Journal of the Korean Institute of Industrial Engineers
Vol. 33, No. 3, pp. 339-348, September 2007.

고객의 행동 변화를 통한 신규고객 세분화와 구매항목 예측

도희정† · 김재련

한양대학교 산업공학과

New Customer Segmentation and Purchase-forecasting Using Changes in Customer Behavior

Hee Jung Do · Jae Yearn Kim

Department of Industrial Engineering, Hanyang University, Seoul 133-791

Since the 1980s, the marketing paradigm has rapidly changed from product-driven marketing to customer-driven marketing. Recently, due to an increase in the amount of information, customer-differentiation strategies have been emphasized more than product-differentiation strategies.
This paper suggests a methodology for new customer segmentation and purchase forecasting using changes in customer behavior. This methodology includes a segmentation method for new customers using existing customer's characteristics and a purchase-forecasting system using the purchase-behavior patterns of existing customers. The proposed methodology not only provides differential services from a segmentation system but also recommends differential items from the purchase forecasting system for new and existing customers.

Keywords: Clustering, Segmentation, Purchase forecasting

1. 서 론

정보의 디지털화 및 네트워크화로 인한 제품 및 서비스의 정보 공유속도와 정보량이 증가하면서 제품 및 서비스간의 차별화 우위 전략이 점차 한계에 이르게 되었고, 이로 인해 제품 위주의 마케팅에서 고객 위주의 마케팅으로 빠르게 변화되고 있다.

오늘날 마케팅 전략에 있어서 고객은 중요한 자원이다. 그러므로 회사는 높은 가치를 가진 고객들을 보유하고 새로운 고객을 획득하는 것이 중요한 문제가 되기 때문에 모든 고객들에게 똑같은 타깃을 둔다던가, 또는 같은 서비스를 제공한다는 것이 무의미해졌다.

따라서 회사는 고객 개개인의 요구 또는 구매 행동을 파악한 후, 각 개인에게 맞는 마케팅 전략을 세우며 제품 차별화 보다는 고객 세분화를 통하여 각 고객들의 욕구에 맞게 차별화할 수 있는 고객 차별화 전략을 강조하고 있다.

기존고객이 신규고객보다 더 이익이 된다(Ha *et al.*, 2002). 마케팅 전문가에 의하면 대부분의 기업들이 연간 평균 25%의 고객을 잃어버리며 기업의 입장에서 새로운 고객을 이끌어 유치하는 것은 기존 고객을 보유하는 것 보다 10배나 더 많은 비용을 초래하기 때문에 기존 고객을 붙잡아 두는 것이 기업 입장에서는 이익이 된다고 한다.

이러한 이유로 이전의 연구들에서 기존고객의 이탈방지 및 유지관리를 위한 연구가 많이 되어 왔다. 대부분의 연구에서는 기존고객들의 일정기간 동안 구매활동을 통해 다른 고객들과의 차별화된 등급을 부여하여 각 고객들의 성향에 맞는 차별화된 마케팅을 함으로써 기존고객들의 이탈방지와 유지관리를 하고자 하였다.

하지만 신규고객의 경우는 구매활동을 처음 시작하는 고객들로 구성되어있기 때문에 구매활동을 통한 다른 신규고객들과의 차별화를 둘 수가 없으므로 신규고객들 모두에게 동일한 마케팅을 하고 있다.

본 논문에서 제안하는 신규고객 세분화는 신규고객이 기존고객이 되기 전 단계로서 신규고객으로 등록되었을 때부터 기존고객들처럼 차별화된 등급을 두어 각 등급의 특성에 맞는

† 연락저자 : 도희정, 133-791 서울특별시 성동구 행당동 17번지 한양대학교 산업공학과, Fax : 02-2296-0471,
E-mail : hyhjd4@hanyang.ac.kr
2006년 07월 접수; 2007년 01월 수정본 접수; 2007년 05월 게재 확정.

차별화된 마케팅을 통해 신규고객들을 유지 관리하고 조기 이탈을 방지함으로써 고객확보 및 유지 비용절감을 이룰 수 있는 효과를 얻고자 한다.

따라서 본 논문에서는 고객 차별화의 일환으로, 시변성을 가지는 기존고객들의 특성을 통해 신규고객을 세분화하여 차별화된 마케팅 전략을 제공하며 또한 기존 고객들의 구매 항목들의 변화를 파악하여 미래의 구매력을 예측하여 신규고객들 뿐만 아니라 기존고객에게 각 등급에 맞는 항목들을 추천한다.

2. 기존 연구

고객 세분화를 이용한 마케팅이나 CRM 등에 적용한 연구가 많이 선행되어 왔다. 세분화를 성공적으로 이끌 수 있는 중요 요소는 세분화 하는데 사용되어지는 변수들의 선택이다. 세분화에 사용되어지는 변수들은 크게 일반적인 변수 즉, 고객정보 또는 라이프스타일과 상품에 관련된 특정 변수 즉, 고객 구매 행동과 의도로 나누어진다(Tsai and Chiu, 2004). 대부분의 연구는 일반적인 변수를 사용한 세분화 방법을 제안했는데 Tsai and Chiu(2004)는 상품에 관련된 특정 변수에 기반을 둔 세분화 방법을 제안했다.

고객의 욕구와 구매행동은 항상 같은 것이 아니라 시간의 흐름에 따라 계속 변화한다. 그러므로 이런 고객 세분화가 한 기간 내에서 이루어진 경우는 그 기간에서만 유효할 뿐 시간이 지나면 무의미해지는 것이 당연한 결과이다. 시간이 지남에 따라 고객들의 세그먼트가 변하는 것을 고려하기 위해 각 고객의 세그먼트의 이동변화를 관찰함으로써 고객의 행동 방향을 파악하여 분석하였다(Ha *et al.*, 2002).

시간 변화에 따른 고객의 패턴에 대한 연구로 Min and Han(2005)은 추천시스템의 성과를 향상시키기 위해 고객의 시간변화에 따른 패턴을 찾는 방법을 제안하였다.

RFM(Recency, Frequency, Monetary)은 고객 세분화 테크닉에 가장 빈번하게 사용되어지는 방법이다. Recency는 고객이 최근 구매한 날로부터 얼마나 지났는지를 측정하는 항목이고, Frequency는 정해진 기간 내에 각 고객이 얼마나 많이 구매했는지를 측정하는 항목이며, Monetary는 각 고객이 구매 시 평균적으로 얼마나 많은 돈을 지불했는지를 측정하는 항목이다(Park *et al.*, 2000). 고객의 RFM 값이 높을수록 회사입장에서는 그 고객의 가치는 높은 의미를 가진다. Ha *et al.*(2002)와 park *et al.*(2000)는 고객 세분화에 RFM을 이용하였다.

데이터마이닝 기법들은 다양한 응용 분야에 이용되어져 왔다. 그 중에서도 클러스터링은 고객 세분화에 주로 사용되어지며, K-means, hierarchical, fuzzy c_means, SOM 등 많은 클러스터링 알고리즘들이 이용되고 있다. K-means clustering은 일반적으로 데이터를 K개의 그룹으로 분할하는 방법이다(Liu and Shih, 2004). 대부분의 기존연구에서 K-means와 SOM을 이용한 세분화 방법들이 제안되었다.

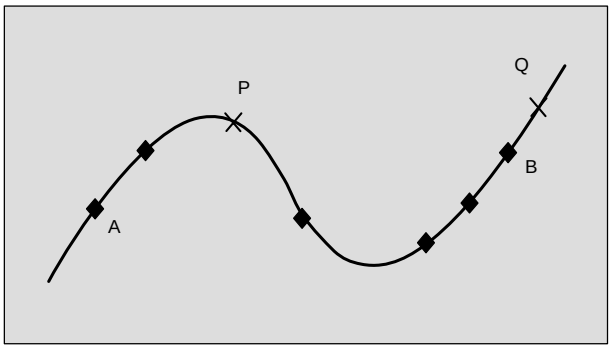

Figure 1. Extrapolation

예측기법으로는 Regression model, Exponential and Box and Jenkins models, Neural networks, Fuzzy systems 등이 있다.

그 중 보외법(Extrapolation method)은 빠르고, 쉽게 사용할 수 있으며 재고와 상품 예측 등에 널리 사용되어지는 방법이다(Scott Armstrong, 2000).

<Figure 1>에서 보는 바와 같이 보외법이란 곡선 위의 2점 A, B와 이 2점으로 한정된 부분 위에 몇 개의 점을 알고 있을 때, A, B로 한정된 부분 위의 다른 점 P의 위치를 추정하는 보간법(interpolation)에 대하여 A, B로 한정된 밖의 부분의 점 Q의 위치를 추정하는 것을 말한다.

3. 방법

본 연구는 고객 차별화의 일환으로, 시변성을 가지는 기존고객들의 특성을 이용하여 신규고객을 세분화하는 방법과 기존고객들의 구매행동패턴을 분석하여 미래의 구매력을 예측하여 기존 고객뿐만 아니라 신규고객에게 차별화된 마케팅 전략을 제공하기 방법들 제안한다.

<Figure 2>는 본 논문에서 제안하는 전체적인 모델로서 각 기간별로 RFM(Recency, Frequency, Monetary)을 변형한 CFM (Continuation, Frequency, Monetary)을 기준으로 고객들을 클러스터링(clustering) 한 후, 각 클러스터링에 등급을 부여하여 시간의 변화에 따른 고객들의 그룹 잔류변화를 파악한다.

고객들의 그룹 잔류변화를 통해 첫째 기존고객들의 특성을 파악하고, 둘째 이를 통해 신규고객 세분화를 위한 신규고객 특성 기준을 찾는다. 그리고 이 특성을 기준으로 신규고객들을 세분화한다. 셋째 고객들의 그룹 잔류변화에 따른 구매 항목들의 변화를 파악하여 미래에 구매되어질 항목을 예측하여 같은 그룹의 신규고객들에게 예측된 항목을 추천한다. 위 세 가지 분석을 통해 고객들에게 차별화된 마케팅 전략을 수립하는데 도움을 주고자 한다.

3.1 기존고객 세분화

기존 고객의 특성을 찾기 위해 RFM(Recency, Frequency,

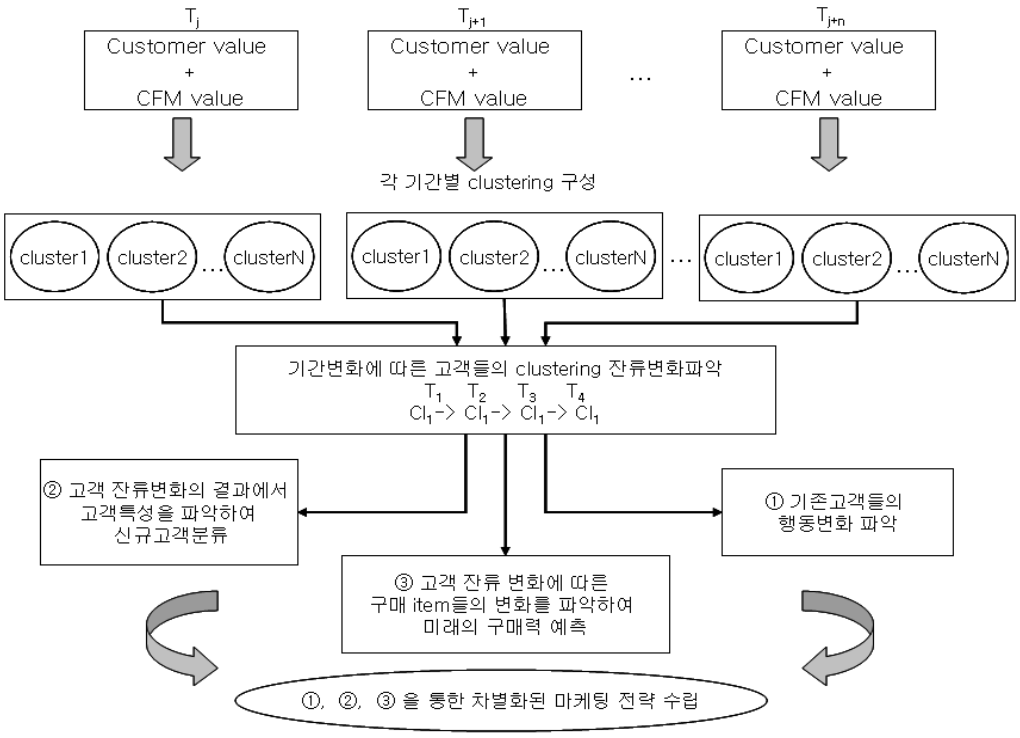

Figure 2. Proposed Model

Monetary)을 이용한 고객 세분화를 수행한다. RFM을 이용함에 있어서 좀 더 정확한 고객의 가치를 측정하기 위해 RFM 중 R(Recency)의 기준을 C(Continuation: 지속성)로 변형한 CFM (Continuation, Frequency, Monetary)값을 기준으로 고객 세분화를 수행하고 그 결과에 따른 각 고객 그룹에 등급을 부여한다.

3.1.1 CFM(Continuation, Frequency, Monetary)

기존의 RFM 기준은 아래와 같다.

- R(Recency): 고객의 최근 구매시점에서부터 경과 한 기간. 따라서 한 달 전에 마지막 구매를 한 고객은 3년 전에 마지막 구매를 한 고객보다 더 높은 점수를 받을 수 있다.
- F(Frequency): 일정기간 동안 고객의 구매 활동 수, 3년 동안 6번 구매활동을 한 고객은 동일한 기간 동안에 1번 구매한 고객보다 높은 점수를 받을 수 있다.
- M(Monetary): 고객의 평균 구매액수, 평균 구매 액수가 100만원인 고객은 20만원인 고객보다 높은 점수를 받는다.

R, F, M 값은 전체 데이터를 같은 크기로 5부분으로 나누어 각 부분을 1(낮은 값)~5(높은 값)로 표현 하며 각 부분에 전체 데이터의 20%를 할당한다. 5부분에는 R, F, M 값의 상위 20%가 차지하고 4부분에는 그 다음 20%가 차지하게 된다. 그러므로 어떤 한 고객이 모든 변수에서 5라는 값을 받았다면(555), 이 고객은 우리 회사에게 아주 중요한 고객이라는 것을 알 수 있으며, 111이라는 점수를 받은 고객은 반대의 의미를 가진다.

R의 기준을 보면 자주 구매를 해 왔지만 최근에 구매가 없는 고객의 경우에는 높은 점수를 받을 수 없게 된다. 비록 최근의 구매가 없지만 자주 구매를 한 고객도 회사 입장에서는 중요한 고객이 될 수 있다.

예를 들면 A 고객의 경우 RFM 값이 R: 5, F: 4, M: 4 이고, B 고객의 RFM 값은 R: 4, F: 4, M: 4 이다. 이 두 고객은 F와 M값은 동일하지만 R값에서 A고객이 높은 점수를 받았다. 하지만 실제 두 고객의 구매활동이 A고객의 경우는 구매활동이 전혀 없다가 최근 3개월 동안의 구매 활동이 많아 R의 값이 5로 측정되었고, B고객은 지속적인 구매활동이 있다가 최근 3개월의 구매활동이 없었기 때문에 R의 값이 4로 측정된 경우이다. 이런 경우 R값은 최근성이 중요하기 때문에 B고객보다 A고객에게 더 높은 점수를 부여하게 되어 RFM값만으로는 A고객이 B고객보다 더 높은 가치를 가지게 된다. 하지만 회사 입장에서는 A고객만큼 B고객 또한 매우 중요할 수 있다.

위에서 제시한 예와 같은 경우로 인해 R의 기준을 동등하게 줄 필요가 있으며 또한 본 논문에서 제안하는 고객잔류 분석에서는 현재 상태의 고객 가치만이 중요한 것이 아니라 과거 구매이력까지도 중요하므로 R의 측정 기준을 변경한다.

그러므로 기존의 R의 측정 기준 대신 지속성(continuation: C)을 제안한다.

지속성(continuation)은 고객의 지속적인 구매 활동 기간을 나타내며 기존과 같이 분류된 R의 등급에 마지막 구매 시점 이전의 구매활동 수를 반영한다.

이렇게 변경된 기준을 C(continuation: 지속성)라고 두었을 때 C는 다음과 같이 정의한다.

C = R * H/R의 전체 등급 수 (1)
R : recency
H : history (구매활동 기간 수)

위의 예에서 A고객의 지속성(C)은 다음과 같다. A고객의 recency(R)는 5이고 구매활동 기간 수(H)는 마지막 달에 한번 있었기 때문에 1이 되며 R은 1에서 5부분으로 나누어졌기 때문에 전체 등급 수는 5가 되므로 식 (1)에 의해 C = 5*1/5 =1 이 된다.

3.1.2 K-means clustering

CFM(continuation, Frequency, Monetary)값을 기준으로 각 기간별로 K-means를 이용한 클러스터링을 통해 고객 세분화를 수행한다.

3.2 고객 등급

K-means clustering에 의해 세분화된 각 클러스터(cluster)내에 있는 고객들의 C, F, M 각각의 평균값을 구하여 각 평균값들의 합을 구한다. 이 평균 합이 제일 큰 값을 가지는 클러스터에 1등급을 부여하고 평균 합에 따라 차례로 각 클러스터에 등급을 부여한다. 즉 CFM값이 가장 높은 고객들로 구성된 1등급 그룹은 Cl_1으로 표시하고 2등급은 Cl_2로 표시한다.

3.3 고객 잔류변화

시간 변화에 따른 고객들의 같은 등급으로의 잔류를 파악하기 위해 고객잔류률(Customer Remain Ratio: CRR)을 구한다.

즉, T_j 기간($j=1,\cdots,n$)에서 클러스터를 형성하고 있는 고객들이 다음 기간(T_{j+1})에 같은 클러스터로 얼마만큼 남아있는지 파악한다. 예를 들면 다음과 같다.

$Cl_i^{T_j}$: T_j 기간의 i등급 클러스터 ($i=1\cdots m,\ j=1\cdots n$)

$$Cl_1^{T_1}\rightarrow Cl_1^{T_2}\rightarrow Cl_1^{T_3},...,Cl_1^{T_n}$$
$$Cl_2^{T_1}\rightarrow Cl_2^{T_2}\rightarrow Cl_2^{T_3},...,Cl_2^{T_n}$$
$$\vdots$$
$$Cl_m^{T_1}\rightarrow Cl_m^{T_2}\rightarrow Cl_m^{T_3},...,Cl_m^{T_n}$$

i 등급 클러스터($i=1,\cdots,m$)에 대한 기간 T_j ($j=1,\cdots,n$)의 고객잔류률(CRR_{ij})은 다음과 같다.

$$CRR_{ij}=p(x\in Cl_i^{T_{(j+1)}}|x\in Cl_i^{T_j})$$

<Table 1>에서 CRR_{11}은 1등급 클러스터인 Cl_1에서 T_1기간에서 T_2기간으로 잔류한 고객잔류률이 0.7임을 나타낸다. 이것은 T_1 기간에 Cl_1에 속해있던 고객들 중 70%가 T_2기간에도 Cl_1에 남아있음을 나타낸다. CRR_{12}는 Cl_1에서 T_2기간에서 T_3기간으로 잔류한 고객잔류률이 0.7로서 T_2기간에 Cl_1에 속해있던 고객들 중 70%가 T_3기간에도 Cl_1에 남아있음을 나타낸다. CRR_{13}은 Cl_1에서 T_3기간에서 T_4기간으로 잔류한 고객잔류률이 0.7로써 T_3기간에 Cl_1에 속해있던 고객들 중 70%가 T_4기간에도 Cl_1에 남아있음을 나타낸다.

Table 1. Customer Remain Ratio

	j = 1	j = 2	j = 3
CRR_{1j}	0.7	0.7	0.7
CRR_{2j}	0.9	0.3	0.9
CRR_{3j}	0.3	0.4	0.5
CRR_{4j}	0.4	0.4	0.4

3.4 잔류 그룹과 이탈 그룹

위에서 구한 고객잔류률을 통해 고객들의 클러스터를 잔류그룹과 이탈그룹으로 구분한다.

잔류그룹이란 같은 등급으로 이동이 많은 그룹을 나타내며 이것은 그 이전 시점에 있던 고객들이 계속 같은 등급으로 유지하고 있다는 의미가 된다.

이탈그룹이란 같은 등급으로의 이동이 적다라는 의미이며 즉, 이탈그룹에 있는 고객들은 이전 시점에서 같은 등급으로의 이동보다는 다른 등급으로의 이동이 많은 경우이다.

그러므로 각 클러스터 별로 사용자가 지정한 임계값(Threshold)을 넘는 고객잔류률(CRR)이 전체 기간의 3분의 2이상이 되는 클러스터를 잔류그룹으로 그 이하의 값을 가지는 클러스터를 이탈그룹으로 정의한다.

여기서 임계값은 전체 고객 잔류률을 통해 사용자의 판단에 의해 정하게 되며 잔류그룹과 이탈그룹의 구분 기준을 전체 고객잔류률의 3분의 2이상으로 두는 것은 한 기간 동안의 고객 잔류률이 아닌 여러 기간에 걸친 고객 잔류률을 통해 잔류그룹과 이탈그룹을 결정하기 때문에 모든 기간 동안의 고객 잔류률이 사용자가 지정한 임계값보다 높은 경우만을 고려하기가 어렵다. 즉, 시간적 개념이 내포되어있기 때문에 어떤 시점에 대한 환경적인 요인으로 인해 고객들의 잔류에 영향을 받게 되어 고객 잔류률이 낮아지는 기간이 생길 수 있으므로 이런 요소를 고려하여 기준을 결정한다.

<Table 1>에서 임계값을 0.5라고 두었을 때, 1등급과 2등급 클러스터는 0.5이상인 CRR이 전체 기간의 2/3이상이 넘는 그룹이므로 잔류그룹으로 구분되고 3등급과 4등급 클러스터는 0.5이상인 CRR이 전체 기간의 2/3이상을 넘지 않기 때문에 이탈그룹으로 구분한다.

3.5 신규고객 세분화

많은 회사에서 기존 고객들에 대한 차별화 전략으로 고객들을 세분화하는 연구가 많이 진행 되어왔다. 이런 고객 세분화

를 통해 기존고객에게는 차별화된 마케팅전략을 세워 고객들의 이탈 방지와 유지 관리를 해왔지만 신규고객에 대해서는 모두에게 일정 기간 동안 똑같은 마케팅을 제공한다.

하지만 신규고객에게도 기존고객과 같이 세분화를 통한 차별화된 마케팅전략을 세운다면 신규고객으로 등록된 고객들을 유지 관리 할 수 있고 조기 이탈을 방지하게 됨으로써 신규고객확보 및 유지 비용절감을 이룰 수 있을 것이다.

제안하는 신규고객 세분화 방법은 기존 고객의 특성을 바탕으로 신규고객을 세분화하여 신규고객에게도 차별화된 서비스를 제공하면서 신규고객의 유지 및 이탈방지를 유도할 수 있도록 하는 것이 목적이다. 즉, 신규고객들을 하나의 그룹으로 파악하던 것을 신규고객의 특성에 따라 신규고객으로 등록한 시점부터 기존 고객과 같이 등급을 부여하여 세분화한다.

그러므로 앞 절에서 제시한 고객 잔류률을 신규고객 세분화에 이용한다. 제안하는 방법으로는 먼저, 신규고객의 클래스(class) 수를 결정하여 기존 고객의 시간 변화에 따른 고객 잔류률을 통해 각 신규고객 클래스에 해당하는 특성을 찾은 다음 거리함수를 이용하여 신규고객을 세분화한다.

3.5.1 각 신규고객 클래스(class) 특성

신규고객을 세분화함에 있어서 신규고객의 클래스 수는 기존 고객의 클러스터 수와 동일하게 둔다. 즉, 앞에서 기존 고객들을 4개의 클러스터로 분류한 경우에는 신규고객 클래스를 4개로 두게 되고 class 1은 1등급에 해당하는 신규고객의 특성으로 구성된다.

신규고객이 들어왔을 때 신규고객의 특성과 유사한 특성을 가진 신규고객 클래스를 찾아 신규고객을 분류하고자 하므로 먼저 각 신규고객 클래스에 따른 특성을 찾는다.

각 신규고객 클래스의 특성을 찾기 위해서 기존 고객들의 시간변화에 따른 고객 잔류률을 이용한다. 즉, 4기간 동안 Cl_1으로 이동한 고객들의 특성을 파악하여 신규고객 class 1의 특징을 찾게 되고 4기간 동안 Cl_2으로 이동한 고객들의 특성을 파악하여 신규고객 class 2의 특징을 찾는다.

신규고객 클래스 특성 결정 방법은 다음과 같다.

앞 절에서 구한 각 클러스터의 고객 잔류률을 통해 나온 잔류그룹과 이탈그룹의 고객 특성을 이용하여 신규고객 클래스 특성을 찾는다.

Step 1. 각 클러스터별로 잔류그룹과 이탈그룹을 정한다.

Table 2. Remain and Secession groups

cluster	CRR			그룹
Cl_1	0.7	0.7	0.7	잔류
Cl_2	0.9	0.3	0.9	잔류
Cl_3	0.3	0.3	0.5	이탈
Cl_4	0.4	0.5	0.4	이탈

<Table 2>에서 보는 봐와 같이 사용자가 지정한 임계값(= 0.5)을 넘는 고객 잔류률이 3분의 2이상이 되는 클러스터를 잔류그룹으로 3분의 2이하가 되는 클러스터를 이탈그룹으로 구분한다.

Step 2. 잔류그룹과 이탈그룹에 해당하는 고객 특성을 찾는다.

1) 잔류 그룹

잔류 그룹은 같은 등급으로의 이동이 많다는 의미이며 이 의미는 그 이전 시점에 있던 고객들이 계속 같은 등급으로 유지하고 있다는 의미가 된다.

잔류그룹의 성격에 의해 각 등급에서 두 시점간의 같은 등급($Cl_1^{T_1} \rightarrow Cl_1^{T_2}$)으로의 잔류변화를 보인 모든 고객들의 공통특성을 통해 신규고객들의 특성을 찾는다.

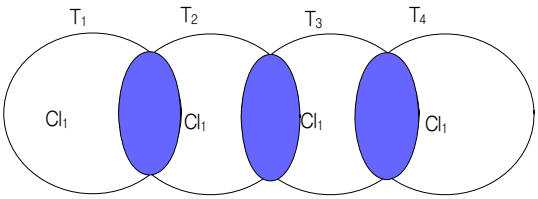

Figure 3. New customer characteristics for remain group.

즉, <Figure 3>에서 보는 봐와 같이 T_1기간의 Cl_1에 있던 고객들 중 T_2기간에 같은 클러스터인 Cl_1에 잔류한 고객들, T_2기간에서 T_3기간으로 같은 클러스터로 잔류한 고객들과 T_3에서 T_4기간으로 같은 클러스터로 잔류한 고객들, 이 모든 고객들의 특성을 찾는다. 각 속성에 대해서 이들 고객들이 가지고 있는 속성 값들 중 가장 많은 수를 차지하는 속성 값, 즉 최빈수(mode)에 해당하는 속성 값을 찾아서 신규고객 클래스의 특성으로 한다. 예를 들어 나이라는 속성의 경우 고객들이 가장 많은 수를 차지하는 나이가 50대라면 나이 = 50 이라는 특성을 발견할 수 있다.

2) 이탈 그룹

이탈그룹은 같은 등급으로의 이동이 낮다는 의미이며 즉, 이탈 그룹에 있는 고객들은 같은 등급으로의 이동보다는 다른 등급으로의 이동이 많은 경우이다.

그러므로 각 기간별 이탈그룹에 속한 모든 고객들의 공통특성을 통해 신규고객 특성을 찾는다.

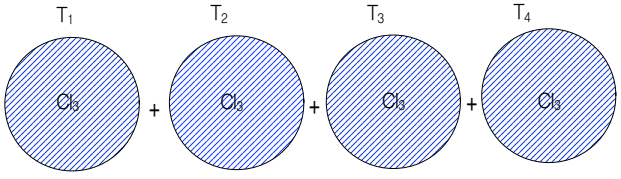

Figure 4. New customer characteristics for secession group

즉, <Figure 4>에서 T_1, T_2, T_3, T_4에 속해있는 모든 고객들의 특성을 찾는다. 잔류그룹과 동일하게 각 속성에 대해 전체 고객들이 가지고 있는 속성 값들 중 가장 많은 수를 차지하는 속성 값을 찾아서 신규고객 클래스의 특성으로 한다.

신규고객 특성을 찾는데 있어서 기간별로 나누어 잔류그룹과 이탈그룹으로 구분하여 찾는 이유는 고객들의 욕구와 구매행동은 항상 같은 것이 아니라 시간의 흐름에 따라 계속 변화하기 때문에 한 시점에서의 고객의 욕구가 다른 시점에서도 같을 수가 없다. 따라서 이러한 이유로 기간별로 고객들의 성향이 어떻게 변화하는지를 파악하여 신규고객의 특성을 찾는다.

잔류그룹의 경우는 계속 같은 등급으로 유지하는 고객들이 많은 경우이기 때문에 고객들의 성향이 일정하다고 볼 수 있으므로 각 등급에서 잔류한 고객들을 대상으로 신규고객의 특성을 찾는다. 하지만 이탈그룹의 경우는 같은 등급으로 유지하기 보다는 다른 등급으로의 이동이 많은 그룹으로 고객들의 성향이 일정하지 않다는 것을 알 수 있다. 그러므로 전체 고객들의 특성으로 신규고객의 특성을 찾는다.

3.5.2 신규고객 세분화

각 신규고객 클래스의 특성이 결정되었으면 이를 토대로 신규고객이 들어왔을 때 신규고객의 특성과 각 신규고객 클래스 특성사이의 거리 값을 이용하여 신규고객의 클래스를 결정한다. 각 속성에 대해 거리함수를 이용하여 속성 값들 중 가장 가까운 한 이웃을 찾게 된다(Berry and Linoff).

신규고객 특징				
Class	나이	수입		신규고객(N)
1	45	500만원		
2	35	200만원	⇦	나이: 46 수입: 400만원
3	30	150만원		
4	30	100만원		

Figure 5. An example of new customer segmentation

<Figure 5>와 같은 데이터가 주어졌을 때 신규고객을 세분화 하는 단계는 다음과 같다.

Step 1. 신규고객이 가지고 있는 속성들 각각과 신규고객 클래스 그룹에 속해있는 속성들 각각에 대한 거리를 구한다.

거리를 구하는 방법은 다음과 같다.

$$d_{norm\_att}(N, Class) = d_{att}(N, Class)/max(d_{att}) \quad (2)$$

$d_{att}(N, Class)$는 속성(att)에 대해 신규고객(N)과 신규고객 각 클래스 속성간의 거리 값을 나타내며, $d_{norm_att}(N, Class)$는 신규고객(N)과 신규고객 각 클래스 속성간의 정규화 된 거리 값을 나타낸다.

예를 들어, 나이에 대한 거리계산은 식 (2)에 의해서

$d_{norm_age}(N, class\ 1) = d_{age}(46, 45)/max(d_{age}(46, 45), d_{age}(46, 35), d_{age}(46, 30), d_{age}(46, 30))$으로 계산되어진다.

나머지 클래스에 대해서도 위와 같은 방법으로 계산하여 나이에 대한 각 클래스의 거리 값을 구한다.

수입에 대해서도 나이와 같은 방법으로 계산한다.

$$d_{norm\_sal}(N, Class) = d_{sal}(N, Class)/max(d_{sal})$$

Step 2. Step 1에서 구한 각 클래스별 속성들의 거리 값들을 합한다.

여기에 각 속성들마다 각기 다른 가중치(weight)를 부여하게 된다. 속성들 중에서 중요 속성을 결정하여 중요속성에는 다른 속성에 비해 작은 가중치를 부여한다.

$$d_{sum}(N, Class) = W_1 \times d_{norm\_age}(N, Class) + W_2 \times d_{norm\_sal}(N, Class) \quad (3)$$

d_{sum} 은 신규고객(N)과 신규고객 클래스 사이의 각 속성들의 거리 값의 합을 나타내며, W는 가중치를 나타낸다.

class 1에 대한 각 속성들의 거리 값의 합은 식 (3)에 의해

$d_{sum}(N, class\ 1) = W_1 \times d_{norm\_age}(N, class\ 1) + W_2 \times d_{norm\_sal}(N, class\ 1)$ 으로 계산된다.

나머지 클래스에 대해서도 위와 같은 방법으로 각 속성들의 거리 값의 합을 구한다.

Step 3. d_{sum} 의 값이 가장 작은 신규고객 클래스에 신규고객(N)을 분류한다.

3.6 구매항목 예측

기존의 추천시스템이나 연관규칙의 경우는 시간의 흐름에 상관없이 한 시점에서 발생한 항목들을 분석하여 이와 비슷한 구매성향을 가진 고객들에게 추천해주거나 마케팅에 활용하였다. 하지만 앞서 설명했듯이 고객의 욕구와 구매행동은 항상 같은 것이 아니라 시간의 흐름에 따라 빨리 변화하므로 시간의 흐름에 따른 고객들의 구매력을 파악하여 이에 맞는 항목들을 추천해 주는 것이 효과적일 것이다.

그러므로 제안하는 방법은 각 클러스터별 빈번하게 구매되어지는 항목들을 시간의 변화에 따른 구매력을 파악하여 미래에 이러한 항목들의 구매정도를 예측하여 신규고객 뿐만 아니라 기존고객에게 추천한다.

$T_j(j = 1,\cdots, n)$기간에 빈번하게 구매되었던 항목들이 T_{j+1}기간에도 빈번하게 구매는 되었지만 이런 빈번 항목들이 계속 증가 추세인지 감소 추세인지를 파악하여 각 항목들의 구매력을 예측한다. 즉, <Table 3>에서 클러스터 1(Cl_1)에서 T_1기간에

서 T_4기간까지 빈번항목이 주어졌을 때 T_5기간에 빈번하게 구매되어질 항목을 예측하기 위해서 빈번 항목들의 시간 변화에 따른 growth rate(GR)를 구한 다음 보외법을 통해 예측한다.

Table 3. An example of frequency items for each cluster

Time	T_1	T_2	T_3	T_4	T_5
Cluster	Cl_1	Cl_1	Cl_1	Cl_1	Cl_1
빈번 항목	A, B, C	B, C, D	B, D	A, C, D	?

구매항목 예측방법은 다음과 같다.

Step 1. 각 클러스터 내에서 각 기간별 빈번하게 구매되어진 항목들을 찾는다.

Step 2. 빈번하게 구매되어진 항목들의 growth rate를 계산한다.

Growth rate(GR)를 구하는 방법은 다음과 같다.

$$C_i^{T_j}=\{x_1,\cdots x_n\},\ C_i^{T_{j+1}}=\{x_1,\cdots x_n\},$$
$$i=1,\cdots,k,\ j-1,\cdots,m$$
$$GR(C_i^{T_j}(x),\ C_i^{T_{j+1}}(x))=\frac{|C_i^{T_j}|}{|C_i^{T_{j+1}}-C_i^{T_j}|}\times\frac{Sup(C_i^{T_{j+1}},x)-Sup(C_i^{T_j},x)}{Sup(C_i^{T_j},x)}$$

$C_i^{T_j}$: T_j 기간의 클러스터 i의 전체 트랜잭션(transaction)수

$C_i^{T_{j+1}}$: T_{j+1}기간의 클러스터 i의 전체트랜잭션(transaction)수

$Sup(C_i^{T_j},x)$: T_j 기간의 클러스터 i에 포함된 x 항목의 개수

$Sup(C_i^{T_{j+1}},x)$: T_{j+1}기간의 클러스터 i에 포함된 x 항목의 개수

만약, 각 기간 클러스터의 트랜잭션 수가 같다면 growth rate를 계산할 때 항목의 Sup만 고려한다.

$$GR(C_i^{T_j}(x),\ C_i^{T_{j+1}}(x))=\frac{Sup(C_i^{T_{j+1}},x)-Sup(C_i^{T_j},x)}{Sup(C_i^{T_j},x)}$$

Growth Rate(GR)의 값이 + 또는 ∞ 이면 항목의 구매가 증가됨을 나타내고, growth rate(GR)의 값이-이면 항목의 구매가 감소됨을 나타낸다.

<Table 4>에서는 4기간 동안 빈번하게 구매 되어진 항목 A, B, C, D, 에 대한 각 growth rate값을 보여준다.

Step 3. Growth rate를 바탕으로 보외법을 이용하여 각 항목의 미래 구매력을 예측한다.

Table 4. An example of growth rate of frequency items

항목	T_1	T_2	T_3	T_4
A	0	-5	-4.5	13.3
B	0	2.5	0	2.2
C	0	10	-6.8	6.6
D	0	5	3	0.83

기간(T)을 X축으로 각 항목의 growth rate를 Y축으로 두어 보외법에 의해 T_5 기간의 growth rate를 구한다.

보외법에 의해 구해진 각 항목들의 growth rate 값이 증가함을 나타내면 T_5 기간에 구매되어질 가능성이 높은 항목임을 알 수 있고, growth rate 값이 감소함을 나타내면 T_5기간에 구매되어질 가능성이 낮은 항목임을 알 수 있다.

이를 통해 growth rate값이 증가한 항목들을 해당 그룹에 추천하는 것은 기존의 추천 시스템과는 달리 그 시점에 맞는 항목들을 추천하게 되는 것이므로 고객들에게 차별화된 정보를 제공해 줄 수 있다. 또한 신규고객에도 기존고객들이 예전에 주로 구매한 항목을 추천해줄 수 있지만 제안하는 구매항목 예측을 통해 나온 결과의 항목들을 추천해 주는 것 또한 차별화된 마케팅 전략이 될 수 있다.

4. 실험결과

4.1 데이터 집합

실험에는 국내 모 백화점 데이터를 사용하였다. 1년간의 데이터를 전체 2개의 데이터 집합으로 구성하였다. 고객번호, 나이, 성별, 주소, 주거형태, 집 평수, 백화점 첫 이용 날짜를 포함하는 508445 명의 고객 정보(customer profile) 데이터 집합과 고객번호, 구매일자, 항목, 구매액수, 결제수단, 첫 구매 시기 등을 포함하는 구매정보(purchasing information)로 이루어진 데이터 집합으로 구성되었다. 구매정보 데이터 집합의 경우는 1년 동안 구매 회수가 30번을 넘지 않은 고객은 제외시켰고, 두 데이터 집합에 있는 고객번호를 통해서 각 고객의 구매정보를 파악할 수 있다.

시간의 변화에 따른 고객들의 행동변화를 파악하기 위해 1년 동안의 데이터 집합을 3개월씩 나누어 4기간으로 부분집합(sub_dataset)구성하였고, 각 기간별 부분집합의 수는 동일하지 않다.

k-means clustering을 통해 4기간의 부분집합에 대해 4개의 클러스터를 구성하였다. 각 기간별로 구성된 클러스터에 CFM 값을 적용하여 등급을 부여하였다. 즉 1등급에서 4등급의 고객 세그먼트가 만들어지게 된다. 또한 각 클러스터에 속하는 고객들의 빈번 구매 항목들을 구한다.

4.2 결과

4.2.1 신규고객 세분화 평가

4기간 동안의 고객잔류률(CRR)을 이용하여 고객정보 데이터 집합에 있는 속성들 중 나이, 주소, 주거형태, 집 평수를 대상으로 각 등급에 맞는 신규고객의 특성들을 찾았다. 이 특성들이 신규고객을 얼마나 잘 구분할 수 있는지를 평가하기 위해 검증 데이터 집합을 구성하였다. 검증 데이터집합은 4기간 중 마지막 기간(10월에서 12월)에 해당하는 고객들 118661명 중 이 기간에 처음으로 백화점을 이용한 고객 356명으로 구성되었다.

제안하는 방법으로 얻은 각 등급의 신규고객 특성과 검증 데이터에서 얻은 각 등급의 고객 특성들을 비교하여 이 특성들이 얼마만큼 일치하는지를 파악하였다. 제안하는 방법으로 찾은 신규고객 1등급에 해당하는 고객들의 특성은 다음과 같이 나타났다.

나이 = 05, 주거형태 = 08, 집 평수 = 04, 주소 = 11380

검증 데이터 집합에서 첫 구매이면서 1등급에 해당하는 100명의 고객들 특성과 위의 특성을 비교하였을 때 100명 중 56명이 위와 같은 특성을 가지고 있었으므로 약 60%가 일치하고 있음을 알 수 있었다.

시간변화를 고려하지 않은 기존 연구에서 수행한 고객 세분화 방법을 통해 기존고객들의 특성을 찾아 신규고객 세분화 결과와 제안하는 방법으로 신규고객을 세분화한 결과를 비교하였다. 기존 연구에서는 기간을 나누지 않고 고객들의 RFM 값을 구하여 클러스터링을 이용해 세분화 하였다. 여기에 제안하는 방법처럼 각 클러스터링에 등급을 부여하여 각 등급에 해당하는 고객들의 특성을 찾아 위에서 수행한 검증방법으로 비교실험을 하였다.

<Table 5>는 기존 모델과 제안하는 모델의 비교 실험결과를 보여준다.

Table 5. Performance result for new customer segmentation(%)

	제안하는 모델	기존 모델
cluster 1(Cl_1)	60	32
cluster 2(Cl_2)	27	21
cluster 3(Cl_3)	30	23
cluster 4(Cl_4)	35	27

기존 고객의 특성을 이용한 신규고객 세분화 방법은 제안하는 모델이 전반적으로 기존 모델에 비해서 좋은 결과를 나타냄을 보여준다. 고객의 구매 욕구나 성향은 계속 변화기 때문에 이런 변화를 고려한 고객 세분화 방법이 고객에 대한 정보를 좀 더 정확하게 제공해 줄 수 있으므로 고객들의 성향에 맞는 서비스 제공 또는 고객 유지 및 유치에 좋은 결과를 얻을 수 있을 것이다.

4.2.2 구매항목 예측 평가

각 클러스터에서 자주 구매되는 항목들 100개씩을 선택하여 예측 실험을 수행하였다.

<Figure 6>은 클러스터 1에서 자주 구매되어진 항목 중의 하나인 품번 120402 항목을 shape-preserving piecewise cubic interpolation을 이용한 예측 결과를 보여준다. 품번 120402 항목을 월별 기준(12달)으로 growth rate값(Y)을 구하여 12개월 이후의 3개월인 13, 14, 15기간의 구매력을 예측한 결과를 보여주고 있다. 예측된 결과에서 품번 120402 항목은 13, 14, 15기간에 구매력이 향상되고 있는 추세를 보여주고 있기 때문에 이 항목을 cluster 1그룹에 추천해 줄 수 있다.

시간(X) <월별>	Growth Rate(Y)
1	0
2	-1.5E + 08
3	-1.7E + 08
4	-2.04E + 08
5	-2.2E + 08
6	-1.6E + 08
7	-7.8E + 07
8	-5.8E + 07
9	-6.6E + 07
10	-7.4E + 07
11	-6.6E + 07
12	-4.9E + 07
13	-2.6824e + 007
14	-7.3699e + 006
15	9.9838e + 005

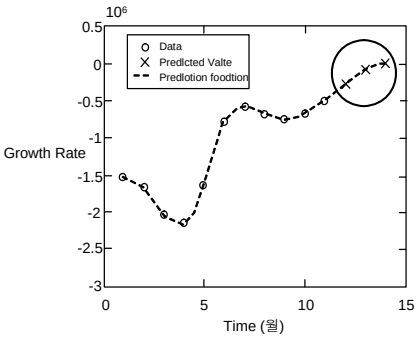

Figure 6. The result of forecasting by extrapolation

보외법에 의한 예측모형의 정확도를 평가하기 위해서 예측모형의 정확도를 측정하는 대표적인 방법인 평균 절대오차(Mean Absolute Error: MAE)를 이용하였다.

$$MAE = (\sum_{t=1}^{N} |e_t|)/N$$

e는 실제 관측치와 예측치와의 편차인 예측오차를 나타내며, N은 예측에 사용한 자료의 수를 나타낸다.

MAE값이 낮을수록 예측 값들이 정확하므로 본 논문에서 보외법을 이용하여 구한 예측 값들의 MAE값을 평가하기 위해 이동평균법(Moving average method)에서 구한 예측 값들의 MAE값과 비교하였다.

이동평균법이란 단순하면서도 많이 사용되는 예측모형의 하나로써, 시계열 자료에 존재하는 확률적 변동이 상쇄되도록 일정기간의 자료를 평균하여 평균값을 구하고, 이 평균값을 바로 다음 기간의 예측치로 사용한다.

각 클러스터별 빈번하게 구매되어진 항목 100개에 대해 7월에서 12월의 실제 growth rate 값과 예측된 값의 평균 절대오차(MAE)값들의 평균값을 구하였다.

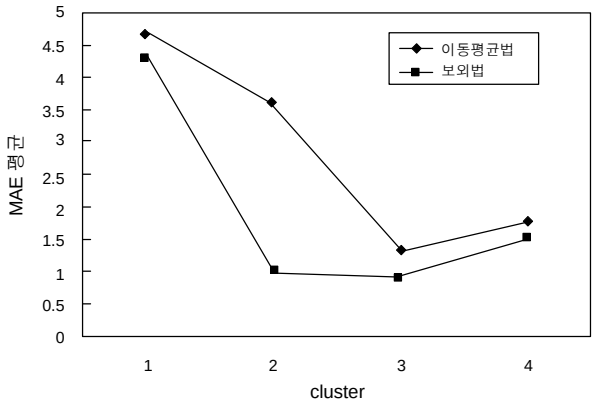

Figure 7. The average of MAE between extrapolation method and moving average method

<figure 7>은 보외법과 이동평균법의 MAE값들의 평균을 보여준다.

실험결과를 통해 두 예측모형의 MAE값들이 cluster 2를 제외하고는 큰 차이를 보이지 않지만 이동평균법 보다는 보외법이 전반적으로 나은 예측결과를 보여주고 있다.

보외법과 이동평균법은 시계열 분석에 사용되는 예측 기법들이지만 과거의 판매 정보를 바탕으로 한 상품 예측에 보외법을 많이 사용하고 있다. 그러므로 본 논문에서 제안하는 구매 항목예측에 있어서는 이동평균법 보다 보외법이 더 나은 결과**를** 가져올 수 있었다.

5. 결 론

본 논문에서는 시변성을 가지는 기존 고객들의 특성을 통해 신규고객을 세분화하고 기존 고객들의 구매 항목들의 변화를 파악하여 미래의 구매력을 예측할 수 있는 방법을 제안하였다. 기존 연구들에서는 기존고객들을 세분화하여 차별화된 마케팅을 적용하고자 하는 시도를 해왔지만 신규고객을 세분화하여 관리하는 것은 간과되었다.

본 논문에서 제안하는 신규고객 세분화 방법은 기존고객들에게 차별화를 두는 것처럼 신규고객들에게도 처음부터 차별화를 두어 각 등급에 맞는 마케팅 전략을 세움으로써 신규고객들을 유지 관리하여 조기 이탈을 방지하고자 한다. 뿐만 아니라, 제안하는 구매항목 예측을 통해서는 각 고객 그룹에서 빈번하게 구매되어지는 항목들을 파악하여 이러한 항목들이 미래에 구매되어질 정도를 예측하여 기존고객 뿐만 아니라 신규고객에게 예측된 항목들을 추천할 수 있다.

또한 기존 연구들에서는 한 기간 동안의 고객들을 세분화하는 방법들을 많이 제안하였다. 하지만 앞서도 언급했듯이, 고객들의 성향은 시간의 흐름에 따라 계속 변화되고 있기 때문에 본 논문에서는 시간의 변화에 따른 고객들의 잔류변화를 통해서 고객들의 성향을 파악한 고객 세분화가 이루어질 수 있고 회사에서 관심이 있는 충성 고객그룹을 명확히 찾을 수 있다. 또한 한 시점의 고객들의 구매력 보다는 시간의 변화에 따른 고객들의 구매력을 파악함으로써 정확한 고객들의 구매 패턴 변화를 파악할 수 있다.

본 논문에서 기존 고객들의 정보를 바탕으로 신규고객의 특성을 찾고자 함에 있어서 몇 가지 한계점을 가지고 있다.

먼저, 고객 속성의 종류에 대한 한계이다. 실제 백화점에서는 고객들의 정보에 대한 속성 종류가 많고 고객에 대한 정보를 얻고자 한다면 고객카드 등을 통해 신규고객 특성을 찾는데 유용한 속성을 찾을 수 있을 것이다. 하지만 본인이 가지고 있는 데이터에서는 이런 속성들이 많이 제한되어 있기 때문에 정확한 고객특성을 찾는데 한계를 가지고 있다.

두 번째로는 신규고객을 세분화하는 것이 이 논문의 목적중의 하나이므로 다른 논문과의 비교실험은 이루어지지 않았다.

향후 연구과제로서는 고객 속성에 대해 다양한 속성 종류와 정확한 속성 값이 주어진 실제 데이터를 바탕으로 평가해보고자 한다.

참고문헌

Berry, M. J. A. and Linoff, G. (1997), Data Mining Techniques: For Marketing, Sales, and customer Support, Wiley, New York, USA.

Chen, M-C.,Chiu, A-L., Chang, H-H. (2005), Mining changes in customer behavior in retail marketing, Expert Systems with Applications **28**, 773-781.

Ha, S. H., Bae, S. M., and Park, S. C. (2002), Customer's time-variant purchase behavior and corresponding marketing strategies: an online retailer's case. Computer & Industrial Engineering, **43**, 801-820.

Hsieh, N-C. (2004), An Integrated data mining and behavioral scoring model for analyzing bank customers, Expert Systems with Applications **27**, 623-633.

Hwang, H., jung, T., and Suh, E. (2004), An LTV model and customer segmentation based on customer value: a case study on the wireless telecommunication industry, Expert Systems with Applications **26**, 181-188.

Lingras P., Hogo, M., Snorek, M., and West, C. (2005), Temporal analysis of clusters of supermarket customers : Conventional versus interval set approach, Information Sciences, **172**, 215-240.

Liu, D-R. and Shih, Y-Y. (2005), Hybrid approaches to product recommendation based on customer lifetime value and purchase preferences, The Journal of Systems and Software, **77**(2), 181-191.

Min, S-H. and Han, I. (2005), Detection of the customer time-variant pattern for improving recommender systems, Expert System with Application **28**,189-199.

Park, S. C., Park, J. H., Ha, S. H., and In, k. H. (2000), From Behavior to Market : Customer Focus in Supply Chain Management, Fifth Conference of the Association of Asian-Pacific Operations Research Societies, Singapore, 5-7.

Scott Armstrong, J. (2000), Extrapolation for Time-Series and Cross-Sectional Data, Principles of Forecasting: A handbook for Research and Practitioners, Kluwer Academic Publishers, Norwell, MA.

Scott Armstrong, J., Morwitz, V. G. and Kumar, V. (2000), Sales forecasts for existing consumer products and services: Do Purchase intentions contribute to accuracy?, International Journal of Forecasting, **16**, 383-397.

Song, H. S. (2001), Mining the change of customer behavior in an internet shopping mall. Expert Systems with Applications, **21**, 157-168.

Soulet, A., Cremilleux, B., and Rioult, F. (2004), Condensed Representation of Emerging Patterns, Pacific-Asia Conference on Knowledge Discovery and Data Mining(PAKDD), 127-132.

Suh, E. H., Noh, K. C., and Suh, C. K. (1999), Customer list segmentation using the combined response model, Expert Systems with Applications **17**, 89-97.

Thomassey, S. and Fiordaliso, A. (2005), A hybrid sales forecasting system based on clustering and decision tree, Decision Support Systems.

Tsai, C.-Y. and Chiu, C.-C. (2004), A purchase-based market segmentation methodology, Expert Systems with Applications **27**, 265-276.

Verhoef P. C. and Donkers, B. (2001), Prediction customer potential value an application in the insurance industry, Decision Support Systems, **32**,189-199.