DOI: 10.3745/KIPSTD.2010.17D.1.1

연관규칙 흥미성 척도의 실용성 향상을 위한 장바구니 크기 효과 반영 방안

김 원 서[†] · 정 승 렬[††] · 김 남 규[†††]

요 약

연관규칙 마이닝은 물품들 간의 동시 구매 패턴 파악에 사용되는 대표적 마이닝 기법 중 하나로, 카탈로그 설계, 교차판매, 매장배치 등 다양한 마케팅 전략 수립에 활용된다. 방대한 데이터로부터 도출된 많은 연관규칙 중 수익성이 있는 규칙만을 식별해 내는 작업은 지나치게 많은 시간 및 비용을 필요로 한다. 따라서 연관규칙들의 흥미성 평가 과정을 신속하고 체계적으로 수행하기 위해 다양한 흥미성 척도들이 고안되어 왔다. 하지만 신뢰도와 지지도를 비롯한 대다수의 척도들은 대상 물품들의 발생 빈도수에만 근거하여 도출되므로, 실제 판매 현상을 정확하게 반영하지 못한다는 한계를 갖는다. 예를 들어, 기존의 척도는 매우 큰 장바구니에서 동시 구매된 한 건의 거래와 작은 크기의 장바구니에서 동시 구매된 한 건의 거래를 동일한 빈도로 측정한다. 그런데 매우 큰 장바구니에서는 서로 연관관계가 없는 물품들이 우연히 동시에 존재할 가능성이 크므로, 이에 대한 보정이 이루어지는 것이 타당하다. 기존의 척도들과 달리, 본 논문에서는 장바구니 크기 효과를 반영한 흥미성 척도를 새롭게 소개한다. 제안하는 척도는 큰 바구니에서 발생한 패턴과 작은 바구니에서 발생한 패턴에 대해 상이한 가중치를 부여하는 방식으로 계산됨으로써, 우연히 발생한 패턴으로 인해 결과가 왜곡되는 현상을 최소화할 수 있을 것으로 기대된다. 또한, 시뮬레이션 데이터 및 실 데이터에 대한 실험을 통해 제안하는 척도와 기존 척도가 다양한 환경 하에서 보이는 정확성과 일관성을 분석하고 그 결과를 제시하였다.

키워드 : 연관규칙 마이닝, 데이터 마이닝, 장바구니 분석, 흥미성 척도

Utilizing the Effect of Market Basket Size for Improving the Practicality of Association Rule Measures

Won Seo Kim[†] · Seung Ryul Jeong[††] · Nam Gyu Kim[†††]

ABSTRACT

Association rule mining techniques enable us to acquire knowledge concerning sales patterns among individual items from voluminous transactional data. Certainly, one of the major purposes of association rule mining is utilizing the acquired knowledge to provide marketing strategies such as catalogue design, cross-selling and shop allocation. However, this requires too much time and high cost to only extract the actionable and profitable knowledge from tremendous numbers of discovered patterns. In currently available literature, a number of interest measures have been devised to accelerate and systematize the process of pattern evaluation. Unfortunately, most of such measures, including support and confidence, are prone to yielding impractical results because they are calculated only from the sales frequencies of items. For instance, traditional measures cannot differentiate between the purchases in a small basket and those in a large shopping cart. Therefore, some adjustment should be made to the size of market baskets because there is a strong possibility that mutually irrelevant items could appear together in a large shopping cart. Contrary to the previous approaches, we attempted to consider market basket's size in calculating interest measures. Because the devised measure assigns different weights to individual purchases according to their basket sizes, we expect that the measure can minimize distortion of results caused by accidental patterns. Additionally, we performed intensive computer simulations under various environments, and we performed real case analyses to analyze the correctness and consistency of the devised measure.

Keywords : Association Rule Mining, Data Mining, Market Basket Analysis, Interest Measures

1. 서 론

정보통신의 발달로 접근 가능한 비즈니스 데이터의 양이 기하급수적으로 증가함에 따라, 데이터의 수집 자체보다는 방대한 양의 데이터에서 유용한 정보와 지식을 추출하여 의

† 준 회 원 : 국민대학교 비즈니스IT전문대학원 석사과정
†† 종신회원 : 국민대학교 비즈니스IT전문대학원 교수
††† 정 회 원 : 국민대학교 비즈니스IT전문대학원 조교수
논문접수 : 2009년 12월 14일
수 정 일 : 2009년 12월 25일
심사완료 : 2009년 12월 25일

사결정에 활용하는 능력이 기업 경쟁력의 핵심으로 대두되고 있다. 이러한 패러다임의 변화는 최근 방대한 데이터로부터 흥미롭고 유용한 지식을 발굴해내는 기법인 데이터 마이닝(data mining)에 대한 관심의 증가를 가져왔으며, 마이닝 분야에서 얻어진 이론적 연구의 성과는 기업경영, 생산관리, 그리고 시장분석에서부터 공학설계와 과학탐구에 이르기까지 광범위한 응용 분야에서 활용되고 있다. 다양한 데이터 마이닝 기법 중 특히 연관규칙 마이닝(association rule mining)은 물품들 간의 동시 구매 패턴 파악에 사용되는 기법으로서, 카탈로그 디자인, 교차판매, 매장배치 등 다양한 마케팅 전략에 활용되고 있다. 하지만 연관규칙 마이닝의 수행이 항상 수익 창출로 직결되지는 않는데, 그 근본 원인은 분석의 결과로 제시되는 연관규칙들의 수가 너무 많다는 것에서 찾을 수 있다. 즉, 방대한 거래 데이터로부터 도출된 물품 집합 간의 연관규칙의 수 또한 방대하기 때문에, 이들 규칙 중 실현 가능하고 수익성이 있는 규칙만을 식별하는 작업은 마이닝의 결과에 대한 마이닝이라고 불릴 정도로 복잡할 뿐 아니라, 시간 및 비용 측면에서 많은 추가 부담을 필요로 한다.

연관규칙 마이닝의 결과로 도출된 연관규칙들 중 어떤 것이 흥미로우며 의미 있는 규칙인지를 체계적이고 객관적으로 평가하기 위해 다양한 흥미성 척도(interesting measure)들이 고안되어 왔다. 각 척도는 도출된 연관규칙들에 대해 척도마다 고유한 방식으로 점수를 계산하고 순위를 부여한 뒤, 높은 순위의 규칙일수록 의미 있는 패턴임을 시사한다. 이들 척도 중 신뢰도(confidence)와 지지도(support)를 비롯한 대다수의 척도들은 대상 물품들의 발생 빈도수에 근거하여 도출되며, 기본적으로 동시에 자주 구매되는 물품 집합들 간에는 강한 연관관계가 존재함을 나타낸다. 이러한 빈도수 기반 척도들은 통계학에 이론적 배경을 두고 있으며, 요약정보를 표현하기 위해 분할표를 사용한다. 예를 들어 <표 1>의 분할표는 기저귀를 구매한 전체 고객 20명 중 맥주를 함께 구매한 고객이 모두 15명임을 나타내고, 따라서 연관규칙 "기저귀 → 맥주" 의 신뢰도는 75%임을 나타낸다.

분할표는 방대한 양의 거래 데이터로부터 관심의 대상이 되는 물품들의 개별 발생 빈도수 및 동시 발생 빈도수를 요약한 표로, 이를 통해 물품들 간의 연관성을 한 눈에 들여다 볼 수 있다는 장점이 있다. 하지만 분할표는 실제 판매 데이터가 내포하는 다양한 정보 중 일부 정보만을 요약해서 보여주기 때문에, 분할표에만 근거해서 도출된 연관규칙은 실제 판매 현상을 정확하게 반영하지 못하는 경우가 많다. 분할표 작성 과정에서 누락되는 대표적인 정보로는 물품의 주관적 가중치, 장바구니의 크기, 개별 물품의 가격 및 수량을 들 수 있다. 이들 정보들은 최근 흥미성 척도 관련 연구에서 매우 중요한 이슈로 다루어지고 있지만, 장바구니의 크기가 연관규칙에 미치는 영향을 분석하기 위한 시도는 거의 찾아보기 어려운 실정이다. 하지만 매우 큰 장바구니에서 동시 구매된 한 건의 거래와 상대적으로 작은 크기의 장바구니에서 동시 구매된 한 건의 거래를 동일한 빈도로 측정하는 기존의 흥미성 척도는 해당 물품 집합들의 순수한 연관관계를 나타낸다고 볼 수 없다. 그 이유는 다음의 <표 2>를 통해 설명될 수 있다.

<표 2>는 평균 장바구니의 크기가 서로 다른 소형 마트와 대형 마트의 거래내역의 예를 보여주고 있다. 위 거래내역에서 "맥주 → 땅콩" 그리고 "맥주 → 샴푸"의 연관규칙을 전통적 척도인 신뢰도를 기준으로 평가한다고 가정하자. 연관규칙 "맥주 → 땅콩"의 신뢰도는 대형 마트의 경우와 소형 마트의 경우 모두 50%로 동일하게 나타난다. 하지만 "맥주 → 샴푸"의 경우 대형 마트에서는 50%의 신뢰도를 보이는 반면 소형 마트에서는 0%의 신뢰도를 나타낸다. 이와 같이 신뢰성을 비롯한 기존의 흥미성 척도가 거래 환경에 따라 상이한 결과를 제시할 수밖에 없는 원인은, 장바구니의 크기가 바구니에 포함된 빈발 패턴의 수에 영향을 미친다는 측면에서 찾을 수 있다. 예를 들어 동시에 자주 구매되는 물품들의 집합을 하나의 패턴이라고 간주할 때, 하나의 바구니에는 하나의 패턴만이 포함되어 있을 수도 있지만 여러 패턴이 하나의 바구니에 동시에 존재할 수도 있는 것이다. 또한 바구니의 크기가 클수록 동시에 존재하는 패턴의 수는 많아질 것이 예상되며, 이 경우 서로 다른 패턴에 각기 속하는 물품들이 동시에 자주 출현하게 되어 마치 새로운 패턴을 형성한 것으로 오인될 가능성이 증가하게 된다. 따라서 물품 개수가 많고 평균 장바구니의 크기가 큰 대형 마트에서 도출된 연관규칙이 상대적으로 물품 개수 및 장바구니의 크기가 적은 소형 마트에서도 적용될 것을 기대하기엔 무리가 있다. 이처럼 기존의 척도가 장바구니 크기 등 분석 환경에 따라 서로 다른 결과를 제시하는 현상을 완화시키기 위해서, 큰 바구니에서 발생한 패턴의 가중치와 작은 바구니에서 발생한 패턴의 가중치를 상이하게 부여하여 흥미성 척도를 보정하기 위한 노력이 필요하다.

이러한 기존 연구의 한계를 극복하기 위해 본 연구에서는 장바구니의 크기 효과를 반영한 척도인 결합력을 소개하고, 결합력과 기존 척도들 간의 우수성을 정확성(correctness) 및 일관성(consistency) 기준에서 평가하고자 한다. 즉 기존

〈표 1〉 분할표의 간단한 예

	맥주	맥주^C	Σrow
기저귀	15	5	20
기저귀^C	5	75	80
Σcol	20	80	100

〈표 2〉 소형 마트와 대형 마트의 거래내역

소형마트		대형마트	
S1	맥주, 땅콩	L1	맥주, 땅콩, 펜, 노트
S2	맥주, 스넥	L2	맥주, 스넥, 샴푸, 린스
S3	식빵, 우유	L3	식빵, 우유, 펜, 노트
S4	식빵, 생크림	L4	식빵, 생크림, 샴푸
S5	샴푸, 린스	L5	샴푸, 린스, 펜
S6	펜, 노트	L6	펜, 노트, 샴푸

의 연구에서는 고려되지 않았던 장바구니 크기 효과를 흥미성 척도 계산 과정에 반영하고, 이를 통하여 다양한 분석 환경에서도 비교적 일관적인 결론을 제시할 수 있는 견고한 척도를 소개하고자 한다. 또한 제안하는 결합력의 정확성 및 일관성 평가를 위해 장바구니의 크기 및 물품 개수를 변화시켜 가면서, 결합력 및 기존 척도의 순위 평가가 변화하는 양상을 살펴보려 한다. 시뮬레이션 데이터뿐 아니라 실 데이터를 활용한 실험을 수행함으로써 제안하는 척도의 현실 적용 가능성 또한 평가하고자 한다.

본 논문의 이후 구성은 다음과 같다. 다음 절인 2절에서는 연관규칙 마이닝과 흥미성 척도에 대한 기존 연구들을 간략하게 소개한다. 또한 3절에서는 장바구니 크기 효과를 반영한 척도인 결합력을 재정의하고, 결합력과 기존 척도와의 우수성 평가를 위한 기준인 일관성 기준을 소개한다. 이에 대한 실험 결과는 4절에 요약하였으며, 마지막 절인 5절은 본 연구의 기여 및 한계, 그리고 향후 연구 방향을 제시한다.

2. 이론적 배경

다양한 마이닝 기법(연관규칙, 분류, 예측, 군집분석 등)에 대한 이론적 고찰[1]이 다년간 이루어져 왔으며, 이러한 기법들을 다양한 비즈니스 데이터에 적용한 사례들[2]도 최근 활발하게 발생하고 있다. 연관규칙 마이닝의 전통적 척도인 신뢰도와 지지도에 대한 개념[3] 및 연관규칙을 구현하기 위한 최초의 알고리즘인 Apriori 알고리즘[4]이 소개된 이래, 이를 개선하기 위한 시도[5-7]가 최근까지도 국내외에서 활발하게 이루어지고 있다. 또한, Apriori 알고리즘에서 증명된 지지도의 하향 폐쇄성(downward closure property)은 이후 많은 흥미성 척도 연구에서 활용되고 개선되었다. 적합한 최소 지지도를 설정하기 위한 시도로는 최소 지지도가 분석의 실행 시점(run-time)에 설정되는 방법을 제안한 연구[8]와, 신뢰도와 향상도(lift)를 활용하여 지지도를 자동으로 계산하는 방법을 제안한 연구[9]를 들 수 있다. 또한 연관규칙 탐사를 위한 최근의 연구로는, 인터벌 이벤트들 사이에 존재하는 인과관계에 대한 연관규칙을 탐사하기 위한 빈발 인터벌 관계 탐사 알고리즘[10]을 들 수 있다.

가중치를 부여한 척도는 하향폐쇄성을 활용할 수 없다는 단점을 개선하기 위한 대안으로서의 비교 연구[11]도 수행되었다. Geng and Hamilton[12]은 연관규칙의 흥미성을 평가하기 위해 9가지 관점을 제시하고 이들을 각각 객관적, 주관적, 의미적 기준으로 분류하였으며, Lenca et al.[13]에서는 흥미성 척도들이 가져야 할 바람직한 기준이 제시되었다. 또한 최근의 연구에서는 사용자의 성향에 따라 최적의 척도를 선정하기 위한 방법론[14]이 제시되었으며, 총 21개의 척도에 대해서 임의로 생성된 10000개의 분할표로부터 도출된 평가 결과를 비교한 연구[15]도 수행되었다. 각 척도들이 동일한 분할표에 대해서도 상이한 순위를 부여하지만, 지지도에 근거한 가지치기와 정규화 과정을 거친 척도들이 부여한 순위는 비교적 일관성이 있는 것으로 실험 결과 확인되었다. 흥미성 척도의 비교 실험에는 실제 데이터 집합이 사용되기도 하지만, 보다 다양한 작업부하(workload) 하에서의 실험을 위해 주로 데이터 합성기(data synthesizer)[16]로 부터 생성된 데이터 집합이 사용되기도 한다. 실생활에서 수집된 데이터가 IBM QUEST 프로그램[16]에 의해 생성된 데이터와 상이한 패턴을 보임을 지적하고, 이 차이를 극복하여 보다 현실성 있는 실험용 데이터 집합을 생성하기 위한 연구[17]도 최근 수행된 바 있다.

장바구니의 크기를 흥미성 평가에 반영한 척도로는 비교적 최근에 고안된 척도인 결합력(cohesion)[18]을 들 수 있다. 결합력은 큰 바구니에서 발생한 패턴에는 작은 바구니에서 발생한 패턴보다 낮은 의미를 부여함으로써, 서로 다른 패턴에 속하는 물품들이 큰 바구니에서 우연히 동시 출현했을 때 이들이 의미 있는 패턴으로 오인되는 부작용을 완화시키기 위한 시도이다. 결합력은 장바구니의 크기에 따라 벌점 및 가점을 부여하여 흥미도를 계산하는데, 결합력의 가중치는 정규화(normalization) 과정을 거쳐 각 거래 별로 동시 출현한 물품 집합에 대해 최대 1, 최소 0로 계산된다. 결합력을 계산하기 위한 정규화 모델은 가점만 고려한 선형정규화, 로그 정규화와 가점, 벌점을 모두 고려한 선형정규화, 로그 정규화의 총 네 가지가 있다. 이 연구에서는 언급한 네 가지 모델에 대해 결합력의 정확도를 측정하기 위한 실험을 수행하였는데, 실험 과정에서 빈발 패턴이 미리 알려져 있다는 가정 하에 가상 패턴을 생성하고, 이 패턴에 속하지 않는 규칙이 분석 결과로 도출된 경우 이를 부정확한 규칙으로 판정하였다. 이러한 방식은 정확도의 기준이 되는 패턴이 미리 존재하지 않을 뿐 아니라, 또한 존재하더라도 미리 알려져 있지 않다는 점에서 매우 비현실적 가정에 근거한 실험이라고 할 수 있다. 또한 이 연구에서는 시뮬레이션 데이터를 대상으로 한 실험만을 수행하였기 때문에, 실 데이터에 대한 적용 가능성을 분석하지 못했다는 한계를 갖는다.

3. 장바구니 크기 효과의 흥미성 척도 반영 방안

3.1 장바구니 크기 효과를 반영한 흥미성 척도 - 결합력(cohesion)

본 절에서는 분할표에서 누락된 장바구니 크기 정보를 고려한 흥미성 척도인 결합력(cohesion)을 소개한다. 결합력은 거래크기가 상대적으로 작은 장바구니에서 나타난 동시 구매 패턴에 대해 큰 장바구니에서 나타난 동시 구매 패턴보다 높은 의미를 부여 하는 것으로 각 거래별 가중치가 장바구니 크기에 따라 상이하게 계산된다. 기존 대부분의 척도들이 각 거래 별로 동시 구매 패턴의 존재 여부에 따라 1 또는 0의 가중치 값을 취하는 이분적인 계산을 수행하는 반면, 결합력은 장바구니의 크기에 따라서 1에서 0사이에 속하는 연속적인 가중치 값을 사용한다. 극단적인 경우의 예로, 장바구니의 크기가 전체 물품의 수와 같은 거래의 경우

는 해당 규칙이 장바구니에 포함되는 것이 자명하므로 가중치는 0이 되며, 장바구니의 크기가 규칙을 구성하는 물품의 수와 일치하는 경우 가중치는 1이 된다. 결합력의 범위를 조절하기 위한 방법으로는 선형정규화와 로그정규화를 비롯한 다양한 방법이 존재한다. 선형 정규화는 계산이 용이하고 직관적인 이해가 쉬운 장점이 있지만, 평균 거래 크기에 비해 매장 물품의 총 개수가 매우 큰 경우 장바구니의 크기 변화가 연관규칙의 결합력에 거의 영향을 주지 못하는 치명적인 단점이 있다. 따라서, 본 논문에서는 거래 크기 효과가 흥미성 척도에 미치는 영향을 더욱 강조하기 위해 상용로그에 기반한 정규화를 수행하였다(그림 1).

(그림 1)의 수식에서 ΔCoh_i는 i번째 장바구니로 인해 증가 또는 감소하는 결합력의 크기이며, Max는 매장에서 판매하는 전체 물품의 수, $Size_i$는 i번째 장바구니에 포함된 물품의 수, $n(Left)$와 $n(Right)$는 각각 연관규칙의 왼쪽에 해당하는 조건부(Precondition)와 오른쪽에 해당하는 결론부(Consequence)에 포함된 물품의 수를 나타낸다. (그림 1)은 가점과 벌점의 두 항목으로 구성되어 있는데, 가점 항목은 관심 물품이 출현했을 경우 결합력을 증가시키기 위한 식인 반면, 감점 항목은 관심 물품이 존재하지 않는 반례 거래 출현 시 결합력을 감소시키기 위한 식이다. 즉 물품 A를 포함하는 임의의 거래가 물품 B를 포함하지 않는 경우, 이 거래가 연관규칙 "A → B"의 결합력을 약화시키는 것으로 판단해서 정해진 벌점을 부여하는 방식이다. 반례 거래의 경우, 장바구니 크기가 큰 경우에는 높은 벌점을, 장바구니 크기가 작은 경우에는 비교적 낮은 벌점을 부여하는 것이 합당하다. 예를 들어 상점의 모든 물건 중 오직 하나의 물품을 제외하고 모두 구매했는데 그것이 반례 거래인 경우 가장 높은 벌점을 부여하고, 반례 거래의 크기가 규칙을 구성하는 물품 수의 합보다도 작은 경우는 규칙이 성립할 가능성이 전혀 없었기 때문에 벌점을 부여하지 않는다. 따라서 가장 높은 벌점인 -1은 장바구니의 크기가 전체 물품의 개수보다 오직 하나 작은 경우에 부여되고, 가장 낮은 벌점인 0은 장바구니의 크기가 규칙을 구성하는 물품 수의 합보다 작은 경우에 부여된다.

$$\Delta Coh_i = \frac{\log(Max) - \log(Size_i)}{\log(Max) - \log(n(Left) + n(Right))}$$

$$-\Delta Coh_i = \frac{\log(Size_i) - \log(n(Left) + n(Right) - 1)}{\log(Max - 1) - \log(n(Left) + n(Right) - 1)}$$

(그림 1) 결합력의 정규화를 위한 로그 기반 모델

3.2 흥미성 척도의 정확성 및 일관성 평가 기준

본 절에서는 제안하는 결합력 척도와 기존 척도의 성능 비교를 위한 기준으로 정확성과 일관성을 정의한다. 본 절에서 정의된 기준에 따른 성능 평가는 다음 절인 4절에서 수행된다.

3.2.1 정확성

본 연구에서 사용하는 성능 지표인 정확성은 훈련/검증 분할 방식을 통해 측정된다. 훈련/검증 분할 방식은 주어진 거래 내역 또는 시뮬레이션을 통해 생성된 거래 내역을 훈련(Training) 데이터와 검증(Validation) 데이터로 분할하고, 훈련 데이터에서 발견된 상위 순위 패턴이 별도의 데이터 집합인 검증 데이터에서도 유의한 것으로 나타나는지 여부를 측정한다. 즉, 훈련 데이터에서 일정 순위까지의 상위 패턴을 추출하고, 이 패턴들이 검증 데이터에서 갖는 순위를 계산한 뒤 각 패턴별로 훈련 데이터와 검증 데이터의 절대값의 차를 계산한다. 전체 패턴에 대해서 절대값의 차의 평균이 작은 경우, 해당 척도의 정확성은 높다고 간주된다. 이러한 방식에 의한 척도의 정확성 비교의 예가 (그림 2)에 나타나 있다. (그림 2)에서 척도 1의 경우 훈련 데이터에서 도출된 순위가 검증 데이터에서는 매우 다르게 나타나는 반면 (차이 평균 = 51.25), 척도 2의 경우 그 차이가 상대적으로 작음을 알 수 있다(차이 평균 = 14.75). 따라서 이 경우 척도 2가 척도 1에 비해 정확성이 우수한 것으로 평가된다.

Ptrn \ 데이터	훈련	검증	| 훈련 - 검증 |
Pattern1	1	32	31
Pattern2	73	65	8
Pattern3	227	143	84
Pattern4	320	238	82

(a) 척도 1의 정확성 측정

Ptrn \ 데이터	훈련	검증	| 훈련 - 검증 |
Pattern1	16	32	16
Pattern2	31	32	1
Pattern3	1	6	5
Pattern4	43	6	37

(b) 척도 2의 정확성 측정

(그림 2) 정확성에 대한 간단한 예

3.2.2 일관성

흥미성 척도가 가져야 할 바람직한 특성 중 하나로 분석 환경에 의해 크게 좌우되지 않는 일관적 판단을 내려야 한다는 점을 들 수 있다. 예를 들어 임의의 척도를 사용하여 A라는 환경에서 발견한 규칙이 B, 또는 C라는 환경에서는 전혀 유효하지 않은 것으로 나타난다면, 해당 척도를 사용한 분석의 결과는 일반화가 어려울 것으로 판단된다. 따라서 본 절에서는 다양한 분석 환경 하에서도 비교적 일관적인 패턴 순위를 산출하는 척도를 보다 유용한 척도로 정의하고, 이러한 기준하에 척도의 일관성(consistency)을 측정하였다. 일관성은 정확성을 보다 일반화한 개념으로 파악될 수 있으며, 이에 대한 예가 (그림 3)에 나타나 있다. (그림 3)의 A, B, C, D 4개의 환경에서 각 패턴들이 보이는 순위의 변화를 살펴보면, 척도 1은 환경이 바뀜에 따라 변화가 큰 반면(표준편차의 평균 = 7.24), 척도 2는 이에 비해 변화가 크지 않음을 알 수 있다(표준편차의 평균 = 2.915). 따라서 이 예의 경우 척도 2가 척도 1에 비해 일관성 측면에서

Ptrn \ 환경	A	B	C	D	표준편차
Pattern1	1	17	13	3	7.72
Pattern2	2	9	14	12	5.25
Pattern3	3	15	10	1	6.45
Pattern4	4	20	20	3	9.54

(a) 척도 1의 일관성 측정

Ptrn \ 환경	A	B	C	D	표준편차
Pattern1	1	2	3	1	0.96
Pattern2	2	4	6	8	2.58
Pattern3	3	3	10	3	3.50
Pattern4	10	10	2	2	4.62

(b) 척도 2의 일관성 측정

(그림 3) 일관성에 대한 간단한 예

우수한 특성을 갖는 것으로 파악된다.

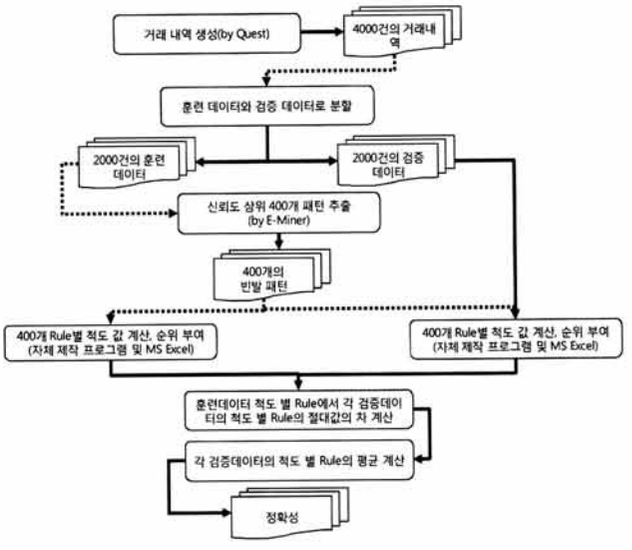
4. 실험 및 분석

4.1 실험 모형 및 환경

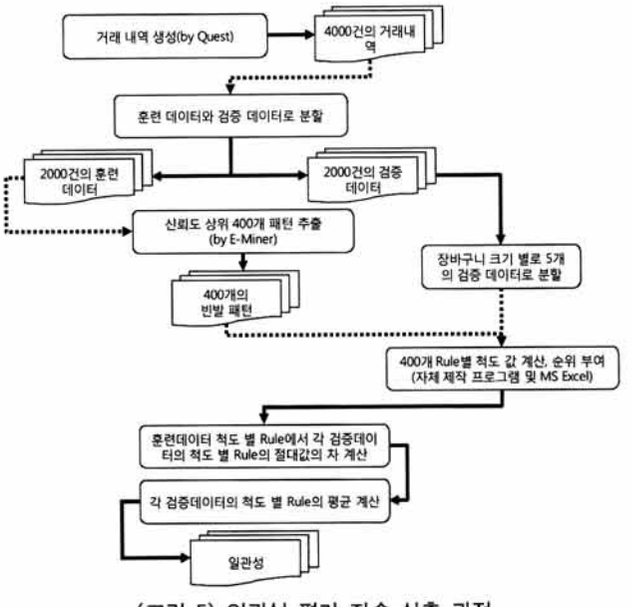
본 실험은 IBM QUEST[16]를 사용하여 생성된 시뮬레이션 데이터 및 실 데이터에 대해 수행되었다. 환경 변화를 위해 각 실험에서 장바구니의 크기 및 거래되는 전체 물품의 개수를 변화시켜가며 정확성과 일관성을 측정하였으며, 비교 척도로는 기존의 척도 중 가장 널리 사용되는 척도인 신뢰도를 채택하였다. 실험에서 장바구니의 평균 크기는 5, 10, 15, 20, 25, 30이, 그리고 물품의 개수는 300, 700, 1000, 1300, 1600, 1900이 사용되었으며, 이를 통해 총 36가지의 환경이 구축되었다. 실 데이터를 대상으로 한 실험에는 국내 한 백화점 식품매장의 판매 데이터 2101 건이 사용되었다. 원 데이터는 총 432 종류의 물품을 포함하고 있었으며, 전처리 후의 물품 종류는 총 283가지가 사용되었다.

시뮬레이션 데이터를 이용한 정확성 분석은 (그림 4)의 과정에 의해 수행된다. (그림 4)의 각 환경에서 4000개의 거래내역이 생성되어 2000개의 훈련 데이터와 2000개의 검증 데이터로 사용되었다. 다음으로는 훈련 데이터로부터 SAS 9.1 Enterprise Miner 4.3의 Association 분석을 사용하여 신뢰도 상위 400개 패턴을 추출하고, 이들 패턴의 신뢰도 및 결합력에 대한 순위를 각각 산출한다. 이들 상위 400개 패턴에 대한 순위 산출 작업을 검증 데이터에 대해서도 수행하고, 각 패턴의 훈련 데이터 상에서의 순위와 검증 데이터 상에서의 순위의 차이를 계산한다. 마지막으로 이렇게 계산된 차이 값들의 평균을 구하여 이 값을 정확성 비교를 위한 값으로 사용한다. 즉 최종 값이 작을수록 훈련 데이터와 검증 데이터에서의 순위 차이가 적고, 따라서 정확성이 높음을 나타낸다.

일관성 실험은 (그림 5)와 같이 검증 데이터 2000개 내의 하위 집합을 사용하여 수행한다. 즉, 상이한 환경 구축을 위해 검증 데이터 2000개에 포함된 거래내역을 장바구니 크기에 따라 V1 ~ V5의 5가지 집합으로 구분한다. 다음으로



(그림 4) 정확성 평가 지수 산출 과정



(그림 5) 일관성 평가 지수 산출 과정

각 패턴에 대해 5가지 집합 각각에서의 신뢰도(및 결합력) 기준 순위를 구하고, 이들 5가지 순위의 표준편차를 구하여 신뢰도(및 결합력)의 일관성 척도로 사용한다. 즉 최종 값이 작을수록 환경 변화에 따른 패턴의 순위 차이가 적고, 따라서 일관성이 높음을 의미한다.

4.2 실험 결과 및 분석

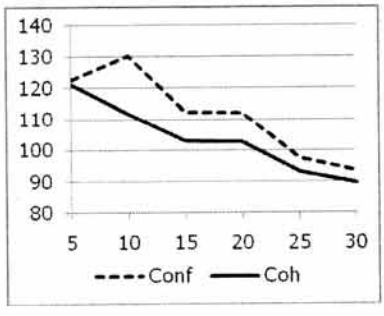
본 부절에서는 장바구니의 크기와 물품 개수에 따라 신뢰도 및 결합력의 정확성과 일관성이 변화하는 양상을 살펴보고자 한다. 시뮬레이션 데이터를 사용한 모든 실험에서 기본 환경변수는 전체물품개수 = 1000, 패턴의 평균길이 = 3, 장바구니의 평균 크기 = 15의 값이 사용되었다.

(그림 6)은 장바구니의 크기 변화에 따른 일관성 분석 결과를 나타낸다. (그림 6) 에서 기존의 척도인 신뢰도(Conf)와 장바구니의 크기 효과를 반영한 결합력(Coh)을 일관성

측면에서 비교할 때 결합력이 신뢰도에 비해 실험된 모든 환경에서 낮은 오차평균, 즉 높은 일관성을 가짐을 알 수 있다. 또한 장바구니의 평균 크기가 커질수록 두 척도의 일관성은 대체로 향상되는 추세를 보이며, 두 척도간의 일관성 차이는 줄어드는 것으로 나타났다. 이와 같이 장바구니 크기가 커질수록 신뢰도와 결합력의 두 척도간 차이가 줄어드는 것은 크기가 작은 장바구니에 비해 상대적으로 큰 장바구니에 여러 패턴이 존재할 가능성이 높기 때문이며, 장바구니의 평균 크기가 5인 실험에서 전체 추세와 다른 결과가 나온 것은 동률 순위를 갖는 패턴이 다수 존재하여 순위 정보가 왜곡되었기 때문인 것으로 사료된다. 이 분석에 대한 통계적 유의성 검정 결과는 <표 3>에 제시되어 있다. <표 3>에서 보는 바와 같이, 장바구니의 평균 크기가 5인 환경을 제외하고는 모든 실험에서 결합력이 신뢰도에 비해 우수한 일관성을 나타내는 것으로 나타났다(유의수준: $p < 0.01$)

이러한 추세는 장바구니 크기의 평균을 변화시켜가며 두 척도간의 정확성을 평가하기 위한 실험인 (그림 7)에서도 동일하게 나타났다. 정확성 역시 일관성과 마찬가지로 모든 실험 구간에서 결합력이 신뢰도에 비해 우수한 것으로 나타났으며, 그 차이는 장바구니의 평균 크기가 커질수록 줄어드는 것으로 나타났다. 이에 대한 통계적 유의성 검정 결과는 (표 4)에 나타나 있으며, 장바구니 크기가 30인 실험을 제외하고는 모든 구간에서 결합력이 신뢰도에 비해 우수한 정확성을 가짐이 통계적으로도 유의한 것으로 나타났다(유의수준 $p < 0.01$).

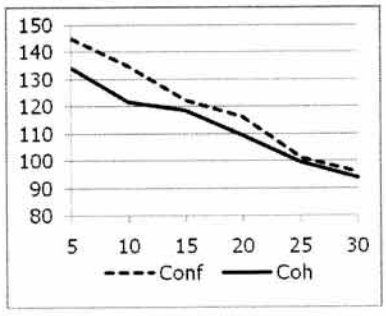
(그림 8)은 전체 물품 수의 변화에 따른 일관성 분석 결과를 나타낸다. (그림 8)에서 일관성은 전체 물품의 수가 증가할수록 값이 증가하는 추세를 보이며, 전체 실험에 걸쳐서 결합력의 순위표준편차평균이 신뢰도에 비해 조금씩 낮게 나타남을 알 수 있다. 즉, 물품 수를 변화시켜가며 수행한 실험의 전체 구간에서 결합력의 일관성이 신뢰도보다 우수하게 나타났다. 물품 수의 증가는 전체 물품 수 대비 장바구니의 평균 크기의 비의 감소를 의미하므로, 장바구니의 평균 크기가 감소하는 것과 같은 결과를 나타내게 됨을 알 수 있다. 이 실험에 대한 통계적 유의성 검정 결과는 (표 5)



(그림 6) 장바구니 크기 변화에 따른 일관성 분석

〈표 3〉 장바구니 크기 별 신뢰도와 결합력의 일관성 대응표본 분석

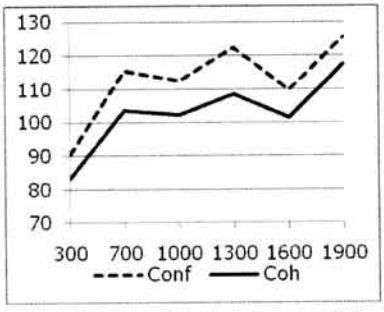
장바구니 크기	대응차		유의확률 (양쪽)
	평균	표준편차	
5	1.57329	25.74465	.222
10	18.56615	21.38820	.000
15	9.43652	11.43942	.000
20	9.54859	12.72079	.000
25	4.34103	7.67813	.000
30	3.88771	24.03722	.001



(그림 7) 장바구니 크기 변화에 따른 정확성 분석

〈표 4〉 장바구니 크기 별 신뢰도와 결합력의 정확성 대응표본 분석

장바구니 크기	대응차		유의확률 (양쪽)
	평균	표준편차	
5	11.02050	43.37120	.000
10	13.20600	26.23069	.000
15	3.62500	15.78666	.000
20	6.80800	14.40250	.000
25	1.72150	12.15416	.005
30	2.21250	18.71323	.019



(그림 8) 물품 개수 변화에 따른 일관성 분석

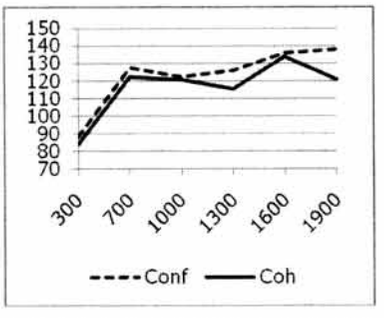
(표 5) 물품 개수 별 신뢰도와 결합력의 일관성 대응표본 분석

물품 개수	대응차		유의확률 (양쪽)
	평균	표준편차	
300	7.02124	10.68046	.000
700	12.00252	15.24209	.000
1000	10.34377	46.54359	.000
1300	14.14886	14.78949	.000
1600	8.52129	9.86206	.000
1900	7.78176	22.73140	.000

에 나타나 있으며, 모든 실험 구간에서 결합력이 신뢰도에 비해 우수한 일관성을 가짐이 통계적으로도 유의한 것으로 나타났다(유의수준 $p < 0.01$).

이러한 추세는 전체 물품의 수를 변화시켜가며 두 척도간의 정확성을 평가하기 위한 실험인 (그림 9)에서도 동일하게 나타났다. 또한 이에 대한 통계적 유의성 검정의 결과는 (표 6)에 제시되어 있는데, 물품 개수가 1000, 1600 인 두 실험을 제외한 나머지 실험에서 두 척도의 정확성 차이가 유의한 것으로 나타났다(유의 수준: $p < 0.01$).

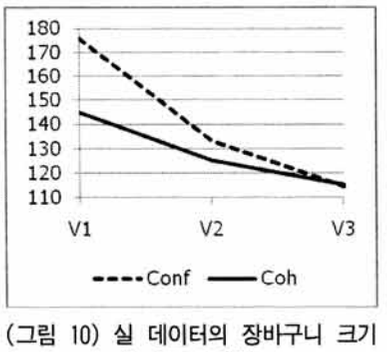
마지막 실험인 실 데이터에 대한 정확성 비교 실험의 결과는 (그림 10)에 나타나 있다. 실험에 사용된 실 데이터에서 장바구니의 평균 크기는 6이며, V1, V2, V3는 검증 데이터를 각 장바구니 크기에 따라 세분화한 것이다 (V1<V2<V3). 이 실험에서도 역시 시뮬레이션 실험과 마찬가지로 신뢰도에 비해 결합력이 정확성 측면에서 우수하게 나타났으며, 정확성의 차이는 장바구니의 크기가 작은 환경인 V1에서 더욱 큰 것으로 나타났다. 이에 대한 통계적 유



(그림 9) 물품 개수 변화에 따른 정확성 분석

(표 6) 물품 개수 별 신뢰도와 결합력의 정확성 대응표본 분석

물품 개수	대응차		유의확률 (양쪽)
	평균	표준편차	
300	4.23900	10.70940	.000
700	5.13850	16.03424	.000
1000	1.65100	59.38692	.579
1300	10.04150	16.78906	.000
1600	2.25250	24.26872	.064
1900	17.48350	30.56689	.000



(그림 10) 실 데이터의 장바구니 크기에 따른 정확성

의성 분석 결과 장바구니의 크기가 작은 V1에서는 결합력이 신뢰도보다 우수함을 입증하였지만 장바구니의 크기가 큰 V2, V3에서는 입증하지 못하였다<표 7>. 즉 장바구니의 크기가 작은 환경에서는 결합력과 신뢰도의 정확성의 차이가 통계적으로도 유의하게 나타나지만, 장바구니의 크기가 커질수록 그 차이는 점차 통계적으로 구별이 어려운 수준으로 감소함을 알 수 있다.

〈표 7〉 실 데이터의 장바구니 크기 별 신뢰도와 결합력의 정확성 대응표본 분석

장바구니 크기	대응차		유의확률 (양쪽)
	평균	표준편차	
5	30.57000	120.20494	.000
10	7.76750	126.15403	.219
15	-.84500	37.85450	.656

5. 결 론

본 연구에서는 연관규칙 마이닝의 다양한 흥미성 척도들이 장바구니 크기 효과를 간과하고 있음을 지적하고, 이를 개선하기 위한 척도인 결합력을 제안하였다. 제안하는 척도는 우연히 발생한 패턴에 대한 과대 평가 현상을 최소화하기 위해, 거래크기가 상대적으로 작은 장바구니에서 나타난 동시 구매 패턴에 대해 큰 장바구니에서 나타난 동시 구매 패턴보다 높은 의미를 부여한다. 또한 결합력과 기존 척도의 우수성 비교를 위해 특정 환경에서 파악된 규칙이 다른 환경에서도 유효한 규칙으로 평가되는지에 대한 기준인 정확성 및 일관성 기준을 제시하였다.

실험 결과 장바구니의 크기를 고려한 척도인 결합력이 기존의 대표적 척도인 신뢰성에 비해 정확성 및 일관성 측면에서 우수한 성능을 보임을 알 수 있었다. 이는 제안하는 척도가 다양한 분석 환경에서도 비교적 일관적인 결론을 제시함을 의미하며, 따라서 특정 환경에서 수행한 분석 결과를 타 환경에 적용하고자 할 때 활용 가치가 높아질 수 있음을 시사한다. 또한 기존의 대부분의 흥미성 척도 관련 실험이 시뮬레이션 데이터 또는 실 데이터 중 어느 한 쪽만을 대상으로 수행된 반면, 본 연구에서는 두 종류의 데이터에 대한 실험을 모두 수행하였다. 두 종류의 실험 결과가 서로 일치하였으므로, 제안하는 척도의 현실 적용 가능성 또한 매우 높음을 알 수 있었다.

본 연구가 갖는 한계는 다음과 같다. 첫째, 제안하는 척도의 결합력 모델에 대한 충분한 이론적 배경이 부족하다. 즉, 결합력의 충분한 이론적 고찰을 통해 정확성 및 일관성을 보다 향상 시킬 수 있는 결합력 모델을 제시하는 방향으로의 연구가 필요하다. 둘째, 실 데이터의 일관성 측정을 위한 다양한 환경 하에서 실험이 수행되지 못했다. 본 연구에서 시뮬레이션 데이터의 경우 다양한 환경 구축을 통한 추세 실험을 수행한 반면, 실 데이터의 경우는 한 가지 사례 데이터만을 가지고 실험을 수행하였다. 제안하는 척도의 다양

한 환경에서의 정확성 및 일관성 변화를 보이기 위해서는, 향후 다양하고 풍부한 사례 데이터에 대한 실험이 반드시 보강되어야 한다.

참 고 문 헌

[1] J. Han and M. Kamber, "Data Mining: Concepts and Techniques," Morgan Kaufmann Publishers, California, 2007.

[2] D. Olson and Y. Shi, "Introduction to Business Data Mining," McGraw-Hill, New York, 2007.

[3] R. Agrawal, 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, pp.207-216, 1993.

[4] R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules," in Proc. 20th International Conference on Very Large Data Bases, Santiago, Chile, pp.487-499, 1994

[5] 한경록, "CRM과 SCM의 전략적 통합을 위한 데이터 마이닝의 활용," LGCNS 엔트루정보기술연구소, 제7권, pp.151-161, 2008.

[6] 한갑수, "연관규칙 탐사 응용을 위한 한 번 읽기에 의한 최대 크기 빈발항목 추정기법," 정보처리학회논문지(D), 제15권, 제4호, pp.475-484, 2008.

[7] 채덕진, 김룡, 이용미, 황부현, 류근호, "한 번의 데이터베이스 탐색에 의한 빈발항목집합 탐색," 정보처리학회논문지(D), 제15권, 제1호, pp.15-30, 2008.

[8] K. Wang, Y. He, and J. Han, "Pushing Support Constraints into Association Rule Mining," IEEE Transactions on Knowledge and Data Engineering, Vol.15, No.3, pp.642-657, 2003.

[9] W. Y. Lin and M. C. Tseng, "Automated Support Specification for Efficient Mining of Interesting Association Rules," Journal of Information Science, Vol.32, No.3, pp.238-250, 2006.

[10] 송명진, 김대인, 황부현. "인터벌이벤트의 영향력관계에 기반한 연관규칙 탐사기법," 한국정보과학회 2009 한국컴퓨터종합학술대회 논문집(C), 제36권, 제1호, pp.96-100, 2009.

[11] B. Barber and H. Hamilton, "Extracting Share Frequent Itemsets with Infrequent Subsets," Data Mining and Knowledge Discovery, Vol.7, pp.153-185, 2003.

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

[13] P. Lenca, B. Vaillant, P. Meyer, and S. Lallich, "Association Rule Interestingness Measures: Experimental and Theoretical Studies," Quality Measures in Data Mining, Chap.3, Springer, pp.51-76, 2007.

[14] P. Lenca, P. Meyer, B. Vaillant, and S. Lallich, "On Selecting Interestingness Measures for Association Rules: User Oriented Description and Multiple Criteria Decision Aid," European Journal of Operational Research, Vol.184, No.2, pp.610-626, 2008.

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

[16] R. Agrawal, M. Mehta, J. C. Shafer, R. Srikant, A. Arning, and T. Bollinger, "The Quest Data Mining System," in Proc. 2nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Portland, Oregon, pp.244-249, 1996.

[17] C. Cooper and M. Zito, "Realistic Synthetic Data for Testing Association Rule Mining Algorithms for Market Basket Databases," in Proc. 11th European Conference on Principles and Practice of Knowledge Discovery in Databases, Warsaw, Poland, pp.398-405, 2007.

[18] 김남규, "장바구니 크기가 연관규칙 척도의 정확성에 미치는 영향," 경영정보학연구, 제18권, 제2호, pp.95-114, 2008.

김 원 서

e-mail : promise_iou@nate.com
2008년 인하공업대학(전문학사)
2009년 평생교육진흥원 학점은행(학사)
2009년~현 재 국민대학교 BIT대학원 석사과정
관심분야 : 데이터 마이닝, ERP, Data Modeling

정 승 렬

e-mail : srjeong@kookmin.ac.kr
1985년 서강대학교(학사)
1989년 Univ. of Wisconsin(석사)
1995년 Univ. of South Carolina(박사)
1995년~1997년 삼성SDS 컨설팅사업부 컨설턴트
1997년~현 재 국민대학교 비즈니스IT학부 교수
관심분야 : 프로세스 설계, 시스템 구현, 프로젝트 관리

김 남 규

e-mail : ngkim@kookmin.ac.kr
1998년 서울대학교(학사)
2000년 한국과학기술원(석사)
2007년 한국과학기술원(박사)
2007년~2009년 국민대학교 비즈니스IT학부 전임강사
2009년~현 재 국민대학교 비즈니스IT학부 조교수
2009년~현 재 한국지능정보시스템학회, 한국CRM학회 이사
관심분야 : 시멘틱 데이터 관리, 데이터베이스 설계, 데이터 마이닝