

데이터마이닝과 다중모형조합기법을 이용한 온라인상점 상품추천시스템 개발

이연경, 김경재

동국대학교 경영대학 정보관리학과

서울 중구 필동 3-26, 100-715

Tel: 02-2260-3299, Fax: 02-2260-8824, E-mail: jenny@dongguk.edu, kjkim@dongguk.edu

초록

온라인상점의 상품추천시스템은 일대일마케팅의 대표적 실현수단으로써의 가치를 인정받고 있다. 대부분의 상품추천시스템은 시시각각 변화하는 소비자의 기호에 따라 상품을 어떻게 추천할 것인가에 대한 문제에 직면해 있다. 본 연구에서는 급변하는 온라인상점 환경에 탄력적으로 대응하기 위하여 데이터마이닝과 다중모형조합기법을 이용한 상품추천시스템 모형을 제안하고자 한다. 제안하는 상품추천시스템은 현재 운영중인 온라인상점 데이터로 프로토타입을 구축하고 실제 소비자에 대한 적용가능성을 검증하였으며, 그 결과 실제 유용할 것으로 확인되었다.

키워드

상품추천시스템, 다중모형조합기법, 연관규칙, 의사결정나무, 인공신경망

1. 서론

전자상거래의 폭발적 증가로 인해 온라인상점 상품추천시스템의 역할이 점차 확대되고 있다. 상품추천시스템에서의 추천방식은 내용기반 상품추천방식과 협업필터링 상품추천방식이 있으며, 두 방식은 각각의 장단점이 있으나 두 방식 모두가 신규고객에 대한 추천의 어려움과 판매자의 의도를 반영치 못하는 문제가 존재한다. 이러한 문제점을 해결하려는 시도로써 두 방식을 적절히 혼합한 하이브리드 상품추천시스템 방식에 대한 연구가 활발히 진행 중이다. 대부분의 상품추천시스템은 시시각각 변화하는 상품을 어떻게 추천할 것인가에 대한 문제에 직면해 있다. 본 연구에서는 문헌연구를 통해 일반적인 상품추천시스템의 문제점을 파악하여 이를 보완할 수 있는 개선된 상품추천시스템을 제시하고자 한다. 제시하는

모형은 데이터마이닝과 다중모형조합기법을 통해 데이터마이닝 분류기법의 과잉학습 단점을 보완하고 실제 값과 예측 결과 값의 차이를 줄여 모형의 정확도를 높일 것으로 기대된다. 또한 본 연구에서는 실제 온라인상점 구매데이터를 적용하여 제안하는 모형의 실제 적용가능성을 검증할 것이다.

2. 선행연구

본 연구에서는 전술한 바와 같이 데이터마이닝과 다중모형조합기법을 통해 기존 상품추천시스템의 문제점을 보완한 모형을 제시하고 한다. 본 장에서는 먼저 상품추천시스템에 대한 선행연구들을 살펴 보고, 본 연구에서 제안하는 상품추천시스템의 기반 기술인 데이터마이닝 기법을 살펴 볼 것이다. 본 연구에서는 신경망, 의사결정나무, 로지스틱 회귀분석과 같은 전통적인 데이터마이닝 기법이 사용되었지만 이에 대해서는 이미 많은 연구에서 소개하고 있으므로 설명을 생략하고 다중모형조합기법과 연관규칙기법에 대해서만 간단히 살펴 본다.

2.1 상품추천시스템

인터넷을 활용한 상거래가 일반화된 현재에는 인터넷을 통해 제공되는 수많은 자료와 정보를 소비자가 어떻게 활용하는지가 매우 중요한 관심사가 되고 있다. 이러한 관점에서 전자상거래 소비자에게 수많은 자료와 정보를 개인화된 방법으로 제공하는 일대일마케팅에 대한 관심은 점점 높아지고 있다. 일대일마케팅은 고객을 단일객체가 아닌 이종객체로 가정하고 각 고객과의 인터랙티브한 관계를 구축하여 고객에게 지속적인 만족을 주어 고객충성도를 유지하는 마케팅 전략 중 하나이다.(Allen et al., 1998; Pine et al., 1995; Peppers et al., 1999) 본 절에서는 대표적인 일대일마케팅의 도구로 주목 받는 상품추천시스템에 관한 기존 연구를 살펴 본다. 상품추천시스템이란 제품이나

서비스의 고객 맞춤 자동화 도구로서 고객의 취향이나 선호를 바탕으로 고객에게 가장 가치 있는 제품이나 서비스를 찾도록 도와주는 가이드를 의미한다.(Schafer et al., 1999; Sawar et al., 2000) 상품추천시스템 추천방식은 현재 가장 일반적인 협업필터링(collaborative filtering) 추천방식과 내용기반(content based) 추천방식으로 나누어 볼 수 있다. 협업필터링 추천방식은 고객의 구전효과를 자동화하여 활용하는 것으로 유사한 특성을 지닌 다른 사람들의 정보를 이용해서 상품을 추천하는 방식이다.(Konstant et al., 1997; Herlocker et al., 1999; Herlocker, 2000) 한편, 내용기반 추천방식은 고객이 구매결정에 영향을 미치는 제품의 특징을 바탕으로 추천하는 방식이다.(Balabanovic and Shoham, 1997; Lawrence et al, 2001) 현재 많은 연구자들이 내용기반과 협업필터링 추천방식을 결합하는 하이브리드 추천방식에 대한 연구를 활발히 진행하고 있다.(Balabanovic and Shoham, 1997; Sawar et al., 1997; Basu et al., 1998; Good et al., 1999; Kim et al., 2002; 안현철, 2002) <표 1>은 선행연구를 토대로 언급한 추천방식을 간단히 정리한 것이다.(Konstant et al., 1997; Herlocker et al., 2000; Herlocker, 2000; Basu et al., 1998; Good et al., 1999; Sawar et al., 1998; Balabanovic and Shoham, 1997; Lawrence et al., 2001; Kim et al., 2002; 김재경 외, 2002; 안현철, 2002)

<표1> 상품추천시스템 종류

	내용기반	협업필터링	하이브리드
정의	고객 선호를 결정하는 제품 특징을 바탕으로 추천하는 방식	유사한 선호를 가진 사람들의 정보를 사용하여 추천하는 방식	내용기반과 협업필터링을 결합하여 추천하는 방식
장점	신제품 추천에 용이	품질 평가 가능 추천 범위의 확대	
단점	품질 평가의 어려움 과도한 특화가능성 다양한 추천의 불가능	저밀도 데이터의 적용 어려움 신제품에 대한 추천 불가능	협업필터링의 단점을 내용기반의 제품 특징을
한계	판매자의 의도 반영 불가능 신규고객에 대한 추천 어려움	판매자 의도 반영 불가능 신규고객에 대한 추천 어려움	이용하여 보완

김재경 외(2002)는 상품추천시스템에 상품분류체계(Product Taxonomy)를 이용하였다. 이 논문에서는 상품분류체계를 기반으로한 연관규칙과 의사결정나무를 사용하여 협업필터링의 단점을 해결할 수 있는 상품추천방법론을 제시하였다.

김종우 외(1999)는 개인화 광고를 제공할 수 있는 추천 기법에서 상품군별 고객의 선호를 나타내는 선호도 트리를 소개하였다. 선호도트리는 상품분류체계의 나무구조 형태를 도입한 것이다. 김종우 외(2000), Kim et al.(2001)는 또한 연관규칙기법, 의사결정나무를 적용한 개인화 광고 추천기법을 제안하였다. 안현철(2002)은 연관규칙과 분류기법을 이용한 상품추천시스템을 제안하였다. 고객 구매행동 예측 대상물 ‘상품 자체’가 아닌 ‘상품군’으로 결정하였다. 이 논문 연구자는 고객 구매행동예측을 위한 데이터마이닝 분류기법으로 로지스틱회귀분석, 의사결정나무, 신경망을 적용하였다. 상품군 마다 각각의 분류모형을 만들어 동일한 데이터에 대한 모형 예측결과가 가장 우수한 기법을 최종적으로 선택하였다.

2.2 다중모형조합기법(Ensemble Method)

다중모형조합기법은 모형의 성능을 향상시키기 위해 다수개의 모형을 조합하여 최종적인 결론을 내리는 기법이다.(Heskes, 1997) 다중모형조합기법은 여러가지 방법이 있는데 대표적인 방법은 다음과 같다.(Breiman, 1996; Heskes, 1997)

Bumping: 가장 낮은 오차의 결과값을 갖는 모형의 결과값을 사용한다.

Bagging: 모든 모형의 결과값 평균을 사용한다. 동일기법의 다수개의 모형 조합시 각 결과값의 가중치는 고려하지 않는다.

Balancing: Bumping 과 Bagging의 중간형태로서 동일기법의 다수개의 모형 조합시 각 결과값에 가중치를 고려한 평균을 사용한다.

2.3 연관규칙기법

연관규칙기법은 두 객체의 발생빈도가 서로 연관되어 있다고 판단될 때 그 연관정도를 파악할 수 있는 분석기법이다. 특히, 연관규칙기법에서 가장 폭넓게 사용하고 있는 장바구니분석은 상품구매시 일련의 제품이 있다고 할 때 같이 구매되는 다른 일련의 제품이 함께 존재할 것인지를 측정하는 체계적 기법이다. 장바구니분석에서는 지지도, 신뢰도, 향상도의 지표를 이용하여 연관관계 규칙을 분석한다. (Berry et al., 1997; Berry et al, 2000; 강현철 외, 2001)

지지도(Support): 전체거래 중 항목 A와 항목 B를 동시에 포함하는 거래의 비율

$$\text{Support}(A \Rightarrow B) = P(A \cap B)$$

신뢰도(Confidence): 항목 A 를 포함하는 거래 중에서 항목 B 가 포함된 거래의 비율

$$\text{Confidence}(A \Rightarrow B) = \frac{P(A \cap B)}{P(A)}$$

항상도(Lift): 항목 A 를 구매한 경우, 그 거래가 항목 B 를 포함하는 경우와 항목 B 가 임의로 구매되는 경우의 비율

$$\text{Lift}(A \Rightarrow B) = \frac{P(A \cap B)}{P(A) \cdot P(B)}$$

지지도는 전체거래 중 규칙이 얼마나 기여했는가를 나타내며 신뢰도와 항상도는 규칙의 정확성을 나타낸다. 연관규칙기법 분석자는 세가지 지표를 고려하여 의미 있는 규칙을 결정해야 한다.

3. 제안 상품추천시스템

상품추천시스템과 관련한 선행연구를 통해서 상품추천방식에 대한 종류와 장점, 단점 그리고 한계성을 살펴보았다. 고찰된 문제점을 완화시켜 개선된 상품추천방식을 제안한다.

3.1 개요

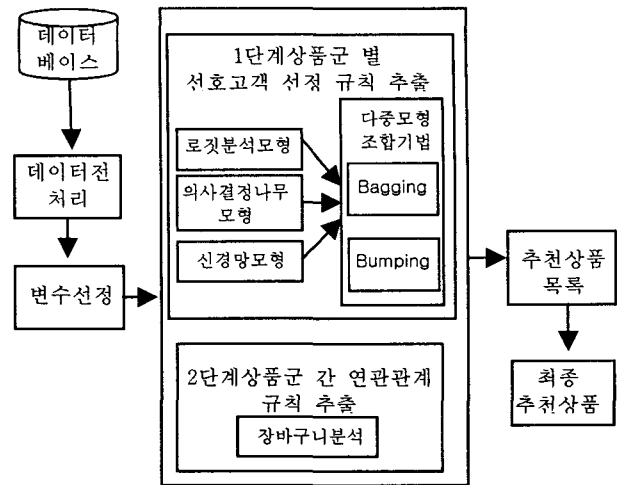
기존연구를 검토해 본 결과 많은 연구자들이 추천 알고리즘을 제안하고 있지만 대부분의 연구가 단일 데이터마이닝 기법을 활용하고 있다. 최근 데이터마이닝 연구들을 살펴 보면 단일기법보다는 다양한 기법의 조합모형이 더 우수한 성과를 보이는 것으로 보고하고 있다.(손소영 외, 1998; 손소영 외, 2000; 이영찬 외, 1999) 또한, 상품의 계층화를 나타내는 상품분류체계를 데이터마이닝 기법과 함께 활용하고 있었으나 그 적용범위가 분류기법에 집중되었다.(김재경 외, 2002; 김종우 외, 1999; 김종우 외, 2000; 안현철, 2002) 선행연구를 통해 나타난 추천 알고리즘의 한계성 및 단점을 보완하여 보다 개선된 추천 알고리즘을 제안하고자 한다.

3.2 제안 상품추천시스템 모형

제안 상품추천시스템 모형은 두 가지 주안점을 두고 개발한다. 첫째, 상품군 마다 고객을 정의함에 있어 로지스틱회귀분석, 신경망, 의사결정나무 분류기법을 이용한다. 세 개의 각 분류모형의 결과를 다중모형조합기법 Bumping 과 Bagging으로 각각 조합한다. 두 개의 조합모형 중 예측력이 우수한 모형을 최종 분류모형으로 선택한다. 둘째, 상품분류체계를 연관규칙기법과 분류기법 양쪽 모두 적용하여 다양하게 활용하고자 한다.

본 연구에서 제안하는 제안 상품추천시스템 모형은

<그림 1>과 같다.



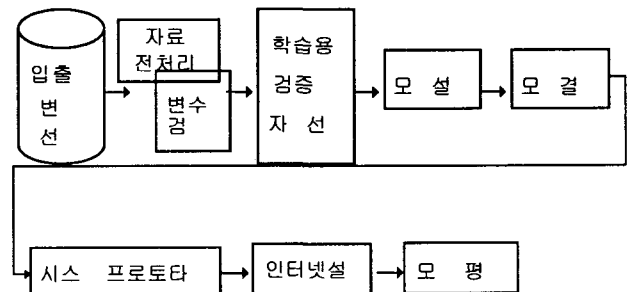
<그림1> 제안 상품추천시스템 모형

4. 실험

전술한 제안 상품추천시스템 모형을 온라인상점 데이터에 적용하여 실제 적용가능성을 검증해 보고자 한다.

4.1 개요

본 연구에서 제안하는 상품추천시스템의 유용성을 확인하기 위하여 실제 데이터를 활용한 간단한 시스템 프로토타입을 구축하고 이를 활용하여 실제와 유사한 환경에서의 추천 효과를 확인한다. <그림2>는 제안 상품추천시스템 구축 과정을 나타낸다.



<그림2> 제안 상품추천시스템 구축절차

4.1.1 실험 대상 기업

이 실험을 위해 실제 인터넷에서 운영되고 있는 온라인상점을 선정하였다. 선정된 I아트샵 온라인상점은 유명 미술관과 박물관의 상품 약 550종을 취급하는 온라인 아트샵이다. 이 온라인상점의 상품은 전체 5개 상품군과 이를 다시 21개 상품군으로 세분화하여 관리되고 있으며 취급상품은 약 550종에 이른다. 5개 상품군은 '패션', '홈데코', '오피스', '포스터', '조각상모형' 이다.

4.1.2 연구데이터

제안 상품추천시스템 모형을 검증하기 위해 사용된 데이터는 I아트샵의 구매데이터로 수집기간은 2001년 10월 20일부터 2003년 9월 25일까지이며 이 기간 중 총 5759건의 구매거래가 이뤄졌다. 실험에 사용하게 될 데이터는 전처리 과정 이전 기준으로 8개 필드, 5759개 레코드이다. 5개 상품군 중 '조각상모형' 군의 레코드는 그 수가 나머지 4개 상품군에 비해 현저히 낮아 실험에서 제외시켰다.

4.2 전처리 및 변수선정

4.2.1 전처리

결측치 및 이상치 처리를 하고 범주형 변수의 경우 통계분석에 적합하도록 더미변수로 변환하는 등의 전처리 과정을 거쳐 최종적으로 58개 필드, 3167개 레코드를 결정하였다. <표2>는 '조각상모형' 상품군을 제외한 I아트샵 상품분류체계이다.

<표2> I아트샵 상품분류체계

1 단계상품군	2 단계상품군	3 단계상품
패션	열쇠고리류	Peace 열쇠고리,,,
	타이류	로댕 넥타이,,,
	셔츠류	짐다인 티셔츠,,,
	우산류	하늘 우산,,,
	가방류	호퍼 솔더백,,,
홈데코	목용용품류	베이비 롱,,,
	인테리어소품류	하링 시계,,,
	자석류	마블 자석,,,
	테이블웨어류	램프란트 컵받침,,,
오피스	카드류	샤갈 카드셋,,,
	메모지류	마티스 큐브,,,
	다이어리류	쿠킹다이어리,,,
	취미용품류	구겐하임 요요,,,
	노트류	고흐 노트,,,
	기타류	설계도면 문진,,,
포스터	포스터류	모네 아침,,,

<표2>에서 보여지는 '1단계상품군'은 고객정의를 위한 '분류모형 적용범위'에 해당되며 '2단계상품군'은 연관규칙추출을 위한 '연관규칙기법 적용범위'가 된다. 최종 상품추천결과는 '2단계상품군' 단위로 생성하게 된다. 즉, 다수개의 상품들로 구성된 추천결과를 생성한다.

4.2.2 변수선정

전처리과정을 거친 데이터셋 변수 중에서 로지스틱회귀분석 분류모형에 사용할 변수의 경우, 연속형변수는 독립표본T검정을 실시하고 범주형변수는 카이제곱검정을 실시한다. 통계검정 실시 결과가 5% 유의수준에서 유의한 변수만을 선정하여 실험을 한다. 상품군별 유의한 변수를 표로 정리하였다. <표3>은 연속형변수 T검정결과를 나타내고 <표4>는 범주형변수 카이제곱검정결과를 정리한 것이다.

<표3> 상품군별 연속형변수 'age'에 대한 T검정결과

var		패션	홈데코	오피스	포스터
age	t	8.354	-7.085	-0.012	-6.517
	df	3165	3165	3165	3165
	Sig(2-tailed)	0.000	0.000	0.991	0.000

<표3>로부터 연속형변수 'age'는 오피스상품군의 구매와는 통계적 차이가 없다는 결과가 도출되었다.

<표4> 상품군별 범주형변수 카이제곱검정결과

var		패션	홈데코	오피스	포스터
Gender	Pearson	1.270	0.151	9.406	9.039
	df	1.000	1.000	1.000	1.000
	Asymp.Sig. (2-sided)	0.260	0.697	0.002	0.000
Address	Pearson	71.567	115.041	81.853	96.332
	df	37.000	37.000	37.000	37.000
	Asymp.Sig. (2-sided)	0.001	0.000	0.000	0.000
Occupation	Pearson	77.895	64.841	38.639	123.454
	df	12.000	12.000	12.000	12.000
	Asymp.Sig. (2-sided)	0.000	0.000	0.000	0.000

<표4>의 범주형변수의 카이제곱검정 결과, 패션상품군과 홈데코상품군에서 'gender' 변수를 제외한 'address, occupation' 변수에 대해 유의하다는 결과가 도출되었고, 오피스상품군과 포스터상품군에서는 'gender, address, occupation' 세 개의 변수 모두가 유의하다는 통계결과를 얻을 수 있었다.

신경망 분류모형의 변수는 독립표본T검정과 카이제곱검정을 단변량분석으로 간주하고 로지스틱회귀분석 분류모형의 결과를 다변량분석으로 간주하여 로지스틱회귀분석 분류모형의 최종 선택된 변수들만을 신경망 분류모형의 입력변수로 활용하였다.

의사결정나무 분류모형 변수의 경우는 소프트웨어가 최적의 입력변수를 선택할 수 있도록 모델을 구축하였기에 변수선택 작업을 하지 않았다.

4.3 모형 구축

제안 상품추천시스템 모형 구축을 위한 실험은 '1단계상품군별 선호고객 분류'와 '2단계상품군 간 연관규칙'으로 나누어진다.

4.3.1 선호고객 규칙 추출

상품군별 선호고객에 대한 규칙을 추출하기 위해서 데이터마이닝 분류기법 중 로지스틱회귀분석, 신경망, 의사결정나무를 적용한다.(손소영 외, 2000; 안현철, 2002) 상품군마다 세 가지 기법의 분류모형을 구축한다. 세 개의 분류모형을 다중모형조합기법으로 조합하여 조합모형 결과값을 비교해 가장 우수한 분류모형을 선택한다. 상품군의 구분은 I아트샵의 상품분류체계를 그대로 도입해 사용하였다.

1단계상품군에 동일 데이터를 적용하여 상품군마다 세 가지 기법의 분류모형을 구축한다. 분류모형 결과값을 Bumping 과 Bagging으로 각각 조합하여 상품군마다 두 개의 조합모형을 만들고 그 결과값을 비교하여 가장 성과가 우수한 분류모형을 선택하고자 한다.

4.3.1.1 데이터셋

모형구축에 사용되는 데이터 구성에 있어서 로지스틱회귀분석과 의사결정나무 분석의 경우, 학습용과 검증용데이터를 8:2비율로 하였으며 신경망 분석의 경우 과적합을 완화하기 위해 테스트용데이터를 포함하여 학습용, 검증용, 테스트용 데이터를 6:2:2의 비율이 되도록

구성하였다. <표 5>는 상품군별 데이터 수를 로지스틱회귀분석과 의사결정나무 기준으로 정리한 것이고 <표 6>은 분류모형별 데이터 분포를 나타낸 것이다.

<표5> 상품군별 데이터셋

상품군	전체 레코드수	학습용 레코드수	검증용 레코드수
패션	1176	941	235
홈데코	405	324	81
오피스	1368	1094	274
포스터	218	174	44

<표6> 분류모형별 데이터셋

분류모형	전체 레코드수	학습용 레코드수	검증용 레코드수	테스트용 레코드수
로지스틱 회귀분석	3167	2533	634	N/A
의사결정 나무	3167	2533	634	N/A
신경망	3167	1901	634	632

4.4.1.2 실험결과

세 개의 분류모형 중에서 각 상품군별 우수모형을 정리해 보면 '패션상품군'에서는 로지스틱회귀분석 분류모형이 가장 좋은 예측결과를 나타냈고, '홈데코상품군', '오피스상품군'에서는 의사결정나무 분류모형 예측률이 좋았으며 '포스터상품군'에서는 신경망 분류모형이 우수했다. 세 분류모형을 조합한 조합모형의 경우는 '포스터상품군'을 제외하고 모든 상품군에서 Bumping조합모형의 예측결과가 높게 나타났으며 다양한 기법의 조합모형 결과가 단일기법의 우수모형 결과보다 고르게 높은 예측률을 보였다. <표7>에서 이를 정리하였다.

<표7> 상품군별 선호고객 분류모형 결과

		패션 (%)	홈데코 (%)	오피스 (%)	포스터 (%)
분류모형	로지스틱 회귀분석	64.67	87.38	56.94	93.06
	의사결정 나무	63.72	87.70	58.04	93.06
	신경망	63.09	87.22	57.89	93.20
조합모형	Bagging	63.81	87.40	57.63	93.10
	Bumping	65.14	88.16	58.13	93.06

4.3.2 상품군 간 연관규칙

선행연구에서는 다수의 연구자가 개별 상품간 연관규칙을 도출하여 상품추천시스템 알고리즘을 구축하였다. 본 연구에서는 상품추천시스템의 한계로 지적된 판매자 의도 반영의 어려움을 해소하고 급격하게 변화하는 환경에 유연하게 대응할 수 있는 방안의 하나로 개별상품 사이의 연관성 보다 상품군 사이의 연관성에 초점을 맞춘 연관규칙 추출을 제안한다. 상품군 간 연관규칙을 도출하여 판매자 의도를 반영할 수 있는 기회를 제공하고 상품구성의 교체와 고객니즈의 변화가 심한 온라인상점에서 탄력적인 상품추천이 가능할 것으로 기대한다.

4.3.2.1 실험결과

연관규칙 결과는 장바구니 분석기법의 세가지 지표 중 지지도(Support)가 높으며 향상도(lift)가 1점 이상인 규칙만을 채택하여 총 규칙 31개를 도출하였다. 이 중 중복되는 규칙을 제거한 후 15개의 규칙이 최종적으로 결정되었다. 15개 규칙 중 11개의 규칙이 '오피스상품군' 내에서 생성되었다. 이는 타 상품군에 비해 비교적 낮은 가격대를 형성하고 있는 '오피스상품군'의 상품을 다수개 구매하는 경우의 빈번함을 추정할 수 있다. 도출된 규칙 15개 중 나머지 4개의 규칙도 '오피스상품군' 과 연관된 것이었다. '오피스상품군과 홈데코상품군'에서 3개의 규칙이 생성되었고 '오피스상품군과 패션군상품군'에서 규칙 하나가 생성되었다. <표8>은 15개 규칙을 정리한 것이다. 도출된 첫번째 규칙은 '오피스상품군의 메모지류를 구입하는 고객은 홈데코상품군의 자석류도 함께 구입한다.'를 의미한다.

<표8> 상품군 간 연관규칙 도출결과

	향상도	지지도 (%)	신뢰도 (%)	해당 건수	도출된 연관규칙
1	2.77	1.84	19.35	24	오피스군 메모지류 → 홈데코군 자석류
2	2.58	2.07	18	27	오피스군 기타류 → 홈데코군 자석류
3	2.54	1.84	16	24	오피스군 기타류 → 오피스군 취미용품류
4	2.51	3.6	23.86	47	오피스군 노트류 → 오피스군 메모지류
5	1.94	1.61	14	21	오피스군 기타류 → 오피스군 카드류
6	1.9	3.3	28.67	43	오피스군 기타류 → 오피스군 노트류

7	1.75	2.68	16.67	35	오피스군 다이어리류 → 오피스군 메모지류
8	1.7	1.61	10.66	21	오피스군 노트류 → 오피스군 취미용품류
9	1.69	1.84	12.18	24	오피스군 노트류 → 오피스군 카드류
10	1.67	4.06	26.9	53	오피스군 노트류 → 오피스군 다이어리류
11	1.57	2.91	25.33	38	오피스군 기타류 → 오피스군 다이어리류
12	1.54	1.69	14.67	22	오피스군 기타류 → 오피스군 메모지류
13	1.5	1.69	10.48	22	오피스군 다이어리류 → 홈데코군 자석류
14	1.45	1.69	10.48	22	오피스군 다이어리류 → 오피스군 카드류
15	1.17	1.76	28.05	23	오피스군 취미용품류 → 패션군 가방류

5. 시스템 프로토타입

5.1 개요 및 구성

본 연구의 제안 상품추천시스템 모형의 추천결과가 어느정도 유용한지를 확인하기 위해 추천결과를 실제 이용자들에게 제공하고 이에 대한 이용자들의 만족도를 평가하기 위해서 시스템 프로토타입을 구축하였다.

시스템 프로토타입은 모두 3단계로 구성된다. 시스템 프로토타입 첫 화면에서 프로토타입에 대한 소개를 하고 화면 하단에 위치한 '추천기반정보입력하기' 버튼을 클릭하게 되면 시스템 프로토타입이 시작된다. 1단계는 상품추천기반정보를 입력하고 2단계는 제안 상품추천시스템 모형의 추천결과를 보여주고 만족도를 평가한다. 마지막 3단계는 무작위로 추천한 결과를 보여주고 만족도를 평가한다. 단, 2단계와 3단계는 랜덤한 순서로 제공된다.

시스템 프로토타입의 개발환경은 다음과 같다.

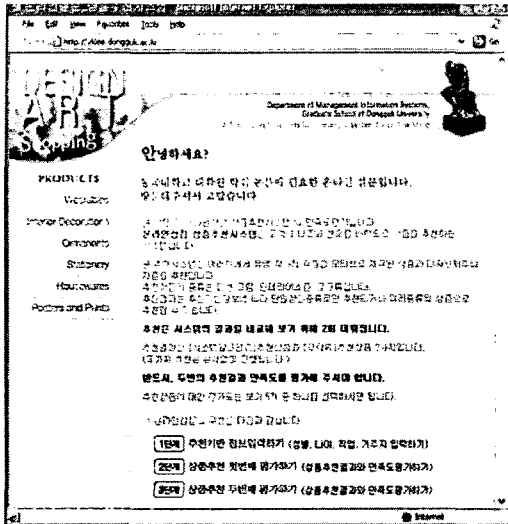
URL: <http://yklee.dongguk.ac.kr>

Operating System: Microsoft Windows X P professional

Language: Microsoft Active Server Page, Javascript

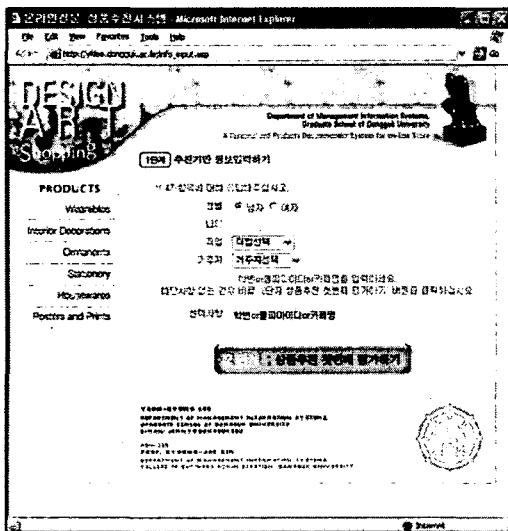
Database Management System: Microsoft Access 2002

<그림3> 은 시스템 프로토타입 첫 화면으로 제안 상품추천시스템에 관한 설명과 3단계로 구성된 시스템 프로토타입에 관한 소개 그리고 온라인설문 절차 및 설문 방법에 대한 소개를 한다.



<그림3> 시스템 프로토타입 개설화면

<그림4>는 상품추천시스템의 추천결과를 생성하기 위한 추천 기반정보를 입력하는 화면이다.



<그림4> 상품추천 기반정보 입력화면

<그림 5>는 이전 화면에서 입력 받은 상품추천 기반정보를 토대로 제안 상품추천시스템 모형에 의한 추천결과를 보여주는 것이다. 상품추천결과는 2 회에 걸쳐 보여지며 두번의 추천결과는 제안 상품추천시스템의 추천결과와 무작위로 추천하는 추천결과이다. 제안 상품추천시스템 모형의

최종결과는 앞서 상품분류체계에서 소개한바 '2 단계 상품군' 형태로 총 20 개의 상품이 추천결과로 보여진다. 1아트샵은 2단계 상품군 마다 최소 80 여 종의 상품을 보유하고 있다. 그렇기 때문에 각 2 단계상품군 마다 '전문가선택', 즉 상품군 별 '도메인 엑스퍼트'가 고객에게 제안하는 상품을 우선순위 기준으로 정한다.

도메인엑스퍼트 제안상품을 우선으로 20 개의 상품을 랜덤하게 추천한다. 상품군에 따라 전문가가 추천한 상품의 수가 20 개 미만일 수도 그 이상일 수도 있다. 비교대상인 '무작위추천'은 전체상품 중 상품군 구분 없이 20 개 상품을 무작위로 추천한다.



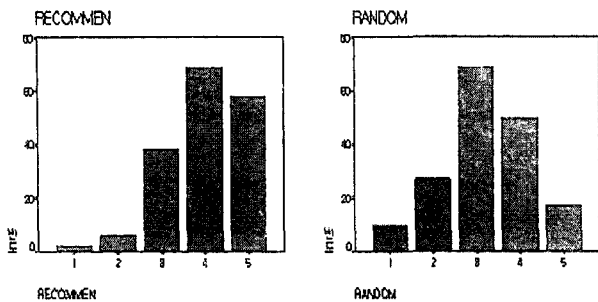
<그림 5> 상품추천 결과화면

5.2 온라인설문

온라인설문은 제안 상품추천시스템 추천결과에 대한 만족도를 평가하는 것이다. 만족도는 5 점 척도로 평가하며 온라인설문대상자는 MSN 메신저사용자, 다음카페회원, P2P 사이트 회원이며 설문기간은 2004년 6월 7일부터 6월 15일까지이고, 총 173 명이 설문에 참여하였다.

5.2.1 온라인설문 결과

총 173 명이 참여한 온라인설문에서 제안 상품추천시스템 추천결과와 무작위 추천결과에 대한 5 점 척도 만족도 조사에서 제안 상품추천시스템 추천결과의 평균은 4.01, 무작위 추천결과의 평균은 3.21 이었다. <그림 6>은 제안 상품추천시스템의 추천결과와 무작위 추천결과에 대한 만족도를 비교한 그래프이다.



<그림 6> 추천결과에 대한 만족도 조사 결과

<그림 6>에서 좌측그래프 'RECOMMEN'은 제안 상품추천시스템 모형이 추천한 결과에 대해 만족도를 평가한 그래프이고 우측그래프 'RANDOM'은 무작위 추천결과에 대한 만족도를 평가한 그래프이다. 'RECOMMEN' 그래프가 'RANDOM' 그래프 보다 우측으로 치우쳐진 형태이다. 이는 만족도 점수의 평균이 높다는 것을 의미한다. 이 그래프를 통해 무작위 추천결과 보다 제안 상품추천시스템 모형의 추천결과가 높은 만족도를 보였음을 직관적으로 판단할 수 있다.

<그림 5>의 결과에 대한 평균값 간의 차이가 통계적으로 유의한지를 확인하고자 대응표본 T검정(Paired Sample T-test)을 실시하였다. 검정결과 유의수준 5%내에서 통계적으로 유의한 차이가 있음을 알 수 있었다. 그 결과는 <표 9>와 같다.

<표9> 온라인설문 대응표본T검정 결과

	t	df	Sig(2-tailed)
RECOMMEN- RANDOM	7.972	173	0.0000

<표9>에서 'RECOMMEN'은 제안 상품추천시스템 모형의 추천결과에 대한 만족도 평가를 나타내고 'RANDOM'은 무작위 추천결과에 대한 만족도 평가를 나타낸다. 두 상품추천 결과에 대한 차이가 유의함으로 제안한 상품추천시스템의 실제 유용성을 확인할 수 있다.

6. 결론

본 논문에서는 선행연구를 통하여 일대일 마케팅의 대표적 실현수단인 상품추천시스템을 고찰하여 기존 상품추천시스템의 보완하기 위해 개선된 상품추천시스템을 제안하였다. 다중모형조합기법으로 고객 분류모형의 정확도를 향상시키고 상품분류체계를 다양하게 활용하였다. 제안 상품추천시스템 모형을 실제 온라인상점 데이터에 적용하여 프로토타입을 구축하여 적용가능성을

검증하였고 제안 상품추천시스템이 추천한 결과에 대해 온라인설문으로 만족도를 평가 하였다. 평가결과 무작위로 추천하는 추천결과에 비해 만족도가 높았다. 이는 제안 상품추천시스템 모형의 유용성을 확인할 수 있는 평가였다.

본 연구는 몇가지 한계점을 가지고 있다. 첫째, 각 단일 분류모형의 결과 차이가 크지 않으므로 보다 정교한 분류모형을 개발할 필요가 있다. 둘째, 각 상품군 별 데이터셋의 구성비율은 동일하나 레코드의 수치가 상품군마다 차이가 있어 데이터셋에 대한 연구가 요구된다. 마지막으로 본 연구의 결과가 보다 일반화 되기 위해서는 다양한 상품군에 대한 적용과 많은 이용자에 대한 만족도 조사가 필요하다.

참고문헌

- [1] 강현철, 한상태, 최종후, 김은석, 김미경(2001), SAS Enterprise Miner 4.0 을 이용한 데이터마이닝:방법론 및 활용, 제3판, 자유아카데미.
- [2] 김재경, 서지혜, 안도현, 조운호(2002), "A Personalized Recommendation Methodology based on Collaborative Filtering," 한국지능정보시스템학회논문지, 8권, 2호
- [3] 김종우, 이경미, 김영국, 유관중(1999), "인터넷 상점에서 개인화 광고 제공 기법", 경영정보학연구, 제9권, 제4호, 12월, pp.107-124.
- [4] 김종우, 이경미(2000), "인터넷 상점에서 개인화 광고를 위한 장비구니 분석 기법의 활용", 경영과학지, 제17권, 제2호, 11월, pp.19-32.
- [5] 손소영, 신형원(1998), "데이터마이닝을 이용한 교통사고 심각도 분류분석," 대한교통학회지, 제16권, 제4호
- [6] 손소영, 이성호(2000), "데이터융합, 앙상블과 클러스터링을 이용한 교통사고 심각도 분류분석," 대한산업공학회/한국경영과학회 2000 춘계공동학술대회 논문집.
- [7] 안현철(2002), "데이터 마이닝을 활용한 인터넷 쇼핑몰의 상품 추천 시스템 개발", 한국과학기술원 석사학위논문, 2002.
- [8] 이영찬, 박수환(1999), "신경망 학습앙상블에 관한 연구," 한국지능정보시스템학회논문지, 5권, 1호
- [9] Allen, C., Kania, D., and Yaekel, B.(1998), Internet World Guide to One to One Web Marketing, New York: Wiley.
- [10] Balabanovic, M and Shoham, Y. (1997), "Fab: Content-Based, Collaborative Recommendation," Communiation of the ACM, Vol. 40, pp.66-72.
- [11] Basu, C., Hirsh, H. and Cohen, W. (1998), "Recommendation as Classification: Using Social and Content-Based Information in Recommendation," Proceedings of the 1998 Workshop on Recommender Systems, pp. 43-52.

- [12] Berry, J., and Linoff, G.(1997), *Data Mining Techniques for Marketing, Sales and Customer Support*, New York: Wiley.
- [13] Berry, J., and Linoff, G.(2000), *Mastering Data Mining, The Art and Science of Customer Relationship Management*, New York: Wiley.
- [14] Breiman, L.(1996), "Bagging predictors," *Machine Learning*, Vol.24, pp. 123-140.
- [15] Good, N., Schafer, J.B., Konstan, J., Borchers, A., Sarwar, B., Herlocker, J., and Riedl, J., Combining Collaborative Filtering with Personal Agents for Better Recommendations. Proceedings of the 1999 Conference of the American Association of Artificial Intelligence (AAAI-99), July, 1999
- [16] Herlocker, J., Konstan, J., and Riedl, J., Explaining Collaborative Filtering Recommendations. Proceedings of the ACM 2000 Conference on Computer Supported Cooperative Work, December 2-6, 2000
- [17] Herlocker, J. Understanding and Improving Automated Collaborative Filtering Systems. Ph.D Dissertation, University of Minnesota, 2000.
- [18] Heskes, T.(1997), "Balancing between bagging and bumping." In Mozer, M. Jordan, and T. Petsche, editors, *Advances in Neural Information Processing System 9*, Cambridge, 1997. MIT Press, pp. 466-472.
- [19] Kim, J.K., Cho, Y.H., Kim, W.J., Kim J.R., Suh, J.H.(2002), "A Personalized Recommendation procedure for internet shopping support," *Electronic Commerce Research and Applications*, pp. 301-313.
- [20] Kim, J.W., Lee, B.H., Shaw, M.J., Chang, H. , Nelson, M.(2001), "Application of Decision Tree Induction Techniques to Personalized Advertisements on Internet Storefront", *International Journal of Electronic Commerce*, Vol. 5, No. 3, Spring, pp. 45-62.
- [21] Konstan, J., Miller, B., Maltz, D., Herlocker, J., Gordon, L., and Riedl, J.(1997), "GroupLens: Applying Collaborative Filtering to Usenet News," *Communication of the ACM*, Vol.40, pp.77-87.
- [22] Lawrence, R., Almasi, G., Kotlyar, V., Viveros, M. and Duri, S.(2001), "Personalization of Supermarket Product Recommendations", *Journal of Data Mining and Knowledge Discovery*, Vol. 5, 2001, pp. 11-32.
- [23] Peppers, D. and Rogers, M. (1999), *The One to One Manager: Real-World Lessons in Customer Relationship Management*, New York: Doubleday.
- [24] Pine B.J., Peppers, D. and Rogers(1995), "Do you want to keep your customers forever?"; Boston:Harvard Business School Press.
- [25] Sarwar, B. M., Karypis, G., Konstan, J. A., and Riedl, J. (2000), "Analysis of recommendation algorithms for e-commerce", *Proceedings of Conference on ACM*, pp. 158-167.
- [26] Sarwar, B., Konstan, J., Borchers, A., Herlocker, J., Miller, B., and Riedl, J. Using Filtering Agents to Improve Prediction Quality in the GroupLens Research Collaborative Filtering System. Proceedings of the 1998 Conference on Computer Supported Cooperative Work. Nov. 1998.
- [27] Shafer, B., Konstan, J. and Riedl, J.(1999), "Recommender Systems in E-Commerce," *Proceedings of Conference on ACM Electronic Commerce*.