 정도를 5점 척도로 평가하였다. [그림 6]은 본 연구에서 제안하는 방법과 무작위 추천방법에 의해 추천된 상품에 대한 사용자의 만족도 평가 결과를 응답점수 별 빈도수로 나타낸 것이다.

두 추천 기법 간 만족도의 차이에 관한 통계적 유의성 검정을 위해 대응표본 t검정을 실시하였으며 그 결과가 <표 2>에 제시되어 있다. 검증 결과, 두 기법 간 결과의 평균에 관한 차이가 유의수준 1% 내에서 통계적으로 유의한 것으로 나타났다.



[그림 6] 추천기법간 만족도의 정도 평가 분포도

* GA: 유전자 알고리즘 기반 추천시스템에 의한 상품추천 결과, Rand: 무작위 상품추천 결과, 단위: 가로축-만족도에 대한 5점 척도, 세로축-평가결과 (점수별 평가자 명수)

<표 2> 추천기법간 만족도 차이에 대한 통계적 검증 (GA: 제안 알고리즘 추천, RAND: 무작위 추천)

				95% Confidence interval of the Difference		t	Sig. (2-tailed)
	Mean	Std. Deviation	Std. Error Mean	lower	upper		
GA-RAND.	.6016	1.42426	.12842	.3474	.8558	5.685	.000

6. 결론

본 연구에서는 협업필터링과 같이 상품추천시스템에 많이 활용되고 있는 기존의 추천방법의 한계점을 지적하고 이 한계점들을 보완할 수 있는 방법으로 규칙기반의 추천기법이 유용할 수 있음을 주장하였다. 그러나 규칙기반 추천기법의 경우에도 선행연구에서 추천규칙 생성의 어려움이 가장 큰 한계점으로 지적되었다. 본 연구에서는 규칙기반 추천기법의 한계점을 보완하고자 전역탐색 데이터 마이닝 기법 중 하나인 유전자 알고리즘을 활용하여 상품 추천을 위한 마케팅 규칙을 생성하고 모델 기반의 추천시스템의 프로토타입을 구축하였다. 본 연구에서 제안한 모형은 규칙기반 추천기법에서 규칙 생성을 용이하게 해 주는 장점이 있다. 또한 협업필터링 등의 한계점인 희박성과 확장성의 문제도 개선하였다는 점이 본 연구의 의의가 될 수 있다.

그러나 본 연구도 몇 가지 한계점을 가지고 있다. 첫째, 유전자 알고리즘에서 추천하는 규칙이 상품구매의 가능성이 매우 높은 규칙을 생성해 주지만, 상품 구매의 가능성이 있으면서도, 확률이 낮은 고객에 대해서는 적합한 추천규칙이 제공하지 못하므로 의사결정나무 분석을 통해 생성된 추천규칙을 보조적으로 사용하는데, 이로 인해 유전자 알고리즘에 기반한 상품 추천의 순수한 효과를 확인할 수 없었다. 둘째, 본 연구에서 사용된 고객 데이터와 거래 데이터가 실제 상거래에서 발생한 것이 아니라 가상의 쇼핑몰을 통해 수집한 실험실 데이터이기에 실제의 상거래 상황과는 차이가 있을 수 있다.

따라서 향후 연구에서는 유전자 알고리즘 기반의 추천시스템에서도 구매 가능성이 낮은 고객에게 추천규칙을 제공해 줄 수 있도록 적합도 함수

의 수정 등을 시도하는 연구가 필요하다. 또한 실제 상거래에서의 데이터 확보와 이에 적용을 통해 실제 상거래 환경에서의 적용 가능성에 대해서 추가적인 분석이 필요하다.

참고문헌

- [1] 김재경, 서지혜, 안도현, 조윤희, "A personalized recommendation methodology based on collaborative filtering", *한국지능정보시스템학회논문지*, 8권 2호(2002), 139-157.
- [2] 김재경, 안도현, 조윤희, "개인별 상품추천시스템, WebCF-PT: 웹마이닝과 상품계층도를 이용한 협업필터링", *경영정보학연구*, 15권 1호(2005), 63-79.
- [3] 김종우, 배세진, 이홍주, "협업 필터링 기반 개인화 추천에서의 평가자료의 희소 정도의 영향", *경영정보학연구*, 14권 2호(2004), 131-149.
- [4] 김종우, 이경미, "인터넷 상점에서 개인화 광고를 위한 장바구니 분석 기법의 활용", *경영과학*, 17권 3호(2000), 19-30.
- [5] 안현철, 한인구, "데이터 마이닝을 활용한 인터넷 쇼핑몰의 상품 추천 시스템 개발", *한국경영정보학회 춘계학술대회 논문집*, 739-748, 2002.
- [6] 조윤희, 박수경, 안도현, 김재경, "재구성된 제품 계층도를 이용한 협업 추천 방법론 및 그 평가", *한국경영과학회지*, 29권 2호(2004), 59-75.
- [7] 조희연, 김영민, "유전자 알고리즘을 이용한 주식투자 수익률 향상에 관한 연구", *한국지능정보시스템학회지*, 12권 2호(2003), 1-20.
- [8] 황성희, 김영지, 이미희, 우용태, "인구통계학적 특성에 따른 협동필터링 알고리즘의 추천 효율 분석", *한국데이터베이스학회 국제학술대회 논문집*, 362-368, 2001.

- [9] Adeli, H. and S. Hung, *Machine learning: neural networks, genetic algorithms, and fuzzy systems*, Wiley, New York, 1995.
- [10] Agrawal, R., T. Imielinski and A. Swami, "Mining associations between sets of items in massive databases", *Proceedings of the ACM-SIGMOD International Conference on Management of Data*, 207-216, 1993.
- [11] Balabanovic, M., "An adaptive Web page recommendation service," *Proceedings of the First International Conference on Autonomous Agents*, 378 - 385, 1997.
- [12] Balabanovic, M. and Y. Shoham, "Fab: Content-based, collaborative recommendation", *Communications of the ACM*, Vol.40, No.3(1997), 66-72.
- [13] Breiman, L., J. Friedman, R. Olshen and C. J. Stone, *Classification and Regression Trees*, Chapman and Hall, New York, 1984.
- [14] Chun, I. and I. Hong, "The implementation of knowledge-based recommender system for electronic commerce using Java expert system library", *Proceedings of IEEE International Symposium on Industrial Electronics*, 1766-1770, 2001.
- [15] Funakoshi, K. and T. Ohguro, "A Content-based collaborative recommender system with detailed use of evaluations", *Proceedings of the 4th International Conference on Knowledge-Based Intelligent Engineering Systems & Allied Technologies*, 253-256, 2000.
- [16] Goldberg, D. E., *Genetic algorithms in search, optimization, and machine learning*, Addison-Wesley, Reading, MA, 1989.
- [17] Hurley, S., L. Moutinho and N. M. Stephens, "Solving marketing optimization problems using genetic algorithms", *European Journal of Marketing*, Vol.29 (1995), 39-56.
- [18] Konstan, J., B. Miller, D. Maltz, J. Herlocker, L. Gordon and J. Riedl, "GroupLens: Applying collaborative filtering to usenet news," *Communication of the ACM*, Vol.40(1997), 77-87.
- [19] Lawrence, R. D., G. S. Almasi, V. Kotlyar, M. S. Viveros and S. S. Duri, "Personalization of supermarket product recommendations", *Data Mining and Knowledge Discovery*, Vol.5, No.1(2001), 11-32.
- [20] Resnick, P., N. Iacovou, M. Suchak and P. Bergstrom, "GroupLens: An open architecture for collaborative filtering of netnews", *Proceedings of the ACM Conference on Computer Supported Cooperative Work*, 175-186, 1994.
- [21] Pazzani, M. J., "A framework for collaborative, content-based and demographic filtering", *Artificial Intelligence Review*, Vol.13, No.5-6(1999), 393-408.
- [22] Sarwar, B., G. Karypis, J. Konstan and J. Riedl, "Application of dimensionality reduction in recommender systems: A case study", *Proceedings of the WebKDD Workshop at the ACM SIGKDD*, 2000.
- [23] Schafer, J. B., J. Konstan and J. Riedl, "Recommender systems in e-commerce", *Proceedings of the ACM Conference on Electronic Commerce*, 158 - 166, 1999.
- [24] Shin K., K. Kim and I. Han, "Financial data mining using genetic algorithms technique: Application to Korean stock market", *Proceedings of Pacific Asia Conference on Expert Systems*, L.A., 1999.
- [25] Shin, K. and Y. J. Lee, "A genetic algorithm application in bankruptcy prediction modeling", *Expert Systems with Applications*, Vol.23, No.3(2002), 321-328.

Abstract

Product Recommender System for Online Shopping Malls using Data Mining Techniques

Kyoung-jae Kim* · Byoung-guk Kim*

This paper presents a novel product recommender system as a tool for differentiated marketing service of online shopping malls. The proposed model uses genetic algorithm, one of popular global optimization techniques, to construct a personalized product recommender system. The genetic algorithm may be useful to recommendation engine in product recommender system because it produces optimal or near-optimal recommendation rules using the customer profile and transaction data. In this study, we develop a prototype of Web-based personalized product recommender system using the recommendation rules from the genetic algorithm. In addition, this study evaluates usefulness of the proposed model through the test for user satisfaction in real world.

Key words : Product recommender system; Genetic algorithm; Customer relationship management; Data mining

* Department of Information Systems, Dongguk University