지능정보연구 제18권 제1호 2012년 3월(pp.23~38)

# 연관규칙 마이닝에서의 동시성 기준 확장에 대한 연구

김미성
국민대학교 BIT전문대학원 석사과정
(rkaaktm@naver.com)

김남규
국민대학교 경영정보학부 조교수,
(ngkim@kookmin.ac.kr)

안재현
KAIST 정보미디어경영대학원 교수
(jahn@business.kaist.ac.kr)

온라인 쇼핑몰은 인터넷을 통해 손쉽게 접근이 가능하기 때문에, 최초 구매의사가 발생한 시점으로부터 이에 대한 실제 구매가 실현되기까지의 기간이 오프라인 쇼핑몰에 비해 비교적 짧게 나타난다. 즉 오프라인 쇼핑몰의 경우 구매 희망 물품을 바로 구매하기 보다는 몇 개의 물품들을 모아서 구매하는 행태가 일반적이다. 하지만, 인터넷 쇼핑몰의 경우 단 하나의 물품만을 포함하고 있는 주문이 전체 주문의 절반 이상을 차지한다. 따라서 온라인 쇼핑몰 데이터의 장바구니 분석에 전통적 데이터마이닝 기법을 그대로 적용할 경우, Null Transaction의 수가 지나치게 많음으로 인해 합리적 수준의 지지도(Support)를 만족시키는 규칙을 찾는 것이 매우 어렵게 된다. 이러한 이유로 온라인 데이터를 사용한 많은 연구는 동시성 기준을 여러 방법으로 확장하여 사용하였는데, 이들 동시성 기준은 명확한 근거나 합의 없이 연구자의 상황에 따라 임의로 선택된 측면이 있다. 따라서 본 연구에서는 온라인 마켓 분석에 적용되는 구매의 동시성 기준을 정확도 측면에서 평가함으로써, 구매의 동시성 기준 선정을 위한 근거를 제시하고자 한다. 또한 동시성 기준의 정확도가 고객의 평균 구매간격에 따라 상이하게 나타나는 것을 파악하여, 향후 고객의 특성에 따른 차별화된 추천 시스템 구축을 위한 기본 방향을 제시하고자 한다. 이를 위해 국내 대형 인터넷 쇼핑몰의 최근 2년간 실제 거래 내역을 대상으로 실험을 수행하였으며, 실험 결과 단골 고객의 구매 추천을 위한 분석의 경우 추천 범위와 분석 데이터의 동시성 기준을 맞추어 연관규칙을 도출하는 것이 바람직하며, 비단골 고객의 경우 대부분의 추천 범위에 대해서 분석 데이터의 동시성 기준을 비교적 길게 설정하여 연관규칙을 도출하는 것이 바람직한 것으로 나타났다.

논문접수일 : 2012년 01월 30일 게재확정일 : 2012년 03월 06일
투고유형 : 국문급행 교신저자 : 김남규

## 1. 서론

웹에 접근할 수 있는 방식이 다양해짐에 따라, 온라인 거래 및 이로 인한 온라인 데이터의 규모는 매우 빠른 속도로 증가하고 있다. 방송통신위원회와 한국인터넷진흥원의 조사(한국인터넷진흥원, 2010)에 따르면, 국내 인터넷 이용자 수는 2000년에 1,904만 명에서 2010년에 3,701만 명으로 2배 가량 증가하였으며, 특히 동일 기간의 인터넷 쇼핑몰 이용률은 12.3%에서 64.3%로 52%가 증가하였다. 즉 스마트폰을 포함한 각종 유무선 장비의 보급으로 인해 형성된 새로운 생태계에서, 사용자들은 정보 검색, 물품 구매, SNS(Social Networking Service) 등의 다양한 활동을 굳이 오래 기다리지 않고 원하

* 이 논문은 2011년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 연구되었음(NRF-2011-327-A2011-0178).

는 즉시 수행할 수 있는 제반 환경을 갖추게 된 것이다.

이처럼 유무선 단말을 통한 인터넷 접속이 더욱 활성화됨에 따라, 온라인 거래를 통해 축적되는 데이터의 종류 및 양의 증가 추세는 더욱 가속화되고 있다. 온라인 거래를 통해 축적된 데이터는 그 규모가 방대할 뿐 아니라 필연적으로 전산화되어 있고, 대부분의 경우 데이터베이스로 구축이 되어 있다는 점에서 많은 연구자들의 분석 대상으로 선호되어 왔다. 이러한 온라인 마켓에 관련한 연구로는 온라인 마켓플레이스(Online Marketplace)의 수익률에 대한 연구(정영조 외, 2009), 온라인 쇼핑몰의 구매패턴에 대한 연구(이종민 외, 2003), 온라인 데이터를 활용한 마케팅 전략 수립 연구(송만석 외, 2008) 등이 있으며, 최근 데이터마이닝의 연관분석을 활용한 추천시스템에 대한 연구(안현철 외, 2006)도 온라인 마켓을 대상으로 수행된 바 있다.

온라인 마켓은 오프라인 마켓에 비해 접근성이 뛰어나기 때문에, 최초 구매의사가 발생한 시점으로부터 이에 대한 실제 구매가 실현되기까지의 기간이 오프라인 쇼핑몰에 비해 매우 짧게 나타날 것으로 예상된다. 즉 오프라인의 경우 구매를 희망하는 물품이 있더라도 일정 기간을 기다렸다가 한꺼번에 여러 물품을 구매하는 경향이 있는 반면, 온라인의 경우 접근의 용이성 때문에 굳이 오랜 기간을 기다릴 필요가 없다는 것이다. 이러한 예상은 본 연구에서 수행한 실험의 결과로도 뒷받침되는데, 국내 한 대형 온라인 쇼핑몰의 최근 2년간 거래 데이터를 분석한 결과 전체 주문 건수 중 단 1건의 물품만 포함하고 있는 주문 건수의 비율이 76.5%나 차지하고 있음을 확인하였다. 이러한 상황에서 온라인 쇼핑몰 데이터의 장바구니 분석에 전통적 데이터마이닝 기법을 그대로 적용할 경우, Null Transaction의 수가 지나치게 많음으로 인해 합리적 수준의 지지도(Support)를 만족시키는 규칙을 찾는 것이 매우 어렵게 된다. 이를 극복하기 위한 대안으로 단 하나의 물품을 포함하는 주문은 아예 분석 대상에서 제외할 경우, 전체 주문의 76.5%에 해당되는 주문을 인위적으로 배제함으로 인해 분석 결과의 정당성을 확보하기가 어렵다는 한계를 갖게 된다.

이러한 현상은 온라인 거래 데이터의 분석을 위해서는 전통적인 데이터마이닝 분석에서 사용되어 온 장바구니 기준, 즉 동시성 기준에 대한 확장된 정의가 필요함을 의미한다. 하나의 장바구니를 정의하기 위해 가능한 기준들은 <그림 1>을 통해 설명 가능하다.

| | 회원번호 | 구매일자 | 주문번호 | 물품번호 |
|---|---|---|---|---|
| ③ | A | 1/1 | 0_01 | P_01 |
| | A | 1/1 | 0_01 | P_02 |
| ② | A | 1/1 | 0_02 | P_03 |
| | A | 1/2 | 0_03 | P_04 |
| ① | A | 1/2 | 0_03 | P_04 |

<그림 1> 동일한 구매 내력에 대한 다양한 동시성 기준

예를 들어 <그림 1>의 기준 ③의 경우 전통적인 장바구니 분석에서 사용된 정의를 나타낸다. 즉 하나의 주문을 하나의 장바구니로 간주하는 것이다. 만약 특정 회원이 하루 동안 주문한 모든 물품을 동시 구매로 간주한다면 이는 기준 ②를 채택한 것이다. 마지막으로 기준 ①은 기간에 관계없이 특정 회원이 구매한 모든 물품을 하나의 장바구니에 담긴 것으로 간주하는 기준을 나타낸다. 어떤 기준이 보다 합리적인가에 대한 판단에 앞선 더욱 중요한 문제는, 이들 기준 중에 어떤 기준을 적용하여 온라인 데이터에 대한 장바구니 분석을 수행할

지에 대한 명확한 가이드라인 자체가 없어서 연구자 임의의 선택에 따라 연구가 수행되고 있다는 것이다. 예를 들어 온라인 마켓을 대상으로 데이터 마이닝을 수행한 국내의 연구 중 일부 연구는(강동원, 이경미, 2001; 하성호, 박상찬, 2002) 전통적인 장바구니의 기준인 기준 ③을 사용했으며, 다른 연구는 특정 회원이 구매한 모든 물품을 하나의 장바구니로 간주(정영수, 강경화, 2004)하기도 하였다. 그 외 다수의 연구들은 기준 ②와 같이 일정 기간 내에 이루어진 주문들을 묶어서 이들을 동시 구매로 간주하기도 하였다.

지금까지와 같이 온라인 마켓에 대해 장바구니 분석이 서로 상이한 장바구니 기준 하에서 수행될 경우, i) 연구자가 선정한 장바구니의 기준에 대한 선정 근거 제시가 어렵고 ii) 임의의 장바구니 기준을 적용함으로써 현실과 동떨어진 결과를 도출할 위험이 있을 뿐 아니라 iii) 개별 연구 성과들을 통합하여 그 기여도를 극대화하지 못한다는 한계를 갖게 된다. 예를 들어 고객이 물품을 장바구니에 담는 동시에 이루어지는 실시간 추천의 경우를 가정해 보자. 이는 하나의 주문을 하나의 장바구니로 간주하는 경우를 나타낸다. 만약 이러한 추천을 위해 기존의 거래 내력에 대한 연관분석을 실시되었는데, 분석 과정에서는 동일 날짜에 구매된 물품들이 모두 동시 구매로 간주되었다고 하자. 이 경우, 즉 하루 기준의 분석을 통해 도출한 연관규칙을 개별 주문 단위의 구매 추천 및 예측에 활용하는 것이 과연 정당한 것인가에 대한 논의가 필요하다.

따라서 본 연구에서는 이러한 한계를 극복하기 위해 다음의 관점에서 논의를 진행하고자 한다. 우선 다양한 동시성 기준을 적용한 연관성 분석 결과를 비교해 봄으로써, 동시성 기준에 따라 연관규칙의 정확도가 어떻게 변하는지 살펴보고자 한다. 즉 온라인 마켓의 경우 동시성 기준이 명확하게 나타나 있지 않은 상황을 인식하고, 이를 다양한 기준으로 실험함으로써 물품을 구매할 때 어떤 동시성 기준이 고객의 성향을 잘 반영하는지에 대해 실험을 통해 확인 하고자 한다. 이를 통해 향후 온라인 마켓의 분석에 대한 연구 및 프로젝트 수행 시 적용될 장바구니 기준에 대한 가이드라인을 제시하고자 한다. 또한 동시성 기준의 정확도는 고객들의 특성에 따라 다르게 나타날 수 있다. 특히 동시성 기준의 정확도는 평균 구매간격, 즉 각 고객의 하나의 주문으로부터 그 다음 주문이 이루어지기까지의 평균 기간에 많은 영향을 받을 것으로 예상될 수 있다. 따라서 고객 개개인의 평균 구매간격에 따라 전체 고객을 여러 고객군으로 나누고, 고객군 별로 동시성 기준에 따라 정확도가 변화하는 양상이 다르게 나타나는지 여부를 파악하고자 한다. 이러한 연구 결과는 고객군별 차별화된 마케팅 전략 수립에 활용될 수 있을 것으로 기대한다. 예를 들어 실시간 추천을 위해 평균 구매간격이 8일인 고객의 경우 동시성 기준을 2주일로 적용하여 연관규칙을 도출하고, 평균 구매간격이 30일인 고객의 경우 동시성 기준을 4주일로 적용하여 연관규칙을 도출해야 한다는 등의 차별화된 전략을 수립할 수 있을 것이다.

본 논문의 이후 구성은 다음과 같다. 다음 장인 제 2장에서는 고객관계관리(CRM : Customer Relationship Management), 데이터마이닝(Data Mining), 흥미성 척도(Interestingness Measures), 그리고 온라인 마켓(Online Market)에 대한 기존 연구들을 간략하게 소개한다. 또한 제 3장에서는 연구 모형 및 절차를 소개하며, 이에 대한 실험 결과는 제 4장에 제시한다. 마지막 절인 제 5장에서는 본 연구의 기여 및 한계 그리고 향후 연구방향을 제시한다.

## 2. 관련 연구

고객의 니즈가 복잡하고 다양해지면서 많은 기업들은 다양한 고객관계관리(CRM : Customer Relationship Management) 기법을 도입하고 있다. CRM이란 용어는 1990년대에 정보기술에 기반을 둔 고객 솔루션 제공업체가 처음으로 사용하여 지금까지 활용되고 있다(Parvatiyar and Sheth, 2001). 한편으로 CRM은 IT분야와 관련된 새로운 기법이라기보다는 지속적이고 적절한 커뮤니케이션을 통해 고객의 행동을 이해하고 고객의 행동에 영향을 미칠 수 있도록 하기 위한 전사적인 접근방법이라고 할 수 있다. 기술적 관점에서 보는 CRM은 다양한 분석 기법을 통해 고객 중에서 잠재적으로 수익성이 높은 고객을 파악하여 내부적인 자원을 적정하게 배분한 후 기업성과를 향상시키는 것을 강조한다(Johnson and Selnes, 2004). 최근에는 데이터마이닝의 풍부한 통계적 기법을 CRM에 적용하기 위한 연구(송만석 외, 2008; 하성호, 이재신, 2003)가 활발히 수행되고 있다.

데이터마이닝은 방대한 데이터로부터 유용한 정보나 패턴을 추출하는 기법으로, 통계적 기법, 인공지능 기법 등을 통해 연관관계(Association), 분류(Classification), 군집화(Clustering) 등의 여러 가지 지식을 창출하는 과정(Han and Kamber, 2007)에 널리 활용되고 있다. 특히 연관관계 분석(Agrawal et al., 1993; Agrawal and Srikant, 1994)은 데이터들의 빈도수와 동시 발생 확률을 이용하여 데이터와 데이터간의 관계를 찾고 이를 규칙으로 표현하는 분석 기법으로, 장바구니 분석, 인터넷 쇼핑몰 추천시스템, 교차판매, 매장배치, 카탈로그 설계, 판촉전략 수립 등 다양한 분야(김남규, 2008; 안현철 외, 2006; 윤성준, 2005; Burke, 2000; Wang et al., 2004; Wang et al., 2007)에서 활용되고 있다. 하지만 분석의 결과로 제시되는 연관규칙들의 수가 지나치게 방대하기 때문에, 이들 규칙 중 실현 가능하고 수익성이 있는 규칙만을 식별해내는 작업은 마이닝의 결과에 대한 마이닝이라고 불릴 정도로 복잡할 뿐 아니라, 시간 및 비용 측면에서 많은 추가 부담을 필요로 한다. 이러한 이유로 방대한 연관규칙들 중 의미 있는 규칙들만을 식별해내는 과정을 지원하기 위해서 다양한 흥미성 척도들(Interestingness Measures)이 고안되어 왔다.

다양한 척도들 중 어떤 척도를 기준으로 정하느냐에 따라 도출되는 규칙의 수 및 규칙들의 순위가 결정되기 때문에 척도의 고안 및 선정 작업은 연관규칙 분석의 성패를 좌우하는 가장 중요한 작업으로 알려져 있다. 따라서 다양한 척도들 간의 이론적, 실무적 성능을 평가하기 위한 많은 연구가 수행된 바 있다. Tan et al.(2002)은 흥미성 척도가 가져야 하는 바람직한 5가지 속성을 제시하고, 이에 기반하여 다양한 척도들의 우수성을 평가하였다. 또한 이 연구에서는 각 척도들에 의해 계산된 규칙들의 순위와, 전문가들의 의견을 통해 도출한 규칙들의 우선순위를 비교함으로써 척도들의 신뢰성을 평가하기 위한 실험도 이루어졌다. 또한 Geng and Hamilton(2006)은 발견된 규칙의 흥미성을 판단하기 위한 관점을 9가지로 제시하였으며, 연관분석과 분류에 각기 사용되는 척도들을 통합하기 위한 분석의 틀을 제시하였다. 본 연구에서는 다양한 흥미성 척도들 중 가장 기본적이고 널리 적용되고 있는 신뢰도(Confidence)와 지지도(Support)(Agrawal and Srikant, 1994)를 기반으로 온라인 마켓 데이터에 대한 연관분석을 수행하고자 한다.

온라인 쇼핑은 고객들에게 인터넷을 통한 신속한 제품 구매와 빠른 서비스를 바탕으로 현재까지 큰 성장을 거듭하고 있으며, 2000에서 2010년까지

인터넷 쇼핑 이용률은 약 52% 증가한 것으로 나타났다(한국인터넷진흥원, 2010). 인터넷 쇼핑몰의 성공 배경은 여러 측면에서 찾을 수 있는데, Jarvenpaa and Todd(1997)은 인터넷 쇼핑몰의 가장 중요한 성공 요인으로 고객이 필요한 정보를 신속히 얻을 수 있도록 접속시간과 반응시간을 관리하는 능력을 꼽은 바 있다. 온라인 쇼핑이 발전을 거듭함에 따라, 온라인 시장과 오프라인 시장 중의 소비자 선택에 관한 연구 등 온라인과 오프라인 쇼핑몰의 특성을 비교한 다양한 연구(박철, 2000; Ward, 2000)가 수행된 바 있다. 특히 온라인 쇼핑몰 데이터의 방대함과 체계성으로 인해 온라인 거래 데이터에 대한 장바구니 분석 결과를 상품 추천에 활용하기 위한 연구가 활발하게 이루어지고 있다(강동원, 이경미, 2001; 정영수, 강경화, 2004; 하성호, 박상찬, 2002). 하지만 이들 연구는 서로 상이한 장바구니 기준을 채택함으로써, 개별 연구 성과들을 통합하여 확장된 결론을 도출하기 어렵다는 한계를 갖는다.

# 3. 연구 모형

## 3.1 연구방법 및 제안모형

본 절에서는 장바구니의 동시성 기준을 구매간격에 따라 다양하게 정의하고, 특정 동시성 기준 하에 도출된 연관규칙의 정확도를 파악하기 위한 방법을 제시한다. 본 연구에서 연관규칙의 정확도는 정보검색(Information Retrieval) 분야에서 주로 사용되는 F-score에 기반하여 측정된다. F-score는 Precision과 Recall의 지표를 통해 도출되며, 자세한 식은 다음과 같다. 아래의 식에서 n({Relevant})는 발견되어야 하는 전체 규칙의 수를, n({Retrieved})는 분석에 의해 발견된 규칙의 수를 의미한다. 즉 Precision이 높을수록 규칙 중 의미 없는 규칙의 수가 적음을 의미하며, Recall이 높을수록 의미 있는 규칙 중 발견에서 누락된 규칙의 수가 적음을 의미한다.

$$\text{precision} = \frac{\text{n}(\{\text{Relevant}\} \cap \{\text{Retrieved}\})}{\text{n}(\{\text{Relevant}\})}$$

$$\text{recall} = \frac{\text{n}(\{\text{Relevant}\} \cap \{\text{Retrieved}\})}{\text{n}(\{\text{Relevant}\})}$$

$$\text{F\_score} = \frac{\text{recall} \times \text{precision}}{(\text{recall} + \text{precision})/2}$$

물론 연관관계 분석은 예측 통계가 아닌 기술통계의 영역에 속하는 분석이므로, 연관관계분석을 통해 나온 결과에 대한 정확성을 직접 평가하는 것은 일반적 방법은 아니다. 따라서 정확성에 대한 의미의 확장이 필요하며, 본 연구에서 정확도는 하나의 표본집합에서 도출된 규칙이 다른 표본집합에서도 여전히 유의한 규칙으로 작용하는지 여부를 측정하는 기준으로 사용된다. 이를 위해, 전체 데이터를 구매 발생 시점에 따라서 특정 시점 T 이전의 구매 내력을 담고 있는 Training Set (TS)와 T 이후에 구매 내력을 담고 있는 Validation Set(VS)로 분리하고, VS에 존재하는 규칙을 TS 분석을 통해 어느 정도 발견할 수 있는지를 측정하고자 한다.

이와 같은 정의에 정확도 도출 예가 <그림 2>에 나타나있다. <그림 2>는 VS의 동시성 기준을 1일(1D)로 고정하였을 때, 상위 100개의 규칙에 대한 발견 능력을 평가하는 가상 실험에 대한 측정값이다. 이 실험에서 TS의 동시성 기준은 1일(1D), 1주일(1W) 그리고 2주일(2W)의 다양한 기준이 사용되었다. 동시성 기준을 2주일(2W)로 설정한 마지막 행을 예로 들면, TS에서 600개, VS에서 100개의 규칙이 각각 도출되었으며, TS에서 도출된 600개의 규칙 중 VS에도 존재하는 규칙은

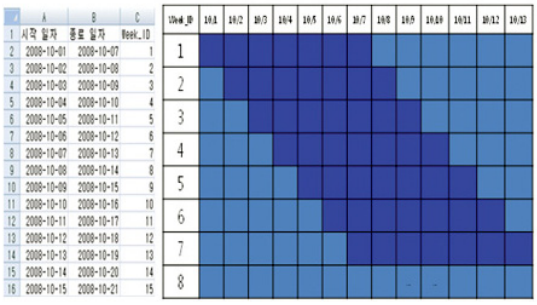
총 40개로 6.67%의 Precision을 나타내고 있다. 한편, VS의 100개 규칙 중 40개만이 TS에서 발견되었으므로 Recall은 40%로 나타났고, 그 결과 F-score는 약 11.43%로 나타남을 보이고 있다. 한편 동시성 기준으로 1W를 사용한 경우의 F-score는 16.67%로 2W에 비해 높게 나타났다. 이 경우 표본집합 VS에서 1W 기준으로 도출한 연관성 규칙이 2W 기준으로 도출한 연관성 규칙에 비해 표본집합 TS의 동시성 기준을 1일로 설정한 연관규칙을 찾아내기에 더욱 적합한 것으로 평가할 수 있으며, 본 연구에서는 이와 같은 경우 동시성 기준 1W가 동시성 기준 2W 보다 더 정확한 것으로 간주한다.

| 기간 | VS | VS ∩ TS | TS | Precision | Recall | F-score |
|---|---|---|---|---|---|---|
| | relevant | relevant and retrieved | retrieved | | | |
| 1D | 100 | 10 | 120 | 0.0833 | 0.1 | 0.0909 |
| 1W | 100 | 25 | 200 | 0.125 | 0.25 | 0.1667 |
| 2W | 100 | 40 | 600 | 0.0667 | 0.4 | 0.1143 |

<그림 2> 정확성 비교를 위한 가상 실험의 측정값

### 3.2 동시성 기준 확장에 따른 장바구니의 재구성

주문 단위로 저장된 거래 내역으로부터 각 동시성 기준에 따른 새로운 장바구니를 정의하기 위해 거래 내역의 확장이 필요하다. 이를 위해 본 연구에서는 슬라이딩 윈도우 기법을 도입하고자 하며, 이 과정이 <그림 3>에 나타나있다. <그림 3>은 윈도우의 크기가 7인, 즉 동시 구매의 기준을 7일로 설정했을 때의 장바구니 ID 할당 과정을 묘사하고 있다. 예를 들어 동시 구매의 기준을 7일로 설정한 경우, 2008년 10월 1일부터 2008년 10월 7일까지 구매한 물품은 모두 동시에 구매된 것으로 간주된다. 만약 단순한 방법으로 10/1~10/7사이에 구매된 물품을 바구니 1로, 10/8~10/14사이에 구매된 물품을 바구니 2로 정의하는 경우를 고려해보자. 이 경우 10월 7일에 구매된 물품과 10월 8일에 구매된 물품은 실제로 하루 간격으로 구매되었음에도 불구하고 서로 다른 바구니에 속하게 되는 이상 현상이 발생한다. 따라서 본 연구에서는 이러한 이상 현상을 제거하기 위해 슬라이딩 윈도우방식을 제안하여 사용하고자 한다. 슬라이딩 윈도우 방식이란 동시 구매의 기준이 n일로 정의되었을 때 (1일~n일), (2일~n+1)일, (3일~n+2일) 등으로 1일 단위로 증가하며 장바구니를 정의하는 방식이다. 이 과정에서 장바구니 ID가 새로 생성되며, 추후 분석에서 이 ID는 주문번호를 대체하여 바구니의 식별자로 사용된다.



<그림 3> 장바구니 확장을 위한 슬라이딩 윈도우의 예(1주일 단위)

## 4. 실험 및 결과분석

### 4.1 실험 대상 및 환경

본 연구에서는 실험을 위해 국내 한 대형 인터넷 쇼핑몰의 최근 2년간 실제 거래 내역을 사용하였다. 대상 기간의 거래 내역 중 약 2만 5천 명의 고객을 표본으로 추출하였으며, 이들 고객이 주문한 전체 내역 약 90만 건이 분석 대상이 되었다. 물품의 경우 물품번호대신 상품명을 식별자로 사

용하였는데, 이는 최소 분석 단위로 지나치게 세분화된 단위를 사용할 경우 의미 있는 분석 결과가 나오기 어렵기 때문이다. 예를 들면 “LG USB 메모리 32GB”라는 물품이 있을 경우, 세부 규격을 제외한 “LG USB 메모리”를 물품에 대한 식별자로 사용하였다. 연관규칙 분석의 전 과정은 SAS Enterprise Miner 4.3 상에서 수행하였다.

앞에서 정의한 바와 같이 본 연구에서 정확도는 하나의 표본집합에서 도출된 규칙이 다른 표본집합에서도 여전히 유의한 규칙으로 작용하는지 여부를 측정하는 기준으로 사용된다. 이를 위해, 전체 데이터를 2009년의 구매 내력을 담고 있는 Training Set(TS)와 2010년의 구매 내력을 담고 있는 Validation Set(VS)로 분리하고, VS에 존재하는 규칙을 TS 분석을 통해 어느 정도 발견할 수 있는지를 측정하였다. 하지만 동일한 고객이 TS와 VS에 동시에 포함되는 경우에는 2009년의 구매 내력이 2010년의 구매에 영향을 끼침으로써 실험의 정확성을 해칠 우려가 있다. 즉 2009년에 냉장고를 구매한 고객의 경우, 이 고객이 2010년에 냉장고를 다시 구매할 가능성은 다른 고객에 비해 매우 낮을 것이다. 이러한 왜곡을 없애기 위해 본 실험에서는 회원번호가 짝수인 회원의 2009년 구매 내역을 TS로 사용하고, 회원번호가 홀수인 회원의 2010년 구매 내역을 VS로 사용하였다.

본 연구의 중요한 목적 중 하나인 고객 유형별 구매의 최적 동시성 기준 비교를 위해, TS와 VS의 모든 고객에 대한 고객 세분화를 수행하였다. 고객 세분화를 위해 고객의 성별, 연령 등의 전통적 기준을 사용하는 방법도 있겠으나, 본 실험에서는 본 연구에서 살펴보고자 하는 고객의 특성과 직접적인 연관이 있는 새로운 기준을 마련하였다. 즉 각 고객에 대해 한 번 구매 후 다음 번 구매가 이루어지기까지의 평균 기간인 평균 구매간격을 계산하였으며, 평균 구매간격이 작을수록 잦은 구매를 하는 고객이고, 그 값이 클수록 구매를 자주 하지 않는 고객임을 의미한다. 평균 구매간격의 계산식은 다음과 같다(단, 대상 기간 중 구매 회수가 1회인 고객은 제외함).

$$(\text{고객 } A\text{의 평균구매간격}) = \frac{\left(\begin{array}{l}(\text{대상 기간중고객 } A\text{의마지막구매일}) \\ -(\text{대상기간중고객 } A\text{의 최초구매일})\end{array}\right)}{(\text{대상기간중고객 } A\text{의구매빈도수}-1)}$$

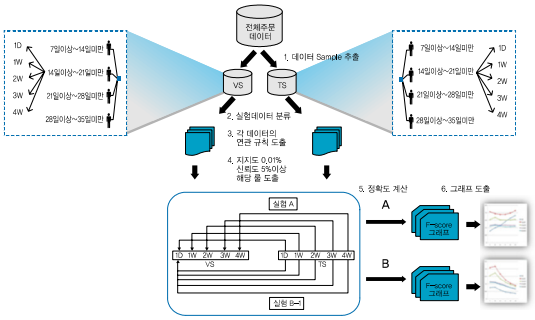
TS와 VS는 각각 평균 구매간격에 따라 여러 고객군으로 나누어지며, 각 고객군의 거래 내역은 구매의 동시성 기준 확장에 따라 슬라이딩 윈도우 기법을 통한 재조직화를 거치게 된다. 본 실험에서는 전체 고객군을 평균 구매간격에 따라 4개의 그룹으로 나누고 5개의 동시성 기준(1일, 1주, 2주, 3주, 4주 단위)을 적용하였으므로, 연관관계 도출의 대상이 되는 집합의 수는 2(TS, VS)×4(고객군 수)×5(동시성 기준 수) = 40개가 된다. 실험에 사용된 두 표본인 TS와 VS의 고객군별 고객의 수와 주문의 수, 고객당 평균 주문수 그리고 확장된 동시성 기준에 따른 주문의 수는 <그림 4>에 요약되어 있으며, 실험 집합 구성을 위한 전체 과정은 <그림 5>에 소개되어 있다.

<그림 5>에 나타난 바와 같이, 본 연구에서는 40개의 집합에 대해 두 가지 유형의 실험을 수행하고자 한다. 본 연구의 핵심 실험은 예측 대상의 동시성 기준, 즉 VS의 동시성 기준이 정해져 있을 때 예측의 정확도를 높이기 위해서 TS의 동시성 기준이 어떻게 설정되는 것이 가장 바람직한가를 살펴보는 것(실험 B)이다. 예를 들면, <그림 5> 실험 B-1는 VS의 동시성 기준을 1일(1D)로 고정한 실험을 나타낸다. 따라서, 만약 이 실험에서 TS의 동시성 기준을 1D로 설정했을 때의 정확도가 다른 동시성 기준을 적용한 경우보다 높게 나타난다면,

| 평균구매간격 (TS) | 고객수 | 주문수 | 고객당 평균주문수 | 확장된 동시성기준 | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | 1D | 1W | 2W | 3W | 4W |
| 7이상~14미만 | 393 | 7926 | 20 | 6652 | 35283 | 55647 | 68983 | 78587 |
| 14이상~21미만 | 527 | 6683 | 13 | 5799 | 33625 | 57147 | 74733 | 88104 |
| 21이상~28미만 | 533 | 4833 | 9 | 4320 | 26186 | 46409 | 62541 | 75697 |
| 28이상~35미만 | 396 | 2804 | 7 | 2548 | 15894 | 28909 | 39916 | 49105 |
| 합계 | 1849 | 22246 | 49 | 19319 | 110988 | 188112 | 246173 | 291493 |

| 평균구매간격 (TS) | 고객수 | 주문수 | 고객당 평균주문수 | 확장된 동시성기준 | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | 1D | 1W | 2W | 3W | 4W |
| 7이상~14미만 | 451 | 16696 | 37 | 13722 | 68870 | 103447 | 122556 | 133544 |
| 14이상~21미만 | 604 | 13505 | 22 | 11483 | 64339 | 105311 | 133085 | 152254 |
| 21이상~28미만 | 526 | 8603 | 16 | 7527 | 44379 | 75678 | 98502 | 115657 |
| 28이상~35미만 | 417 | 5228 | 13 | 4615 | 27976 | 49064 | 65599 | 78652 |
| 합계 | 1998 | 44032 | 88 | 37347 | 205564 | 333500 | 419742 | 480107 |

<그림 4> 고객군별 고객 수 및 주문 수(동시성 기준 확장 적용 후)




<그림 5> 동시성 기준의 정확도 도출 모형(Target = 1D)

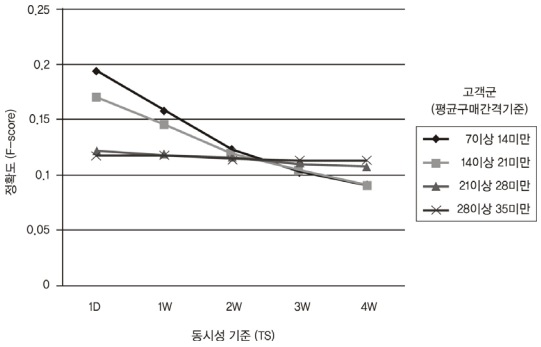
동시성기준 1일의 동시구매에 대한 예측은 1일(1D)의 동시성 기준을 적용하여 분석한 결과를 토대로 이루어지는 것이 바람직함을 의미한다.

실험 B는 세부 실험 5가지(실험 B-1~실험 B-5)로 구성되며, 각각 예측 대상의 동시성 기준(Target)을 1D, 1W, 2W, 3W, 4W로 설정했을 때의 정확성 평가 결과를 나타낸다. 실험 A는 실험 B를 위한 사전 실험으로, TS와 VS의 동시성 기준을 서로 동일하게 설정한 상태에서 수행된다. 이는 TS와 VS의 동시성 기준을 일치시켰을 때, 각 고객군에 따른 예측 정확성이 어떻게 나타나는지를 살펴보기 위한 실험이다.

## 4.2 실험 결과 및 해석

### 4.2.1 실험 A - 동일한 동시성 기준 하의 고객군별 예측 정확도 비교

실험 A는 TS와 VS에 동일한 동시성 기준을 적용했을 때 각 고객군별 예측 정확도가 어떻게 나타나는지 살펴보기 위한 실험이다. <그림 6>은 이 실험의 결과를 보여주며, 이 실험에서 신뢰도는 5%, 지지도는 0.01%의 임계값이 사용되었다. 실험 A 결과 각 고객군 내에서는 다양한 동시성 기준에 따른 정확도의 차이는 거의 없는 것으로 나타났다. 하지만 고객군에 따른 정확도의 차이는 명확하게 나타났는데, 즉 평균 구매간격이 작은 고객군일수록 평균 구매간격이 큰 고객군에 비해 정확도가 높게 나타난 것이다. 이는 가끔(약 30일 간격) 쇼핑몰을 이용하는 고객의 경우 고객의 해당 쇼핑몰에 대한 구매 의존도가 낮으므로, 분석에 사용된 데이터가 해당 고객의 구매 특성을 충분히 반영하고 있지 않기 때문인 것으로 해석될 수 있다. 반면 주문을 자주(약 1.5주 간격) 발생시킨 고객군의 경우, 분석 데이터에 구매 특성이 충분히 분석에 반영되었기 때문에 예측 정확도가 다른 고객군에 비해 높게 나온 것으로 판단된다.



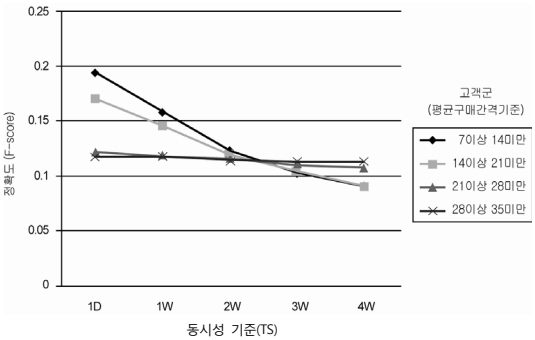

<그림 6> 실험 A - 동일한 동시성 기준 하의 고객군별 예측 정확도 비교

#### 4.2.2 실험 B-동시성 기준에 따른 정확도 비교 실험

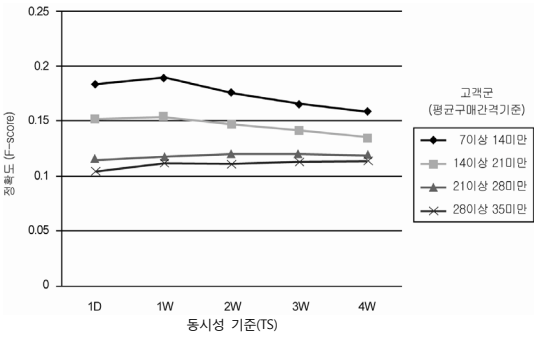
실험 B는 VS의 동시성 기준을 고정하였을 때, TS의 각 동시성 기준에 의해 발견된 연관규칙과 VS에서 발견된 연관규칙이 어느 정도 정확하게 일치하는지를 측정하기 위한 실험이다. 예를 들어 VS의 동시성 기준을 1D(1일)로 고정한 실험 B-1의 경우, TS에서 1주일 단위로 도출한 규칙이 VS에서 도출한 규칙과 20% 일치하고, TS에서 2주일 단위로 도출한 규칙이 VS에서 도출한 규칙과 10% 일치한다면, 각 1D 단위의 연관성 예측을 위한 분석은 2주일 보다는 1주일 단위로 이루어지는 것이 바람직함을 의미한다. 실험 B-1~실험 B-5에서도 실험 A와 마찬가지로 신뢰도 5%, 지지도 0.01%의 임계값이 사용되었으며, 각 실험의 결과는 <그림 7>~<그림 11>에 나타나 있다.

실험 B를 통해 확인할 수 있는 현상은 다음과 같다. 우선 실험 A와 마찬가지로 동시성 기준에 따른 정확도 비교 실험에서도 전반적으로 주문간격이 짧은 고객군이 대체적으로 높은 정확도를 보이는 것으로 나타났다. <그림 7>~<그림 11>의 그래프를 보면 5가지 동시성 기준 모두에서 평균 구매간격이 7이상 14미만인 고객군의 정확도가 가장 높게 나타났다. 이러한 현상은 앞에서 언급한 것과 같이 구매간격이 큰 고객군일수록 구매특성이 충분히 분석 데이터에 반영되었다고 보기 어렵기 때문에 예측 정확도가 비교적 낮게 나타난 것으로 판단된다. <그림 7>의 경우 평균 구매간격이 7이상 14미만인 고객군의 정확도는 동시성 기준이 2W 미만인 구간에서는 가장 높게 나타났지만, 동시성 기준이 길어짐에 따라 정확도가 급격히 떨어져서 2W 이상의 동시성 기준 구간에서는 다른 고객군에 비해 낮은 정확도를 보이는 것으로 나타났다. 이러한 현상은 다음과 같은 이유로 설명될 수 있다.

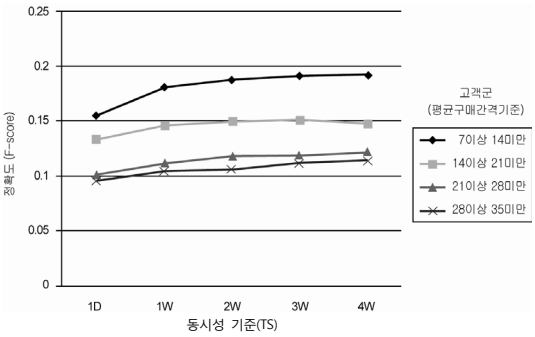
직관적으로, 예측 대상의 환경과 가장 유사한



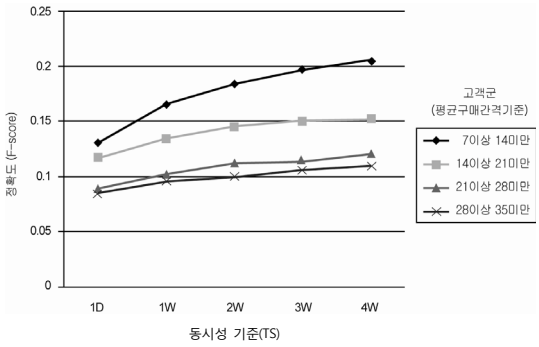

<그림 7> 동시성 기준에 따른 정확도 비교
(예측 대상의 동시성 기준 = 1D)



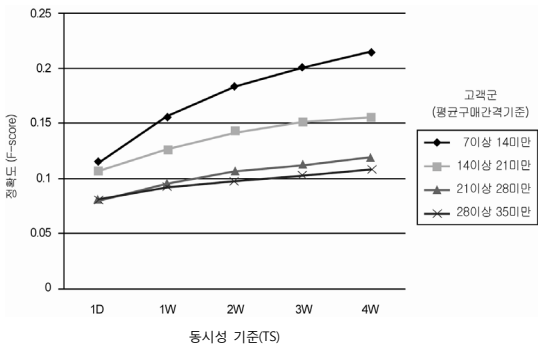

<그림 8> 동시성 기준에 따른 정확도 비교
(예측 대상의 동시성 기준 = 1W)




<그림 9> 동시성 기준에 따른 정확도 비교
(예측 대상의 동시성 기준 = 2W)




<그림 10> 동시성 기준에 따른 정확도 비교
(예측 대상의 동시성 기준 = 3W)




<그림 11> 동시성 기준에 따른 정확도 비교
(예측 대상의 동시성 기준 = 4W)

환경에서의 분석을 통해 도출된 모델의 정확도가 다른 경우에 비해 높게 나올 것으로 예상할 수 있다. 이는 환경의 변화에 따라서 발견되는 규칙들의 평균 신뢰도와 지지도가 크게 달라지기 때문이다. 이는 TS에서 사용된 동시성 기준이 VS의 동시성 기준과 유사할수록 높은 정확도가 나올 것으로 예상됨을 의미한다. 이러한 예상은 실제 실험 결과에서도 유사하게 나타났다. <그림 7>, <그림 8>, 그리고 <그림 11>은 예측 대상의 동시성 기준을 각각 1D, 1W, 4W로 설정한 실험의 결과이며, 이들 실험에서 모든 고객군들에 대해 정확도가 가장 높은 TS의 동시성 기준은 각각 1D, 1W, 4W인 것으로 나타났다. 또한 <그림 9>, <그림 10>의 경우 예측 대상의 동시성 기준은 각각 2W, 3W이며, 정확도가 가장 높은 TS의 동시성 기준은 각각 3W, 4W, 즉 예측 대상의 동시성 기준과 인접한 구간인 것으로 나타났다. 결론적으로 주어진 VS의 동시성 기준에 대해서, 이와 동일하거나 유사한 동시성 기준을 적용한 TS의 분석을 통해 도출된 연관규칙이 보다 정확함을 알 수 있다.

이러한 추이는 모든 고객군에 대해 동일하게 나타나지만, 정확도가 변하는 정도는 고객군에 따라 매우 다르게 나타나는 것으로 파악되었다. 즉 예측 대상의 동시성 기준과 분석 대상의 동시성 기준의 차이에 따른 정확도의 변화는 구매간격이 짧은 고객군에서는 매우 두드러지게 나타나는 반면, 구매간격이 긴 고객군은 큰 차이를 보이지 않는 것으로 나타났다. 즉 <그림 7>~<그림 11>에서 평균 구매간격이 21일 미만인 두 고객군의 경우 각 실험이 진행됨에 따라 우하향 패턴에서 우상향 패턴으로 명확하게 변화함을 알 수 있었다. 하지만 동일한 실험에서 평균 구매간격이 21일 이상인 두 고객군의 경우 모든 실험에서 매우 완만한 기울기를 가지며, 대부분의 실험에서 4W의 동시성 기준을 적용한 경우에 가장 높은 정확도를 갖는 것으로 나타났다. 이 결과에 따르면 구매간격이 짧은 고객군의 경우에는 예측 대상의 동시성 기준과 일치하는 기준을 사용한 분석을 통해 연관규칙을 도출해야 하지만, 구매간격이 긴 고객군의 경우에는 예측 대상의 기준과 무관하게 4W의 동시성 기준에서 연관규칙을 도출하는 것이 바람직한 것으로 나타났다.

이상의 실험에 대한 결과 및 해석을 요약하면 다음과 같다. 우선 전반적인 실험에서 동시성 기준별 연관규칙의 정확도는 단골 고객(평균 구매간격이 짧은 고객)이 비단골 고객에 비해 높게 나타났다.

이는 구매가 너무 뜸한 비단골 고객의 경우, 분석 데이터가 해당 고객의 구매 특성을 충분히 반영하기 어렵기 때문인 것으로 판단된다. 또한 단골 고객의 경우 예측 대상의 동시성 기준과 분석 대상의 동시성 기준이 동일하거나 유사할 때 연관규칙의 정확도가 비교적 우수하게 나타나는 것을 알 수 있었다. 이는 단골 고객의 구매 추천을 위한 분석의 경우, 추천 목적 및 범위에 맞추어 연관규칙을 도출하는 것이 바람직함을 의미한다. 하지만 비단골 고객의 경우 연관규칙의 정확도는 예측 대상의 동시성 기준의 영향을 크게 받지 않았으며, 가장 긴 동시성 기준을 적용하였을 때 전반적으로 높은 정확도를 나타냄을 알 수 있었다. 이는 비단골 고객의 구매 추천을 위해서는 동시성 기준을 비교적 길게 설정한 상태에서 연관규칙을 도출하는 것이 바람직함을 의미한다.

## 5. 결론

온라인 쇼핑몰은 인터넷을 통해 손쉽게 접근이 가능하기 때문에, 최초 구매의사가 발생한 시점으로부터 이에 대한 실제 구매가 실현되기까지의 기간이 오프라인 쇼핑몰에 비해 매우 짧게 나타난다. 이러한 이유로 온라인 쇼핑몰은 단 하나의 물품만을 포함하고 있는 주문이 전체 주문의 절반 이상을 차지하고 있다. 따라서 온라인 쇼핑몰 데이터의 장바구니 분석에 전통적 데이터마이닝 기법을 그대로 적용할 경우, Null Transaction의 수가 지나치게 많음으로 인해 합리적 수준의 지지도(Support)를 만족시키는 규칙을 찾는 것이 매우 어렵게 된다. 이로 인해 온라인 데이터를 대상으로 한 많은 연구는 저마다 동시성 기준을 확장하여 사용하였는데, 이러한 동시성 기준은 명확한 근거나 합의 없이 연구자의 선택에 따라 임의로 선택된 측면이 있다. 따라서 본 연구에서는 온라인 마켓 분석에 적용되는 구매의 동시성 기준을 정확도 측면에서 평가함으로써, 구매의 동시성 기준 선정을 위한 근거를 제시하였다. 또한 구매의 동시성 기준의 정확도가 고객의 평균 구매간격에 따라 상이하게 나타나는 것을 파악하여, 향후 고객의 특성에 따른 차별화된 추천 시스템 구축을 위한 기본 틀을 제시하였다.

본 연구의 실험에는 국내 한 대형 인터넷 쇼핑몰의 최근 2년간 실제 거래 내역이 사용되었다. 대상 기간의 거래 내역 중 약 2만 5천 명의 고객을 표본으로 추출하였고, 이들 고객이 주문한 전체 내역 약 90만 건을 분석 대상으로 사용하였다. 전반적인 실험에서 동시성 기준 별 연관규칙의 정확도는 단골 고객(평균 구매간격이 짧은 고객)이 비단골 고객에 비해 높게 나타났다. 또한 단골 고객의 경우 예측 대상의 동시성 기준과 분석 대상의 동시성 기준이 동일하거나 유사할 때 연관규칙의 정확도가 비교적 우수하게 나타나는 것을 알 수 있었다. 이는 단골 고객의 구매 추천을 위한 분석의 경우, 추천 목적 및 범위에 맞추어 연관규칙을 도출하는 것이 바람직함을 의미한다. 하지만 비단골 고객의 경우 연관규칙의 정확도는 예측 대상의 동시성 기준의 영향을 크게 받지 않았으며, 가장 긴 동시성 기준을 적용하였을 때 전반적으로 높은 정확도를 나타냄을 알 수 있었다. 이는 비단골 고객의 구매 추천을 위해서는 동시성 기준을 비교적 길게 설정한 상태에서 연관규칙을 도출하는 것이 바람직함을 의미한다.

본 연구의 후속연구에서 반드시 다루어져야 할 사항은 다음과 같다. 우선 연관분석은 예측의 영역이 아닌 기술의 영역에 속하는 분석이므로, 연관분석을 통해 도출된 규칙의 우수성을 평가하기 위한 추가 노력이 필요하다. 본 연구에서는 성능의 척도

로 F-score에 근거한 정확도를 채택하였으나, 보다 다양하고 정밀한 평가 기준이 고안될 필요가 있다. 또한 본 실험에 사용된 동시성 기준 중 가장 간격이 큰 기준은 4주(4W)였다. 따라서 동시성 기준이 4주보다 길어질 경우에는 비단골 고객의 동시성 기준 정확도가 어떻게 나타날지는 현재 실험 결과로는 예측하기 어렵다. 즉, 보다 넓은 범위의, 그리고 보다 세분화된 동시성 기준에 대한 추가 실험이 반드시 필요하다. 마지막으로 본 실험에서는 고객의 특성 중 동시성 기준에 가장 크게 영향을 끼칠 것으로 예상된 평균 구매간격만이 고객 분류의 기준으로 사용되었다. 추후 연구에서는 고객 분류 과정에서 보다 많은 변수가 채택되어야 하며, 이를 통해 실질적으로 차별화된 추천시스템을 설계할 수 있을 것으로 기대된다.

## 참고문헌


강동원, 이경미, "인터넷 쇼핑몰에서 원투원 마케팅을 위한 장바구니 분석 기법의 활용", *컴퓨터산업교육학회논문지*, 2권 9호(2001), 1175~1182.

김남규, "장바구니크기가 연관규칙 척도의 정확성에 미치는 영향", *경영정보학연구*, 18권 2호(2008), 95~114.

박 철, "인터넷 정보탐색가치가 인터넷 쇼핑 행동에 미치는 영향에 관한 연구 : 쇼핑몰 방문빈도와 구매의도를 중심으로", *마케팅연구*, 5권 1호(2000), 143~162.

송만석, 박종환, 김삼원, 조윤재, "프로야구구단의 효율적인 CRM을 위한 데이터마이닝 기법의 적용", *한국스포츠산업경영학회지*, 13권 2호(2008), 205~222.

안현철, 한인구, 김경재, "연관규칙기법과 분류모형을 결합한 상품추천시스템 : G인터넷 쇼핑몰의 사례", *Information System Review*, 8권 1호(2006), 181~201.

윤성준, "데이터마이닝 기법을 통한 백화점의 고객 이탈예측 모형 연구", *한국마케팅저널*, 6권 4호(2005), 45~72.

이종민, 정홍, 김진상, "신경망과 연관규칙을 이용한 구매패턴 분류시스템의 구현", *퍼지 및 지능시스템학회*, 8권 5호(2003), 530~538.

정영수, 강경화, "데이터마이닝 기법을 이용한 인터넷 쇼핑몰 사이트의 CRM 사례분석", *경영경제연구*, 27권 1호(2004), 139~156.

정영조, 장대철, 안병훈, "판매자간 경쟁과 구매자간 경쟁을 고려한 온라인 마켓 플레이스의 수수료 구조 분석", *한국경영과학회지*, 34권 1호(2009), 85~100.

하성호, 박상찬, "인터넷 쇼핑몰에서의 지능화된 마케팅과 상품화 계획 기법", *경영정보학연구*, 12권 3호(2002), 71~88.

하성호, 이재신, "데이터마이닝을 활용한 동적인 고객분석에 따른 고객관계관리 기법", *한국지능정보시스템학회논문지*, 9권 3호(2003), 23~47.

한국인터넷진흥원, "2010년 인터넷 이용 실태 조사", 한국인터넷진흥원, 2010, (available at : http://www.kisa.or.kr).

Agrawal, R., T. Imielinski, and A. Swami, "Mining association Rules between Sets of Items in Large Databases", in Proc. ACM SIGMOD International Conference on Management of Data, Washington D. C., (1993). 207~216.

Agrawal, R. and R. Srakant, "Fast Algorithms for Mining Association Rules", *International Conference on Very Large Data Bases, Santiago, Chile*, (1994), 487~499.

Burke. R, "Knowledge-based recommender systems", *Encyclopedia of Library and Information Systems*, Vol.69(2000).

Geng, L. and Hamilton, H. J., "Interestingness Measures for Data Mining : A Survey", *ACM*

*Computing Surveys*, Vol.38, No.3(2006).

Han, J. and M. Kamber, "Data Mining : Concepts and Techiques, Morgan Kaufmann Publishers California, 2007.

Srikka L. Jarvenpaa and Peter A. Todd, "Consumer Reaction to Electronic Shopping on the World Wide Web", *International Journal of Electronic Commerce*, Vol.1, No.2(Winter, 1997), 59~88.

Johnson, M. D. and F. Selnes, "Customer Portfolio Management : Toward a Dynamic Theory of Exchange Relationships", *Journal of Marketing*, Vol.68(2004), 1~17.

Parvatiyar, A. and J. N. Sheth, "Conceptual Framework of Customer Relationship Management", Customer Relationship Management - Emerging Concepts, Tools and Applications, New Delhi, India : Tata/Mc-Graw-Hill, (2001), 3~25.

Tan, P. N., V. Kumar, and J. Srivastava, "Selecting the Right Interestingness Measure for Association Patterns", 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Alberta, Canada, (2002), 32~41.

Wang, W. F., Y. L. Chung, M. H. Hsu, and A. C. Keh, "A Personalized Recommender System for the Cosmetic Business", *Expert Systems with Applications*, Vol.26, No.3(2004), 427~434.

Wang, W. F., Chung, Y. L., Hus, M. H. and Keh, A. C., "A Personalized Recommender System for the Cosmetic Business", *Expert Systems with Applications*, Vol.26, No.3(2007), 427~434.

Ward, M. R., "Will Online Shopping Compete more with Traditional Retailing of Catalog Shopping?", *Working Paper*, Univ. of Illinois, Urban-Champaign, 2000.

Abstract

# An Investigation on Expanding Co-occurrence Criteria in Association Rule Mining

Kim, Misung* · Kim, Namgyu** · Ahn, Jae-Hyeon***

There is a large difference between purchasing patterns in an online shopping mall and in an offline market. This difference may be caused mainly by the difference in accessibility of online and offline markets. It means that an interval between the initial purchasing decision and its realization appears to be relatively short in an online shopping mall, because a customer can make an order immediately. Because of the short interval between a purchasing decision and its realization, an online shopping mall transaction usually contains fewer items than that of an offline market. In an offline market, customers usually keep some items in mind and buy them all at once a few days after deciding to buy them, instead of buying each item individually and immediately. On the contrary, more than 70% of online shopping mall transactions contain only one item. This statistic implies that traditional data mining techniques cannot be directly applied to online market analysis, because hardly any association rules can survive with an acceptable level of Support because of too many Null Transactions.

Most market basket analyses on online shopping mall transactions, therefore, have been performed by expanding the co-occurrence criteria of traditional association rule mining. While the traditional co-occurrence criteria defines items purchased in one transaction as concurrently purchased items, the expanded co-occurrence criteria regards items purchased by a customer during some predefined period (e.g., a day) as concurrently purchased items. In studies using expanded co-occurrence criteria, however, the criteria has been defined arbitrarily by researchers without any theoretical grounds or agreement. The lack of clear grounds of adopting a certain co-occurrence criteria degrades the reliability of the analytical results. Moreover, it is hard to derive new meaningful findings by combining the outcomes of previous individual studies.

In this paper, we attempt to compare expanded co-occurrence criteria and propose a guideline for selecting an appropriate one. First of all, we compare the accuracy of association rules discovered according to various co-occurrence criteria. By doing this experiment we expect that we can provide a guideline for selecting appropriate co-occurrence criteria that corresponds to the purpose of the analysis.

* Master's Course, Graduate School of BIT, Kookmin University
** Assistant Professor, School of MIS, Kookmin University
*** Professor, Graduate School of Information and Media Management, KAIST

Additionally, we will perform similar experiments with several groups of customers that are segmented by each customer's average duration between orders. By this experiment, we attempt to discover the relationship between the optimal co-occurrence criteria and the customer's average duration between orders. Finally, by a series of experiments, we expect that we can provide basic guidelines for developing customized recommendation systems.

Our experiments use a real dataset acquired from one of the largest internet shopping malls in Korea. We use 66,278 transactions of 3,847 customers conducted during the last two years. Overall results show that the accuracy of association rules of frequent shoppers (whose average duration between orders is relatively short) is higher than that of causal shoppers. In addition we discover that with frequent shoppers, the accuracy of association rules appears very high when the co-occurrence criteria of the training set corresponds to the validation set (i.e., target set). It implies that the co-occurrence criteria of frequent shoppers should be set according to the application purpose period. For example, an analyzer should use a day as a co-occurrence criterion if he/she wants to offer a coupon valid only for a day to potential customers who will use the coupon. On the contrary, an analyzer should use a month as a co-occurrence criterion if he/she wants to publish a coupon book that can be used for a month. In the case of causal shoppers, the accuracy of association rules appears to not be affected by the period of the application purposes. The accuracy of the causal shoppers' association rules becomes higher when the longer co-occurrence criterion has been adopted. It implies that an analyzer has to set the co-occurrence criterion for as long as possible, regardless of the application purpose period.

**Key Words** : Data Mining, Online Market Analysis, Market Basket Analysis, Association Rule Mining

# 저 자 소 개

**김미성**

서경대학교 컴퓨터공학과에서 학사 학위를 취득하였으며, 현재 국민대학교 비즈니스IT전문대학원 비즈니스IT전공 석사 과정에 재학 중이다. 주요 관심분야는 데이터베이스, 데이터 관리, 데이터마이닝 등이다.

**김남규**

현재 국민대학교 경영정보학부에서 조교수로 재직 중이다. 서울대학교 컴퓨터공학과에서 학사 학위를 취득하고, KAIST 테크노경영대학원에서 Database와 MIS를 전공하여 경영공학 석사 및 박사학위를 취득하였다. 한국지능정보시스템학회 이사, JITAM 편집위원 및 한국정보시스템학회, 한국경영정보학회 종신회원으로 활동 중이다. 주요 관심분야는 시멘틱 데이터 관리, 데이터베이스 설계 및 데이터마이닝 등이다.

**안재현**

현재 KAIST 정보미디어 경영대학원 교수로 재직 중이며, 서울대학교 산업공학과에서 학사 및 석사, Stanford University에서 Management Science and Engineering 분야 박사학위를 취득한 후 AT&T Bell Lab에서 senior researcher로 근무하였다. 관심분야는 IT 및 미디어 산업의 마케팅 및 전략 분석, Eye-tracking을 이용한 digital contents 평가, Social commerce 등이다.