Journal of the Korean Data & Information Science Society
2009, **20**(6), 1129–1143

한국데이터정보과학회지

추천시스템을 이용한 이메일 효율성 제고에 관한 연구[†]

김연형[1] · 이석원[2]

[12]전주대학교 여론정보통계학과

접수 2009년 9월 10일, 수정 2009년 11월 21일, 게재확정 2009년 11월 24일

요 약

인터넷 쇼핑몰은 그 특성상 직접 상품을 살펴보기 힘들고 판매자와의 상호작용이 어렵다. 그래서 소비자들은 인터넷상품 구매 시 의사결정에 확신이 부족하거나 절차를 간소화하기 위하여 상품 평이나 추천을 고려한다. 추천의 정교화 및 성과를 높이기 위하여 수 많은 연구가 진행되었으나, 이러한 연구들은 목적을 선정하지 않고 상품간, 사용자간, 협업적 연관성을 바탕으로 진행되어 비슷한 유형을 나열하는 것에 그치고 있다. 그러므로 목적성을 가지는 기업의 캠페인에 바로 적용하기에는 어려움이 존재하였고, 부가적으로 정보를 가공하여 로지스틱회귀모형 등 모형 작업을 실시하는 것이 일반적이었다. 본 논문에서 제안하는 목적성을 고려한 추천은 개인마다 점수를 부여하여 개인화에 따른 추천이 가능토록 하였으며, S주식회사 쇼핑몰의 이메일 캠페인에 적용하여 개봉율, 클릭율, 구매율에 대하여 그 우수성을 증명하였다.

주요용어: 개인화, 로지스틱회귀모형, 목적성, 연관성, 캠페인.

1. 서론

최근 기업들은 그 동안 정보수집에서 얻은 방대한 양의 데이터를 통해 마케팅 수단의 확대와 가치 있는 고객 정보로의 전환을 꾀하고 있다. 하지만 그 활용에 있어 고객의 심리변화와 사회의 다변화로 마케팅 접점을 찾기란 또 많은 시간과 비용을 투자해야 하는 현실적인 어려움이 존재하고 있다. 때문에 가장 접근하기 쉬운 형태인 고객 성향과 행동 패턴을 몇 개의 그룹으로 자사 환경에 맞게 세분화하고 기업의 수익지표를 통해 기여도 측면의 분석을 중심으로 활용하고 있으나 이익집단에만 마케팅 역량을 집중한 나머지 고객의 피로도 상승, 캠페인효과에 대한 반감, 재 구매 유도의 한계점 등 여러 가지 부작용이 나타나게 되었다. 이를 극복하기 위해 포인트 제도를 도입하여 고객을 유입하고, 단기적 매출증가에도 성공을 이루었으나 장기적으로는 수많은 포인트 적립금이 부채로 남을 수 있어 포인트 소진을 위한 적절한 유도가 병행되어야 하는 과제가 존재한다.

따라서, 양질의 데이터와 수익을 극대화 하는 방법으로 기업들이 추구하는 방향은 기존 CRM (Customer Relationship Management) 캠페인 진행시 군집 위주에서 개인별 점수화에 의한 맞춤 개인화 서비스가 주요 쟁점으로 부각되고 있다.

개인화 서비스를 위한 데이터마이닝 방법론들은 연관성 분석이 주로 사용되며, 규칙 탐색의 가장 대표적인 알고리즘으로는 Apriori 방법론이 사용된다 (Agrawal과 Srikant, 1994). 이는 후보 상품집단들의 발생 빈도수를 계산한 후 일정 반응의 지지도를 설정한 상품집단을 계산하는 것이다.

[†] 이 논문은 2009년 전주대학교 학술연구비 지원에 의해 이루어졌음.
[1] (560-759) 전북 전주시 완산구 효자동3가 1200, 전주대학교 여론정보통계학과, 교수.
[2] 교신저자: (560-759) 전북 전주시 완산구 효자동3가 1200, 전주대학교 여론정보통계학과, 겸임교수.
E-mail: leeseokwon@hanmail.net

Apriori 방법론 이후 연관성 분석 연구는 주로 빈도 수량에 가중치를 부여하는 방법, 수행 속도를 줄이기 위한 트랜잭션 (transaction) 감소 기법 그리고 시간 순서로 된 거래 데이터 집합에서 규칙을 찾아내는 순차패턴 연관성 분석이 있다.

가중치를 부여하는 방법은 사용자가 중요하다고 생각되는 상품에 더 높은 확률값을 부여하는 방법(Cai 등, 1998; Yue 등, 2000; Ramkumar, 1997)이다. 이는 주관적 측면이 강하고 사용자에 따라 다른 결과 값을 가질 수 있다는 한계가 존재한다. 그러므로 주관적 측면을 배제하기 위한 연구 방법으로 인공지능을 이용한 퍼지 개념과 Apriori 방법을 합쳐서 연관규칙을 탐색하는 방법도 연구되어 왔다(Shragai와 Schneider, 2001; Hong 등, 1999). 김진규 (2002)는 상품에 가중치를 임의로 부여한 후 수량정보는 퍼지함수를 이용하여 최소가중치 지지도를 만족하는 상품을 연관규칙에 생성하는 정보로 사용하였다.

Apriori 방법의 처리 속도를 줄이기 위한 처리 감소 기법은 Agrawal과 Strikant (1994), Han과 Fu (1995)에 의해 설명되었다. 현재의 알고리즘 유형은 컴퓨터 과부하의 원인인 엄청난 양의 빈발상품들을 자료의 손실 없이 유용한 패턴을 찾아내 예측하고 고객에게 사용할 것인가를 고민하고 있다. 이에 Xin, 등. (2005)은 빈발상품 집합 축약을 위해 군집화기반 접근법을 제시하였고, 이희춘 (2009)는 협력적 필터링 기법을 이용한 선호도 예측 과정에서 이웃의 수와 선호도 예측 정확도와의 관계를 분석하였다.

순차패턴 연관성 분석 알고리즘은 시간의 개념을 적용하여 의미 있는 지식을 탐사하기 위한 방법이다(Agrawal, 1995). 순차패턴이란 일정한 시간 안에서 시간의 흐름에 따라 순차적으로 발생되는 현상을 말하며 시계열성의 자료에서 일반적인 연관성 규칙에서와 같이 지지도, 신뢰도 등에서 우수한 규칙을 찾는 것이 목적이다. 순차패턴 연관성 분석을 위해서는 시간에 대한 정보가 사용되며, 고객의 거래 내역을 추적하기 위하여 고객의 계좌번호, 주민등록번호 등 고객의 정보를 인식할 수 있는 정보와 고객의 거래 정보 등이 필요하다.

이러한 연구결과를 바탕으로 최근에는 정보통신의 발달과 더불어 과거 구매 패턴 위주 연구에서 정보검색, 네트워크 침입 탐지 등으로 다양하게 응용되고 있으며, 인터넷을 통한 쌍방향 의사소통으로 고객의 특성을 파악하고 개개인의 생애가치와 개성을 고려한 개인화 추천 방식이 연구되어지고 있다.

그러나 현재까지 연구되어진 방법론들은 목적성 (target)을 선정하지 않고 상품간의 연관성, 사용자간의 연관성, 협업적 연관성을 바탕으로 선호도 추천을 진행함으로써 비슷한 유형을 나열함에 따라 의미있는 정보를 찾기 위해 재차 정보를 가공하는 단점을 내포하고 있었다 (이석원, 2008).

본 논문에서는 상품추천의 정교화 및 성과를 높이기 위해 목적성을 고려한 개인화 위주의 연관성 분석 기법인 추천시스템을 제안한다. 추천시스템은 고객의 구매주기, 웹 행동, 이메일 등 채널에 대한 수신과 반응, 구매 상품에 대한 행동 데이터만을 사용하며, 반응주기 기반, 반응율 기반, 최근 반응 기반과 통합정보로 구성된다. 이는 기존 수량적인 정보와 시간적인 개념을 반영하지만 반응변수에 목적형 변수를 선정하지 않으면서 개인별 점수화를 도출할 수 있으며, 기존 고객 뿐 만 아니라 신규고객에도 즉각적으로 적용이 가능하다. 또한 고객의 선호도와 정확도가 자동으로 계산되어지기 때문에 모델링에서처럼 일정한 기간마다 모델의 갱신, 수정을 반복적으로 수행하지 않아도 되며, 갱신된 고객 정보를 바탕으로 고객에게 상품을 적극적으로 노출하고, 개인화된 지표를 통하여 기업이 실시하는 마케팅 행위에 실시간으로 적용이 가능하다. 이에 본 논문에서 제안하는 추천방법론과 Apriori 방법, 그리고 로지스틱회귀모형과의 캠페인 결과를 비교하여 제안시스템의 효율성을 보여주는 데 그 목적이 있다.

2. Apriori 방법론과 타겟팅을 고려한 연관성규칙

2.1. Apriori 방법

Apriori 방법은 연관성규칙 탐색의 가장 대표적인 알고리즘으로 데이터베이스에서 매번 모든 자료를 읽어 들여 순차적으로 수행하며 이를 통하여 빈도수를 계산한다. L_k는 빈발 k-상품 집합을 말하며, C_k는 L_k를 생성하기 위한 후보 k-상품 집합이다. 첫 번째 상품들 (L_1)을 도출하기 위하여 C_1의 집합에서 상품 집합의 빈도수를 계산하여 사전 지지도에 일치하는 값만을 도출한다. L_2부터는 도출된 L_{k-1} * $L_k - 1$의 자기조합 (self-join)과 (k-1)에서의 가지치기 (prune) 작업을 거치게 되며, 이때 (k-1)부분 집합이 $L_k - 1$에 존재하지 않으면 그 대상인 후보 C_k집합을 삭제한다.

일회 구매 시 상품목록은 집합 $I = \mathrm{i}_1, \mathrm{i}_2, \cdots, \mathrm{i}_\mathrm{m}$로 주어진다. 임의의 상품집합 T에 대하여 $T \subseteq I$이며, 이 때 T에는 고유한 식별자 TID를 보유하게 된다. 임의의 상품 집합 X가 $X \subseteq T$이면 트랜잭션 T는 X를 포함하며, 또한 지지한다고 한다.

(k-1)번째 발견된 빈발 상품 집합 L_{k-1}은 Apriori-gen 함수를 사용하여 상품 집합 C_k를 생성하는데 이용된다. 다음에 데이터베이스가 읽혀지고 C_k에 있는 후보들의 지지도가 계산되어진다. 그림 2.1은 Apriori 알고리즘과 Apriori-gen 함수 (join과 Prune 단계)를 도식화한 것이다. 예를 들어 총 4회에 걸쳐 100, 200, 300, 400의 고유한 TID를 가지고 상품의 구매가 일어났다고 하자. 각 TID의 구매 이력은 다르게 나타나지만 전체 상품 집합으로 구성할 수 있다. 사용자는 데이터베이스에서 자료를 불러들이기 전에 최소지지도를 설정하게 된다. 본 논문에서는 최소지지도를 임의로 50%(2)를 기준으로 하였다. 이제 첫 번째 데이터베이스에서 자료를 불러들여 한 번이라도 나타난 상품에 대하여 지지도를 계산하고 그 중에서 최소지지도를 만족하는 상품만을 L_1에 진입시킨다. C_1에서는 $\{1\}, \{2\}, \{3\}, \{4\}, \{5\}$의 상품집합을 만들며, 개별 빈발 상품에 대하여 지지도를 계산한다. 각 상품집합의 지지도는 2, 3, 3, 1, 3으로 나타나고 있다. 여기서 상품 {4}의 지지도는 1이므로 최소지지도인 50%(2)를 만족하지 못한다. 따라서 상품 {4}를 제거하고 L_1에는 $\{1\}, \{2\}, \{3\}, \{5\}$ 4개의 상품만을 진입시킨다. L_1의 상품으로 자기조합 단계와 가치치기 단계를 거치는데 {4}를 제거하고 두 번째 상품 집합 C_2를 만든다. C_2는 L_1상품인 $\{1\}, \{2\}, \{3\}, \{5\}$로 자기조합하여 $\{1,2\}, \{1,3\}, \{1,5\}, \{2,3\}, \{2,5\}, \{3,5\}$를 생성한다. 그리고 다시 데이터베이스를 검색하여 최소지지도 50%를 만족하는 상품들 $\{1,3\}, \{2,3\}, \{2,5\}, \{3,5\}$을 L_2로 진입시킨다. 세 번째 상품집단 (L_3)의 과정은 $\{1,3\}, \{2,3\}, \{2,5\}, \{3,5\}$로 자기조합한다. 이 때 $\{1,3\}$의 경우 L_1에서 상품 {4}가 50% 미만으로 제거되었기 때문에 L_3과정에서는 자기조합에 포함될 수 없다. 그리고 $\{2,3\}, \{2,5\}, \{3,5\}$를 가지고 원소가 세 개인 집합을 생성할 경우 C_3는 $\{2,3,5\}$가 생성될 수 있다. 이 때 C_3의 최소지지도가 50%(2)로 나타나므로 L_3에 진입시킨다. 그림 2.2에서 L_4는 $\{2,3,5\}$를 가지고 자기조합하여 $\{1,2,3,5\}$를 얻을 수 있으나 최소 지지도 50%(2)를 만족하지 못한다. 결국 $\emptyset$이므로 알고리즘의 진행을 종료한다. 마지막으로 사용자가 얻을 수 있는 빈발 상품 집합은 L_1 ={1}, {2}, {3}, {5}이며 L_2 ={1,3}, {2,3}, {2,5}, {3,5}, 그리고 L_3 ={2,3,5}가 된다.

2.2. 타겟팅을 고려한 연관성규칙

기존 연구에서 Apriori 방법론은 상품에 대한 관계만을 규명하고 있다. 그리고 수량을 고려하거나 시차를 고려하여 파생된 방법론에 대입하여 Apriori 방법론과 비교하는 연구와 연관성 분석시 시간적 제약을 개선시키기 위한 일련의 작업들이 병행되어지고 있다. 연관성 분석은 목표를 선정하지 않고 상품과의 연관성, 사용자간의 연관성, 협업적 연관성으로 고객에 대한 선호도 추천을 고려한다는 점에서는 우수하나 사용자에게 꼭 필요한 정보를 제공하는 것이 아니라 비슷한 유형을 나열함으로써 불필요한 정보가 포함되어 재차 정보를 가공하여 유용한 정보를 찾아내야 하는 단점이 존재한다. 따라서 이러한 수

1. Apriori 함수
$L_1 = \{large\ 1-itemset\}$
for $(k=2;\ L_{k-1} \neq \varnothing;\ k^{++})$ do begin
$C_k =$ Apriori-gen(L_{k-1}) //new candidate (새로운 후보 상품 집합)
forall transactions $t \in DB$ do begin
$C_t =$subset(C_k, t); //candidates contained on t (후보상품이 빈발 상품집단에 포함)
forall candidate $c \in C_t$ do
c.count++ end
$L_k = (c \in C_k \mid c.count \geq {}^{p}\_\min)$ // 최소지지도를 만족
end
answer= $U_k\ L_k$ // U_k 는 순차적으로 반환되어지는 L_k 집합들

2. Apriori-gen 함수 (join과 Prune 단계)
2.1) 자기조합(join) 단계
insert into C_k
select p.item$_1$, p.item$_2$, …, p.item$_{k-1}$, q.item$_{k-1}$
from L_{k-1} p, L_{k-1} q
where p.item$_1$ = q.item$_1$, … , p.item$_{k-2}$ = q.item$_{k-2}$, p.item$_{k-1}$ < q.item$_{k-1}$

2.2) 가지치기(prune) 단계
forall itemset $c \in C_t$ do forall (k-1)-subset s of c do
if$(s \notin L_{k-1})$ then delete c from C_k

그림 2.1 Apriori 알고리즘과 Apriori-gen 함수 (join과 prune 단계)

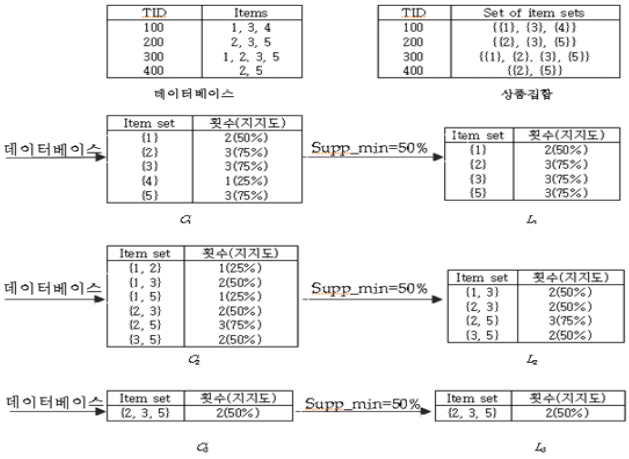

그림 2.2 Apriori 방법론 사례

량적인 정보를 포함하면서 시간적인 개념을 효율적으로 사용하고 소비자를 대상으로 한 마케팅 행위에 실시간으로 적용할 수 있는 방법론을 적용시키고자 한다.

본 논문에서 제안하는 추천시스템 방법은 3가지 방법으로 이루어지며 첫째, 반응주기 기반, 둘째 반응율 기반, 셋째 최근 반응 기반이다. 이들을 적당한 가중치를 두어 설계하고 최종적으로 정확도를 예측하여 고객에게 타겟팅을 전개한다. 이 때 선호도는 일정기간 동안 상품에 대하여 개봉, 클릭, 장바구니, 구매 등 한번이라도 관심을 보인 고객의 수를 전체 고객의 수로 나눈 것이다. 그림 2.3은 각 내용기반에 따른 일련의 처리과정을 정리한 것이다.

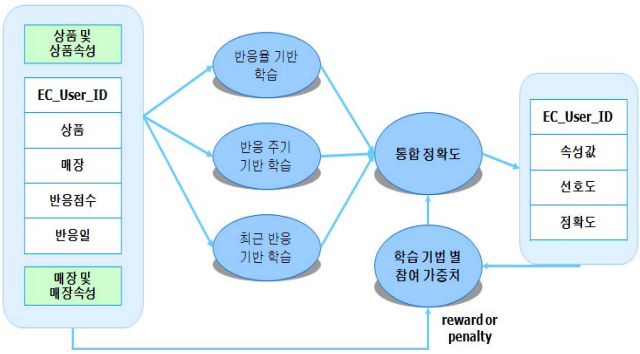

그림 2.3 타겟팅을 위한 기반 학습

2.2.1. 반응주기 기반

반응주기 기반은 속성별 반응주기를 이용한 고객의 선호도 규칙이며 캠페인 집행 시 반응주기에 따라 정확도를 계산한다. 반응주기 기반은 고객의 반응 시점 간격의 평균과 표준편차를 이용하여 생성하며, 반응주기의 정확도가 최소 정확도 미만인 경우는 생성대상에서 제외한다.

반응가능 구간이란 최근 반응시점과 반응의 주기, 반응주기의 표준편차를 고려하여 예측하고 향후 고객의 반응 가능 시점 구간을 말하며, 반응 가능 구간 계산 시 구간의 길이는 반응 주기 표준편차의 2배수로 정한다. 반응주기에 사용되는 수식을 정리하면 그림 2.4와 같다.

1) 반응 주기 계산 : 반응일자 간격의 평균 : $\overline{X_c}$
$\overline{X_c}=(\sum X_j/$웹클릭(방문회수)$-1)$, 여기서 c는 임의의 고객, $j=\{1,2,...,n\}$, n=상품분류

2) 반응 주기의 정확도 계산
- 반응 일자 간격의 표준편차 : σ_c
- 정확도 : $(1-\sigma_c)$ / $\overline{X_c}$ 의 절대값

3) 반응 예상 구간에 사용되는 설정 값
- 최소 반응 구간(Pm) : 표준편차 값이 너무 적을 경우 대비한 최소 반응 구간 값(전체 고객의 반응주기 표준편차의 평균값 설정)
- 반응 기간 배율(Vd) : 반응 구간 산출 시 사용되는 표준편차의 2배수 값.

4) 반응 예상 구간 계산
- 반응 예상 구간 시작 일자 : 최근 반응 일자 + 반응주기 − MAX(Pm/2, Vd*σ_c)
- 반응 예상 구간 종료 일자 : 최근 반응 일자 + 반응주기 + MAX(Pm/2, Vd*σ_c)

그림 2.4 반응주기에 사용되는 수식

예를 들어 표 2.1처럼 웹클릭에서 특정 상품에 대한 임의의 고객 71440에 대하여 패션잡화 상품에 대한 웹클릭 데이터에서 최초 반응일과 최종 반응일 사이 고객 행동 정보가 있다면 다음과 같이 나타낼 수 있으며 각 분류별 반응 일자의 평균을 구할 수 있고, 반응주기에 따라 정확도를 산출할 수 있다.

관심도는 최초 반응일 기준 1회를 (0, 0)으로 설정한 후 2회, 3회 4회를 좌표화 한다면 1회차의 2006년 11월 16일과 2회차의 2006년 11월 28일의 간격은 12일된다. 반응주기 평균($\bar{X}$)은 41.7일이며 또한 반응주기 표준편차 (σ_c)는 36.9로 나타나고 있다. 정확도는 0.86이다. 이 때 1에서 표준편차를 빼 주는 이유는 매일 방문한 경우 분자의 값이 0이 되는 것을 방지하기 위함이다. 반응주기와 그 표준편차를 구하였다면 마케팅 행위를 어느 기간 내에 진행하는 것이 바람직한지 그 구간 값을 구해야 한다.

표 2.1 추출된 자료 (웹클릭)를 근거로 작성된 반응주기

user_id	상품	반응유형	반응일	반응횟수	반응일자구간(X)	반응주기($\bar{X}$)	정확도
71440	패션잡화	웹클릭	2006년 11월 16일	1			
71440	패션잡화	웹클릭	2006년 11월 28일	1	12	41.7	0.86
71440	패션잡화	웹클릭	2007년 2월 20일	1	83		
71440	패션잡화	웹클릭	2007년 3월 22일	1	30		

표 2.2 반응주기에 따른 정확도 예시

user_id	상품	반응유형	반응주기	최초반응일	최종반응일	반응가능시작일	반응가능종료일	정확도
71440	컴퓨터	웹클릭	15.57	2006년 12월 24일	2007년6월 12일	2007년6월 27일	2007년7월 28일	0.55
71440	의류	웹클릭	23.17	2006년10월 11일	2007년3월 15일	2007년4월 3일	2007년5월 21일	0.37
71440	자동차	웹클릭	66.67	2007년2월 20일	2007년6월 2일	2007년4월 26일	2007년10월 2일	0.35
71440	생활가전	웹클릭	39.40	2007년3월 22일	2007년7월 3일	2007년6월 9일	2007년8월 29일	0.18

표 2.2에서 71440인 고객이 패션잡화, 가구/생활용품, 자동차, 생활가전 상품 분류군에 대하여 각각 반응 유형은 웹클릭을 자주하는 것으로 나타났다. 반응주기는 컴퓨터인 경우 15.57로 가장 짧고, 자동차 관련 웹클릭 주기는 66.67로 가장 길게 나타나고 있다. 컴퓨터의 최초 반응일은 2006년 12월 24일이며, 가장 최근 반응일인 최종 반응일은 2007년 6월 12일로 나타나고 있다. 이에 따라 반응주기를 고려한 컴퓨터의 웹클릭 기간은 2007년 6월 27일부터 2007년 7월 28일 사이에 일어날 것으로 예측되며 그 정확도 점수는 0.55로 상당히 높다. 정확도는 0~1사이의 값을 가지며 정확도는 캠페인 진행시 고객별 점수화로 활용된다. 그러므로 예측되는 기간에 컴퓨터 상품 구매에 대한 캠페인을 진행한다면 구매 확률은 상당히 높을 것으로 기대된다.

2.2.2. 반응율 기반

반응율 기반 규칙은 상품 속성 및 캠페인 정보 속성에 대한 고객의 반응율 (이메일 노출 대비 반응횟수)을 기반으로 고객의 선호도를 예측하는 기법으로 그림 2.5는 반응율 기반 규칙에 필요한 수식을 정리한 것이다.

1) 정확도 계산에 사용되는 설정 값
- 월 최소 가중치(Pmw) : 반응에 포함된 기간 중 가장 첫 달 (현재로부터 가장 먼 달)의 가중치. 속성에 관계없이 정해진 값 (임의 설정 값).
- 월 크기(Sm) : 반응이 포함되는 기간의 월의 크기. 속성에 관계없이 정해진 값.

2) 정확도 계산에 사용되는 계산 값
- 월 가중치 증가 비율 (Gm) : 최근 달에 대한 월 단위 가중치 증가 비율. (1−Pmw)/Sm
- 속성 노출 비율 (PA) : 각 속성의 노출 비율. 속성별로 그 값이 다름. 설계의 정확도는 노출 횟수 대비 반응 횟수
- 고객 노출 횟수(Nu) : 이메일 등의 캠페인에 고객이 노출된 횟수. 각 고객마다 값이 다름.

3) 정확도 계산
- 가중치된 반응 횟수 (Crw) : ((Sm − (현재달−반응달))*Gm+Pmw)
- 개인화된 노출 횟수 (Cna) : Nu * PA, − 최종 정확도 : Crw/Cna

4) 선호도 계산(web−log 정보) : 1~5점 사이(ex. 로그인하여 상품카테고리 깊이에 따라 반응점수화)

그림 2.5 반응율 기반 규칙에 필요한 수식

이는 상품별 또는 속성별 반응 강도 학습을 통해 고객의 반응 강도를 추출하게 되며 쇼핑몰 일반에 대한 고객의 반응 강도를 살펴볼 수 있다. 또한 상품 개별에 대한 속성에 대해 개개인의 반응 강도를 추출하는 데 의미가 있다.

반응율 계산 시 월 단위로 최근 반응에 대해 높은 가중치를 부여할 수 있으며 노출이 빈번히 행해지

는 인포메이션 메일 등은 반응율 강도를 약하게 조절할 수 있다. 여기서는 최근 3개월 내에는 가중치를 0.6, 최근 4개월 내에는 0.4, 6개월이 지난 이후에는 0.05로 가중치를 일률적으로 조정하였다.

표 2.3 추출된 자료 (이메일)를 근거로 작성된 반응율 예

user_id	상품	반응유형	반응일	반응횟수	월단위 가중값	노출빈도(Pa)	반응율
5845	패션잡화	이메일	2006년 11월 16일	1	0.05	0.32	0.33
5845	패션잡화	이메일	2006년 11월 28일	0	0.05	0.25	0
5845	패션잡화	이메일	2007년 2월 20일	1	0.4	0.17	0.33
5845	패션잡화	이메일	2007년 3월 22일	1	0.6	0.31	0.25

설정된 월의 크기는 최근 6개월을 기준하였으므로 최대 6이 된다. 표 2.3에서는 월 마다 전부 반응을 하였으므로 월 단위 크기를 계산하면 3이 된다. 또한 노출 비율은 각 이메일에 대한 쇼핑몰의 노출 비중을 백분율로 계산한 것이다. 이 때 월 가중치 증가 비율 값은 0.1333 ((1-0.6)/3)이 된다. 속성 노출 비율은 0.31, 고객 노출 횟수는 0.25로 설정되었다면 패션잡화에 대한 최종 정확도는 4.5가 된다. 그리고 월 단위 감소 값은 M+1월 달에 계산 값과 M월에 계산 값의 차이를 말하며 월 단위 감소 값은 증가 또는 감소될 수 있다. 월단위 가중치를 달리 두는 이유는 반응 최근성에 무게를 두기 위함이며, 노출횟수는 고객마다 다르기 때문에 개인별 노출과 캠페인 유무에 따라 달라지게 된다.

표 2.4에서 고객 번호 (user-id)가 5845인 고객이 각 상품 카테고리별 선호도와 정확도를 나타낸 것이다. 반응율에 따라 가방/지갑/벨트의 선호도는 2.857, 정확도는 0.034이다. 이 때 선호도 값은 평균적으로 상품에 대한 클릭은 자주하고 있으나 장바구니와 구매까지는 미흡하다는 것을 의미한다. 그렇지만 반응횟수가 7회로 나타나고 있어 관심은 매우 높다고 볼 수 있고, 월 단위 감소 값도 0.095로 크게 줄지 않고 있어 이 고객에게 가방/지갑/벨트에 대한 캠페인을 진행 시 할인쿠폰이나 적립금 두 배 등의 가격 조건을 두게 된다면 구매로 쉽게 전환될 것으로 판단된다. 가전/통신기기 군의 경우 월 단위 감소 값도 크게 줄고 있고, 반응횟수도 3회이며, 선호도 값도 1로 크지 않아 이미 구매를 하였거나 관심이 없는 것으로 판단할 수 있다.

표 2.4 반응율에 따른 정확도 예시

user_id	상품	선호도	정확도	반응횟수	월 단위 감소값
5845	가방/지갑/벨트	2.857	0.034	7	0.095
5845	패션	2.750	0.019	4	0.054
5845	가전/통신기기	1.000	0.297	3	0.615

2.2.3. 최근 반응 기반과 통합 정보

최근 반응 기반 규칙은 이전 통합 반응 예측 값 계산 시 반응주기로부터 산출된 값을 이용했던 고객, 신규로 반응주기가 산출된 값이 선호도 계산에 이용될 고객, 최근 반응정보가 통합 반응 예측 값의 계산에 포함된 고객을 모두 대상으로 추출한다. 그리고 직전 최근 특정기간 내 반응한 고객과 최근 특정기간 내 반응한 고객 모두를 추출한다.

임의의 컨텐츠가 N개의 속성 $A_1, A_2 \cdots, A_n$을 가지고, 속성 A_i는 m개의 서로 다른 속성 값 $a_{i1}, a_{i2}, \cdots, a_{im}$을 가질 때 속성 A_i에서의 속성 내 가중치 R_i는 r_{ij}을 가지고 속성 $R_i = r_{i1}, r_{i2}, \cdots, r_{im}$로 정의 된다. 이 때 각 속성 값 a_{ij}에 대응하는 속성 내 가중치 r_{ij}는 다음과 같다.

$$r_{ij} = \frac{k_{ij}}{\sum_{p=1}^{m} k_{ij}}, \quad \text{여기서}, I = 1, 2, \cdots, n, \;\; j = 1, 2, \cdots, m$$

k_{ij}는 사용자가 속성 값 a_{ij}를 열람한 횟수, r_{ij}는 사용자가 열람한 전체 컨텐츠 중에서 각 속성 A_i의 속성 값 a_{ij}가 출현한 상대도수를 나타낸다.

속성간 가중치 d_i는 속성 집합 A_i들과 속성 집합 A_j에 대한 상관관계이며 $D_i = \mathrm{d}_1,\ d_2, \cdots, d_n$이다. 속성 A_i에 대한 속성 가중치 d_i는 다음과 같다. 이 때

$$d_i = \mid y_i - \frac{1}{m}\sum_{j=1}^{m} r_{ij} \mid$$

여기서 $j = 1, 2, \cdots, m$, $i = 1, 2, \cdots, n$, m은 속성수, y_i는 속성내 가중치의 최대값으로 표시할 수 있으며, 각 상품에 대한 빈도수에 따라 속성 내 가중치를 계산하였고, 속성 간 가중치를 구하였다. 즉, 상품분류 디지털/MP3/휴대폰에서 각 상품은 MP3, 휴대폰, 디지털카메라 상품에 대하여 각각 빈도수가 14, 4, 2로 나타났다고 하면, 속성 내 가중치는 전체 빈도수에 각 상품의 빈도수를 나누어 속성 내 가중치를 구하게 된다. 디지털/MP3/휴대폰의 전체 빈도수는 20이므로 MP3의 속성 내 가중치는 0.7로 나타나게 되고, 각 상품에 대한 속성 내 가중치는 0.37로 계산된다.

표 2.5 컨텐츠 분류간 가중치 계산

분류	Cust_id 행동			
컨텐츠 속성	속성값(a_{ij})	빈도수(k_{ij})	속성내 가중치(R_i)	속성간 가중치(D_i)
디지털/MP3/휴대폰	MP3	14	0.7	0.37
	휴대폰	4	0.2	
	디지털카메라	2	0.1	
컴퓨터/주변기기/게임	컴퓨터	10	0.5	0.17
	주변기기	6	0.3	
	게임	4	0.2	
패션잡화	신발	12	0.6	0.10
	가방	8	0.4	

분류별 가중치가 계산되어지면, 다음과 같이 개개인의 최근 반응 기반 정확도를 계산한다. 고객의 상품 구매 접근도에 따라 이메일 개봉은 1점, 캠페인 페이지로의 이동은 2점, 상세 상품정보 클릭은 3점, 장바구니는 4점, 구매는 5점으로 점수를 부여한다.

$$\sum_{p=1}^{m} D_i \times \text{점수},$$

여기서, 점수는 반응정보 (1∼점)의 평균 점수 값.

신규 정확도에 대한 기존 정확도의 적용율은 기존 정확도에서 캠페인 실시 결과인 반응횟수/타겟 횟수로 나타나며 이를 가중치화 한다. 그리하여 갱신된 가중치를 구하게 된다. 갱신된 가중치를 구하는 방법은 다음과 같다.

$$\frac{(Ppr + Pcr * w)}{(1 + w)}$$

여기서, Ppr은 기존 정확도, Pcr은 신규 정확도, w는 가중치.

마지막으로, 반응주기 기반, 반응율 기반, 최근 반응에 기반한 통합정보를 구성하였다.

통합정보는 정보들을 효과적으로 하나의 지표로 생성하여 고객마다의 고유한 점수화를 부여하기 위함이다. 통합정보는 3가지 정확도 값에 각각의 w를 구할 수도 있으나 본 논문에서는 평균값을 사용한다. 통합정보에 사용되는 계산 값은 그림 2.6과 같다.

1) Ppfr : 반응율 기반 정확도
2) Prpl : 반응 주기 기반 정확도
3) Prrl : 최근 반응 기반 정확도
4) 최종 정확도 : (Ppfr + Prpl + Prrl)/3

그림 2.6 통합 정보 산출 정확도 계산에 사용되는 계산 값

표 2.6은 통합정보에 의한 정확도 예시이다. 고객 번호 (user-id)가 7324인 고객의 가전/통신기기, 주방용품, 가구/생활용품 상품 분류군에 대하여 선호도와 정확도를 나타낸 것이다. 통합정보에 따른 선호도는 2.134이며, 정확도는 0.080이다. 또한 반응횟수는 67회로 매우 높다. 그러나 주방용품의 선호도는 3으로 가전/통신기기 보다 관심 측면에서는 높게 나타나고 있으나 실제 반응 횟수는 9로 낮은 수치를 보이고 있다. 이에 캠페인을 수행한다면 이 고객에게는 가전/통신기기 상품군에 먼저 구매를 유도하는 이메일 등을 보내게 된다.

표 2.6 통합 정보에 의한 정확도 예시

user_id	속성값	선호도	정확도	반응횟수
7324	가전/통신기기	2.134	0.080	67
7324	주방용품	3.000	0.010	9
7324	가구/생활용품	2.750	0.009	8

3. 실증 분석

본 연구에서는 국내 S사의 쇼핑몰에서 2005년 1월에서 2005년 12월까지의 자료를 사용하여 추천시스템에 필요한 자료를 분석하였다. 그리고 본 논문에서 제안하는 추천시스템의 효율성 검증을 위하여 2006년 2월부터 몇 차례에 걸친 이메일 캠페인 진행 결과를 통해 기존 Apriori 방법론과 로지스틱 회귀모형의 이메일 발송 후 개봉율, 클릭율, 구매율에 대한 비교 분석을 실시하였다. 또한 1년이 지난 2007년 2월 이후 몇 차례에 걸친 캠페인 결과를 동일하게 비교하였다.

S사는 국내의 온·오프라인 제휴 가맹점에서 고객이 사용한 금액에 따라 포인트 점수를 부여받고 자사 사이트에 포인트를 통합하여 일정점수에 도달하면 해당 가맹점에서 점수만큼 상품을 무료로 구입하거나 다양한 제휴사에서 서비스를 이용할 수 있다. 또한 자체 쇼핑몰에서도 활용이 가능한 특징이 있으며 상품분류는 다음과 같다.

본 논문에서 제안한 반응주기 기반, 반응율 기반, 최근 반응 기반의 3가지 방식으로 생성한 추천시스템과 Apriori 방법과의 이메일에 대한 개봉율, 클릭율, 구매율 차이를 검증하기 위하여 대분류 중 가전/통신 (2월 7일, 4월 4일), 가구/생활 (2월14일, 4월13일)에 대하여 각각 2회씩 총 4차에 걸쳐 이메일 발송을 시행하였다.

비용과 마케팅 효과에 따라 발송할 인원을 미리 정하고 추출방식은 효과를 확실하게 증명하기 위하여 Apriori 방법을 먼저 시행한다. 이후 나머지 인원을 대상으로 추천시스템에서 추출하는 형태를 취하였다. 추출조건 (2월 7일, 2월 14일)은 최근 3개월간 이메일 개봉율이 0%초과 인원을 대상으로 통합 정

표 3.1 상품 분류

대분류	중분류	소분류
가전/통신기기	생활가전, 사무가전, 계절가전, 영상가전 등	면도기, 다리미, 커피메이커, 복사기, 선풍기 등
가구/생활용품	가구, 침구/수예, 인테리어, 가정용 등	책장, 침대, 조명기구, 책상, 액자, 정리함, 공구류 등
패션잡화	시계, 신변잡화, 신발, 가방, 귀금속 등	가방, 지갑, 핸드백, 귀걸이, 목걸이 등
의류	여성, 남성, 아동, 스포츠, UNISEX 등	캐주얼, 정장, 남아, 여아, 유아용 의류 등
스포츠/레저/취미	스포츠용품, 레저용품, 헬스용품,	스키, 자전거, 물놀이, 등산, 골프, 낚시 등
주방용품	조리용품, 주방도구, 용기, 주방잡화 등	보관용기, 팬, 냄비, 주전자, 수저, 제과 등
출산/육아/아동	육아용품, 완구, 육아수유, 출산용품 등	캐리어, 브랜드 완구, 학습완구, 발육기구 등
자동차	전기제품, 인테리어, 편의용품, 시트/매트 등	공기청정기, 타이어, 컵홀더, 매트리스 등
컴퓨터	주변기기, 게임기, PC부품, 소모품 등	용지, 메인보드, 마우스, 소프트웨어 등
천냥하우스	천냥하우스	천냥하우스
기타	기타	기타

표 3.2 이메일 캠페인 추출 인원 및 결과

개봉/클릭율	방법론		
	추천시스템	Apriori 방법	공통
평균 이메일추출(백만)	10.0	11.8	2.0
평균 이메일발송성공(만명)	8.2	10.9	1.9
평균개봉율	19.9%	38.2%	47.9%
평균클릭율a(클릭인원수/발송성공인원수)	6.4%	5.5%	12.9%
평균클릭율b(클릭인원수/개봉인원수)	32.3%	14.3%	26.9%
평균 거래금액(백만)	5.6	2.4	3.4
평균 구매율(%)	0.13	0.05	0.28

확도 순이며, 이후 추출 방법 (4월 4일, 4월 13일)에서는 개봉율 0%인 휴면고객도 포함시켰다. 이는 추천시스템이 기존 반응자 중에서만 효과가 나타나는지 여부와 함께 휴면고객에 대한 활성화에도 기여하는지 판단하기 위해서다. 이메일 수신을 발송성공 인원수와 발송통수 대비로 구분하고 분석 반응기간은 이메일 발송 이후 5일을 기준으로 하였다. 이는 이메일 발송 후 5일 이후에는 이메일 개봉율이 0.1%정도로 저조하게 나타나고 있기 때문이다.

발송 결과를 살펴보면 본 논문에서 제안한 추천시스템이 기존추출방식인 Apriori 방법론에 비하여 평균클릭율a (클릭인원수/발송성공인원수)에서는 6.4%, 평균클릭율b (클릭인원수/개봉인원수)는 추천시스템이 32.3%로 기존 Apriori 방법론 보다 높게 나타나고 있으며, 이는 본 논문에서 제안하는 추천시스템이 '관심'을 갖고 있거나 '거래 가능성'이 높은 구매 가망고객들을 효율적으로 추출한다는 것을 알 수 있다. 또한 구매율 기준으로 살펴보았을 때 본 논문에서 제안한 추천시스템이 약 2.4∼3배 정도의 구매효과가 더 높은 것으로 판단되며, 구매금액은 약 2.3∼3배 정도의 차이가 발생하고 있다.

포인트몰은 고객들이 쇼핑몰에서 물품을 구매하거나 온·오프라인 제휴사에서 모은 포인트를 소진할 수 있도록 별도의 상품을 구성한 것이다. 표 3.3에서 1월∼4월 까지는 기존 Apriori 방법을 사용한 결과이고, 5월은 본 논문에서 제안하는 추천시스템만을 사용한 결과이다. 결과에서 살펴보면 5월의 경우 발송횟수는 24회로 이전 4개월 평균치 정도이며, 총 발송통수는 2,219,834로 가장 적게 이메일 발송을 한 1월에 비해서도 약 30%정도 줄어든 수치이다. 하지만 본 논문에서 제안한 추천시스템의 평균 클릭율 (클릭건수/개봉자수)은 38.3%로 이전 4개월 평균 28.9%보다 9.4%포인트 높게 나타나고 있다.

구매율을 살펴 본 결과 본 논문에서 제안하는 추천시스템만을 사용한 5월은 발송 대비 구매율이 기존 Apriori 방법을 사용한 이전 1월∼4월 평균구매율 보다 0.25%포인트 증가하였다. 또한 구매금액이 약 8.2억원으로 나타나고 있지만 이메일 발송이 평균 57% 감소된 것을 고려한다면 약 6∼7억원 정도 구매금액을 상승시킬 수 있을 것이라 생각된다.

표 3.3 포인트몰 월별 결과

발송월	발송횟수	총발송통수	평균개봉율	평균클릭율	거래금액(억)	거래인원(천)	구매율(%)
1월	17	3,046,542	28.3	32.3	8.6	1.70	0.21
2월	24	4,860,348	23.1	26.6	7.2	1.48	0.14
3월	27	5,379,167	24.3	30.0	7.9	1.66	0.13
4월	24	4,906,044	23.2	26.6	8.4	1.80	0.20
1월∼4월(평균)	23	4,548,025	24.7	28.8	8.0	1.66	0.16
5월	24	2,219,834	12.6	38.3	8.2	1.76	0.41

본 논문에서 제안하는 추천시스템은 기존 Apriori 방법론과의 이메일 클릭율, 구매율에서 우수한 효과를 보이고 있는 것으로 판단된다. 그러나 기존 데이터마이닝에서 구축하는 모델과의 이메일 반응 차이를 살펴보기 위하여 로지스틱회귀모형을 사용하여 모형화 작업을 실시하였다. 사용된 변수는 Kim과 Lee (2008)의 변수를 따라 진행하였다. 사용한 변수는 S사의 고객 기본 정보, 고객 속성 정보, 카드 사용 정보, 매출 정보 등 약 100여개 변수이다.

로지스틱회귀모형에서 목적성 변수는 구매여부 (target)로 구매자는 1, 비구매자는 0의 값을 가지며, 설명변수 중 범주형 변수와 연속형 변수 중 구간을 나눈 변수들로 더미 변수를 생성하였다 (고봉성 외, 2009). 종속변수가 1과 0값만 취하는 이항 변수이므로 로지스틱회귀모형을 자료에 적합 시키기 위해서 분석용 데이터에 PROC LOGISTIC 문을 사용하였다. 로지스틱모형의 구성은 "종속변수 (Target) = 독립변수들" 이며 옵션은 내림차순으로 종속변수가 1인 확률모형을 추정하기 위한 것이다. 설정된 모형에 포함된 많은 변수들 중에서 유의한 변수를 선택하기 위해서 단계별 선택법을 사용하였고 Slentry와 Slstay 옵션은 변수 추가 또는 제거 시 유의수준을 정하는 것으로 0.1로 정하였다. 단계별 선택법은 각 단계마다 변수를 모형에 추가적으로 포함시키거나 제외시키기 위해서 전진 선택법과 후진 선택법을 혼합하여 사용하는 방법으로 각 단계마다 모형에 추가 혹은 제외될 변수, 각 변수에 대한 카이제곱 검정통계량 값과 대응되는 P값을 확인하였다. 추정된 회귀계수를 이용하여 구매 고객이 될 확률이 큰 고객을 다음과 같이 예측하였다.

$$\begin{aligned}\widehat{Y} = &-4.276 + (0.000006961 * \text{가용포인트}) + (-0.2911 * \text{서울}) + (-0.2127 * \text{부산}) \\ &+ (-0.0119 * \text{아파트사이즈}) + (-0.00000255 * \text{누적포인트}), \cdots, +(-0.2273 * \text{대전})\end{aligned}$$

예측된 로지스틱회귀모형에 따라 본 논문에서 제안한 추천시스템과 로지스틱회귀모형을 통해 도출된 최종 모형, 그리고 Apriori 방법을 이용하여 기존의 쇼핑몰 이용 고객들을 대상으로 이메일 개봉율, 클릭율, 구매율에 대한 비교 평가를 2차 (3월 25일, 5월 17일)에 걸쳐 실시하였다. 그리고 향후 1년이 지난 후 1차 평가를 실시하였다 (3월 20일).

1차 캠페인 꾸러미 몰은 일정한 금액은 정해져 있으며 고객이 3개의 상품을 선택하여 구매하는 몰이다. 본 논문에서 제안한 추천시스템은 이메일 발송을 위해 추출한 고객 10.8만 명 중 모델과 약 30% 정도 중복되는 것으로 나타났다. 각 방식별 고객 점수는 0.39 이상의 분포를 보이고 있어 추천시스템과 모델 모두 개별 스코어 생성에서 높은 점수대의 대상고객을 추천하고 있는 것으로 나타났다.

이메일 발송결과로부터 제안하는 추천시스템이 Apriori 방법, 모델보다 개봉율, 클릭율, 구매율에서 높은 결과를 보이고 있다. 또한 전체 쇼핑몰 구매율에서도 추천시스템은 0.23%로 나타나 여타 모델보다 약 4∼5배 차이를 보이고 있다.

다음으로 추천시스템과 Apriori방법, 추천시스템과 모델, Apriori방법과 모델, 그리고 추천시스템, Apriori방법과 모델의 중복고객에 대한 결과를 살펴보면 3가지 방법의 중복고객이 개봉율 60.3%, 클릭율 17.1%, 꾸러미 몰 구매율 0.88%, 전체쇼핑몰 구매율 0.41%로 개봉율과 클릭율은 가장 높게 나타나고 있다. 추천시스템과 모델의 중복고객과 비교하면 꾸러미 몰의 구매율과 전체쇼핑몰 구매율에서 각각

표 3.4 1차 캠페인 - 꾸러미 몰

세분화	고객수(명)	구성비(%)	모델 점수	추천시스템 점수	개봉율	클릭율	꾸러미몰 구매율	전체쇼핑몰구매율
추천시스템	69,261	25.7	0.29	0.46	27.4	6.2	0.59	0.23
Apriori	87,615	32.5	0.21		18.3	3.2	0.13	0.07
모델	67,468	25.0	0.63		0.3	0.1	0.48	0.04
추천&Apriori 방법	5,080	1.9	0.29	0.44	20.1	7.2	0.28	0.16
추천&모델	31,636	11.7	0.65	0.40	57.4	17.0	0.98	0.46
Apriori&모델	6,873	2.5	0.63		56.2	15.3	0.79	0.32
추천&Apriori&모델	1,929	0.7	0.64	0.39	60.3	17.1	0.88	0.41
합계	269,862	100						

0.98%와 0.46%의 결과를 보이고 있다. 이 결과에서 살펴보았듯이 결국 기업 입장에서 수익을 가져다주는 것은 구매율이므로 추천시스템을 주축으로 모델에서 나온 고객들을 보조화하여 사용하는 것이 적절할 것으로 판단된다.

2차 캠페인 할인클럽 몰은 일정한 금액은 정해져 있으며 고객이 2개의 상품을 선택하여 구매할 수 있는 몰이다. 본 논문에서 제안한 추천시스템은 이메일 발송을 위해 추출한 고객 6.7만 명 중 모델과 약 18% 정도 중복되는 것으로 나타났다. 각 방식별 고객 점수는 0.39 이상의 분포를 보이고 있으며, 꾸러미 몰 보다는 각각 0.04∼0.07%포인트 낮게 고객 점수화가 형성되고 있다.

표 3.5 2차 캠페인 - 할인클럽

세분화	고객수(명)	구성비(%)	모델 점수	추천시스템 점수	개봉율	클릭율	할인클럽몰 구매율	전체쇼핑몰구매율
추천시스템	38,590	11.7	0.30	0.45	39.0	13.2	0.63	0.42
Apriori	206,839	62.9	0.21		20.1	4.3	0.18	0.09
모델	46,860	14.3	0.67		0.2	0.1	0.76	0.06
추천&Apriori	16,353	5.0	0.31	0.41	36.0	11.3	0.43	0.29
추천&모델	8,591	2.6	0.68	0.32	69.2	29.1	1.71	1.00
Apriori&모델	8,083	2.5	0.67		65.0	23.0	1.16	0.59
추천&Apriori&모델	3,351	1.0	0.67	0.32	68.2	27.9	1.28	0.60
합계	328,667	100						

이메일 발송 결과 개봉율은 추천시스템이 39.0%로 Apriori 방법 20.1%, 모델 0.2% 보다 높게 나타나고 있으며 클릭율은 추천시스템 13.2%, Apriori 방법 4.3%, 모델 0.1%이며 할인클럽 몰에 대한 구매율은 각각 0.63%, 0.18%, 0.76%를 보이고 있다. 결국 추천시스템이 Apriori 방법, 모델보다 개봉율, 클릭율은 높게 나타나고 있으나 모델에 비해서는 약 0.13%포인트 낮은 결과이다. 그러나 전체 쇼핑몰 구매율에서는 추천시스템이 0.42%로 나타나 약 4.3∼7배 차이를 보이고 있다.

다음으로 추천시스템과 Apriori방법, 추천시스템과 모델, Apriori방법과 모델, 그리고 추천시스템, Apriori방법과 모델의 중복고객에 대한 결과를 살펴보면 추천시스템과 모델의 중복고객이 개봉율 69.2%, 클릭율 29.1%, 할인클럽 몰 구매율 1.71%, 전체쇼핑몰 구매율 1.00%로 가장 높게 나타나고 있다. 결국 할인클럽 몰에서는 다음 캠페인 진행시 개봉율, 클릭율, 구매율에서 효율성을 높이기 위해 추천시스템과 모델 중복고객을 1차 대상으로 추출하는 것이 바람직할 것으로 판단된다.

3차 캠페인 꾸러미몰에서 제안한 추천시스템은 이메일 발송을 위해 추출한 결과 9.9만명 중 모델과 약 19% 정도로 중복율이 11% 포인트 정도 줄어든 것으로 확인되었다. 이는 모델의 노후화 (고봉성 외, 2009)로 중복율이 적어진 경우로 판단된다.

이메일 발송 결과 추천시스템이 Apriori 방법, 모델보다 개봉율, 클릭율, 구매율에서 월등히 높은 결과를 보이고 있는 것으로 판단되며, 또한 전체 쇼핑몰 구매율에서도 추천시스템은 0.51%로 나타나 여타 모델보다 약 5∼10배 차이를 보이고 있다. 그러므로 지속적인 매출을 유지하기 위해서는 추천시스템을

표 3.6 3차 캠페인 - 꾸러미 몰

세분화	고객수(명)	구성비(%)	모델 점수	추천시스템 점수	개봉율	클릭율	할인클럽몰 구매율	전체쇼핑몰구매율
추천시스템	68,342	16.2	0.43	0.51	40.2	16.5	0.62	0.51
Apriori	259,845	61.7	0.23		17.9	3.9	0.13	0.10
모델	53,308	12.7	0.45		0.1	0.1	0.32	0.05
추천&Apriori	14,843	3.5		0.49	34.6	10.1	0.35	0.53
추천&모델	10,364	2.5		0.42	50.4	23.2	1.85	0.77
Apriori&모델	8,983	2.1			34.6	18.4	1.23	0.37
추천&Apriori&모델	5,432	1.3	0.61	0.43	61.9	34.2	1.08	0.79
합계	421,117	100						

보조할 수 있는 주기적인 모델의 갱신이 필요할 것으로 생각된다.

4. 결론

제안한 타겟팅을 고려한 추천시스템을 실제 적용한 결과 기존 Apriori 방법론으로 뽑은 대상자에 비해서 이메일 반응율은 상대적으로 약 2배 정도 높게 나타나고 있으며 클릭율은 약 2.7배 그리고 구매율은 약 5배 정도의 차이를 보이고 있다. 기존 데이터마이닝 모델 기법 중 로지스틱회귀모형을 통한 추천시스템과의 비교에서는 추천시스템 고객 집단이 모델 고객 집단 보다 해당몰에서 반응율이 매우 높게 나타나고 있었다. 그러나 해당몰만 비교한다면 추천시스템이 우수하지만 이메일 대상자가 전체몰에서 구입하는 구매 반응율은 모델쪽이 우수하게 나타나고 있었다.

상품추천, 이미지 추천을 하기 위한 연관성 분석의 효율성 제고뿐만 아니라 개별 목적에 따라 데이터마이닝 모델링을 통해 모델을 생성해야 하는 번거로움, 모델 생성 마다 소요되는 장기적 시간 (약 3~6개월), 모델생성을 위한 전문가 집단 의뢰에서 자유로울 수 있다. 이에 따라 환경이 수시로 바뀌는 기업 캠페인 활동에 유연하고, 즉각적 실행이 가능하며 결과의 분석을 통한 향후 캠페인의 설계가 가능하다. 또한 개인화 점수에 따른 다각적 마케팅 활용이 가능할 것으로 사료된다.

그러나 제안한 타겟팅을 고려한 추천시스템도 기존의 알고리즘 보다 세부적인 정보를 제공하기 때문에 시간적 측면은 고려하지 못하였다. 또한 초기 사이트 분석을 통하여 가중치 및 기간에 대한 표준을 정해야 하는 부분이 상존하고 있으며, 상관관계를 통하여 두 품목 간 관계성을 바탕으로 다른 상품에 대한 유추는 어느 정도 가능하겠지만 직접적인 세 품목 간, 네 품목 간 등으로 이어지는 관계 규명은 향후 연구되어 져야 할 과제이다.

참고문헌

고봉성, 이석원, 허정 (2009). 생명보험사 텔레마케팅 효율성 제고에 관한 연구. <한국데이터정보과학회>, **4,** 673-684.

김진규 (2002). <수량과 가중치를 고려한 퍼지 연관규칙 탐색 방법>. 석사학위논문, 한양대학교, 서울.

이석원 (2008). <데이터마이닝의 association rule 기법에 관한 연구>. 박사학위논문, 전주대학교, 전북.

이희춘 (2009). 협력적 필터링 추천기법에서 이웃 수를 이용한 선호도 예측 정확도 향상. <한국데이터정보과학회>, **3,** 505-514.

Agrawal, R. and Srikant, R. (1994). Fast algorithms for mining association rules. *Proceedings of the 20th VLDB Conference*, Santiago, Chile.

Agrawal, R. and Srikant, R. (1995). Mining sequential patterns, *In Proc. 1995 Int. Conf. Data Engineering (ICDE'95)*, 3-14, Taipei, Taiwan, Mar.

Cai, C. H., Fu, W. C., Cheng, C. H. and Kwong, W. W. (1998). Mining association rules with weighted items. *In proceedings of the international Database Engineering and Applications Symposium*, 68-77.

Han, J. and Fu, Y. (1995). Discovery of multiple-level association rules from large database. *In Proc. 1995 Int. Conf. Very Large Data Bases (VLDB'95)*, 420-431, Zurich, Switzerland, Department.

Hong, T. P., Kuo, C. S. and Chi, S. C. (1999). Mining association rules from quantitative data. *Intelligent Data analysis*, **5**, 363-376.

Kim, Y. H. and Lee, S. W. (2008). An empirical study on telemarketing business. *Journal of the Korean Data & Information Science Society*, **3**, 877-891.

Ramkumar, G. D., Ranka, S. and Tsur, S. (1997). Weighted association rules: Model and algorithm, http://www.cs.ucla.edu/ czdemo/tsur/.

Shragai, A. and Schneider, M. (2001). Discovering quantitative association in databases. *In proceedings of IFSA World Congress and 20th International Conference on NAFIP*, 423-428.

Xin, D., Han, J., Yan, X. and Cueng, H. (2005). Mining compressed frequent-pattern sets. *In Proc. 2005 Int. Conf. Very Large Data Bases (VLDB'05)*, 709-720, Trondheim, Norway, Aug.

Yue, J. S., Tsang, E., Yeung, D. and Shi, D. (2000). Mining fuzzy association rules weighted items. *In proceedings of International Conference on Systems, Man and Cybernetics,* **3,** 1906-1911.

Journal of the Korean Data & Information Science Society
2009, **20**(6), 1129–1143

한국데이터정보과학회지

A study on email efficiency on recommendation system[†]

Yon-Hyong Kim[1] · Seok-Won Lee[2]

[12]Department of Public Survey and Applied Statistics, Jeonju University

Received 10 September 2009, revised 21 November 2009, accepted 24 November 2009

Abstract

This paper proposes a recommendation system (Association Rule System for Targeting) which considers target which is not considered by previous Logistic Regression system, and proves that the efficiency of the recommendation system is better than that of the current and previous Apriori algorithm system. Also this study shows that the click and purchasing rate of the proposed Association Rule System for Targeting is much higher than those of current Apriori algorithm system after the purchasing campaign even though the open rate of the former is lower than that of the latter. In comparison with Logistic Regression methodology, this paper proves with experimental data that the purchasing effect of the proposed system for specific items is much higher in accuracy than that of current Apriori algorithm system even though the purchasing rate of current Apriori algorithm system is higher in whole shopping malls than that of the proposed Association Rule System for Targeting.

Keywords: Apriori algorithm, association rule, campaign, regression model.

[†] This research was supported by the research fund of Jeonju University, 2009.

[1] Professor, Department of Public Survey and applied Statistics, Jeonju University, Jeonju 560-759, Korea.

[2] Corresponding author: Adjunct Professor, Department of Public Survey and applied Statistics, Jeonju University, Jeonju 560-759, Korea. E-mail: leeseokwon@hanmail.net