CRM연구
제4권 제2호(2011. 12)

온라인 연관관계 분석의 장바구니 기준에 대한 연구

김미성
국민대학교 BIT전문대학원
(misung.km@gmail.com)
김남규
국민대학교 경영정보학부
(ngkim@kookmin.ac.kr)

오프라인 쇼핑몰에 비해 온라인 쇼핑몰은 빠르게 접근이 가능하기 때문에 처음 구매의사를 생성하고 실제 구매가 이루어지기까지의 기간이 오프라인 쇼핑몰에 비해 매우 짧게 나타난다. 즉 오프라인 쇼핑몰의 경우 구매 희망 물건을 바로 구매하기 보다는 몇 개의 물건들을 모두 모아서 구매하는 행태가 일반적이다. 하지만, 인터넷 쇼핑몰의 경우 단 하나의 물품만을 포함하고 있는 주문이 전체 주문의 절반이상을 차지한다. 이러한 차이는 온라인 쇼핑몰 거래데이터의 분석을 위해서는 데이터 마이닝 분석에서 사용되어 온 장바구니의 정의에 대한 확장이 필요함을 의미한다. 하지만 현재까지 온라인 데이터를 대상으로 한 장바구니 분석 연구는, 장바구니의 기준 즉 동시구매의 기준에 대한 명확한 근거나 합의 없이 연구자의 선택에 따라 서로 다른 기준으로 수행되어왔다. 따라서 본 연구에서는 온라인 쇼핑몰 분석에 적용되는 동시에 구매되는 물건들에 대한 기준을 고찰해보고 연구모형을 마련하고자 한다.

※ 주제어: 온라인쇼핑몰, 데이터 마이닝, 장바구니 분석, 연관규칙분석

Ⅰ. 서 론

온라인 거래를 통해 축적된 데이터는 그 규모가 방대할 뿐 아니라 필연적으로 전산화되어있고, 대부분의 경우 데이터베이스로 구축이 되어있다는 점에서 많은 연구자들의 분석 대상으로 선호되어 왔다. 이러한 온라인 데이터를 활용한 분석으로는 온라인 쇼핑몰플레이스(Online Marketplace)의 수익률에 대한 연구, 온라인 쇼핑몰의 구매패턴에 대한 연구, 온라인 커뮤니티의 지식생성에 대한 연구, 온라인 데이터를 활용한 마케팅 전략 수립 연구 등이 있으며, 최근 데이터 마이닝의 장바구니 분석을 활용한 연구도 온라인 쇼핑몰을 대상으로 많이 수행되고 있다. 온라인 쇼핑몰을 대상으로 한 장바구니 분석의 경우, 전통적인 데이터 마이닝 분석 기법을 수정 없이 그대로 사용하는 것이 일반적이다.

하지만 오프라인 쇼핑몰에 기반하여 수립된 전통적인 데이터 마이닝 분석 기법을 온라인 쇼핑몰의 데이터 분석에 그대로 적용하는 것이 합당한지에 대해서는 신중하게 살펴볼 필요가 있다. 이는 온라

논문접수일: 2011년 8월 10일 게재확정일: 2011년 11월 5일

인 쇼핑몰과 오프라인 쇼핑몰은 많은 면에서, 특히 접근성 면에서 매우 상이한 특성을 보이기 때문이다. 예를 들어 데이터 마이닝의 가장 고전적 사례인 월마트 사례의 경우, 고객들은 매장에 1주일 혹은 1달에 한 번씩 방문하게 되며, 한 번 방문 시 그 동안 구매하고자 했던 다수의 물품들을 한꺼번에 구매한다는 가정을 기저에 내포하고 있다. 즉 오프라인 쇼핑몰의 경우 구매 결정 후 실제 구매가 실현되기까지 소요되는 기간이 비교적 길다는 특성을 갖는다. 이와 대조적으로 온라인 쇼핑몰의 경우 접근성이 매우 뛰어나기 때문에, 구매 결정을 내린 고객은 구매 실현을 굳이 지연시킬 필요 없이 인터넷을 통해 곧바로 주문을 생성하는 것이 일반적이다.

이처럼 온라인 쇼핑몰과 오프라인 쇼핑몰의 구매의 실현이 지연되는 기간의 차이는 데이터 마이닝의 장바구니 분석 과정에도 영향을 끼치게 된다. 즉 오프라인 쇼핑몰에서 하나의 장바구니에 포함될 물건들이 온라인 구매의 경우에는 여러 주문으로 나뉘어져 발생하게 될 가능성이 높다는 것이다. 직관적으로도 대형 마트에서 단 하나의 물품을 들고 계산대에 서 있는 고객을 찾기는 쉽지 않지만, 인터넷 쇼핑몰에서 단 하나의 물품을 구매하는 고객을 찾기는 그다지 어렵지 않음을 알 수 있다. 온라인 쇼핑몰 데이터에 대한 장바구니 분석에 전통적인 데이터 마이닝 기법을 적용할 경우, Null Transaction의 수가 지나치게 많음으로 인해 합리적인 지지도(Support)를 만족시키는 연관규칙을 찾기가 어려움을 의미한다. 이에 대한 대안으로 단 하나의 물품을 포함하는 주문을 아예 분석 대상에서 제외할 경우, 전체 주문의 반 이상 해당되는 주문을 인위적으로 배제함으로 인해 분석 결과의 정당성을 확보하기가 어렵게 된다.

이러한 현상은 온라인 쇼핑몰 거래 데이터의 분석을 위해서는 전통적인 데이터 마이닝 분석에서 사용되어 온 장바구니의 정의에 대한 확장이 필요함을 암시하고 있다. 하나의 장바구니를 정의하기 위해 가능한 기준들은 〈그림 1〉을 통해 설명 가능하다. 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 기준 ②를 채택한 것이다. 마지막으로 기준 ①은 기간에 관계없이 특정 회원이 구매한 모든 물품을 하나의 장바구니에 담긴 것으로 간주하는 상황을 나타낸다. 어떤 기준이 더 합리적인가에 대한 판단에 앞선 더욱 중요한 문

〈그림 1〉 장바구니에 대한 다양한 기준

회원번호	구매일자	주문번호	물품번호
A	1/1	O_01	P_01
A	1/1	O_01	P_02
A	1/1	O_02	P_03
A	1/2	O_03	P_04
B	1/1	O_04	P_05

제는, 이들 기준 중에 어떤 기준을 적용하여 온라인 데이터에 대한 장바구니 분석을 수행할 지에 대한 명확한 가이드라인이 없어서 임의의 방식대로 연구가 이루어지고 있다는 것이다. 예를 들어 온라인 쇼핑몰을 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 쇼핑몰에 대해 장바구니 분석이 서로 상이한 장바구니 기준 하에서 수행될 경우, i) 연구자의 의도에 따라 장바구니의 기준을 임의로 선정할 우려가 있고 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. 따라서 본 연구에서는 이러한 한계를 극복하기 위해 다음의 2가지 목표 하에 연구를 수행하고자 한다.

목표 1: 온라인 쇼핑몰 분석에 적용되는 구매의 동시성 기준을 마련하고, 기준 선정에 대한 타당성을 제시한다.

온라인 쇼핑몰에서의 구매를 실현시키는데 지연되는 기간을 여러 가지로 가정하고, 이들 각각에 대한 연관성 분석을 통해 합리적인 장바구니 기준을 도출한다. 합리성의 근거로는 흥미성 척도의 일관성을 사용하고자 하며, 구매를 실현시키는데 지연되는 기간을 어떻게 설정했을 때 가장 일관성 있는 결과를 도출하는지를 측정하고자 한다. 이를 통해 향후 온라인 쇼핑몰의 분석에 대한 연구 및 프로젝트 수행 시 적용될 장바구니 기준에 대한 가이드라인을 제시하고자 한다. 즉 다양한 데이터 마이닝 분석 기법들은 전통적인 오프라인 쇼핑몰에 기반을 두고 수립되어 있고, 실제로 많은 연구 및 프로젝트가 온라인 데이터를 대상으로 분석을 수행하고 있기 때문에 분석 도구와 대상간의 간격이 존재하고 있는 상황을 인식하고, 전통적인 장바구니 분석 기법을 온라인 데이터에 적용하기 위한 방안을 제시하고자 한다.

목표 2: 전체 고객군 별로 성별, 연령 등 고객신상정보와 관련된 기준의 고객군과 쇼핑몰에서 물건을 자주 구매하는지에 평균 구매간격에 대한 거래 데이터를 가지고 실험을 진행할 것이고 이를 상품추천시스템 구축 등의 마케팅 전략 수립에 활용한다.

구매를 실현시키는데 지연되는 기간은 고객들의 특성에 따라 다르게 나타날 수 있다. 즉, 인터넷 쇼핑몰의 사용빈도가 높은 어떤 고객군은 구매실현 지연기간이 매우 짧게 나타날 것이고, 반대로 사용빈도가 낮은 고객군은 그 기간이 길게 나타날 것이다. 따라서 성별, 연령에 따라 고객군을 분류한 뒤 각 고객군별 실험을 수행하고 결과를 비교하여, 그 해석을 마케팅 전략 수립에 활용하고자 한다. 또한 고객의 거래데이터를 이용하여 구매를 자주하는 고객과 그렇지 않은 고객군으로 평균구매간격에 따라 나누어 고객군 별로 실험을 수행할 것이다.

II. 이론적 배경

데이터 마이닝은 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 & Kamber 2007]에 널리 활용되고 있다. 특히 연관관계 분석[Agrawal et al. 1993; Agrawal & 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 가져야하는 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 & 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 & Srikant 1994]를 기반으로 온라인 쇼핑몰 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 있으며, Jarvenpaa & 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 장바구니의 기준을 설정한 근거를 특별히 제시하지 않았을 뿐 아니라, 사용한 기준 간에도 서로 차이가 있어서 개별 연구 성과들을 통합하여 의미있는 결론을 내리기 어렵다는 한계를 갖는다.

Ⅲ. 연구방법 및 내용

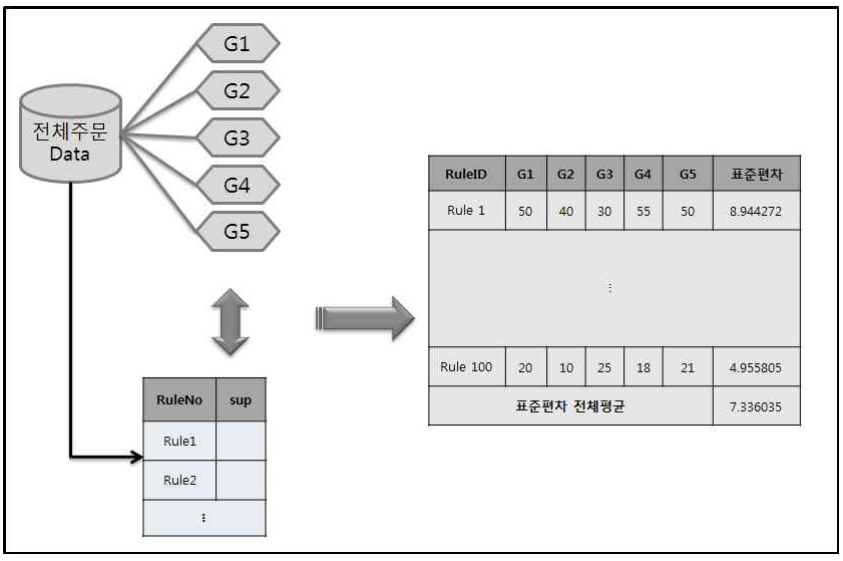
3.1 연구방법 및 제안모형

본 연구의 첫 목표는 온라인 쇼핑몰에서의 구매의 실현을 지연시키는 기간을 여러 가지로 가정하고, 이들 각각에 대한 연관성 분석을 통해 합리적인 장바구니 기준을 도출하는 것이다. 합리성의 근거로는 흥미성 척도의 일관성을 사용하고자 하며, 구매의 실현을 지연시키는 기간을 어떻게 설정했을 때 가장 일관성 있는 결과를 도출하는지를 측정하고자 한다. 흥미성 척도의 비일관성은 다음의 식에 의해 계산되며, 자세한 설명은 〈그림 2〉를 통해 가능하다. 다음의 식에서 $Measure$는 1주일 기준 지지도, 1달 기준 지지도, 1주일 기준 신뢰도, 1달 기준 신뢰도 등의 척도를 나타내며, N은 전체 데이터로부터 도출해 낸 규칙의 수를 의미한다. 또한 $Rule_i$는 i번째 규칙의 해당 척도값을 나타낸다. $\sigma(Rule_i)$는 i번째 규칙이 갖는 흥미성 척도값의 표준편차를 의미하며, 이는 〈그림 2〉에서 설명하도록 한다. 결론적으로, 계산된 값이 낮을수록 비일관성이 낮으며 우수한 척도임을 나타낸다.

$$\text{Inconsistency}(Measure) = \frac{\sum_{i=1}^{N}\sigma(Rule_i)}{N}$$

〈그림 2〉에서 비일관성 계산의 첫 단계는 전체 데이터에 대한 연관분석을 통해 규칙을 도출하는

〈그림 2〉 흥미성 척도의 비일관성 도출 모형

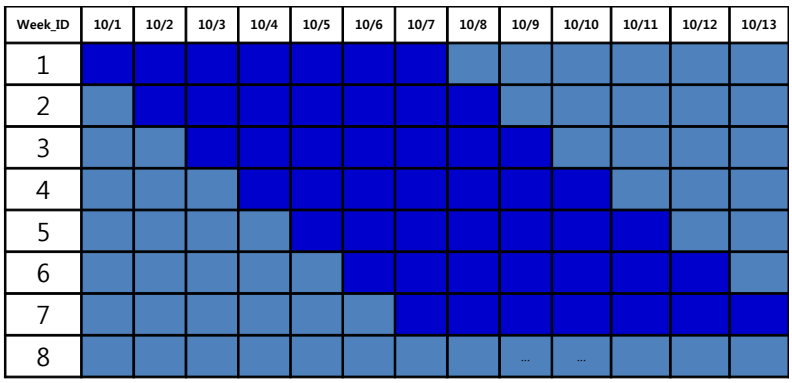


과정이다. 본 예에서는 흥미성 척도로 지지도를 사용하며, 지지도 상위 규칙 100개를 도출하는 것으로 가정한다. 또한 전체 데이터를 임의의 개수의 그룹으로 나누는데, 본 예에서는 G1 ~ G5의 다섯 개의 그룹으로 나누는 것을 가정하였다. 다음 단계는 이미 도출한 100개의 규칙 각각에 대해, 다섯 개 그룹 각각에서 해당 규칙이 갖는 지지도 값을 조사하는 것으로 〈그림 2〉의 우측에 나타나 있다. 다음 단계는 각 규칙이 나타낸 다섯 개의 지지도 값의 표준편차를 구하는 단계이다. 마지막으로 이렇게 계산된 표준편차 100개의 평균을 구하고 이 값을 해당 흥미성 척도의 비일관성 값으로 사용한다. 이 값이 낮을수록, 즉 규칙들의 표준편차의 평균이 낮을수록 규칙들이 각 그룹에 대해 일관적인 지지도를 가짐을 의미하며, 이는 곧 해당 분석에 사용된 구매실현지연기간이 적절하게 설정되었음을 나타낸다.

다음으로, 주문 단위로 저장된 거래 내역으로부터 각 구매의 실현을 지연시키는 기간에 따른 새로운 장바구니를 정의하기 위해 추가적인 변환 작업이 필요하다. 이 작업을 위해 본 연구에서는 슬라이딩 윈도우 기법을 도입하고자 하며, 이는 〈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 간주되어야 한다. 만약 단순한 방법으로 10/1 ~ 10/7 사이에 구매된 물품을 바구니 1로, 10/8 ~ 10/14 사이에 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. 동일한 장바구니를 표시하기 위해 장바구니 ID가 새로 생성되며, 추후 분석에서 이 ID는 주문번호를 대체하여 식별자로 사용된다.

〈그림 3〉 장바구니 확장을 위한 슬라이딩 윈도우의 예 (일주일 단위)

	A	B	C
1	시작 일자	종료 일자	Week_ID
2	2008-10-01	2008-10-07	1
3	2008-10-02	2008-10-08	2
4	2008-10-03	2008-10-09	3
5	2008-10-04	2008-10-10	4
6	2008-10-05	2008-10-11	5
7	2008-10-06	2008-10-12	6
8	2008-10-07	2008-10-13	7
9	2008-10-08	2008-10-14	8
10	2008-10-09	2008-10-15	9
11	2008-10-10	2008-10-16	10
12	2008-10-11	2008-10-17	11
13	2008-10-12	2008-10-18	12
14	2008-10-13	2008-10-19	13
15	2008-10-14	2008-10-20	14
16	2008-10-15	2008-10-21	15



3.2 실험계획

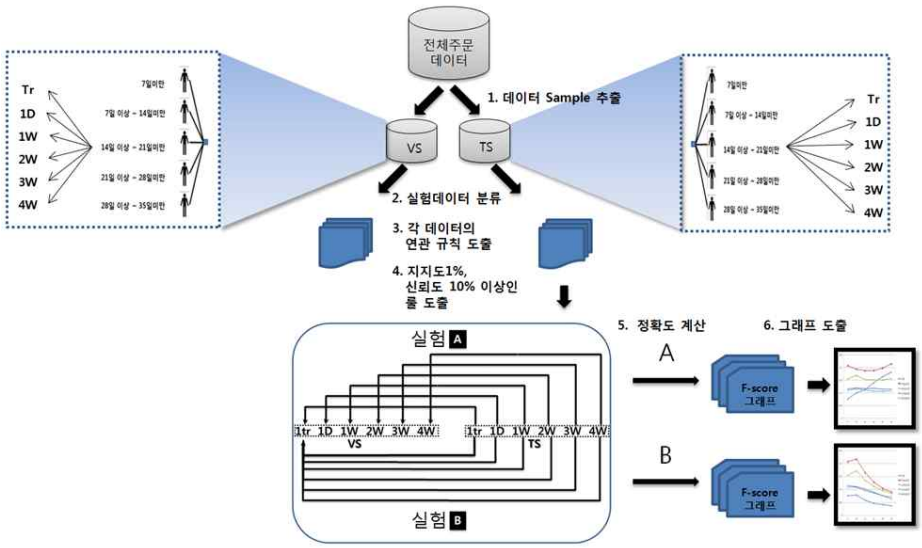
본 연구의 실험을 위해 DATA로는 최근 2년치의 인터넷 쇼핑몰 데이터를 사용하고자 한다. 2년치의 데이터는 각각 1년씩 나누어 전년도(VS)와 후년도(VS)로 나누어 룰을 도출할 것이다. 본 연구에서 결정한 실험의 규모 및 사용 데이터는 다음과 같으며, 전체 실험 개요도는 〈그림 4〉에 제시되어 있다.

〈그림 4〉는 앞에서 부분적으로 설명한 연구 모형을 종합한 전체 개요도이다. 〈그림 4〉의 전체 모형에 대해 본 연구에서는 실험 A와 실험 B의 두 가지 방법으로 정확도를 측정하고자 한다. 실험 A는 TS와 VS의 동시성 기준을 동일하게 설정한 상태에서 수행되는 실험이다. 즉 어떤 동시성 기준이 서로 다른 거래 내역에 대해서 일관성 있는 연관규칙을 도출하는지를 측정하기 위한 실험이다. 한편 실험 B는 VS의 동시성 기준을 주문단위(Tr)로 고정한 상태에서 TS의 각 동시성 기준별 연관규칙이 VS의 연관규칙을 얼마나 정확하게 발견하는지를 측정하기 위한 실험이다. 실험 B의 경우 주문이 이루어지는 순간의 실시간 추천을 위해 어떤 동시성 기준으로 과거 구매 이력에 대한 분석을 수행해야 하는지를 결정하기 위한 근거를 마련하기 위한 목적으로 수행된다.

Ⅳ. 결론

본 연구에서는 온라인 마켓 분석에 적용되는 구매의 동시성 기준을 마련하고, 기준 선정에 대한 근거를 제시하기 위한 분석의 틀을 제시하였다. 또한 구매의 동시성 기준이 성별, 연령 등 고객의 특성에 따라 상이하게 나타내는 것을 파악함으로써, 향후 차별화된 추천 시스템 구축을 위한 기본 방향

〈그림 4〉 평균구매간격 고객군의 실험 개요도



을 제시하였다. 본 연구에서 제안된 연구모형은 향후 온라인 마켓의 분석에 대한 연구 및 프로젝트 수행 시, 성별이나 연령별로 다른 동시성기준을 적용하여 분석해야 함을 실제 실험을 통해 확인함으로써 향후 고객에게 일정한 기준의 분석으로 제품을 추천하는 것이 아니라 성별 연령별로 다른 동시성 기준을 적용하여 추천함으로써 향후 매출 증대에 기여할 수 있을 것으로 기대된다. 즉 인터넷 쇼핑몰에서 실시간 추천을 할 경우 추천 상품을 어떤 동시성 기준으로 과거 구매 이력을 참조하여 추천할 지에 대한 틀을 마련하였다는 점에서 의의를 갖는다.

본 연구의 후속 연구에서는 성별과 연령 외의 다른 성격의 고객군을 나누어 동시성 기준을 더 세밀하게 분석함으로써 연관성 기반 상품 추천시스템의 적중률을 더욱 높일 수 있을 것으로 기대된다. 특히 각 고객에 대해 특정 기간에 이루어진 주문의 수를 파악함으로써 고객별 평균 주문 간격을 도출하고, 이 간격에 따라 고객군을 나누어서 고객군별의 동시성 기준을 파악하기 위한 연구를 현재 진행 중에 있다.

참고문헌

강동원, 이경미(2001), "인터넷 쇼핑몰에서 원투원 마케팅을 위한 장바구니 분석 기법의 활용," **컴퓨터산업교육학회논문지**, 제2권, 제9호, pp. 1175-1182, 2001.

김남규(2008), "장바구니 크기가 연관규칙 척도의 정확성에 미치는 영향," **경영정보학연구**, 제18권, 제2호, pp. 95-114, 2008.

박철(2000) "인터넷정보탐색가치가 인터넷 쇼핑 행동에 미치는 영향에 관한 연구: 쇼핑몰 방문빈도와 구매의도를 중심으로," **마케팅연구**, 제5권, 제1호, pp. 143-162, 2000.

송만석, 박종환, 김삼원, 조윤재(2008), "프로야구 구단의 효율적인 CRM을 위한 데이터 마이닝 기법의 적용," **한국스포츠산업경영학회지**, 제13권, 제2호, pp. 205-222, 2008.

안현철, 한인구, 김경재(2006), "연관규칙기법과 분류모형을 결합한 상품추천시스템 : G인터넷 쇼핑몰의 사례," **Information System Review**, 제8권, 제1호, pp. 181-201, 2006.

윤성준(2005), "데이터 마이닝 기법을 통한 백화점의 고객 이탈예측 모형 연구," **한국마케팅저널**, 제6권, 제4호, pp. 45-72, 2005.

정영수, 강경화(2004), "데이터 마이닝 기법을 이용한 인터넷 쇼핑몰 사이트의 CRM 사례분석," **경영경제연구**, 제27권, 제1호, pp. 139-156, 2004.

하성호, 박상찬(2002), "인터넷 쇼핑몰에서의 지능화된 마케팅과 상품화 계획 기법," **경영정보학연구**, 제12권, 제3호, pp. 71-88, 2002.

하성호, 이재신(2003), "데이터 마이닝을 활용한 동적인 고객분석에 따른 고객관계관리 기법," **한국지능정보시스템학회논문지**, 제9권, 제3호, pp. 23-47, 2003.

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

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

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

Burke. R(2000), "Knowledge-based recommender

systems," Encyclopedia of Library and Information Systems, Vol. 69, 2000.

Geng, L. and Hamilton, H. J.(2006), "Interestingness Measures for Data Mining: A Survey," *ACM Computing Surveys*, Vol. 38, No. 3, 2006.

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

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

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

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

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

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

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

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

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

Kim, Misung* · Kim, Namgyu**

Abstract

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.

* Master's Course, Graduate School of BIT, Kookmin University
** Assistant Professor, School of MIS, Kookmin University

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. 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.

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