「정보시스템연구」 제22권 제3호
한국정보시스템학회
2013년 9월, pp. 25~42

http://dx.doi.org/10.5859/KAIS.2013.22.3.25

구매의도 생성 순서와 구매실현 순서의 역전 현상을 감안한 확장된 순차분석 방법론

김민석** · 김남규***

〈목 차〉

Ⅰ. 서론

컴퓨터 기술의 발전과 이에 기반한 서비스의 확대로 인해, 최근 생성, 공유, 저장되는 데이터의 양은 매우 빠른 속도로 증가하고 있다. 2011년 한 해 동안 전 세계에서 생성된 디지털 정보량은 약 1.8ZB에 이르며, 이러한 디지털 정보량은 2년마다 약 2배씩 증가하고 있는 것으로 알려져 있다. 이러한 현상은 데이터의 양 자체가 문제의 일부분이 되는(O'Reilly Radar Team, 2011) 빅 데이터(Big Data) 분석 기술에 대한 수요와 관심을 증대시키고 있으며, 이와 동시에 이러한 데이터에 대한 분석을 통해 의미 있는 결과를 도출하는 과정을 더욱 어렵고 복잡하게 만들고 있다. 이와 동일한 맥락에서 방대한 양의 데이터로부터 기존에 알려지지 않았던 의미 있는 지식을 창출하기 위한 방법론(Han & Kamber, 2007)인 데이터 마이닝(Data Mining) 기술에 대한 수요가 최근 증가함과 동시에, 기존의 분석 사례를 통해 나타난 전통적 마이닝 기법의 한계를 극복하기 위한 다양한 시도가 이루어지고 있다. 특히 최근에는 기존에는 쉽게 다룰

* 이 논문은 2012년 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (NRF-2012S1A5A2A01015135).
** 국민대학교 비즈니스IT전문대학원 석사과정, sm3236@nate.com
*** 국민대학교 경영정보학부 부교수(교신저자), ngkim@kookmin.ac.kr

수 없었던 비정형 텍스트 데이터에 대한 마이닝을 통해 새로운 지식을 발굴하기 위한 시도(유은지 외, 2012; 이영재 & 이성수, 2011)가 많은 분야에서 이루어지고 있다.

데이터 마이닝이 방대한 데이터의 분석을 통해 다양한 비즈니스 의사결정을 지원할 수 있는 효과적 기술임에도 불구하고, 마이닝에 대한 회의적 시각은 여전히 존재하고 있다. 마이닝에 대한 비판적 시각은 데이터 마이닝 분석을 위해 투입되는 비용, 시간, 인력 등의 자원에 비해서, 실제로 분석 결과를 통해 얻을 수 있는 성과가 두드러지게 나타나지 않는다는 점에 기인한다. 특히 데이터 분석을 희망하는 모든 조직이 타 조직에 대해 기 수행된 마이닝 성과를 전혀 활용하지 못하고 항상 새로운 분석을 시도하고 있으며, 이는 곧 데이터 분석 프로젝트 비용을 더욱 증가시키는 원인으로 작용하고 있다. 만약 신뢰할 수 있는 조직의 데이터에 대해 신뢰할 수 있는 분석가가 데이터 마이닝 분석을 수행하고 이러한 분석 결과를 유사 업종에 속한 타 조직의 데이터 분석에 재사용할 수 있다면, 마이닝 분석 성과의 활용도가 높아질 뿐 아니라 궁극적으로 분석의 품질도 점차 향상되는 선순환이 이루어질 수 있을 것이다.

하지만 실제로는 동일 업종에 속하는 두 조직이라도 각 조직에서 데이터가 생성되는 환경이 서로 다르기 때문에, 한 조직의 데이터 마이닝 분석 결과를 다른 조직에 그대로 적용하는 것은 바람직하다고 할 수 없다. 예를 들어 대형 마트 내 식료품 매장의 연관성 분석(Association Rule Mining) 결과를 토대로 소규모 수퍼마켓의 판매 전략을 수립하는 일 또는 도심에 위치한 스포츠 매장의 연관성 분석 결과를 토대로 시골에 위치한 스포츠 매장의 판매 전략을 수립하는 일 등은 모두 합리적 의사결정이라고 보기 어려울 것이다. 즉 동일 업종의 매장들이라고 하더라도 고객 수, 고객의 재구매율, 취급 물품의 수, 거래당 평균 구매물품의 수 등 각 매장의 상황이 서로 다르기 때문에, 한 매장의 데이터 분석 결과를 다른 매장에 곧바로 적용하기에는 어려움이 있다. 이러한 어려움의 원인 중 한 가지는, 연관성 분석을 위해 고안된 신뢰도(Confidence), 지지도(Support), 향상도(Lift)를 비롯한 수많은 기존의 흥미성 척도들(Interestingness Measures)(Geng & Hamilton, 2006)에는 각 매장의 상황을 반영할 수 있는 요소가 전혀 포함되어있지 않다는 것이다. 만약 각 매장의 상황을 표현하는 변수 중 구매 패턴에 영향을 줄 수 있는 변수들이 식별된다면, 이러한 변수에 의해 매장의 상황을 보다 객관적으로 규정할 수 있을 것이다.

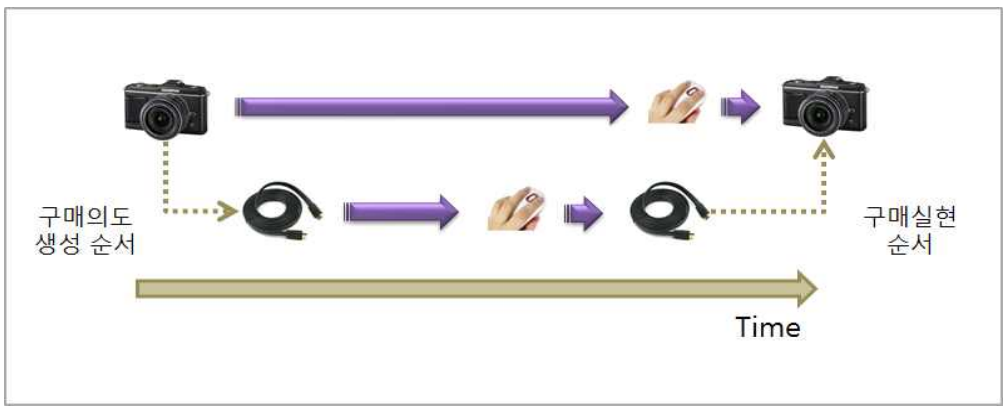
본 연구에서는 데이터 마이닝 분석 성과의 확대 재생산을 위한 상황 변수 중, 동시 구매와 순차 구매를 구분하는 거래 시간 간격에 초점을 맞추고자 한다. 임의의 시간 간격을 두고 발생한 두 건의 구매는 동시성 기준에 따라 동시 구매로 간주될 수도 있고 순차 구매로 간주될 수도 있다. 예를 들어 물품 A의 구매가 이루어진 뒤 1초 후에 물품 B의 구매가 이루어졌을 때, 이를 (A → B)의 순차 구매로 간주하는 것은 다소 불합리한 측면이 있다. 한편 어떤 고객이 물품 A를 구매한 뒤 1주일 후에 물품 B를 구매했다면, 이는 (A → B)의 순차 구매로 파악하는 것이 합당할 것이다. 그렇다면 물품 A와 물품 B의 구매 사이에 5시간의 시간 간격이 존재하는 경우는 어떻게 해석해야 할까? 이 경우엔 동시성의 기준을 어느 정도의 시간으로 설정하느냐에 따라

해석이 달라진다. 예를 들어 1시간 이내에 이루어진 여러 건의 구매를 동시 구매로 정의한다면 5시간 간격으로 이루어진 구매는 순차 구매로 간주하게 되고, 12시간 이내에 이루어진 여러 건의 구매를 동시 구매로 정의한다면 5시간 간격으로 이루어진 구매는 동시 구매로 간주하게 된다.

많은 연구에서 이러한 동시성 기준에 대한 엄밀한 고찰 없이 단순히 원 데이터에 나타난 식별자에 의존하여 구매의 순서를 규정함으로써, 실세계의 현상에 비해 물품의 구매 순서를 지나치게 엄격하게 반영하여 순차 분석을 수행하는 경우를 볼 수 있다. 이러한 분석의 부작용의 예는 <그림 1>을 통해 설명할 수 있다. 어떤 고객이 디지털 카메라를 구매하기로 결정하였으나, 1주일 후 해당 제품의 할인 판매가 있을 것을 예상하여 그 구매를 실현하지 않고 지연시키는 상황을 가정해 보자. 다음날 이 고객이 쇼핑몰 검색 중 해당 디지털 카메라에 사용 가능한 케이블을 발견하고 이를 즉시 구매하였으며, 추후 할인 판매를 통해 해당 디지털 카메라에 대한 구매를 실현했다고 하자. 이 경우 전통적인 순차분석은 구매실현 기준인 (케이블 → 디지털 카메라)의 순차 관계에 집중하게 되며, 이는 구매실현을 유발한 구매의도의 생성 순서인 (디지털 카메라 → 케이블)의 순차 관계와는 반대로 나타나게 된다.

<그림 1>의 논의는 임의의 시간 간격으로 실현된 두 구매에서 구매의도 생성 순서와 구매실현 순서의 역전 현상이 나타날 수 있으며, 이 경우 전통적인 순차분석은 실현된 구매 순서만을 반영할 뿐 구매의도가 생성된 순서는 반영하지 못하는 한계를 가짐을 암시한다.

본 연구는 이러한 현상에 대한 보다 엄밀한 고찰을 목적으로 한다. 구체적으로 본 연구의 방향은 다음의 두 가지로 요약된다. 먼저 구매의도와 구매실현 간 역전현상을 고려함으로써, 분석를 통해 도출된 규칙의 정확도가 변화하는 양상을 파악하고자 한다. 또한, 역전현상이 발생하기에 충분할 정도의 매우 짧은 시간 간격을 다양한 기준에서 정의하고, 어떤 기준 하에서 규칙의 정확도가 가장 우수하게 나타나는지 측정함으로써 특정 환경에서 수행되는 구매 행위에서 "동시"로 간주하기에 적합한 시간의 간격을 발견하



<그림 1> 구매의도 생성 순서와 구매실현 순서의 역전 예

고자 한다. 요약하면 “동시”로 간주하기에 적합한 시간 간격을 발견하고 이 기준을 분석에 적용함으로써, 순차분석 규칙의 정확도를 향상시키는 것이 본 연구의 목적이라고 할 수 있다.

본 논문의 이후 구성은 다음과 같다. 다음 절인 제 2절에서는 연관 분석 및 순차 분석, 그리고 구매의도에 대한 기존 연구를 소개한다. 또한, 제 3절에서는 본 연구에서 제안하는 방법론을 소개하며, 이의 실험 결과는 제 4절에서 다루도록 한다. 마지막 절인 제 5절에서는 본 연구의 기여 및 한계 그리고 향후 연구방향을 제시한다.

Ⅱ. 관련 연구

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. 이처럼 다양한 데이터 마이닝 기법 중 특히 연관 분석과 순차 분석이 본 연구에서 다루는 주제와 밀접한 관계가 있으므로, 본 절에서 이 두 가지에 대해 보다 자세히 소개하도록 한다.

연관관계 분석(Agrawal et al., 1993; Agrawal & Srikant, 1994)은 데이터들의 빈도수와 동시 발생 확률을 이용하여 데이터와 데이터간의 관계를 찾고 이를 규칙으로 표현하기 위한 분석 기법이다. 규칙은 항목 집합 A와 B에 대해 (A → B)의 형태로 표현하며, 항목 집합 A가 나타날 때 항목 집합 B도 함께 나타나는 경향이 있음을 의미한다. 예를 들어, 슈퍼마켓의 판매 데이터에서 (빵, 버터 → 우유)와 같은 연관규칙이 발견된다고 하면 이는 "빵과 버터를 구매하는 고객은 우유도 함께 구매하는 경향이 있다"는 의미를 갖는다. 한편 순차패턴(Agrawal & Srikant, 1995)은 시간적 순서에 따라 나타나는 항목 집합 간의 관계를 반영한다. 만약 항목 집합 A, B에 대해 (A → B)의 순차관계가 존재한다면, 항목 집합 A가 나타난 후 항목 집합 B가 나타나는 경향이 있음을 의미한다. 예를 들어, 비디오 대여점의 대여 기록에서 (베를린 → 신세계)와 같은 순차패턴이 발견된다고 하면 이는 "베를린을 대여한 사람은 조만간 신세계를 대여한다"와 같이 해석될 수 있다. 연관 분석과 순차 분석은 장바구니 분석, 인터넷 쇼핑몰 추천시스템, 교차판매, 매장배치, 카탈로그 설계, 판촉전략 수립 등 다양한 분야(김재경 외, 2005; 안현철 외, 2006; 정영수 & 강경화, 2004; 하성호 & 박상찬, 2002; Burke, 2000; Wang et al., 2004)에서 활용되고 있다.

연관 분석과 순차 분석은 두 사건의 발생 간 시간 간격에 따라 명확하게 구분되어야 함에도 불구하고, 많은 선행연구들은 명확한 기준이 없는 상태에서 두 기법을 선택적으로 적용해 온 경향이 있다. 예를 들어, 어떤 거래가 1시에 발생하고 추후 다른 거래가 2시에 발생했다고 가정하자. 이러한 상황에 대해 어떤 연구는 1시간의 간격을 두고 발생한 순차 구매로 간주하는 반면, 다른 연구는 동일 날짜에 발생한 모든 거래를 하나의 바구니로 묶어서 이 두 거래를 동시 구매로 간주하기도 한다.

이처럼 장바구니 기준을 어떻게 정하느냐에 따라 연관분석과 순차분석의 경계가 모호해지는 현상에 대해서는 상대적으로 많은 연구가 수

행되지 않았으나, 최근 이러한 현상과 관련이 있는 동시성 기준을 다룬 연구(김미성 외, 2012)가 수행된 바 있다. 이 연구에서 지적한 바와 같이, 연관분석과 순차분석을 구분하여 장바구니 분석을 수행하기 위한 명확한 기준이 없기 때문에, 많은 경우 원 데이터(Raw Data)의 형태 또는 연구자의 선택에 따라 연관분석과 순차분석의 경계가 모호하게 적용되고 있다. 특히 이 연구는 동시 구매에 대한 기준, 즉 동시성 기준에 대한 확장된 정의를 제시하고, 다양한 동시성 기준을 적용하여 수행된 연관분석의 결과를 통해 동시성 기준에 따라 연관규칙의 정확도가 어떻게 변하는지를 살펴보았다. 이 연구는 매장의 상황 또는 고객의 특성에 따라 어떤 동시성 기준을 적용할지에 대한 가이드라인을 제시하였으며, 상황에 따라 차별화된 마케팅 전략을 수립할 수 있는 방안을 제시하였다는 점에서 의미를 갖는다.

고객의 니즈가 복잡하고 다양해지면서 많은 기업들은 다양한 고객관계관리(CRM: Customer Relationship Management) 기법을 통해 고객가치를 극대화하기 위한 방안을 모색하고 있다(이현규 & 박영식, 2006). CRM이란 용어는 1990년대에 정보기술에 기반을 둔 고객 솔루션 제공업체가 처음으로 사용하여 지금까지 활용되고 있다(Parvatiyar & Sheth, 2001). 한편 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 & Selnes, 2004). 최근에는 데이터 마이닝의 풍부한 통계적 기법을 CRM에 적용하기 위한 연구(송만석 외 2008; 하성호 & 이재신, 2003)가 활발히 수행되고 있다.

하지만 CRM의 핵심이 고객의 니즈 형성에 초점을 맞춘 것과 달리, 데이터 마이닝을 통한 CRM 분석은 대부분 니즈 형성 시점이 아닌 구매 시점의 데이터에 근거하여 수행되고 있다. 구매는 결국 니즈 형성의 결과로 나타나는 현상이라는 점을 감안하면, 구매 뿐 아니라 니즈 형성 시점에 대한 분석이 반드시 수행되어야 함을 알 수 있다. 이는 곧 영수증 상의 구매 순서, 즉 구매실현에 대한 순서만을 분석에 활용하는 전통적인 순차분석의 한계를 극복하고, 구매행위를 구매의도 생성과 구매실현이라는 2가지 단계로 구분해야 함을 의미한다. 그럼에도 불구하고 대부분의 데이터 마이닝 연구가 구매의도 생성 시점이 아닌 구매실현 시점에 대해 수행된 것은, 분석 가능한 데이터의 관리 및 확보의 한계성에 기인한다. 즉 구매실현 시점의 순서는 판매 데이터를 통해 파악 가능하지만, 구매의도 생성 시점을 나타내는 데이터는 명확하게 관리되기 어렵기 때문이다. 이는 곧 구매실현 시점의 순서를 나타내는 판매 데이터를 활용하되, 구매의도 생성 순서와 구매실현 순서의 역전 가능성까지 포함하여 순차분석을 확장하는 방식으로 연구가 수행되어야 함을 의미한다.

Ⅲ. 확장된 순차분석 방법론

3.1. 구매의도 생성과 구매실현의 구분

본 장에서는 순차분석 과정에서 구매의도 생성 순서와 구매실현 순서 간의 역전 현상을 고려해야 하는 근거를 제시하고, 분석 데이터에 대한 가공을 통해 이러한 현상을 고려한 순차분석을 수행하는 방안을 제시하고자 한다.

순차분석으로 도출된 규칙은 다양한 용도로 활용할 수 있지만, 특히 과거의 구매 패턴을 분석하여 앞으로 구매를 예측하기 위한 추천 시스템의 일부로 활용되는 경우가 많다. 즉 과거 구매 내력에서 물품 A를 구매한 고객은 이후 물품 B도 구매할 가능성이 높은 것으로 파악되었다면, 앞으로 물품 A를 구매하는 고객에게 물품 B도 추천함으로써 물품 B의 매출을 증대시킬 수 있을 것이다. 즉 이 과정은 물품 A를 구매하는 고객에게 물품 B에 대한 정보를 제공하거나 유리한 구매 조건을 제시함으로써, 물품 B에 대한 구매의도가 생성되게끔 유도하고 그 결과로 물품 B에 대한 구매가 실현될 가능성을 높이는 과정으로 볼 수 있다. 이처럼 구매라는 행위를 구매의도 생성과 구매실현이라는 두 단계로 분리하였을 때, 기존의 순차분석은 구매실현 순서만을 대상으로 할 뿐 구매의도 생성 순서는 충분히 고려하지 않고 있음을 알 수 있다. 이러한 한계는 구매실현이 구체적으로 측정할 수 있고 정확한 데이터로 저장되는 것과 달리, 구매의도 생성은 그 시점 자체가 모호할 뿐 아니라 구체적으로 측정하는 방법이 없다는 현실적 한계에서 기인하는 것으로 볼 수 있다. 즉 구매의도 생성은 측정 자체가 불가능했기 때문에 전통적 순차분석은 구매의도 생성 시점에 대한 고려 없이 구매실현 시점만을 대상으로 구매의 순차 관계를 분석했다고 할 수 있다.

물론 실현된 구매의 데이터만을 토대로 구매의도의 생성 시점을 파악하는 것은 현실적으로 한계가 있는 것으로 판단된다. 따라서 본 연구에서는 구매의도 생성 시점을 직접 파악하는 대신, 매우 짧은 간격을 두고 실현된 구매 건에 한해서 구매의도 생성 순서와 구매실현 순서 간에 역전 현상이 발생했을 가능성을 인정하는 우회적인 방법을 사용하고자 한다. 즉 어떤 고객이 물품 A를 구매한 후 매우 짧은 시간 간격 이후에 물품 B를 구매했다면, 실제로 발생한 (A → B)의 순차 관계는 물론 순서가 뒤바뀐 (B → A)의 잠재 순차 관계도 가상으로 발생시켜 분석에 포함시키는 것이다. 이처럼 잠재적인 순차 관계를 가상으로 발생시키는 방법으로 데이터를 가공시킨 후, 가공된 데이터에 대해 순차분석을 수행하면 전통적인 순차분석에서는 찾을 수 없었던 새로운 규칙이 도출될 수 있다. 본 연구에서는 이와 같은 방식으로 확장된 순차분석을 수행하고, 그 결과로 도출된 규칙과 전통적 방식에 따라 도출된 규칙의 정확도를 비교함으로써 제안하는 확장된 순차분석 방법론의 당위성을 보이고자 한다.

3.2. 구매의도 생성 순서와 구매실현 순서의 역전 현상

특정 시간 간격으로 실현된 두 구매에서 구매의도 생성 순서와 구매실현 순서가 서로 다르게 나타나는 역전 현상이 나타날 수 있다. 이는 순차분석 수행 과정에서 실제로 실현된 구매 순서뿐 아니라 그 반대의 순서로 구매의도가 생성되었을 가능성도 고려해야 함을 나타낸다. 이러한 주장은 구매 간 시간 간격이 매우 짧은 경우에는 그 순서를 지나치게 엄격하게 관리할 필요가 없

다는 측면에서 타당한 것으로 받아들여질 수 있다. 하지만 구매 간 시간 간격이 충분히 길게 나타난 경우, 위의 논리를 적용하기에는 무리가 따른다. 예를 들면 물품 A와 B가 10분 간격으로 구매되었을 경우 물품 B에 대한 구매의도가 물품 A에 대한 구매의도보다 오히려 먼저 생성되었을 가능성이 충분히 있지만, 물품 C와 물품 D가 1달 간격으로 구매된 경우라면 물품 D에 대한 구매의도가 물품 C에 대한 구매의도보다 먼저 생성되었을 가능성은 매우 낮을 것으로 예상된다. 즉 구매의도 생성 순서와 구매실현 순서의 역전 현상이 발생할 가능성은 구매 간 시간 간격에 따라 다르게 나타나게 되며, 이는 <그림 2>를 통해 자세히 설명된다.

<그림 2>는 구매의도와 구매실현 간 역전현상이 발생하기에 매우 짧은 구매 간격을 각각 1일, 5일, 8일로 가정했을 경우의 예를 보여준다. "1일 기준" 행은 1월 1일부터 1월 13일 사이의 다섯 날짜에 각각 A, B, C, D, E의 물품이 구매된 사실을 보여주고 있다. 만약 5일이라는 시간을 매우 짧은 시간이라고 정의한다면, 특정 날짜를 기준으로 5일 이내의 기간 내에 이루어진 구매를 동시구매로 간주할 수 있음을 의미한다. 즉 <그림 2>의 "5일 기준" 행에서 1월 7일에 해당하는 셀에는 1월 3일 ~ 1월 7일 사이에 구매된 모든 물품이 기록된다. 같은 원리로 "8일 기준" 행에서 1월 10일에 해당되는 셀에는 1월 3일 ~ 1월 10일 사이에 구매된 모든 물품이 기록된다. 이처럼 구매의도와 구매실현 간 역전현상이 발생하기에 매우 짧은 구매 간격을 각각 1일, 5일, 8일로 가정했을 경우, 이러한 각 가정이 분석 결과에 미치는 영향을 물품 C와 D에 대해 살펴보도록 하자. 실제 구매 기록은 "1일 기준" 행에 나타나며, 이 경우 (C → D)의 순차 관계만 존재한다. 하지만 "5일 기준"의 경우 (C → D)의 순차 관계와 함께 (C, D)의 동시구매로 말미암은 연관관계도 존재하게 된다. 마지막으로 "8일 기준"의 경우 물품 C와 D 간에는 (C → D)의 순차 관계, (C, D)의 연관관계뿐 아니라 (D → C)의 순차 관계도 존재하게 된다. 이는 구매의도와 구매실현 간 역전현상이 발생하기에 매우 짧은 시간 간격을 어떻게 정의하느냐에 따라, 순차분석에서 추가로 고려해야 할 순차 관계의 대상이 달라짐을 보여준다.

본 연구는 이러한 현상에 대한 보다 엄밀한 고찰을 목적으로 한다. 구체적으로 본 연구의 방향은 다음의 두 가지로 요약된다. 먼저 구매의도와 구매실현 간 역전현상을 고려함으로써, 순차분석에서 도출된 규칙의 정확도가 향상되는지를 살펴보고자 한다. 또한, 역전현상이 발생하기

구매일	1월 1일	1월 4일	1월 7일	1월 10일	1월 13일
1일 기준	A	B	C	D	E
5일 기준	A	A, B	B, C	C, D	D, E
8일 기준	A	A, B	A, B, C	B, C, D	C, D, E

<그림 2> 구매 간격 기준 변화에 따른 순차 관계의 변화

에 매우 짧은 시간 간격을 다양하게 정의하고, 어떤 정의 하에서 규칙의 정확도가 가장 우수하게 나타나는지 측정함으로써 구매 행위에서 "동시"로 간주하기에 적합한 시간의 간격을 가늠하고자 한다. 요약하면 "동시"로 간주하기에 적합한 시간 간격을 발견하고 이 기준을 분석에 적용함으로써, 순차분석 규칙의 정확도를 향상하는 것이 본 연구의 목적이라고 할 수 있다.

3.3. 장바구니 확장을 통한 순차분석의 확장

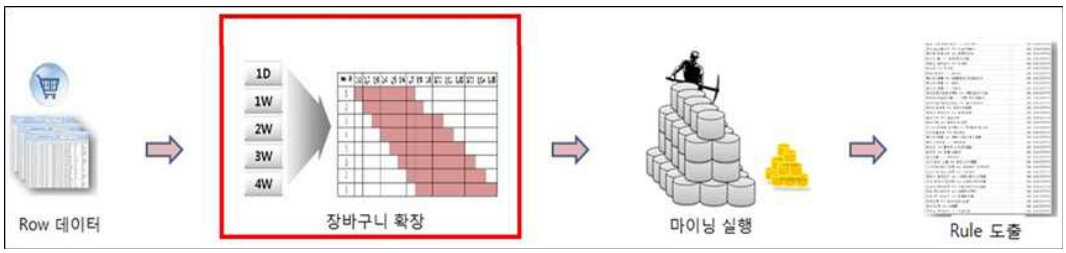
매우 짧은 시간 간격을 두고 발생한 구매실현에 그 반대순서의 잠재 구매를 가상으로 발생시키는 과정은 <그림 2>에서 간략히 소개한 바 있다. 예를 들어 매우 짧은 시간 간격을 a라고 정의한다면, T 시점의 구매를 의미하는 장바구니 Basket(T)을 (T-a+1) 시점에서 T 시점 사이에 구매된 모든 물품을 포함하는 것으로 정의하는 것이다(단 a >= 1). 예를 들어 <그림 2>에서 a를 5일로 정의한 경우 Basket(1월 7일)은 1월 3일부터 1월 7일 사이에 구매된 모든 물품을 포함하는 것으로 정의된다. 이와 같은 데이터의 가공 과정을 장바구니 확장(김미성 외, 2012)이라고 정의하며, 그 과정은 <그림 3>에 보다 자세히 나타나 있다.

<그림 3>은 슬라이딩 윈도우 기법을 활용하여 잠재 순차 관계를 생성하는 과정(김미성 외, 2012)을 보여준다. 예를 들어 <그림 3>에서 1번 장바구니는 1월 1일 ~ 1월 7일 사이에 실현된 구매를, 2번 장바구니는 1월 2일 ~ 1월 8일 사이

주문 시작날자	주문 종료날자	New_ID
1월 1일	1월 7일	1
1월 2일	1월 8일	2
1월 3일	1월 9일	3
1월 4일	1월 10일	4
1월 5일	1월 11일	5
1월 6일	1월 12일	6
1월 7일	1월 13일	7
1월 8일	1월 14일	8
1월 9일	1월 15일	9
1월 10일	1월 16일	10

New_ID	1/1	1/2	1/3	1/4	1/5	1/6	1/7	1/8	1/9	1/10	1/11	1/12	1/13	1/14	1/15
1	■	■	■	■	■	■	■								
2		■	■	■	■	■	■	■							
3			■	■	■	■	■	■	■	■					
4				■	■	■	■	■	■	■	■				
5					■	■	■	■	■	■	■	■			
6						■	■	■	■	■	■	■	■		
7							■	■	■	■	■	■	■	■	
8								■	■	■	■	■	■	■	■

<그림 3> 슬라이딩 윈도우 기법을 활용한 잠재 순차 관계 생성 (1주일 단위)



<그림 4> 확장된 순차분석 모형

에 실현된 구매를 포함하는 장바구니로 정의된다. 이와 달리 1번 장바구니를 1월 1일 ~ 1월 7일 사이에 실현된 구매로, 2번 장바구니를 1월 8일 ~ 1월 14일 사이에 실현된 구매로 서로 겹치는 구간 없이 정의하는 경우를 고려해보자. 즉 한 시점의 구매가 하나의 장바구니에만 포함되게끔 장바구니를 구성하게 되면 1월 7일의 구매와 1월 8일의 구매는 단 하루 간격으로 실현된 구매임에도 서로 다른 장바구니에 속하게 되는 이상 현상을 발생시키게 된다. 따라서 본 연구에서는 이러한 이상 현상의 발생을 막기 위해 슬라이딩 윈도우 기법을 사용하여 장바구니를 확장하였다. 이러한 장바구니 확장 과정을 포함한 제안 방법론의 전체 과정은 <그림 4>와 같이 간략하게 도식화된다.

Ⅳ. 실험 및 결과분석

본 연구는 관련 연구에서 소개한 김미성 외(2012)의 후속 연구의 성격을 가지며, 연구 수행 과정에서 김미성 외(2012)에서 제안된 장바구니 확장 개념을 사용한다. 앞서 설명한 바와 같이, 본 연구의 핵심 목표 중 하나는 구매의도 생성과 구매실현 순서 간의 역전 현상이 발생하기에 충분할 정도의 짧은 구매간격을 식별하는 것이다. 짧은 구매간격은 선행연구에서 다룬 동시성 기준과 직결되므로, 선행연구의 성과를 충분히 활용할 수 있을 것으로 판단된다. 하지만 본 연구는 궁극적으로 순차분석의 정확성을 향상시키는 것을 목적으로 하므로, 연관분석을 수행한 선행연구와 달리 순차분석 기반의 실험이 이루어져야 한다. 또한 분석 기법이 달라짐에 따라 정확성 평가를 위한 기준도 새로 마련되어야 한다. 본 실험 절에서는 동시성 기준의 확장을 통한 순차분석을 수행하고, 각 기준의 정확성 비교 평가를 위한 기준을 마련하며, 그 실험 결과를 제시하고자 한다.

4.1. 실험 개요

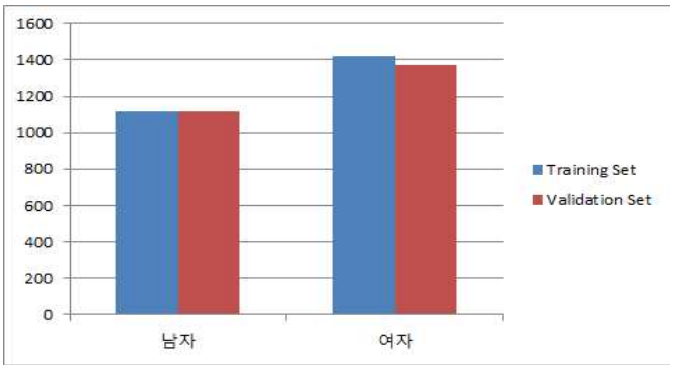
본 장에서는 본문에서 제시한 확장된 순차분석의 당위성을 평가하기 위한 실험을 수행하고 그 결과를 분석하고자 한다. 실험은 국내 한 대형 온라인 쇼핑몰의 최근 2년 거래 데이터를 사용하여 수행되었다. 전체 24,615명의 고객 중 약 10%를 추출하여 학습 데이터(TS: Training Set)로 사용하고, 또 다른 10% 고객을 추출하여 검증 데이터(VS: Validation Set)로 사용하였다. 추출된 데이터에 대해 품목별 비율을 살펴본 초기 분석 결과, 티셔츠의 주문 비율이 전체 주문의 4.75%를 차지하여, 다른 품목에 비해 매우 높게 나타났다. 이는 해당 쇼핑몰의 상품 구성 및 고객층의 특성에 기인한 것으로 판단된다. 즉 해당 쇼핑몰에서의 티셔츠의 구매는 일반적인 오프라인 마트에서의 비닐봉지에 해당할 정도로 거의 모든 거래에 포함되어 있기 때문에, 티셔츠로 인해 의미있는 다른 규칙이 과소평가되는 현상을 막기 위해 티셔츠 항목은 분석에서 제외하였다. 또한 표본추출 과정에서는 각 거래가 아닌 각 고객에 대한 임의추출을 수행한 뒤 해당 고객이 발생시킨 모든 주문을 표본에 포함시킴으로써, 각 고객이 발생시킨 주문의 일부만 표본에 포함되는 부작용을 미연에 방지하였다. 그 결과 TS에서는 87,106건의 주문이, VS에서는 83,063건의 주문이 분석에 사용되었다. 분석

데이터에 포함된 고객 수, 주문 건수, 그리고 각 고객당 평균 주문 수에 대한 내용이 <그림 5>에 요약되어 있다. 또한, TS와 VS를 구성하고 있는 고객의 성별 및 연령에 대한 분포는 <그림 6>과 <그림 7>에 요약되어있다. 그리고 TS와 VS의 주문에 포함된 물품의 분포는 <그림 8>에 나타나 있다.

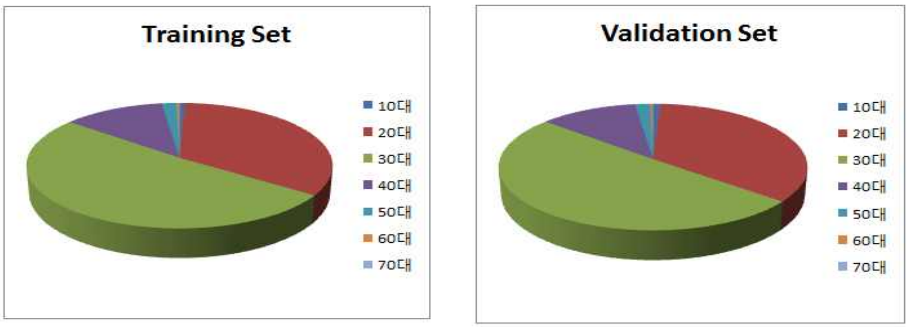
TS	고객수	주문수	고객당 평균 주문수	확장된 구매영향 지속성					
				TS	1D	1W	2W	3W	4W
	2,539	87,106	34	87,106	87,106	146,664	210,115	270,780	329,192

VS	고객수	주문수	고객당 평균 주문수	확장된 구매영향 지속성					
				TS	1D	1W	2W	3W	4W
	2,492	83,063	33	83,063	83,063	132,383	185,861	236,562	285,806

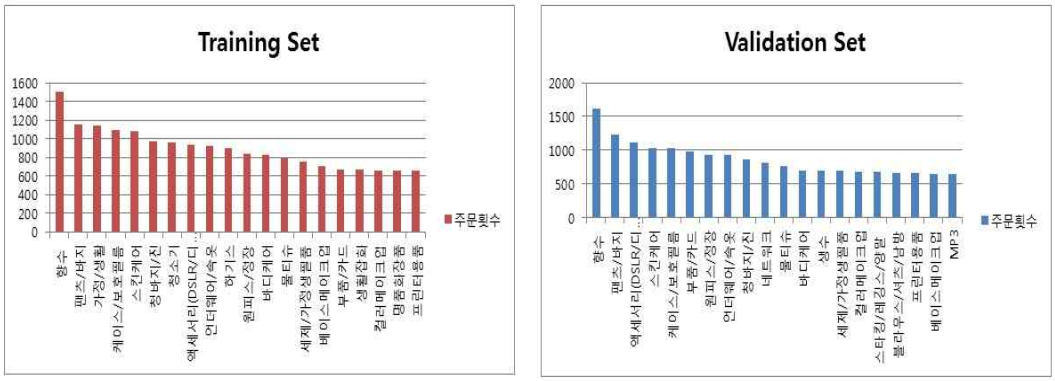
<그림 5> 실험 데이터 특성 요약



<그림 6> TS와 VS의 성별 분포

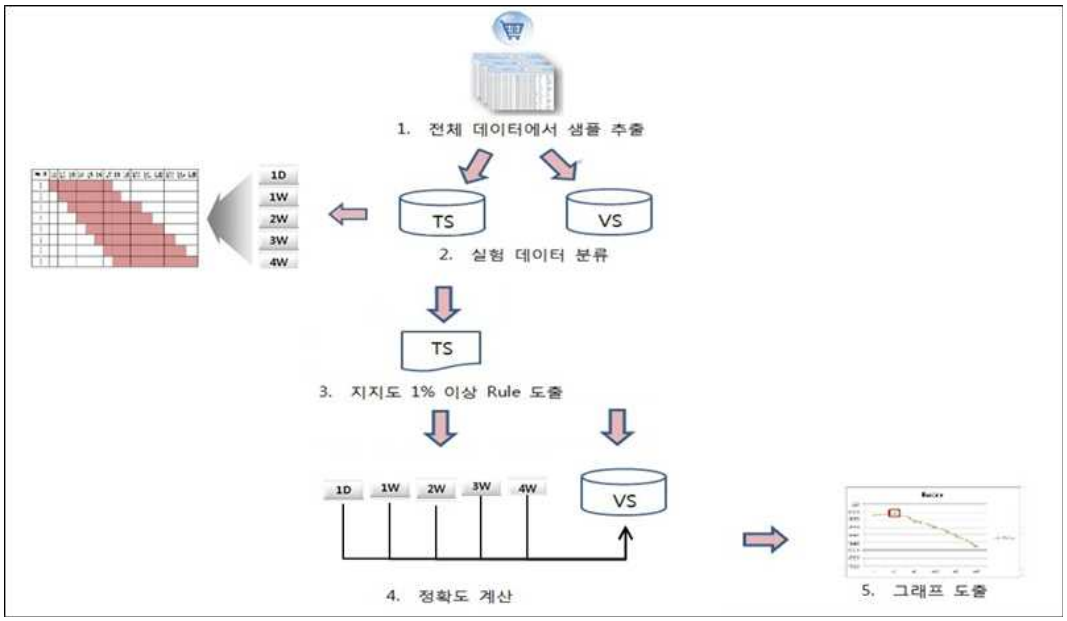


<그림 7> TS와 VS의 연령 분포



<그림 8> TS와 VS의 주문 물품 분포

이상 소개한 데이터를 사용하여 제안하는 방법론의 정확도를 평가하기 위한 실험의 전체 과정은 <그림 9>에 요약되어 있다. 전체 데이터에서 표본을 추출하고 이를 TS와 VS로 나누는 과정이 1단계와 2단계에 소개되어 있다. TS의 초기 데이터는 1D로 명명하였는데, 이는 초기 데이터에서 각 거래의 최소 단위가 하루로 정의되어 있음을 의미한다. 1W, 2W, 3W, 그리고 4W는 초기 데이터에 대해 1주일, 2주일, 3주일, 그리고 4주일의 시간 간격을 적용하여 장바구니를 확장한 결과로 나타난 데이터 집합을 의미한다. 이렇게 준비된 5개의 데이터 집합에 대해 각각 순차분석을 수행하며, 그 결과로 도출된 상위 규칙만을 추려서 VS에서의 정확도를 비교 분석한



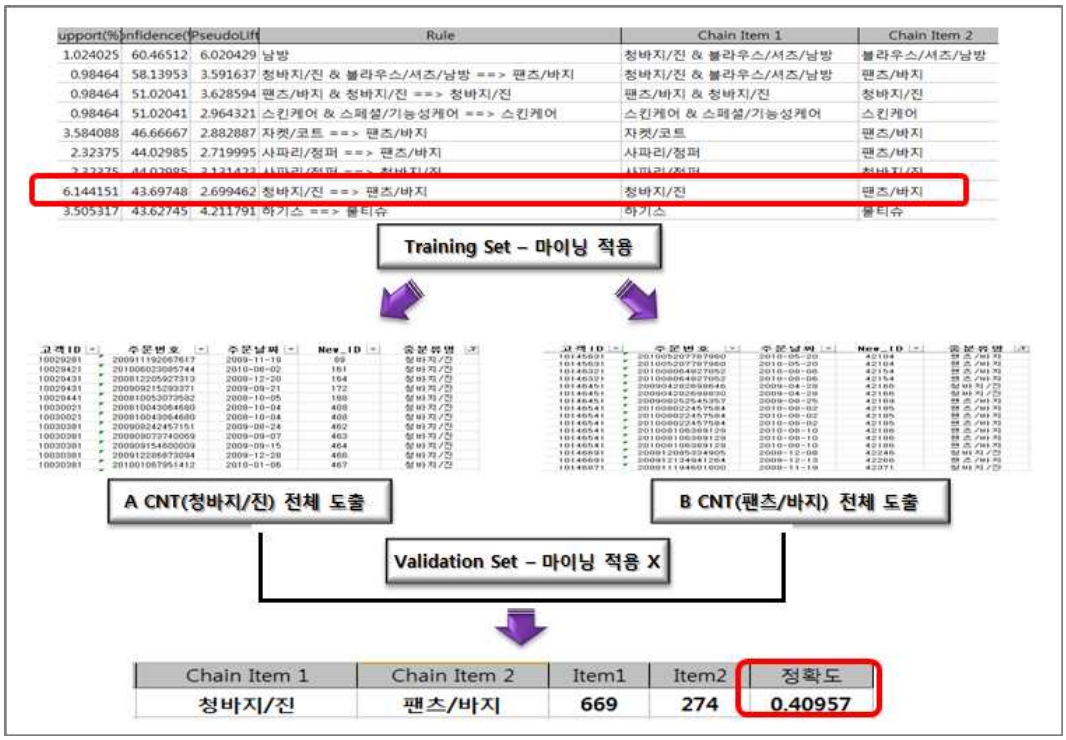
<그림 9> 정확도 평가 실험 모형

다. 즉, 지지도(Support) 1%이상의 규칙 중 신뢰도(Confidence) 기준 상위 500개, 1,000개, 1,500개의 규칙을 추출하여 평균 정확도를 비교하였다. 순차분석의 전 과정은 SAS Enterprise Miner Workstation 7.1을 사용하여 수행하였다.

<그림 9>의 마지막 단계는 1일 ~ 4주의 각 시간 간격에 대해 확장된 TS로부터 순차규칙을 도출하고, 그 규칙이 VS에서 나타내는 정확도를 분석하는 과정을 의미한다. 예를 들어 TS에서 (A → B)라는 순차규칙이 도출되었다고 가정하자. VS에서 A를 포함한 주문의 집합을 OrderA로, 그리고 OrderA의 개수를 Count(A)로 정의하였다. OrderA가 발생한 이후 해당 주문을 발생시킨 고객이 B를 한 번이라도 구매하였다면 해당 규칙이 적중한 것이며(Hit 수 증가), 해당 주문을 발생시킨 고객이 해당 주문 이후에 B를 단 한 번도 구매하지 않았다면 해당 규칙이 실패한 것이다(Fail 수 증가). 이때, 규칙 (A → B)의 정확도는 Hit / (Hit + Fail)로 계산된다. 1D, 1W, 2W, 3W, 그리고 4W 데이터 집합에서 도출된 규칙의 평균 정확도는, 이처럼 계산된 모든 규칙의 정확도의 평균값으로 정의된다. 이러한 방식으로 정확도를 계산하는 예가 <그림 10>에 나타나 있다.

4.2. 실험 결과 및 해석

본 절에서는 앞에서 소개한 실험 과정을 통해 도출된 결과를 소개하고 이를 분석하고자 한다. 우선 TS의 1D ~ 4W에 대해 최소 지지도를 만족

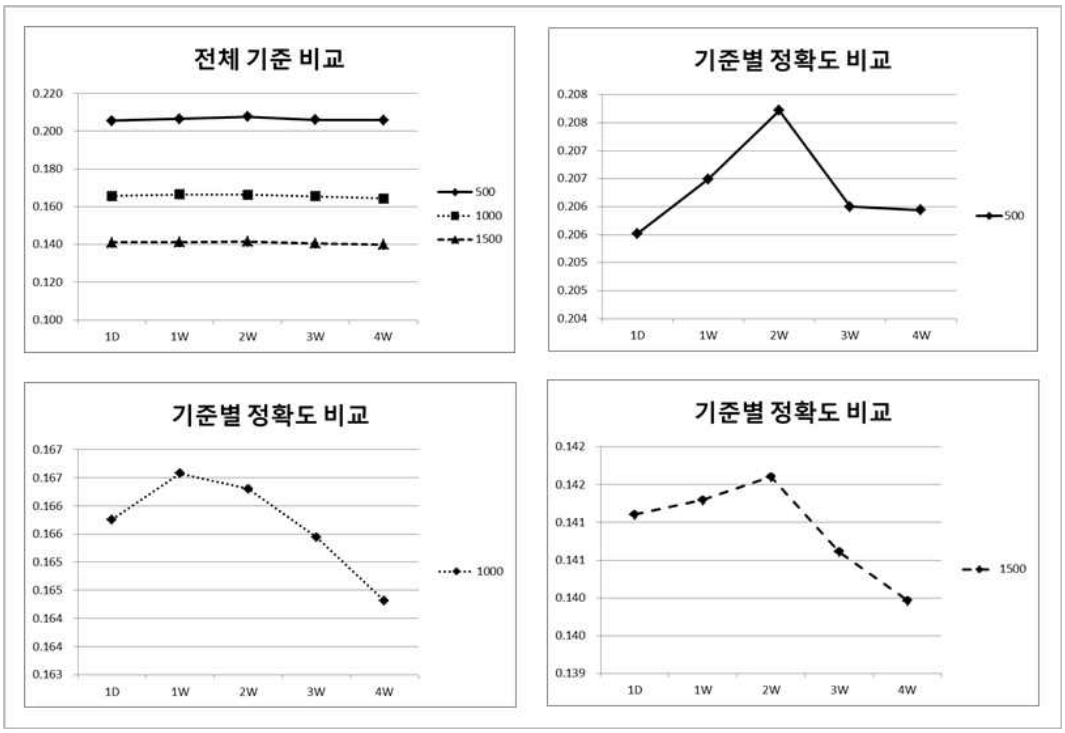


<그림 10> 각 규칙의 정확도 계산 예

	상위 규칙 수		
동시 구매 기준	500	1000	1500
1D	0.2055	0.1658	0.1411
1W	0.2065	0.1666	0.1413
2W	0.2077	0.1663	0.1416
3W	0.2060	0.1654	0.1406
4W	0.2059	0.1643	0.1400

<그림 11> 상위 규칙들의 정확도 평균 비교(표)

하는 규칙 중 신뢰도가 높은 상위 규칙들을 도출하였으며, 이들 규칙이 VS의 데이터에 대해 갖는 평균 정확도를 <그림 11>에 요약하였다. 또한 <그림 12>는 이 결과를 도식화한 그래프이다. <그림 12>에서 좌측 상단의 그래프는 신뢰도 기준 상위 500개, 1,000개, 1,500개의 규칙에 대한 VS에서의 평균 정확도를 동시에 보여주고 있다. 전체적인 정확도의 추이는 500개 > 1,000개 > 1,500개의 순으로 높게 나타나며, 이는 상위 규칙만을 적용했을 때 평균 정확도가 높게 나타난다는 측면에서 당연한 결과라고 할 수 있다. <그림 12>의 나머지 세 그래프는, 좌측 상단



<그림 12> 상위 규칙들의 정확도 평균 비교(그래프)

그래프에 나타난 500개, 1,000개, 1,500개의 규칙에 대한 정확도를 상세하게 나타낸 것이다. 각 그림에서 1D는 장바구니 확장을 적용하지 않은, 즉 초기 데이터를 그대로 사용하여 전통적인 순차분석을 수행했을 때의 평균 정확도를 보여주며, 본 실험에서 대조군으로 사용된다. 한편 1W ~ 4W는 각각 1주 ~ 4주의 시간 간격을 기준으로 장바구니를 확장한 경우의 평균 정확도를 보여준다. 모든 그래프에서 1주에서 2주 사이의 평균 정확도가 가장 높게 나타남을 볼 수 있는데, 이는 구매의도 생성 순서와 구매실현 순서의 역전 가능성을 감안하여 확장된 순차분석을 수행하였을 때 전통적인 순차분석에 비해 더 높은 정확도를 갖는 규칙을 도출할 수 있음을 의미한다. 하지만 3주 ~ 4주의 기준을 적용한 3W, 4W의 경우 평균 정확도가 점차 하락하는 현상을 보였다. 이는 1주에서 2주 사이에서는 상대적으로 짧은 시간 간격으로 간주되기에 충분하여 역전현상이 발생할 가능성이 높았지만, 3주 이상의 시간 간격을 두고 실현된 구매는 역전현상이 발생할 가능성이 상대적으로 낮기 때문에 나타난 현상으로 파악된다.

<그림 12>의 결과를 통해 얻을 수 있는 결론은 다음과 같다. 우선 신뢰도를 기준으로 상위 500개, 1,000개, 1,500개의 규칙을 도출하였을 때, 이들이 VS에서 갖는 정확도 평균은 상위 500개 > 1,000개 > 1,500개 순으로 나타났다. 이는 높은 신뢰도를 갖는 규칙이 VS에서도 높은 정확도를 나타내는 것으로 해석될 수 있다. 또한, 대부분의 실험에서 전통적인 순차분석(1D)에 비해 1주에서 2주 사이를 기준으로 확장된 순차분석(1W~2W)을 통해 도출된 규칙의 평균 정확도가 높게 나타났으며, 평균 정확도는 3W, 4W로 기준 기간이 늘어날수록 점점 낮아지는 현상을 보였다. 즉, 위 실험을 통해 구매의도 생성 순서와 구매실현 순서의 역전현상이 발생하는 최대 시간 간격은 1주에서 2주 사이임을 알 수 있으며, 1주~2주를 기준으로 한 장바구니 확장을 통해 구매 내력 원 데이터를 가공함으로써 더 높은 정확도를 갖는 순차 관계 규칙을 도출할 수 있음을 알 수 있다.

Ⅴ. 결론

최근 다양한 마이닝 분석을 통해 의미 있는 지식을 발견하기 위한 수요가 꾸준히 증가하고 있지만, 이와 반대로 데이터 마이닝 분석의 결과가 비즈니스 문제 해결의 성과로 항상 직결되지는 않는다는 비판도 증가하고 있다. 본 연구에서는 이러한 한계의 주요 원인 중 하나로 전통적인 순차분석에서 구매 간 선후관계가 지나치게 엄격하게 반영되는 현상을 지적하고, 이를 완화함으로써 도출되는 규칙의 정확도를 향상하게 시킬 수 있음을 보였다. 즉 매우 짧은 시간 간격을 두고 실현된 구매 간에는 구매의도 생성과 구매실현 간 순서 역전 현상이 발생할 수 있으므로, 실제로 실현된 구매 순서뿐 아니라 실현된 구매의 역순으로 구성된 잠재 구매를 가상으로 발생시키는 방법으로 확장된 순차분석을 수행하는 방법론을 제안하였다.

구매의도 생성과 구매실현 간 순서 역전 현상을 보이기 위해 국내 한 대형 온라인 쇼핑몰의 실제 구매 데이터에 대한 실험을 수행하였으며, 실험 결과 이러한 현상이 실제로 발생할 뿐 아니라 확장된 순차분석을 통해 더욱 정확도가 높은

순차관계 규칙을 발굴할 수 있음을 보였다. 분석 대상이 된 온라인 쇼핑몰의 경우 이러한 역전 현상이 발생할 수 있는 최대 시간 간격은 1주에서 2주 사이로 파악되었으며, 이를 통해 1주~2주의 시간 기준에 대해 장바구니 확장 기법을 적용함으로써 더욱 높은 정확도를 갖는 순차 관계 규칙을 도출할 수 있음을 파악하였다.

하지만 본 연구는 다음의 측면에서 앞으로 보완이 이루어져야 할 것으로 판단된다. 우선 제안하는 방법론의 성능을 평가하기 위한 실험이 하나의 온라인 쇼핑몰에 대해서만 수행되었다는 한계를 가진다. 구매의도 생성과 구매 실현 순서의 역전 현상은 분석 대상이 되는 쇼핑몰의 규모, 접근성 등의 특성에 따라 다르게 나타날 것으로 예상한다. 따라서 더욱 엄밀한 평가를 위해서는 둘 이상의 쇼핑몰, 그리고 오프라인 쇼핑몰에 대해서도 유사한 실험을 수행하고 그 결과를 분석할 필요가 있을 것으로 판단된다. 또한 추후 실험에서는 1일, 1주, 2주, 3주, 4주 단위가 아닌 보다 세분화된 시간 간격 기준에 대해 분석을 수행함으로써, 시간 간격의 증가에 따른 정확도 변화 추이의 일관성을 파악하기 위한 과정이 포함되어야 할 것이다.

참고문헌

김미성, 김남규, 안재현, “연관규칙 마이닝에서의 동시성 기준 확장에 대한 연구,” 지능정보연구, 제18권, 제1호, 2012, pp. 23-38.

김재경, 안도현, 조윤호, “개인별 상품추천시스템, WebCF-PT: 웹 마이닝과 상품계층도를 이용한 협업필터링," 경영정보학연구, 제15권, 제1호, 2005, pp. 63-79.

송만석, 박종환, 김삼원, 조윤재, “프로야구구단의 효율적인 CRM을 위한 데이터 마이닝 기법의 적용,” 한국스포츠산업경영학회지, 제13권, 제2호, 2008, pp. 205-222.

안현철, 한인구, 김경재, “연관규칙기법과 분류모형을 결합한 상품추천시스템 : G인터넷 쇼핑몰의 사례,“ Information Systems Review, 제8권, 제1호, 2006, pp. 181-201.

유은지, 김정철, 이춘열, 김남규, “시맨틱 텍스트마이닝을 위한 온톨로지 활용 방안,” 정보시스템연구, 제21권, 제3호, 2012, pp. 137-161.

이영재, 이성수, “텍스트마이닝 기반의 인적재난사고사례 신뢰도 측정연구,” 정보시스템연구, 제20권, 제3호, 2011, pp. 63-79.

이현규, 박영식, “고객가치 극대화를 위한 전자상거래 구매의사결정 요인에 관한 연구,” 정보시스템연구, 제15권, 제1호, 2006, pp. 121-144.

정영수, 강경화, “데이터마이닝 기법을 이용한 인터넷 쇼핑몰 사이트의 CRM 사례분석," 경영경제연구, 제27권, 제1호, 2004, pp. 139-156.

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

하성호, 이재신, “데이터 마이닝을 활용한 동적

인 고객분석에 따른 고객관계관리 기법," 한국지능정보시스템학회논문지, 제9권, 제3호, 2003, pp. 23-47.

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

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

Agrawal, R. and Srikant, R., "Mining Sequential Patterns," in Proceedings of the 11th International Conference on Data Engineering, 1995, pp. 3-14.

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

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

O'Reilly Radar Team, "Big Data Now: Current Perspectives from O'Reilly Radar," O'Reilly Media, 2011.

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

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

김민석(Kim, Minseok)

한국정보통신기능대학에서 이동통신공학을 전공하였고, 평생교육진흥원에서 정보통신공학을 전공하였다. 국민대학교 비즈니스IT전문대학원에서 비즈니스IT전공 석사학위를 취득하였으며, 현재 한국특허정보원 정보통신팀에서 선행기술조사원으로 재직 중이다. 주요 관심분야는 데이터 마이닝, 빅 데이터 분석 등이다.

김남규(Kim, Namgyu)

현재 국민대학교 경영정보학부에서 부교수로 재직 중이다. 서울대학교 컴퓨터공학과에서 학사 학위를 취득하고, KAIST 테크노경영대학원에서 Database와 MIS를 전공하여 경영공학 석사 및 박사학위를 취득하였다. 한국정보기술응용학회 부회장, 한국경영정보학회 이사, 한국지능정보시스템학회 이사, 한국CRM학

회 이사, 한국인터넷정보학회 편집위원, JITAM 편집위원을 역임하였으며, 한국경영정보학회, 한국지능정보시스템학회, 한국정보시스템학회 종신회원 및 한국생산성본부 자문위원으로 활동 중이다. 주요 관심분야는 시맨틱 데이터 관리, 데이터베이스 설계 및 데이터 마이닝 등이다.

<Abstract>

An Investigation on Expanding Traditional Sequential Analysis Method by Considering the Reversion of Purchase Realization Order

Kim, Minseok · Kim, Namgyu

Recently various kinds of Information Technology services are created and the quantities of the data flow are increase rapidly. Not only that, but the data patterns that we deal with also slowly becoming diversity. As a result, the demand of discover the meaningful knowledge/information through the various mining analysis such as linkage analysis, sequencing analysis, classification and prediction, has been steadily increasing. However, solving the business problems using data mining analysis does not always concerning, one of the major causes of these limitations is there are some analyzed data can't accurately reflect the real world phenomenon. For example, although the time gap of purchasing the two products is very short, by using the traditional sequencing analysis, the precedence relationship of the two products is clearly reflected. But in the real world, with the very short time interval, the precedence relationship of the two purchases might not be defined. What was worse, the sequence of the purchase intention and the sequence of the purchase realization of the two products might be mutually be reversed.

Therefore, in this study, an expanded sequencing analysis methodology has been proposed in order to reflect this situation. In this proposed methodology, the purchases that being made in a very short time interval among the purchase order which might not important will be notice, and the analysis which included the original sequence and reversed sequence will be used to extend the analysis of the data. Also, to some extent a very short time interval can be defined as the time interval, so an experiment were carried out to determine the varying based on the time interval for the actual data.

Keywords: Association Rule Mining, CRM, Data Mining, Sequential Pattern Analysis

* * 이 논문은 2013년 7월 29일 접수하여 1차 수정을 거쳐 2013년 9월 9일 게재 확정되었습니다.