경영정보학연구
제18권 제2호
2008년 6월

장바구니 크기가 연관규칙 척도의 정확성에 미치는 영향

김 남 규*

Effect of Market Basket Size on the Accuracy of Association Rule Measures

Kim, Namgyu

Recent interests in data mining result from the expansion of the amount of business data and the growing business needs for extracting valuable knowledge from the data and then utilizing it for decision making process. In particular, recent advances in association rule mining techniques enable us to acquire knowledge concerning sales patterns among individual items from the voluminous transactional data. Certainly, one of the major purposes of association rule mining is to utilize acquired knowledge in providing marketing strategies such as cross-selling, sales promotion, and shelf-space allocation. In spite of the potential applicability of association rule mining, unfortunately, it is not often the case that the marketing mix acquired from data mining leads to the realized profit. The main difficulty of mining-based profit realization can be found in the fact that tremendous numbers of patterns are discovered by the association rule mining. Due to the too many patterns, data mining experts should perform additional mining of the results of initial mining in order to extract only actionable and profitable knowledge, which exhausts much time and costs.

In the literature, a number of interestingness measures have been devised for estimating discovered patterns. Most of the measures can be directly calculated from what is known as a contingency table, which summarizes the sales frequencies of exclusive items or itemsets. A contingency table can provide brief insights into the relationship between two or more itemsets of concern. However, it is important to note that some useful information concerning sales transactions may be lost when a contingency table is constructed. For instance, information regarding the size of each market basket (i.e., the number of items in each transaction) cannot be described in a contingency table. It is natural that a larger basket has

* 국민대학교 경상대학 비즈니스IT학부 전임강사

a tendency to consist of more sales patterns. Therefore, if two itemsets are sold together in a very large basket, it can be expected that the basket contains two or more patterns and that the two itemsets belong to mutually different patterns. Therefore, we should classify frequent itemset into two categories, inter-pattern co-occurrence and intra-pattern co-occurrence, and investigate the effect of the market basket size on the two categories. This notion implies that any interestingness measures for association rules should consider not only the total frequency of target itemsets but also the size of each basket.

There have been many attempts on analyzing various interestingness measures in the literature. Most of them have conducted qualitative comparison among various measures. The studies proposed desirable properties of interestingness measures and then surveyed how many properties are obeyed by each measure. However, relatively few attentions have been made on evaluating how well the patterns discovered by each measure are regarded to be valuable in the real world. In this paper, attempts are made to propose two notions regarding association rule measures. First, a quantitative criterion for estimating accuracy of association rule measures is presented. According to this criterion, a measure can be considered to be accurate if it assigns high scores to meaningful patterns that actually exist and low scores to arbitrary patterns that co-occur by coincidence. Next, complementary measures are presented to improve the accuracy of traditional association rule measures. By adopting the factor of market basket size, the devised measures attempt to discriminate the co-occurrence of itemsets in a small basket from another co-occurrence in a large basket. Intensive computer simulations under various workloads were performed in order to analyze the accuracy of various interestingness measures including traditional measures and the proposed measures.

Keywords : Association Rule Mining, Correctness of Interestingness Measures, Data Mining, Market Basket Analysis, Market Basket Size

Ⅰ. 서 론

데이터 마이닝(data mining)에 대한 최근 관심의 증가는, 비즈니스 데이터의 양적 팽창 및 데이터로부터 유용한 정보와 지식을 추출하여 의사결정에 활용하고자 하는 비즈니스 요구의 확산에 기인한다. 또한 정보통신의 발달로 인해 데이터의 저장, 가공 및 지식의 생성이 빠르고 정확하게 수행될 수 있게 됨에 따라, 그 동안 마이닝 분야에서 얻어진 이론적 연구의 성과를 실제 비즈니스 사례에서의 수익으로 연결시키기 위한 시도가 활발하게 이루어지고 있다. 특히, 방대한 양의 판매 데이터로부터 물품들 간의 의미 있는 연관성을 발견하기 위한 연관규칙 마이닝(association rule mining) 또는 장바구니 분석(market basket analysis)은, 그 분석 결과를 교차판매, 판촉전략 수립, 그리고 매장 배치 등의 다양한 마케팅 전략에 활용할 수 있다는 점에서 많은 연구의 대상이 되어왔다. 하지만 연관규칙 마이닝이 갖는 이러한 잠재적 유용성에도 불구하고, 마이닝을 통해 최적의 마케팅 조합(marketing mix)을 구성하고 이를 통해 수익을 증진시킨 실제 사례는 많지 않다. 연관규칙 마이닝의 수행이 수익 증진과 직결되지 못하는 가장 근본적인 원인은, 분석의 결과로 제시되는 연관규칙들의 수가 너무 많다는 것에서 찾을 수 있다. 즉, 방대한 거래 데이터로부터 도출된

물품집합(itemset) 간의 연관규칙의 수 또한 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.

사용자가 결과로 도출된 방대한 연관규칙들 중 의미 있는 규칙들만을 식별해내는 과정을 지원하기 위해서 다양한 흥미성 척도들(interestingness measures)이 고안되어왔다. 각 척도는 추출된 연관규칙들에 대해 고유한 방식으로 점수를 계산하고 점수 순으로 순위를 부여한 뒤, 높은 순위의 규칙일수록 관심을 가질만한 가치가 높은 흥미로운 패턴임을 시사한다. 이들 척도 중 신뢰도(confidence)와 지지도(support)를 비롯한 대다수의 척도는 대상 물품집합들의 발생 빈도수에 근거하여 도출되며, 기본적으로 동시에 자주 구매되는 물품집합들 간에는 강한 연관관계가 존재함을 나타낸다. 이러한 빈도수 기반 척도들은 통계학에 이론적 배경을 두고 있을 뿐 아니라, 규칙 생성 과정에서 알고리즘에 적용되기가 용이하다는 장점으로 인해 다른 접근 방법을 따르는 척도들에 비해 널리 사용되어왔다. 신뢰도와 지지도를 비롯한 대부분의 빈도수 기반 척도들은, 방대한 양의 거래 데이터로부터 관심의 대상이 되는 물품집합들의 개별 발생 빈도수 및 동시 발생 빈도수를 요약한 표인 분할표(contingency table)를 사용한다.

분할표는 빈도수 기반 척도 계산을 위한 요약정보를 제공할 뿐만 아니라 물품집합들 간의 연관성에 대한 간략하고 직관적인 초견을 제시할 수 있다는 장점을 갖는다. 하지만, 분할표는 실제의 판매 데이터로부터 얻어진 모든 정보를 누락없이 표현할 수는 없다는 한계를 갖기 때문에, 분할표로부터 계산된 규칙들의 순위가 실제 사례로부터 경험적으로 얻어지는 순위와는 동떨어지게 나타나는 경우가 많다. 거래 데이터에 포함되어있는 중요 정보임에도 불구하고 분할표에 표현되지 않는 대표적인 항목으로는 개별 물품의 가격, 수량, 그리고 주관적인 가중치 등을 들 수 있으며, 이들은 비교적 최근의 흥미성 척도 관련 연구에서 중요하게 다루어져 왔다. 하지만 이 외에도 거래 관련 정보 중 과거의 연구들에서 간과된 요소들이 있는데, 그 중 대표적인 것이 장바구니의 크기, 즉 하나의 거래에 포함된 물품의 개수가 흥미성 척도에 미치는 영향에 대한 연구이다. 즉, 기존의 척도들은 어떠한 물품집합들이 자주 함께 구매되었는지를 밝히는 과정에서, 이들 물품집합들을 동시에 구매한 거래의 크기인 장바구니의 크기를 중요하게 고려하지 않았다. 하지만 장바구니의 크기에 대한 고려 없이 단순히 분할표로부터 도출되는 척도는, 실제 현상과 동떨어진 이론적인 수치만을 제공할 우려가 있다.

흥미성 척도가 장바구니의 크기를 감안해야 한다는 논리적 근거는, 하나의 장바구니가 둘 이상의 빈발 패턴을 포함할 수 있다는 것에서 찾을 수 있다. 즉, 동시에 자주 구매된 물품들이 실제로 하나의 패턴에 소속되어 있을 수도 있지만, 반면 물품들이 서로 다른 패턴에 각기 속하고, 이들 패턴들이 동시에 자주 출현하는 것일 가능성도 있는 것이다. 전자인 패턴 내 동시출현(intra-pattern co-occurrence) 물품들은, 물품들 자체가 의미적 연관성을 가지므로 장바구니의 크기에 큰 영향을 받지 않고 동시에 구매되는 성향이 있다. 하지만 후자인 패턴간 동시출현(inter-pattern co-occur-

T1	빵, 우유	2
T2	빵, 우유	2
T3	빵, 시리얼	2
T4	빵, 콜라	2
T5	펜, 지우개	2
T6	맥주, 스넥, 베이컨	3
T7	비누, 샴푸, 향수	3
T8	펜, 지우개, 비누, 샴푸, 향수	5
T9	펜, 지우개, 맥주, 스넥, 베이컨	5
T10	펜, 지우개, 맥주, 스넥, 베이컨, 비누, 샴푸, 향수	8

<그림 1> 12개의 물품들에 대한 거래 내역

rence) 물품들은, 큰 장바구니에는 여러 구매 패턴이 함께 존재하므로 동시에 출현할 가능성이 크지만, 작은 장바구니에는 많은 수의 패턴이 나타나기 어려우므로 함께 구매되는 빈도수가 현저하게 낮아진다는 특성이 있다. 이러한 현상을 살펴보기 위해 총 12개의 물품에 대해 10건의 거래로 구성된 다음의 판매 데이터베이스를 생각해보자.

12개 물품에 대한 10건의 거래 내역과 각 거래의 장바구니 크기가 <그림 1>에 제시되어 있다. 이 때, 이들 물품들에 대해 4가지의 패턴 {빵, 우유}, {펜, 지우개}, {맥주, 스넥, 베이컨}, 그리고 {비누, 샴푸, 향수} 가 존재하는 것으로 미리 알려져 있다고 가정하자. 4가지 패턴은 과거의 거래 내력, 또는 전문가의 자문으로부터 얻어진 것으로 가정한다. <그림 1>의 거래 내역을 미리 정의된 빈발 패턴의 관점에서 살펴보면, T8 ~ T10는 각각 2개, 2개, 그리고 3개의 패턴으로 구성되어 있음을 알 수 있다. 본 예에서는, 총 10건의 거래에 대해서 2가지의 연관규칙 "빵 → 우유"와 "펜 → 맥주" 만을 전통적 척도인 신뢰도와 지지도 측면에서 살펴보고자 한다. 간단한 계산에 의해서, 두 연관규칙 모두 20%의 지지도와 50%의 신뢰도를 가짐을 알 수 있다. 즉, 전통적인 척도는 "빵 → 우유"와 "펜 → 맥주" 규칙의 흥미성을 동일하게 평가하고 있음을 알 수 있다. 하지만, 이러한 평가는 실제 상황을 정확하게 반영하고 있지 못함을 알 수 있다. 예를 들어, 한 손에 빵을 든 고객이 다른 손에 우유를 들었을 가능성과, 한 손에 펜을 든 고객이 다른 손에 맥주를 들었을 가능성 중 어떤 쪽이 더 높은지를 생각해보면, 전통적인 척도에 의한 평가가 실제 경험, 또는 직관적인 예상과 일치하지 않음을 알 수 있다. 이러한 현상이 발생하는 이유는, 전통적인 척도의 계산 과정에서 장바구니 크기의 영향이 전혀 고려되지 않았기 때문이다. 따라서 이러한 부작용을 개선하기 위해서는, 두 번째 규칙의 신뢰도 향상에 기여한 T8와 T10의 경우 펜과 맥주 이외에도 많은 수의 물품을 포함하고 있으므로, 첫 번째 규칙의 신뢰도 향상에 기여한 T1과 T2에 비해 연관성에 대한 기여도가 낮게 책정되는 방향으로 척도 계산법의 수정이 이루어져야 한다.

위의 예는 패턴 내 동시출현과 패턴간 동시출현 사이에 존재하는 연관성의 질적 차이를 암시한다. 즉 패턴간 동시출현 비율이 높은 물품집합의 경우, 높은 신뢰도와 지지도에도 불구하고 작은 크기의 장바구니에서 이 물품들이 함께 구매될 가능성은 높지 않을 수 있기 때문에, 신뢰도와 지지도만으로 마케팅 전략을 세우는 것은 매우 위험하다고 할 수 있다. 한편, 위의 예에서 보는 바와 같이 빈발 패턴이 미리 알려져 있는 경우는 매우 드물다. 따라서 패턴들이 미리 정확하게 알려져 있지 않은 상황에서, 패턴 내 동시출현과 패턴간 동시출현의 차이를 흥미성 척도에 반영할 수 있는 방향의 연구가 필요하다. 본 논문에서는 패턴간 동시출현의 가능성이 패턴 내 동시출현의 가능성에 비해 장바구니 크기의 영향을 많이 받는다는 점에 착안하여, 전통적인 흥미성 척도의 개선 방향을 제시하고자 한다. 즉, 기존의 대표적인 흥미성 척도들에 장바구니의 크기 효과를 반영하였을 경우, 척도들의 성능이 개선되는 양상을 살펴보고자 한다.

연관규칙 마이닝에 대한 기존의 많은 연구에서, 다양한 척도들이 제안되었을 뿐 아니라 제안된 척도들에 대한 비교 분석도 이미 활발하게 이루어졌다. 척도들의 성능에 대한 비교를 위해 기존의 연구들이 주로 사용한 방법은, 흥미성 척도가 가져야 할 바람직한 특성들을 정의하고, 각 척도가 이러한 특성 중 몇 가지를 만족시키는지를 분석하는 것이었다. 하지만 각 척도에 의해 흥미도가 높은 것으로 평가된 연관규칙들이 실제로도 얼마나 의미 있게 간주되는지에 대한 정량적인 연구는 상대적으로 매우 부족하다. 본 논문에서는 척도들의 성능을 비교하기 위한 정량적 기준으로서 정확성(accuracy)을 정의하고, 이에 근거하여 기존의 척도들과 본 논문에서 제안하는

개선된 척도들에 대해 정확성 측면에서의 성능을 평가하고자 한다. 물론 Geng and Hamilton [2006]에서 논의된 바와 같이 어떤 패턴이 흥미로운 패턴인가에 대한 관점은 매우 다양해서, 규칙의 발생 빈도(frequency), 포함 범위(coverage), 의외성(surprisingness), 그리고 응용 가능성(applicability) 등의 다양한 기준이 흥미성 평가에 적용될 수 있다. 본 논문에서는 흥미성에 대한 이들 다양한 관점 중, 동시에 함께 구매된 빈도수가 많은 패턴이 흥미성이 높은 패턴이라는 관점에 기반하여 논리를 전개하고자 한다.

본 논문의 이후 구성은 다음과 같다. 다음 절인 제 Ⅱ절에서는 연관규칙 마이닝과 흥미성 척도에 대한 기존의 연구들에 대해 간략하게 소개한다. 본 논문에서 제안하는 흥미성 척도의 성능 평가를 위한 정확성 기준, 그리고 척도의 정확성을 개선하기 위해 장바구니의 크기를 활용하는 방안은 제 Ⅲ절에서 제시되며, 이에 대한 실험 및 결과는 제 Ⅳ절에 나타나 있다. 마지막으로 제 Ⅴ절에서는 본 연구의 기여 및 한계, 그리고 향후 연구 방향을 제시한다.

Ⅱ. 관련 연구

연관규칙 마이닝은 Agrawal *et al.*[1993]에서 신뢰도 및 지지도에 대한 개념과 함께 소개되었고, 이 개념의 실제 구현을 위한 Apriori 알고리즘은 Agrawal and Srikant[1994]에서 제시되었다. Agrawal and Srikant[1994]에서 증명된 지지도의 하향 폐쇄성(downward closure property)은 향후 많은 흥미성 척도 연구에서 활용되고 개선되었다. Cai *et al.*[1998]은 가중치가 부여된 물품(weighted item)들에 대한 연관규칙을 생성하기 위해 지지도에 가중치를 부여하였으며, Hu and Chen[2006]과 Liu *et al.*[1999]에서는 희귀 물품 문제(rare item problem)을 해결하기 위한 다중 최소 지지도 (multiple minimum support)에 대한 연구가 이루어졌다. 적합한 최소 지지도를 설정하기 위한 또 다른 시도로는, 최소 지지도가 분석의 실행 시점(run-time)에 설정되는 방법을 제안한 Wang *et al.*[2003]과, 신뢰도와 향상도(lift)를 활용하여 지지도를 자동으로 계산하는 방법을 제안한 Lin and Tseng[2006]을 들 수 있다. 연관규칙 마이닝을 비롯한 분류, 예측, 그리고 군집분석 등, 다양한 마이닝 기법에 대한 이론적 고찰은 Han and Kamber[2007]에 자세히 소개되어있으며, 마이닝의 여러 기법들을 다양한 비즈니스 데이터에 적용한 사례들은 Olson and Shi[2007]에서 찾을 수 있다.

연관규칙의 적합한 평가를 위해 다양한 흥미성 척도들이 고안되었을 뿐 아니라, 척도들에 대한 비교 연구도 활발하게 수행되었다. Carter *et al.*[1997]은 각 물품에 대해 가중치를 부여하여 흥미성 척도를 계산하였다. 가중치를 부여할 경우 전통적인 지지도의 하향폐쇄성을 활용할 수 없으므로, 빈발 패턴의 추출을 위한 다른 대안이 Barber and Hamilton[1994]에서 제시되었다. 이외에도 Brin *et al.*[1997], Chen *et al.*[1996], 그리고 Tao *et al.*[2003] 등에서 흥미성 척도를 개선하기 위한 시도가 다양한 시각에서 이루어졌다. 흥미성 척도의 우수성을 규정하기 위한 시도로는, 연관 규칙의 흥미성을 평가하기 위해 9가지 관점을 제시하고 이들을 각각 객관적, 주관적, 의미적 기준으로 분류한 Geng and Hamilton[2006]을 들 수 있다. 또한 Lenca *et al.*[2007]과 Vaillant *et al.*[2004]은 척도들이 가져야 할 바람직한 기준을 제시하고, 이 기준에 근거하여 총 20개의 척도와 10종류의 데이터 집합에 대해 수행한 비교 실험 결과를 제시하였다. 또한, Lenca *et al.*[2008]은 사용자의 성향에 따라 최적의 척도를 선정하기 위한 방법론을 제시하였으며, Tan *et al.*[2002]는 총 21개의 척도에 대해서 임의로 생성된 10000개의 분할표로부터 도출된 평가 결과를 비교하였다. 각 척도들이 동일한 분할표에 대해서도 상이한 순위를 부여하지만, 지지도에 근거한 가지치기와 정규화 과정을 거친 척도들이 부여한 순위는 비

교적 일관성이 있는 것으로 실험 결과 확인되었다. 흥미성 척도의 비교 실험에는 실제 데이터 집합이 사용되기도 하지만, 보다 다양한 작업부하(workload) 하에서의 실험을 위해 주로 Agrawal *et al.*[1996]와 같은 데이터 합성기(data synthesizer)로 부터 생성된 데이터 집합을 사용한다. 보다 현실성 있는 실험용 데이터 집합을 생성하기 위한 최근의 시도는 Cooper and Zito[2007]에서 찾을 수 있다.

Ⅲ. 장바구니의 크기를 활용한 연관규칙 척도의 정확성 향상 방안

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

본 부절에서는 기존의 분할표 기반 흥미성 척도들의 한계점을 보완하기 위한 시도로, 개별 장바구니의 크기 효과를 반영하여 연관규칙 척도를 개선하는 방안을 제시하고자 한다. 기본적인 개선 방향은 어떤 물품들이 크기가 작은 장바구니에서 동시에 구매된 경우, 크기가 큰 장바구니에서 동시에 구매된 경우에 비해 보다 높은 연관성을 암시하는 것으로 평가하는 것이다. 제안하는 개념을 통해 물품 간의 의미적 연관성을 결합력(cohesion)이라고 명명하였으며, 결합력은 기존 척도들의 한 요소로 내재되어 사용될 수 있다. 결합력의 가장 중요한 철학은 장바구니의 크기를 반영하여 흥미도을 계산한다는 것이며, 계산 과정은 기술적인 측면에서 더욱 세분화될 수 있다. 본 절의 나머지 부분에서는 자세한 설명을 위해 <그림 1>의 거래 내역을 참조하기로 한다.

우선 결합력은 척도의 범위를 조절하기 위한 정규화(normalization) 과정에서 두 가지로 세분된다. 본 논문에서 정규화는, 각 거래 별로 동시 출현한 물품 집합에 대해 최대 1, 최소 0의 결합력을 부여하기 위해서 수행한다. 전통적인 척도의 경우, 관심 물품 집합들이 동시에 구매된 거래에 대해서는 1을, 동시에 구매되지 않은 거래에 대해서는 0을 부여하는 이분적인 계산을 수행하지만, 결합력의 경우 장바구니의 크기에 따라서 1에서 0사이의 임의의 값을 가질 수 있다. 결합력의 범위를 최대 1, 최소 0으로 한정하기 위한 정규화의 가장 간단한 방법은 선형 정규화이다. 즉, 관심 패턴이 발생하기 위한 장바구니의 최소 크기에서 실제로 그 패턴이 발생했을 경우에 결합력을 1 증가시키는 것이다. 또한 어떤 장바구니에 그 매장에서 판매되는 모든 물품이 포함되어 있다면, 이 장바구니에는 관심 패턴 역시 포함되어 있는 것이 너무 자명하다. 즉 관심 패턴에 속한 물품집합 간에 의미적 연관성을 부여할 수 없으므로 결합력을 0으로 부여한다. <그림 3>에는 선형 정규화를 위한 일반식과 함께, 단일 항목 두 개 사이의 간단한 연관규칙에 적용하였을 경우의 특수 식이 제시되어 있다. <그림 3>을 포함한 이후의 계산식에서 사용되는 표기법에 대한 정의는 <그림 2>에 나타나있다.

Max	거래되는 전체 물품의 개수
T_i	i 번째 개별 거래
$Size_i$	i 번째 개별 거래의 장바구니 크기
ΔCoh_i	i 번째 개별 거래에 의한 결합력의 변화량
$Left$	연관규칙의 좌측 물품 집합
$n(Left)$	$Left$에 속한 물품의 개수
$Right$	연관규칙의 우측 물품 집합
$n(Right)$	$Right$에 속한 물품의 개수

<그림 2> 결합력의 계산에 사용되는 표기법에 대한 정의

결합력에 기반한 신뢰도는, <그림 3>에서 계산된 결합력을 연관규칙의 좌측 물품 집합이 구매된 거래의 건수로 나눔으로써 도출할 수 있다. <그림 1>의 거래 내역 중 연관규칙 "팬 → 맥주"에 대해 결합력 기반 신뢰도를 계산하여 보자. <그림 1>에서 거래된 물품의 개수는 총 12가지

일반식:

$$\Delta Coh_i = \frac{Max - Size_i}{Max - (n(Left) + n(Right))} \quad ,(\forall T_i, Left \in T_i, Right \in T_i)$$

특수식: $n(Left) = 1, n(Right) = 1$

$$\Delta Coh_i = \frac{Max - Size_i}{Max - 2} \quad ,(\forall T_i, Left \in T_i, Right \in T_i)$$

<그림 3> 결합력의 선형 정규화를 위한 일반식과 특수식

일반식:

$$\Delta Coh_i = \frac{\log(Max) - \log(Size_i)}{\log(Max) - \log(n(Left) + n(Right))} \quad ,(\forall T_i, Left \in T_i, Right \in T_i)$$

특수식: $n(Left) = 1, n(Right) = 1$

$$\Delta Coh_i = \frac{\log(Max) - \log(Size_i)}{\log(Max) - \log(2)} \quad ,(\forall T_i, Left \in T_i, Right \in T_i)$$

<그림 4> 결합력의 로그 정규화를 위한 일반식과 특수식

이므로, *Max* = 12, *n(Left)* = 1, *n(Right)* = 1임을 알 수 있다. 따라서 연관규칙 "펜 → 맥주" 의 결합력은 1.1 (= 7/10 + 4/10)이며, 결합력에 기반한 신뢰도는 27.5% (= 1.1/4)이다. 한편 같은 방식으로 연관규칙 "빵 → 우유"의 결합력 기반 신뢰도는 50%로 계산된다. 즉, 두 연관규칙의 흥미도를 50%로 동일하게 평가한 전통적인 신뢰도와는 달리, 결합력에 기반한 신뢰도는 연관규칙 "빵 → 우유" 를 "펜 → 맥주" 보다 흥미로운 규칙으로 평가함을 알 수 있다. 선형 정규화를 통한 결합력의 도출 방식은 계산이 용이하고 직관적 이해가 쉽다는 장점을 갖지만, 일반적인 장바구니의 크기에 비해 거래되는 물품의 총 개수가 매우 큰 경우 장바구니의 크기 변화가 연관규칙의 결합력에 거의 영향을 주지 못한다는 한계를 갖는다.

선형 정규화의 한계를 극복하고 장바구니 크기의 효과가 흥미성 척도에 미치는 영향을 더욱 강조하기 위해, 결합력의 선형 정규화 대신 로그 정규화를 수행할 수 있다. 로그 정규화 역시 관심 패턴이 발생하기 위한 장바구니의 최소 크기에서 실제로 그 패턴이 발생했을 경우에 결합력을 1 증가시키고, 장바구니가 매장에서 판매되는 모든 물품을 포함하는 경우 결합력을 0으로 부여한

다는 점은 선형 정규화와 같다. 로그 정규화에 대한 일반식 및 특수식은 <그림 4>에 나타나 있으며, 이 식에 의하면 연관규칙 "빵 → 우유"와 "펜 → 맥주"는 각각 50%와 18%의 결합력 기반 신뢰도를 갖게 된다(단, 로그는 상용로그를 사용). 선형 정규화와 로그 정규화의 성능 평가 결과는 다음 절의 실험을 통해 제시된다.

정규화 방식 이외에 결합력 도출 모델을 세분화하기 위한 또 하나의 기준은, 반례(counter example) 거래 출현 시 결합력에 벌점(penalty)를 부여할 지의 여부이다. 즉, 연관규칙 "A → B"의 결합력을 계산하는 과정에서, 물품 A는 포함하지만 물품 B를 포함하지 않는 거래를 어떻게 처리할 것인가 하는 문제이다. 여기에는 크게 두 가지 대안이 있다. 첫 번째 대안은 이러한 거래에 대해서는 아무 벌점을 부여하지 않고 단순히 무시하는 것으로, 기존의 지지도, 신뢰도, 그리고 향상도 등이 이 방법을 따른다. 또 다른 대안은 이러한 거래가 출현할 경우, 이 거래가 연관규칙 "A → B"의 흥미성을 약화시키는 것으로 판단해서 정해진 벌점을 부여하는 방식이다. 벌점을 감안하여 결합력을 계산하는 경우, 반례 거래의 장바구니 크기가 작은 경우에는 비교적 낮은 벌점을, 반례 거래의 장바구니 크기가 큰 경우에는 높은 벌점을 부여하는 것이 합당하다. 가장 높은 벌점인 -1은, 반례 거래의 크기가 상점에서 거래하는 물품의 개수보다 오직 하나 작은 경우에 부여되고, 가장 낮은 벌점인 0은 반례 거래의 크기가

일반식:

$$-\Delta Coh_i = \frac{Size_i - (n(Left) + n(Right) - 1)}{(Max - 1) - (n(Left) + n(Right) - 1)} \quad ,(\forall T_i, Left \in T_i, Right \notin T_i)$$

특수식: $n(Left) = 1, n(Right) = 1$

$$-\Delta Coh_i = \frac{Size_i - 1}{Max - 2} \quad ,(\forall T_i, Left \in T_i, Right \notin T_i)$$

<그림 5> 결합력의 벌점을 정규화하기 위한 선형 계산식

일반식:

$$-\Delta Coh_i = \frac{\log(Size_i) - \log(n(Left) + n(Right) - 1)}{\log(Max - 1) - \log(n(Left) + n(Right) - 1)} \quad ,(\forall T_i, Left \in T_i, Right \notin T_i)$$

특수식: $n(Left) = 1, n(Right) = 1$

$$-\Delta Coh_i = \frac{\log(Size_i)}{\log(Max - 1)} \quad ,(\forall T_i, Left \in T_i, Right \notin T_i)$$

<그림 6> 결합력의 벌점을 정규화하기 위한 로그 계산식

패턴을 구성하는 물품 수의 합보다 하나 작은 경우에 부여된다. 즉, 상점의 모든 물건 중 오직 하나의 물품을 제외하고 모두 구매했는데 그것이 반례 거래일 경우 가장 높은 벌점을 부여하고, 반례 거래의 크기가 패턴을 구성하는 물품 수의 합보다도 작은 경우는 패턴이 발생할 가능성이 원래 없는 것이므로 벌점을 부여하지 않는 것이다. 가점과 벌점을 동시에 감안한 결합력 역시 선형 혹은 로그 정규화 과정을 거치게 되며, 이를 위한 계산식은 <그림 5>와 <그림 6>에 각각 나타나있다. 예를 들어 <그림 5>와 <그림 6>의 일반식에서 장바구니의 크기가 $(Max - 1)$인 경우의 벌점은 -1로 계산되고, 장바구니의 크기가 $(n(Left) + n(Right) - 1)$인 경우의 벌점은 0으로 계산된다. 가점 계산식은 <그림 3>과 <그림 4>에서 이미 소개되었으므로, <그림 5>와 <그림 6>에서는 벌점에 대한 계산식만을 제시한다. 가점과 벌점을 동시에 감안하는 경우, 결합력의 선형 정규화에 의해 "빵 → 우유"의 결합력 기반 신뢰도는 45%, "펜 → 맥주"의 결합력 기반 신뢰도는 15%로 나타났다. 한편 가점과 벌점을 동시에 감안한 결합력의 로그 정규화에 의해, 두 연관규칙의 신뢰도는 각각 36%와 -6%로 나타남을 알 수 있다.

본 논문에서 제안한 결합력의 정규화 모델 4가지가 <그림 7>에 요약되어 있다. 지금까지의 논의에서, 결합력은 독자적으로 사용된 것이 아니라 기존의 신뢰도 척도를 개선하기 위해 장바구니의 크기를 반영하는 형태로 사용되었다. 결합력은 신뢰도 이외의 전통적인 다양한 흥미성 척도의 개선에도 사용될 수 있으며, 신뢰도, 지지도, 및 향상도에 결합력을 반영한 모형과 그 성능에 대한 평가를 제 Ⅳ절에서 다루었다.

3.2 흥미성 척도의 정량적 평가를 위한 정확성 기준

최근 15년 동안 서로 다른 목적에 부합하는 흥미성 척도들이 많이 고안되어왔지만, 척도들의 성

결합력 구분	정규화 모델	
선형 정규화 (가점만 고려)	$\Delta Coh_i = \frac{Max - Size_i}{Max - (n(Left) + n(Right))}$	$,(\forall T_i, Left \in T_i, Right \in T_i)$
로그 정규화 (가점만 고려)	$\Delta Coh_i = \frac{\log(Max) - \log(Size_i)}{\log(Max) - \log(n(Left) + n(Right))}$	$,(\forall T_i, Left \in T_i, Right \in T_i)$
선형 정규화 (가점, 벌점 고려)	$\Delta Coh_i = \frac{Max - Size_i}{Max - (n(Left) + n(Right))}$	$,(\forall T_i, Left \in T_i, Right \in T_i)$
	$-\Delta Coh_i = \frac{Size_i - (n(Left) + n(Right) - 1)}{(Max - 1) - (n(Left) + n(Right) - 1)}$	$,(\forall T_i, Left \in T_i, Right \notin T_i)$
로그 정규화 (가점, 벌점 고려)	$\Delta Coh_i = \frac{\log(Max) - \log(Size_i)}{\log(Max) - \log(n(Left) + n(Right))}$	$,(\forall T_i, Left \in T_i, Right \in T_i)$
	$-\Delta Coh_i = \frac{\log(Size_i) - \log(n(Left) + n(Right) - 1)}{\log(Max - 1) - \log(n(Left) + n(Right) - 1)}$	$,(\forall T_i, Left \in T_i, Right \notin T_i)$

<그림 7> 결합력의 계산을 위한 4가지 정규화 모델

능에 대한 비교 연구는 아직 충분히 이루어지지 않았다. 물론 척도들의 상대적 우수성을 비교하기 위한 시도도 다수 있었으나, 이들 연구의 대부분은 흥미성 척도가 가져야 할 바람직한 특성들을 정의하고, 각 척도들이 정의된 특성을 얼마나 가지고 있는지를 평가하는 정성적인 비교 연구가 대부분이다. 즉, 각 척도들이 연관규칙들에 대해 얼마나 적절한 점수를 부여하고 있는지에 대한 정량적 평가는 찾아보기가 매우 힘들다. 본 절에서는 척도들의 성능 평가를 위한 객관적인 기준으로서 정확성을 정의하고, 각 척도들의 정확성을 계량화할 수 있는 방법론을 제시한다.

연관규칙 마이닝에서 도출된 빈발 패턴은, 실제로 존재하는 의미 있는 패턴일 수도 있고, 둘 이상의 패턴이 우연히 동시에 발생하여 형성된 것일 수도 있음은 이미 논의한 바 있다. 어떤 척도에 의해 발굴된 빈발 패턴 중 실제로 존재하는 패턴의 비율이 높은 경우, 우리는 이 척도의 정확성이 높다고 간주할 수 있다. 따라서 정확성은 다음과 같이 정의될 수 있다;

$ACC_n(M)$ = **(척도 M에 의해 발견된 빈발 패턴 상위 n개 중 실제로 존재하는 패턴의 개수)/n.**

즉, 척도 M의 n-정확성($ACC_n(M)$)은, 척도 M이 가장 흥미로운 것으로 평가한 n개의 패턴 중, 실제로 존재하는 패턴의 비율을 나타낸다. 실제로 존재하는 패턴 m개가 이미 알려져 있고, $n \leq m$ 이라고 가정하자. 이 때, 상위 n개의 패턴 모두가 실제 패턴 m개에 포함된다면, $ACC_n(M_i) = 100\%$로 계산된다. 현실적으로는 임의의 척도에 의해 발견된 빈발 패턴 가운데에는 실제로는 존재하지 않는 패턴이 포함되어 있으므로, 대부분의 척도는 100% 미만의 값을 갖게 된다. 정확성을 근거로 하여 척도의 성능을 평가할 수 있는 관점은 크게 두 가지이다. 첫 번째는, 대상 척도가 갖는 전반적인 정확성 값에 대한 평가이다. 물론, 정확성 값이 100%에 가까울수록 해당 척도는 정확성 측면에서 바람직한 속성을 갖고 있는 것으로 평가된다. 또 다른 관점으로는, n값이 작아질수록 정확성의 값이 높아지는 척도가 바람직한 척도라고 할 수 있다. 척도에 의해 부여 받은 순위가 높은 패턴일수록 실제로 존재하는 패턴일 비율이 높게 나왔다면, 해당 척도는 실제 패턴에 높은 점수를, 우연히 조합된 가상 패턴에 낮은 점수를 부여하고 있는 것으로 볼 수 있기 때문이다. 극단적인 반대의 경우, n값이 변화하더라도 정확성의 값에 변화가 전혀 없다면, 해당 척도가 패턴에 부여한 순위는 실제로 존재하는 패턴을 구별해내는 데에 아무 역할을 하지 못하고 있는 것이므로, 바람직한 특성을 가졌다고 보기 어렵다.

정확성을 활용한 척도들의 성공적인 성능 평가를 위해 반드시 전제되어야 하는 것은, 실제 존재하는 의미 있는 패턴들이 사전에 주어져야 한다는 것이다. 이를 위한 첫 번째 대안은, 실제 데이터에 대해 수 많은 경험적인 분석을 수행하고, 그 결과에 대해 전문가의 검증을 거치는 과정을 반복함으로써, 해당 데이터 집합에 실제로 존재하는 패턴을 수동적으로 정의하는 것이다. 이러한 접근 방법은 실제 사례로부터 패턴이 정의되었으므로 현장감이 매우 뛰어나다는 장점을 갖지만, 패턴을 찾기 위한 시간과 비용이 많이 소모될 뿐 아니라 모두가 합의할 수 있는 패턴을 찾는 일이 매우 어렵다는 측면에서 실제로 활용되기에는 많은 한계를 갖는다. 존재하는 패턴을 사용하여 척도의 정확성을 평가하기 위한 두 번째 대안은 컴퓨터 시뮬레이션이다. 즉, 컴퓨터를 사용하여 사전에 가상 패턴을 생성하고, 고객, 물품, 거래에 대한 통계적 분포들과 이 패턴을 함수적으로 활용하여 임의의 거래 내역을 생성한 뒤 정확도를 평가하는 것이다. 이 방법의 경우 다양한 작업부하 하에서 반복 실험을 할 수 있다는 큰 장점이 있지만, 생성된 패턴 및 거래 자체가 왜곡되었을 수도 있다는 한계를 갖는다. 따라서 후자의 방법을 택하는 대부분의 연구에서는 잘 알려진 공신력 높은 데이터

		지지도	신뢰도	향상도
전통적 척도		*Support*	*Confidence*	*Lift*
결합력 반영 척도	선형 정규화 (가점만 고려)	*Coh_Lnr_P_Sup*	*Coh_Lnr_P_Conf*	*Coh_Lnr_P_Lift*
	로그 정규화 (가점만 고려)	*Coh_Log_P_Sup*	*Coh_Log_P_Conf*	*Coh_Log_P_Lift*
	선형 정규화 (가점, 벌점 고려)	*Coh_Lnr_PN_Sup*	*Coh_Lnr_PN_Conf*	*Coh_Lnr_PN_Lift*
	로그 정규화 (가점, 벌점 고려)	*Coh_Log_PN_Sup*	*Coh_Log_PN_Conf*	*Coh_Log_PN_Lift*

<그림 8> 실험 대상 흥미성 척도들의 명칭

집합 및 데이터 합성 프로그램을 사용하며, 본 연구에서도 이러한 취지에서 Agrawal *et al.*[1996]에서 고안되어 현재까지도 널리 사용되는 데이터 합성기 중의 하나인 QUEST 시스템을 사용하여 다음 절의 실험을 수행하였다.

Ⅳ. 성능 평가

4.1 실험 모형 및 환경

본 절에서는 다양한 흥미성 척도의 정확성을 평가하기 위한 실험 과정 및 결과를 제시한다. 즉, 대표적인 전통적 척도에 결합력이 반영되었을 때 척도의 정확성이 변화하는 양상을 측정하고자 한다. 본 실험에서 평가 대상이 되는 척도는 전통적 척도 3가지(지지도, 신뢰도, 향상도)와, 이들 척도에 결합력을 반영한 12가지의 척도이며, 이들 15가지 척도들의 명칭이 <그림 8>에 요약되어있다. 실험에 사용된 데이터 집합은 Agrawal *et al.*[1996]의 Quest 시스템을 사용하여 생성되었으며, 데이터 생성에 사용된 인수들의 종류와 값은 <그림 9>에 요약되어있다. <그림 9>에서 모든 실험에 동일하게 사용된 인수는 "고정 값" 항목에 기록하였고, 작업부하에 따른 정확성 변화의 추이를 살피기 위해 값이 변경된 인수는 "조절 값" 항목에 요약되어있다. 장바구니의 평균 크기에 대한 패턴의 평균 길이의 비율이 낮을수록 하나의 장바구니가 많은 수의 패턴을 포함하는 경향이 있다. 이러한 경향이 정확도에 미치는 영향을 살피기 위해, "조절 값" 항목에서는 패턴의 평균 길이와 장바구니의 평균 크기를 변화시켰다. "조절 값" 항목에서 기본 값으로 사용된 값들에는 밑줄을 표기하여 다른 값들과 구분하였다.

고정 인수

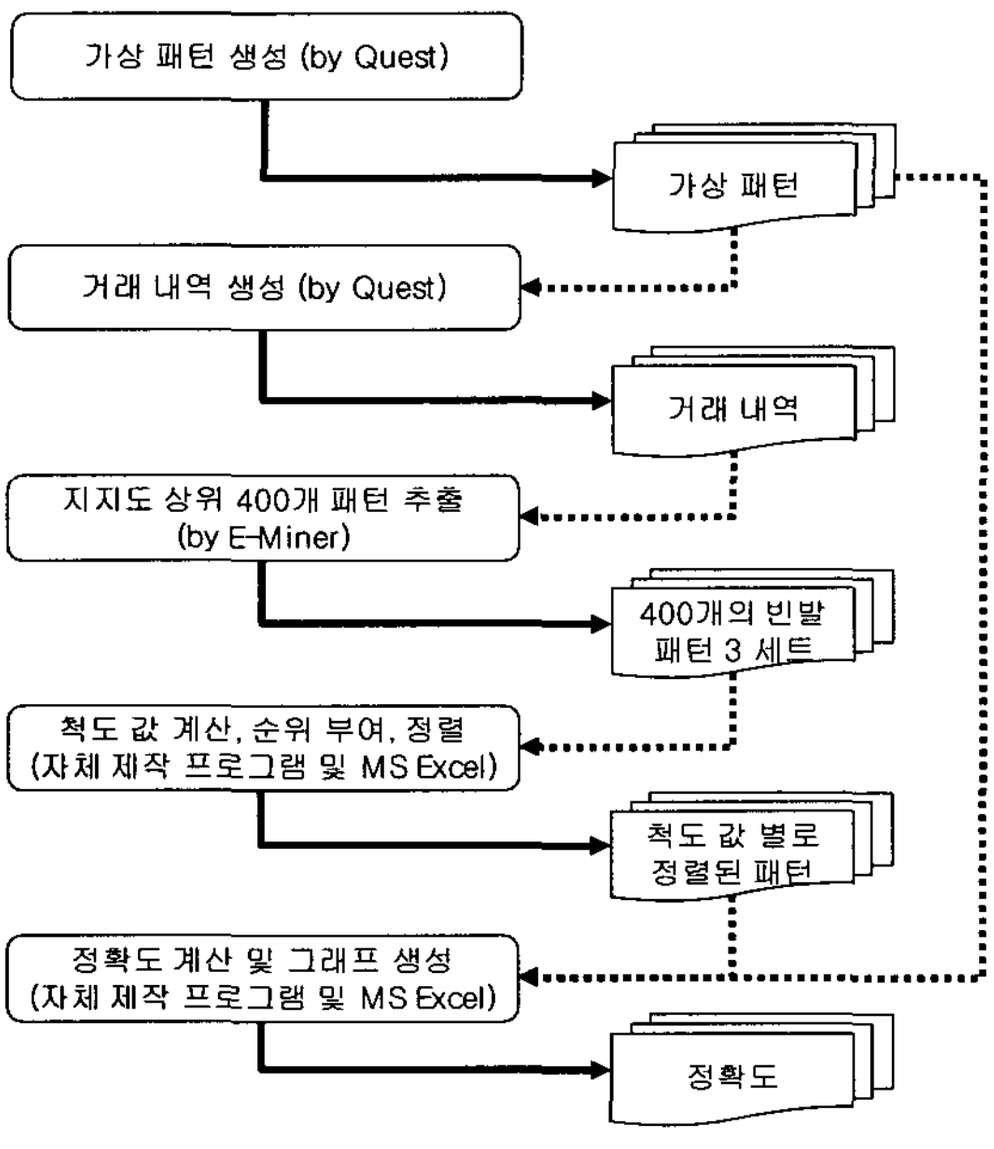
연속 패턴 간의 상관 계수	= 0.25
각 규칙의 평균 신뢰도	= 0.75
신뢰도의 변화율	= 0.1
거래의 총 수	= 2000
물품의 총 수	= 1000
조합된 패턴의 총 수	= 200

가변 인수

패턴의 평균 길이 (LP)	= *Short(2)*, Long(3)
장바구니의 평균 크기 (LT)	= Small(5), Medium(10), *Large(30)*

<그림 9> 패턴 및 거래 데이터 생성에 사용된 인수 값

한 회의 실험에서 정확성을 평가하기 위한 일



<그림 10> 각 실험에 대한 정확성 평가 과정

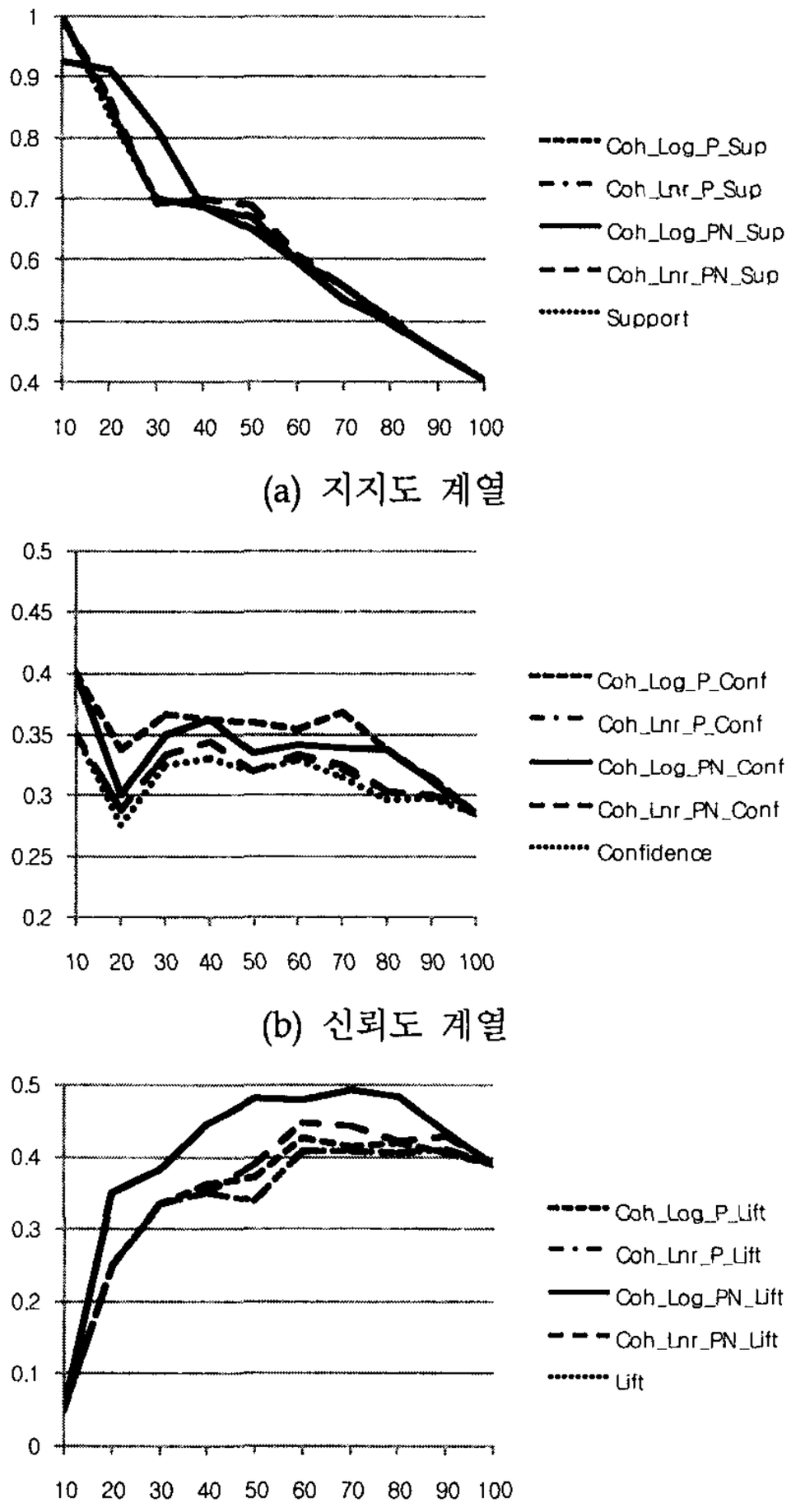
련의 과정이 <그림 10>에 간략하게 나타나있으며, 각 과정을 소개하면 다음과 같다. 우선 Quest 시스템을 사용하여 패턴을 생성하고, 이 패턴을 통계적으로 적용하여 실험용 거래 내역을 생성한다. 생성된 가상 거래 데이터로부터 빈발 패턴을 추출한다. 이 때, 패턴을 추출하기 위한 제약으로 패턴 내 항목의 개수가 4개 이하인 패턴만 추출하도록 하였고, 최소 지지도는 1%의 값이 사용되었다. 다음으로, 추출된 패턴에 대해 지지도, 신뢰도, 향상도 순으로 각각 정렬하여 상위 400개의 패턴만을 각각의 시트에 따로 저장한다. 지지도 기준으로 추출된 400개의 패턴에 대해서 <그림 8>에 소개된 지지도 계열 5가지 척도(*Support, Coh_Lnr_P_Sup, Coh_Log_P_Sup, Coh_Hnr_PN_Sup, Coh_Log_PN_Sup*)의 값을 각각 계산하고 계산된 값에 따라 순위를 부여한 뒤, 5개의 시트로 각각 저장하여 정렬한다. 이후, 개별 시트에 대하여 각 척도 별 정확도를 계산하는데, 시각적 파악이 용이하도록 하기 위해서 개별 패턴이 아닌 40개씩 패턴을 집단화하여 정확도를 계산한다. 즉, 해당 척도를 기준으로 하여 순위를 부여했을 때, 순위 1위 부터 40위, 41위 부터 80위 등으로 400개의 패턴을 총 10개의 집단으로 분화하여, 각 집단에 속한 40개의 패턴 중 실제로 존재하는 패턴 집합에서 발견되는 패턴의 비율을 계산한다. 계산된 정확도는 그래프로 시각화되며, 다양한 작업부하 하에서 수행된 실험의 결과 그래프가 다음 부절에서 결과 분석에 사용된다. 신뢰도 및 향상도에 대한 분석 과정도 위와 유사하므로, 자세한 설명은 생략하기로 한다.

실험에 사용된 응용 프로그램에 대한 정보는 다음과 같다. 패턴 및 거래 내역의 생성에는 Quest가 사용되었고, 거래 내역으로부터 지지도, 신뢰도, 향상도 기준 상위 400개의 패턴 3세트를 추출하는 과정에는 SAS 9.1의 Enterprise Miner 4.3이 사용되었다. 각 척도의 값과 정확도는 Visual C++ 6.0 컴파일러를 사용하여 직접 제작한 프로그램에서 계산되었으며, 각 척도에 대한 패턴의 순위 선정 및 정렬, 그리고 결과 그래프의 생성 과정에서는 Microsoft Excel 2007이 사용되었다. 모든 실험은 Microsoft Windows XP 운영체제 상에서 수행되었다.

4.2 실험 결과 및 해석

본 부절에서는 패턴의 평균 길이와 장바구니의 평균 크기를 변화시켜가며 각 척도들의 정확성이 어떻게 변화하는지 살펴보기로 한다. <그림 11>은 <그림 8>에 소개된 15개의 척도에 대해서, 패턴의 평균 길이(LP) = 2, 장바구니의 평균 크기(LT) = 30인 작업부하 하에서 정확성 분석 실험을 수행한 결과이다. <그림 (a)>는 지지도 계열 척도 5개, (b)는 신뢰도 계열 척도 5개, 그리고 (c)는 향상도 계열 척도 5개에 대한 실험 결과이다.

위 실험에서 파악할 수 있는 사항은 크게 두 가지이다. 우선, 지지도 계열 척도들이 신뢰도나 향상도 계열 척도들에 비해 정확성 측면에서 좋은 성능을 보임을 알 수 있다. 지지도 계열 척도



(a) 지지도 계열

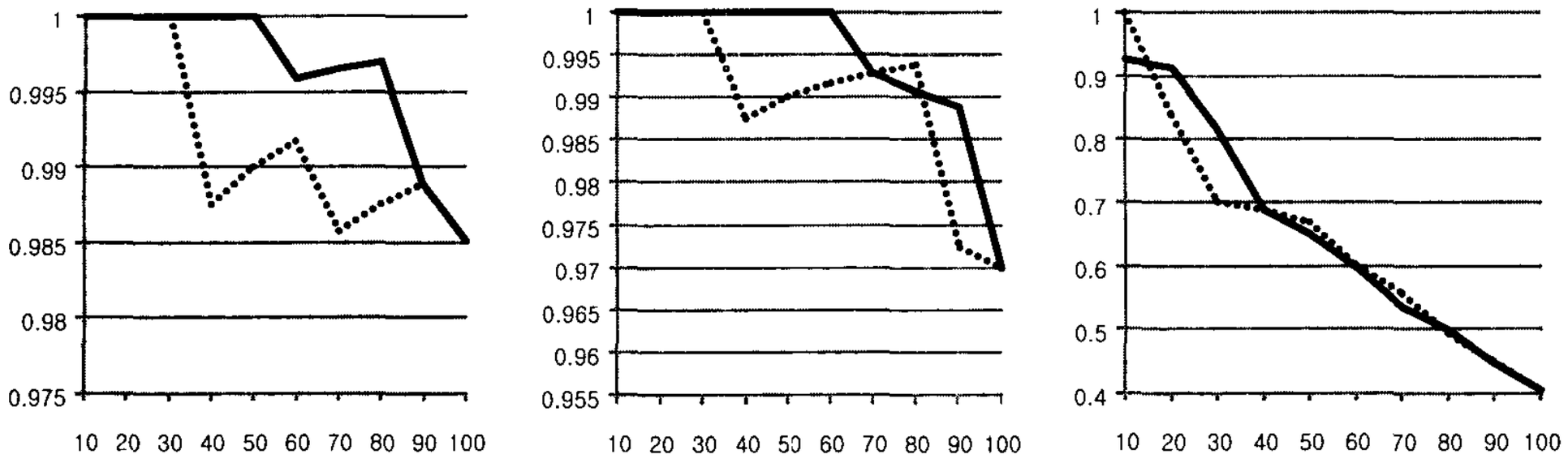
(b) 신뢰도 계열

(c) 향상도 계열

<그림 11> 기준 환경 하에서의 15개 척도들에 대한 정확성 평가

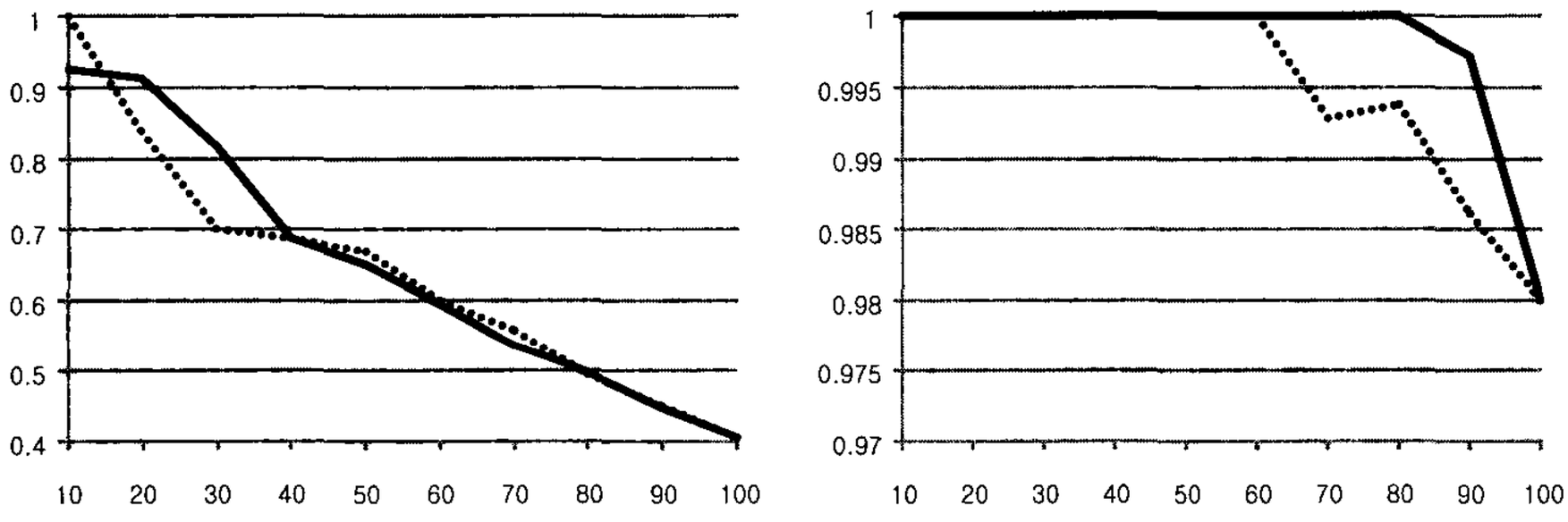
들은 전체적으로 정확성이 높게 나타날 뿐 아니라, 상위 랭크에 속한 패턴일수록 정확도가 높게 나타나므로, 의미있는 패턴에 비교적 높은 점수를 부여하고 있음을 알 수 있다. 한편, 신뢰도와 향상도 중에서는 신뢰도가 다소 높은 정확도를 보이며, 특히 향상도의 경우 상위 랭크에 속한 패턴의 정확성이 우수하게 나타나지 않으므로 정확성 측면에서 패턴을 올바르게 평가하고 있다고 보기 어렵다. 실험에서 파악할 수 있는 두 번째 특성은, 본 논문에서 고안한 척도들이 고전적 척도들에 비해 정확성이 다소 높게 나타난다는 것이다. 즉, 정규화 모델에 따라 정도의 차이는 있지만, 장바구니의 크기를 감안한 척도가 그렇지 않은 척도에 비해 높은 정확성을 보임을 알 수 있다. 정확성은 실험 수행 환경에 의해 상이하게 나타날 수 있으므로 다양한 작업부하 하에서의 추가 실험이 필요하다. 특히 다양한 인수를 변형해가면서 수행한 실험 결과 장바구니의 크기의 평균 크기와 패턴의 평균 길이의 비율에 따라 정확성이 가장 많이 영향을 받는 것으로 나타났다. 따라서 본 부절에서는 패턴의 평균 길이를 고정한 상태에서 장바구니의 평균 크기가 정확도에 미치는 영향과, 장바구니의 평균 크기를 고정한 상태에서 패턴의 평균 길이가 정확도에 미치는 영향을 분석하였다. 모든 실험에서 결합력을 반영한 척도들이 전통적인 척도들보다 대체적으로 높은 정확도를 보였다. 하지만 <그림 11>에서 보는 바와 같이 5개 척도들의 정확성을 하나의 그래프에서 나타내는 데에 어려움이 있으므로, 그래프의 가독성을 높이기 위해 제안된 척도들 중 전반적으로 가장 높은 정확도를 보이는 Coh_Log_PN 계열 척도와 전통적인 척도만을 그래프에서 비교하였다.

지지도 계열 척도의 정확도를 보다 상세하게 분석하기 위해 다양한 작업부하 하에서 실험을 수행하였으며, 그 결과는 <그림 12>에 나타나있다. <그림 12>는 패턴의 길이를 Short (LP = 2)로 고정한 상태에서 장바구니 평균 크기를 Small, Medium, 그리고 Large로 변화시켰을 때 나타나는 지지도 계열 척도의 정확성 변화를 나타낸 그래프이다. 모든 환경에서 정확도는 상위 랭크 패턴이 높게 나타났으며, 장바구니의 평균 크기가 클수록 정확도가 낮게 나타남을 알 수 있다. 이는, 장바구니의 평균 크기가 커질수록 한 바구니 내에 존재하는 패턴의 개수가 많아지게 되고, 따라서 패턴간 동시출현의 비율이 높아지기 때문인 것으로 사료된다. <그림 13>은 장바구니의 평



(a) LT = Small **(b) LT = Medium** **(c) LT = Large**

<그림 12> 장바구니 평균 크기의 변화에 따른 지지도 계열 척도의 정확성 변화

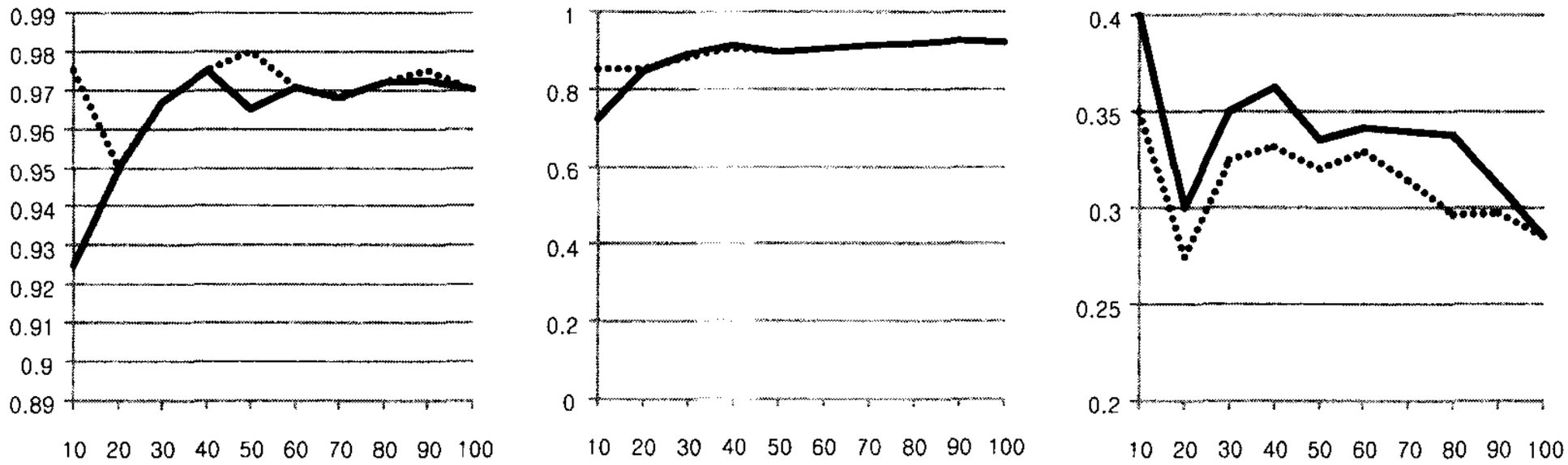


(a) LP = Short **(b) LP = Long**

<그림 13> 패턴의 평균 길이 변화에 따른 지지도 계열 척도의 정확성 변화

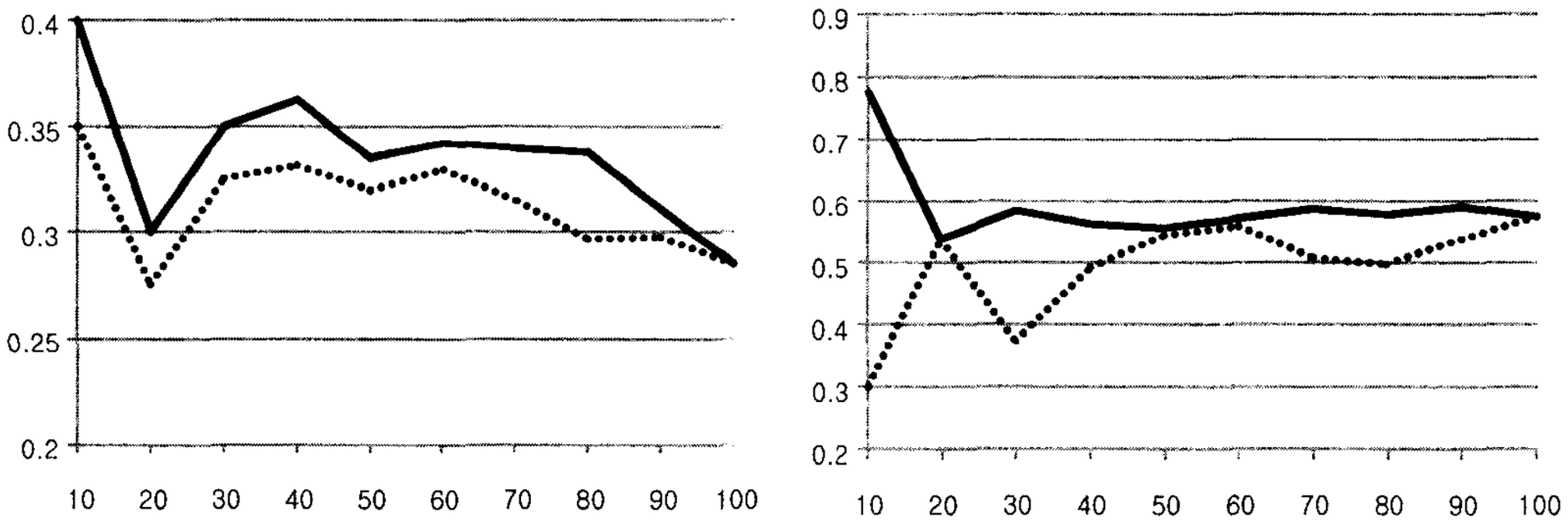
균 크기를 Large로 고정한 상태에서, 패턴의 평균 길이가 지지도 계열 척도의 정확성에 미치는 영향을 나타낸 그래프이다. 모든 환경에서 정확도는 상위 랭크 패턴이 높게 나타났으며, 패턴의 평균 길이가 큰 오른쪽 그림이 왼쪽 그림보다 정확도가 높게 나타남을 알 수 있다. 이것은 패턴의 평균 길이가 커질수록 한 바구니 내에 존재하는 패턴의 개수가 줄어들게 되고, 따라서 패턴간 동시출현의 비율이 낮아지기 때문인 것으로 사료된다.

다음으로, 신뢰도 계열 척도의 정확도를 보다 상세하게 분석하기 위해 다양한 작업부하 하에서 실험을 수행하였으며, 그 결과는 <그림 14>에 나타나있다. <그림 14>는 장바구니 평균 크기의 변화에 따른 신뢰도 계열 척도의 정확성 변화를 나타낸 그림이다. 지지도의 경우와 마찬가지로, 장바구니의 평균 크기가 클수록 정확도가 낮게 나타남을 알 수 있다. 이는, 장바구니의 평균 크기가 커질수록 한 바구니 내에 존재하는 패턴의 개수가 많아지게 되고, 따라서 패턴간 동시출현



(a) LT = Small **(b) LT = Medium** **(c) LT = Large**

<그림 14> 장바구니 평균 크기의 변화에 따른 신뢰도 계열 척도의 정확성 변화

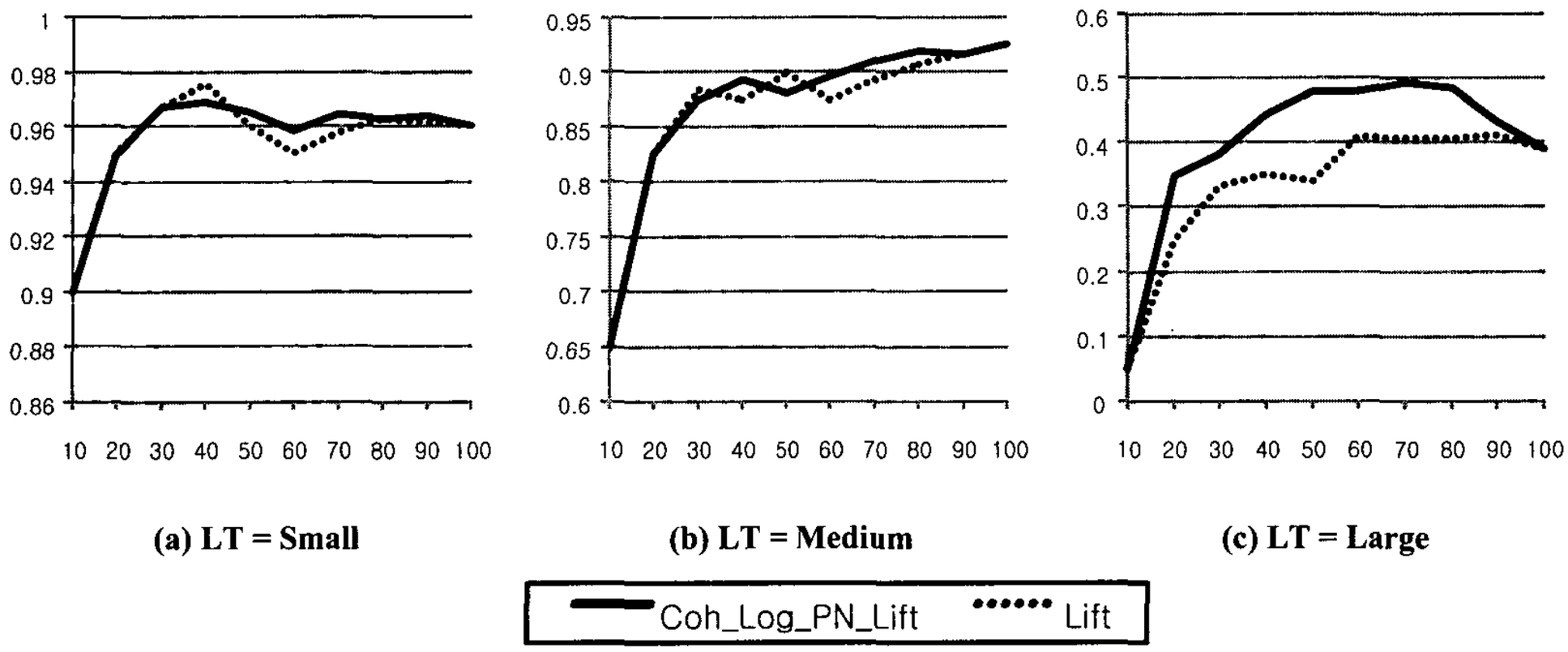


(a) LP = Short **(b) LP = Long**

<그림 15> 패턴의 평균 길이 변화에 따른 지지도 계열 척도의 정확성 변화

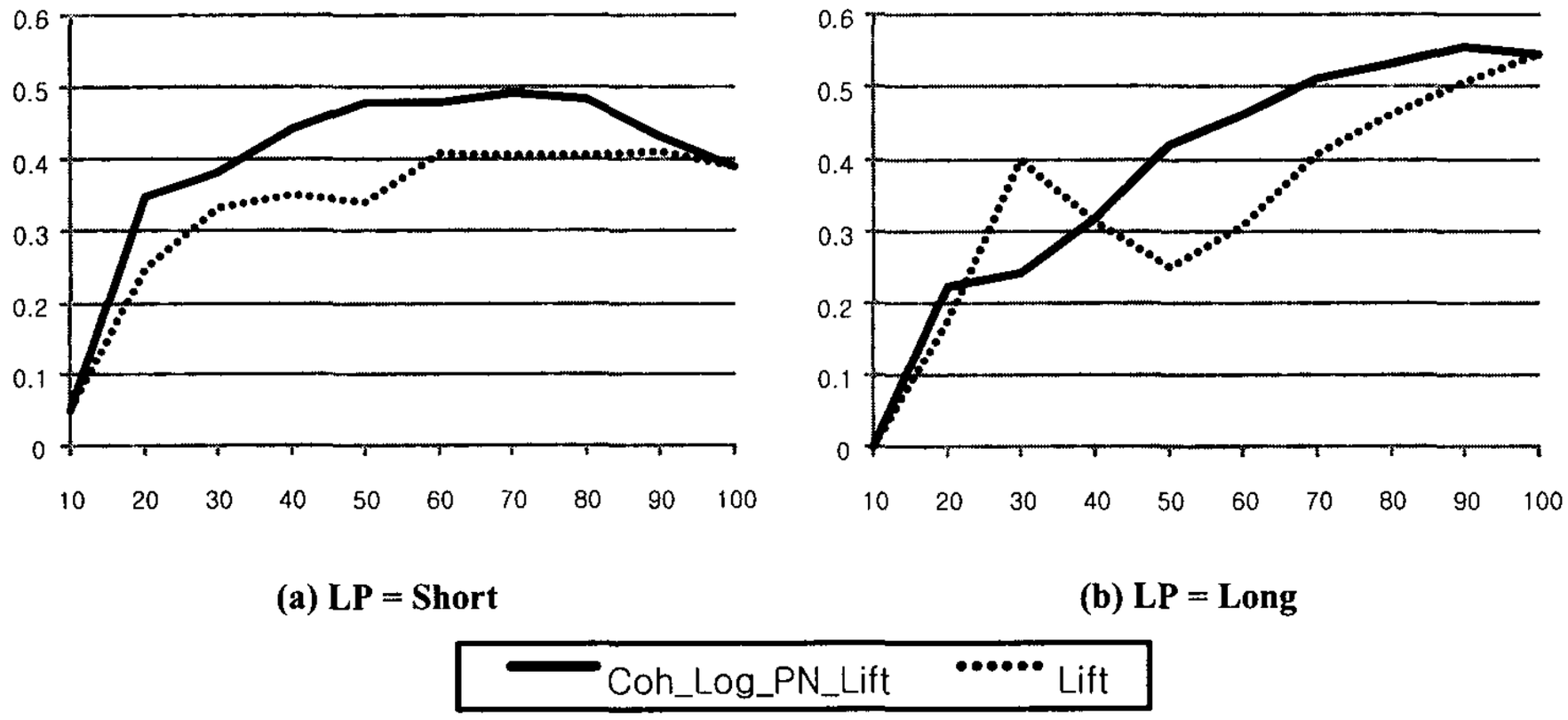
의 비율이 높아지기 때문인 것으로 사료된다. <그림 15>은 장바구니의 평균 크기를 고정한 상태에서, 패턴의 평균 길이가 신뢰도 계열 척도의 정확성에 미치는 영향을 나타낸 그림이다. 역시 지지도와 마찬가지로, 패턴의 평균 길이가 큰 우측 그림에서 정확도가 높게 나타남을 알 수 있다. 이것은 패턴의 평균 길이가 커질수록 한 바구니 내에 존재하는 패턴의 개수가 줄어들게 되고, 따라서 패턴간 동시출현의 비율이 낮아지기 때문인 것으로 사료된다. 마지막으로 장바구니의 평균 크기와 패턴의 평균 길이가 향상도의 정확성에 미치는 영향이 <그림 16>과 <그림 17>에 각각 나타나있다. 향상도의 경우도 장바구니의 크기가 클수록 정확도가 낮아지는 현상을 보였지만, 패턴의 평균 길이에 대해서는 의미있는 변화를 보이지 않은 것으로 나타났다.

본 부절에서는 흥미성 평가를 위한 전통적 척도인 지지도, 신뢰도, 그리고 향상도와, 이들 척



(a) LT = Small **(b) LT = Medium** **(c) LT = Large**

<그림 16> 장바구니의 평균 크기 변화에 따른 향상도 계열 척도의 정확성 변화



(a) LP = Short **(b) LP = Long**

<그림 17> 패턴의 평균 길이 변화에 따른 향상도 계열 척도의 정확성 변화

도에 결합력 개념을 접목시킨 다양한 척도들에 대한 정확성 측면에서의 성능을 분석해보았다. 실험 결과 이들 전통적 척도 3가지 중, 정확성 측면에서 가장 바람직한 성능을 보이는 것은 지지도인 것으로 나타났다. 또한 이들 세 가지 척도 모두 결합력 개념의 적용을 통해 정확도를 크게 개선할 수 있음을 볼 수 있었다. 특히 선형 정규화를 거친 모델보다는 로그 정규화를 거친 모델이 정확도의 향상에 더욱 기여하는 것으로 나타났다. 로그 정규화 모델 중에는 가점만 고려한 모델보다는 가점과 벌점을 모두 고려한 모델이 실험 전체에 걸쳐서 골고루 우수한 성능을 보이는 것으로 나타났다. 실험 전반에 걸친 일관된 결론은, 장바구니의 크기가 클수록 그리고 패턴의 평균 길이가 짧을수록 정확성이 떨어진다는 것이다. 이는 패턴의 평균 길이에 대한 장바구니의 평균 크기의 비율이 클수록 한 바구니 내에 여러 패턴이 존재하게 되어 패턴간 동시출현의 비율이 높아질 것이라는 예상을 뒷받침하는 결과이다.

V. 결 론

본 논문에서는 연관규칙의 흥미성을 평가하기 위한 기존의 척도들이 실제로 존재하는 의미 있는 패턴과 우연히 조합되어 발생한 무의미한 패턴을 식별하지 못하는 한계를 지적하고, 이 한계를 극복하기 위해 패턴의 발생 빈도뿐 아니라 장바구니의 크기까지 고려한 결합력 기반 흥미성 척도를 제시하였다. 제안하는 척도는 큰 장바구니에서 발생한 패턴의 경우 작은 장바구니에서 발생한 패턴에 비해 낮은 가중치를 부여함으로써, 실제로 의미 없는 패턴이 전통적 척도들에 의해 가치 있는 규칙으로 과대평가되는 부작용을 완화할 수 있을 것으로 사료된다.

또한 기존의 흥미성 척도들에 관한 비교 연구들이 대부분 정성적 접근방법만을 취하고 있다는 한계를 극복하기 위해서, 본 논문에서는 척도들의 정량적 성능 비교를 위한 방안으로서 정확도의 기준을 제안하였다. 정확도는 각 척도에 의해 흥미성이 높은 것으로 평가 받은 패턴들 중 실제로 존재하는 의미 있는 패턴의 비율로 정의되며, 실험 결과 기존의 대표적 척도 중 지지도, 신뢰도, 그리고 향상도의 순으로 정확도가 높은 것으로 나타났다. 또한 이들 전통적 척도들에 장바구니의 크기를 고려한 결합력의 개념을 접목시킴으로써 현저한 정확도의 개선을 가져올 수 있음을 실험을 통해 입증하였다.

본 논문에서는 결합력을 적용하여 기존의 흥미성 척도들을 보완할 수 있는 가능성을 제시하였으며, 이러한 가능성을 구체화하기 위해 다양한 측면에서의 후속 연구를 진행하고 있다. 후속 연구의 한 가지 방향은, 이론적인 배경을 충분히 갖춘 결합력 모델을 제시하는 것이다. 본 논문에서 정규화 방법 및 벌점 감안 여부에 따라 네 가지의 결합력 모델을 제시했지만, 충분한 이론적 고찰을 통해 정확도를 보다 향상시킬 수 있는 새로운 결합력 모델이 고안될 수 있을 것으로 기대한다. 후속 연구의 또 다른 방향은 보다 다양한 작업부하 하에서의 충분한 실험을 통해 결합력의 효과 및 성능 평가 기준으로서의 정확도의 활용 가능성을 검증하는 것이다. 또한 본 연구에서 비교 대상으로 선정된 전통적인 세 가지 척도 이외에도 비교적 최근에 고안된 다른 흥미선 척도들도 많이 존재하므로, 이들에 대한 비교 연구도 다양한 작업부하 하에서 수행되어야 한다.

〈참 고 문 헌〉

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

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

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

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

[5] Brin, S., Motwani, R., and Silverstein, C.,

"Beyond Market Baskets: Generalizing Association Rules to Correlations," *in Proc. ACM SIGMOD International Conference of Management of Data*, Tucson, Arizona, 1997, pp. 265-276.

[6] Cai, C.H., Fu, A.W.C., Cheng, C.H., and Kwong, W.W., "Mining Association Rules with Weighted Items," *in Proc. 10th International Symposium on Database Engineering and Applications*, Wales, U.K., 1998, pp. 68-77.

[7] Carter, C.L., Hamilton, H.J., and Cercone, N., "Shared Based Measures for Itemsets," *in Proc. 1st European Symposium on the Principles of Data Mining and Knowledge Discovery*, Trondheim, Norway, 1997, pp. 14-24.

[8] Chen, M.S., Han, J., and Yu, P.S., "Data Mining: An Overview from a Database Perspective," *IEEE Transactions on Knowledge and Data Engineering*, Vol. 8, 1996, pp. 866-883.

[9] Cooper, C., and Zito, M., "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, 2007, pp. 398-405.

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

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

[12] Hu, Y.H. and Chen, Y.K., "Mining Association Rules with Multiple Minimum Supports: A New Mining Algorithm and a Support Tuning Mechanism," *Decision Support Systems*, Vol. 42, 2006, pp. 1-24.

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

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

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

[16] Liu, B., Hsu, W., and Ma, Y., "Mining Association Rules with Multiple Minimum Supports," *in Proc. 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Diego, California, 1999, pp. 337-341.

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

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

[19] Tao, F., Murtagh, F., and Farid, M., "Weighted Association Rule Mining using Weighted Support and Significance Framework," *in Proc. 9th ACM SIGKDD International Conference on Knowledge Discovery and Data*

Mining, Washington D.C., 2003, pp. 661-666.

[20] Vaillant, B., Lenca, P., and Lallich, S., "A Clustering of Interestingness Measures," *in Proc, 7th International Conference on Discovery Science*, Padova, Italy, 2004, pp. 290-297.

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

◆ 저자소개 ◆

김남규 (Kim, Namgyu)
현재 국민대학교 경상대학 비즈니스IT학부에서 전임강사로 재직 중이다. 서울대학교 컴퓨터공학과에서 학사 학위를 취득하고, KAIST 테크노경영대학원에서 Database와 MIS를 전공하여 경영공학 석사 및 박사학위를 취득하였다. The Journal of Systems and Software, Expert Systems with Applications, The Journal of Computer Information Systems, 경영정보학연구지, 정보과학회논문지 등의 학술지에 논문을 다수 게재하였으며, 주요 관심분야는 Data Mining 및 Data Modeling이다.

◈ 이 논문은 2008년 02월 29일 접수하여 1차 수정을 거쳐 2008년 06월 02일 게재확정되었습니다.