한국빅데이터학회지
제7권 제1호, 2022, pp. 213-224 https://doi.org/10.36498/kbigdt.2022.7.1.213

순차적 레이어 필터링을 이용한 상품 판매 연관도 분석*

Association Analysis of Product Sales using Sequential Layer Filtering

방선호 · 이강현 · 장지영 · Tsatsral Telmentugs · 신광섭†

인천대학교 동북아물류대학원

요 약

물류와 유통에서 장바구니 분석(MBA: Market Basket Analysis)은 주요 판매 상품 간의 연관성을 분석하고, 내부 운영 효율성을 높이기 위한 중요한 수단으로 활용된다. 특히, 장바구니 분석의 결과는 상품 구매예측, 상품 추천 및 매장의 상품 전시 구조 등 의사결정 과정에 중요한 참고자료로 활용된다. 최근 전자상거래의 발전으로 하나의 유통 및 물류 기업이 취급하는 품목의 수가 급격하게 증가하면서 기존의 분석기법인 Apriori와 FP-Grwoth 등의 방법은 계산량의 기하급수적 증가로 인한 속도저하와 실제 비즈니스에 적용하기 위한 중요한 연관규칙을 살피기에는 한계가 있다. 본 연구에서는 이러한 한계를 극복하기 위해, 상품의 최상위 분류체계인 Main-Category 수준에서는 상품의 판매량을 함께 고려할 수 있는 utility item set mining 기법을 활용하여 주로 함께 판매된 상품군을 우선 선별하였다. 그 후, sub-category 수준에서는 FP-Growth를 활용하여 함께 판매되는 상품 유형을 식별하였다. 이렇게 순차적 레이어 필터링 기법을 활용하여 불필요한 연산을 줄일 수 있어 현실적으로 활용가능한 결과를 제시할 수 있다.

■ 중심어 : 장바구니분석, 높은 유틸리티 항목집합 마이닝, 유통물류

Abstract

In logistics and distribution, Market Basket Analysis (MBA) is used as an important means to analyze the correlation between major sales products and to increase internal operational efficiency. In particular, the results of market basket analysis are used as important reference data for decision-making processes such as product purchase prediction, product recommendation, and product display structure in stores. With the recent development of e-commerce, the number of items handled by a single distribution and logistics company has rapidly increased, And the existing analytical methods such as Apriori and FP-Growth have slowed down due to the exponential increase in the amount of calculation and applied to actual business. There is a limit to examining important association rules to overcome this limitation, In this study, at the Main-Category level, which is the highest classification system of products, the utility item set mining technique that can consider the sales volume of products together was used to first select a group of products mainly sold together. Then, at the sub-category level, the types of products sold together were identified using FP-Growth. By using this sequential layer filtering technique, it may be possible to reduce the unnecessary calculations and to find practically usable rules for enhancing the effectiveness and profitability.

■ Keyword : Market Basket Analysis, High Utility Itemset Mining, Distribution and Logistics

2022년 05월 16일 접수; 2022년 05월 31일 수정본 접수; 2022년 06월 04일 게재 확정.

* 이 논문은 2022년도 정부(산업통상자원부)의 재원으로 한국산업기술진흥원의 지원을 받아 수행된 연구임 (P0008691, 2022년 산업혁신인재성장지원사업)

† 교신저자 (ksshin@inu.ac.kr)

Ⅰ. 서 론

물류와 유통에서 장바구니 분석(MBA: Market Basket Analysis)은 주요 판매 상품 간 연관성을 파악하고, 내부 운영 효율성을 높이기 위한 중요한 수단으로 활용된다. 특히, 장바구니 분석의 결과는 상품 구매예측, 상품 추천 및 매대 진열, 판촉 상품 구성 등과 같은 의사결정 과정에 참고자료로 활용된다.

전통적으로 Apriori와 FP-Growth 알고리즘을 이용한 장바구니 분석은 소비자 유형과 계절 별, 카테고리 별 등의 소비자 패턴을 분석하여 재고관리와 상점 내 상품 배치, 물류센터 내 제품 적재 배치 결정 등의 효율성 향상을 위해 개발된 분석기법이다. 이 기법을 사용하여 판매자는 소비자의 요구의 대한 이해와 소비 패턴을 파악할 수 있다. 이는 판매자가 기존 제품의 보유 수량을 높이거나 새로운 소비자 확보를 위한 제품들의 정렬에 도움을 준다 (S. S. Khedkar & S Kumari, 2021).

최근 전자상거래의 발전으로 하나의 유통 물류기업이 취급하는 상품의 수가 급격하게 증가하게 되었다. 이러한 상황에서 기존의 알고리즘은 상품 수가 많을수록 계산량이 기하급수적으로 증가하며 현실적으로 활용 가능한 범위의 연관규칙을 발견하기 어렵게 되었다. 이는 Tree 구조를 도입하여 기존 대비 계산속도 증가와 데이터베이스 스캔 횟수를 감소시키는 방법을 적용한 FP-Growth에서도 동일한 문제가 발생한다. (황정희, 2020)

이에 본 연구에서는 연관상품들이 묶여있는 카테고리를 하나의 레이어로 설정하고 각 항목의 가중치와 발생빈도를 고려할 수 있는 높은 유틸리티 항목집합 마이닝(High Utility Itemset Mining) 기법을 사용하여 일정 임계치 이상의 유의미한 값들을 추출한다. 이후 추출된 항목들의 하부 레이어에서 FP-Growth를 이용한 연관도 분석하여 기존 방식과는 다르게 유틸리티를 이용한 항목 필터링 방법을 제시한다.

몽골은 최근 글로벌 전자상거래와 국가 간 무역이 활발해지며 수출입 물동량이 꾸준한 상승세를 보이고 있다. 또한 몽골 내 온라인 커머스 분야의 경쟁은 점차 치열해지고 화장품 시장은 2016년도부터 지속적으로 성장하고 있다. (Nandintsatsral Amarsanaa, 2020) 따라서, 본 연구에서 제안하는 기법을 활용하여 전체 매장에서 주로 판매되는 상품들 사이의 연관성을 파악하고, 향후 판매량을 예측하기 위한 모델을 개발하거나 오프라인 유통 채널의 개선을 위한 근거를 확보하고자 한다. 분석에 사용된 데이터는 몽골 내 유통 기업이 2014년부터 2021년까지 오프라인 매장을 통해 판매하는 미용용품 및 개인용품의 월별 판매 수량과 판매총액으로 구성되어 있다.

본 논문의 나머지는 다음과 같이 구성된다. II 장에서는 본 연구와 관련된 기존 연구를 분석하고, 한계점을 설명하였다. III 장에서는 본 연구에서 제안하는 분석기법에 대해 간략하게 설명하였으며, 분석 결과는 IV장에서 정리하였다. 마지막으로 V장에서는 본 연구가 가지는 의의와 한계점을 설명한다.

Ⅱ. 관련 연구

정병수 외 3인 (2009)은 기존 유틸리티 패턴마이닝 중 빈발 패턴(Frequent Pattern) 마이닝에서 Apriori 규칙을 적용 시 성능이 현저히 저하되었던 문제점을 Prefix-tree를 사용하여 개선하였다. 기존 Apriori와 FP-Growth 기법은 모든 항목에 대해 동일한 중요도(Weight)를 가지고 하나의 트랜잭션에서 각 항목이 이진수 형태로 나타나는 문제점을 가지고 있다. 실생활에서 제품들은 각자의 중요도와 가격이 모두 다르기에 총 수익에 대한 중요 지분을 가진 요소를 찾기에는 어려움이 있었다. 이에 Prefix-tree 구조를 기반으로 각 노드

에 TWU 및 빈도수를 저장하여 높은 유틸리티 패턴을 빠르게 찾아낼 수 있었다.

Maria E. Garcia-Diaz dt al (2021) 는 Orange Canvas Tool과 FP-Growth 알고리즘을 사용하여 가전제품, 컴퓨터용품, 가구, 스포츠용품 등 생활용품 판매회사 위주의 장바구니 분석을 실행하였다. Orange Canvas Tool을 사용하여 거래 데이터베이스에서 연관규칙분석 기법을 적용하면 소매업체가 매출을 증가시킬 수 있는 제품을 식별할 수 있고, 특정 상품을 판촉가격으로 홍보하거나 묶음 판매 상품 등을 배치하는 등 마케팅 전략 수립에 도움을 줄 수 있다고 설명하였다.

김진형 외 1인 (2019)은 Apriori 알고리즘과 FP-Growth 알고리즘의 효율성을 분석하였다. Apriori 알고리즘은 후보 집합 생성 시에 아이템의 개수가 많아지면 계산 복잡도가 증가하게 된다. 또한 패턴을 찾기 위해서 DB를 스캔하는 횟수가 최대로 가장 긴 트래잭션의 아이템의 수만큼 발생할 수 있다. 해당 논문에선 트리와 연결리스트 자료구조를 사용하여 데이터 베이스 스캔 횟수를 줄이고 후보 집합을 생성하지 않는다는 점에서 FP-Growth 알고리즘은 Apriori 알고리즘의 단점을 효율적으로 개선된 알고리즘이라고 말하고 있다.

Chun-Wei Lin 외 2인 (2011)은 Apriori 알고리즘과 FP-Growth 알고리즘의 문제를 개선하기 위해 HUP-Tree (high Utility Pattern tree)를 제안했다. Apriori와 FP-Growth의 접근 방식은 데이터베이스의 모든 항목을 이진변수로 취급하여 거래에서 항목을 구매했는지 여부만 고려한다. 이 경우 빈번한 항목 집합은 거래에서 항목 집합의 발생 중요도를 나타낼 뿐 가격이나 이익과 같은 다른 암묵적 요소를 반영하지 않는다. 또한 빈도만으로는 수익성이 높은 항목을 식별하기에는 충분하지 않다. 그리하여 FP-Tree 구조와는 유사하지만 데이터베이스 검색시간을 줄여주고 제안된 알고리즘의 성능이 2상 알고리즘보다 더 빠르게 실행되며 특정 항목을 포함하는 가능한 항목 집합이 동시에 생성된다고 말하고 있다.

Jyothi Pillai (2011)은 고전적인 연관규칙 마이닝 접근 방식과 관련된 문제를 개선하기 위하여 비즈니스 분석가가 편리하게 사용할 수 있도록 장바구니 분석에 대한 사용자 중심 접근 방식을 제안했다. ARM(Associate Rule Mining)의 고전적인 문제로는 희귀 아이템 항목 집합을 추출할 때 어려움을 겪는 것이다. 제안된 사용자 중심 접근 방식(user-centric approach)은 관리자들에게 항목 효용 값이 시간에 따라 동적으로 허용되는 희귀 항목 집합 효용 마이닝에 대한 체계적인 지침과 제안을 제공하며 슈퍼마켓과 온라인 스트림 마이닝의 소매 마케팅을 위한 효과적인 계획을 이끌 수 있는 통찰력을 생성한다고 말하고 있다.

Le wang 외 3인 (2020)은 고효율 패턴 마이닝 알고리즘의 최적화 접근 방식에 초점을 맞추고 새롭게 구분된 비후보 항목을 반복적으로 제거하여 마이닝 프로세스의 검색공간을 줄이고 마이닝 효율성을 높이는 개선안을 제안했다. 제안된 개선안은 EFIM(Efficient high-utility Itemset Mining) 알고리즘에 적용되었으며 개선된 알고리즘이 후보 수를 효과적으로 감소시키고 시간 효율성 면에서 EFIM을 능가할 수 있음을 보여준다고 말하고 있다.

황정희 (2020)는 유틸리티-리스트 구조를 이용하여 항목의 Join count에 따른 연산 비용을 감소시키기 위해 유틸리티-리스트 구조에 Prefix 항목에 유틸리티를 포함시키고 높은 유틸리티 항목집합이 될 가능성을 검사하기 위해 비트연산을 이용하는 알고리즘을 제안했다. 제안된 알고리즘인 UL-HUM은 1-item 수가 많은 경우, 1-item 수가 많을수록 조인 연산 비용을 줄이는 데 더 큰 효과가 있었고, 이것은 같은 항목을 포함하는 트랜잭션의 검사를 위해 비트연산을 이용하는 것이 효율적임을 입증하였으며 제안된 알고리즘은 항목 종류의 수가 많은 경우, 높은 유틸리티 항목이 많

은 경우에 매우 효과적이라고 말하고 있다.

Anshul Bhargav 외 2인 (2014)은 장바구니 분석에 관한 문제점을 지적하고 이를 개선하기 위하여 인공 신경망 기술 사용을 제안하였다. 장바구니 분석에는 여러 가지 문제점이 있으며 첫 번째의 문제점으로는 고객의 니즈가 계절과 시간에 따라 계속 바뀌며 이 때문에 장바구니 분석 결과는 계절과 시간에 전적으로 의존하기 때문에 반복적으로 수행해야 한다. 두 번째로는 후보 집합과 빈번한 항목 집합을 찾기 위해 고객 트랜잭션의 전체 데이터 베이스를 반복적으로 스캔해야 하는 Apriori 알고리즘과 관련이 있다. 이러한 문제점을 개선하기 위하여 인공 신경망 기술을 사용하는 데 초점을 맞추었으며 데이터베이스의 반복 스캔에 걸리는 시간을 줄이고 알고리즘의 효율성을 높이는 단일 레이어 피드-포워드 부분 연결 신경망 기술을 제안했다.

Run-Qing Liu 외 2인 (2018)은 국내 온라인 쇼핑몰의 VIP 거래 데이터 및 데이터 마이닝 기법(K-평균 클러스터링 및 연결 규칙)을 활용하여 고객을 VIP와 Non-VIP로 세분화하고 구매 패턴을 파악하여 CRM 전략을 제안하였다. 이 연구는 시장점유율이 낮은 중소기업에 도움이 되며 SME가 VIP 고객 그룹을 올바르게 식별하여 귀중한 고객을 보유할 수 있도록 지원하며 시장점유율을 개선하고, 시장 위치를 확고히 하는 전략이라고 말하고 있다.

Hidayat 외 5인 (2019)은 2018년 11월 Breiliant Store 내 화장품 월 매출 거래 데이터를 이용하여 제품 간의 조합을 형성할 수 있는 시스템의 필요성을 느껴 Apriori 알고리즘과 FP-Growth 알고리즘을 이용하여 데이터를 분석하였다. 상점 주인의 수익을 증가시키기 위해 판매 전략과 상품 홍보의 측면에서 구매 품목의 상관관계를 얻을 수 있고 상관관계, 제품 조합의 결과를 얻는 데에 FP-Growth 알고리즘이 Apriori 알고리즘보다 더 빠르게 결과가 도출되었다. 분석 과정에서 Apriori 알고리즘은 제품 + 제품의 이름을 사용하고 높은 정확도 값의 규칙을 생성하는 반면 FP-Growth 알고리즘은 Apriori 알고리즘보다 낮은 규칙을 가진 제품 + 제품 코드를 사용한다고 말하고 있다.

Rajendra Prasad (2017)는 주로 높은 유틸리티 항목 세트의 효과적인 생성을 위한 연관 마이닝 방법에 초점을 두고 HUIM(High-Utility itemset mining)의 향상된 버전인 FHM(Fastest high-utility mining method)를 이용하여 Optimal FHM (OFHM)을 제안했다. FHM은 항목 집합 생성 시 조인 작업 횟수를 줄여 HUIM보다 빠르며 두 방법론 모두 큰 데이터 세트를 처리할 때 비용이 매우 비싸진다. 이 때문에 제안된 방법은 저수익 항목 집합을 제거하기 위해 가지치기 기반 유틸리티 동시 발생 구조(PEUCS)를 구축하여 이 문제를 해결하기 위해 최적의 높은 효율 항목 집합만 처리하므로 Optimal FHM(OFHM)으로 명칭하였으며 실험 결과 OFHM은 계산 필요량이 줄어들기 때문에 벤치마킹된 대규모 데이터 세트를 처리할 때 기존의 다른 방법론들보다 효율적이라고 말하고 있다.

Shish Kumar Dubey외 3인 (2021)은 장바구니 분석(MBA) 연관성을 찾고 소비자의 과거 구매를 기반으로 소비자에게 제품을 추천하는 데 얼마나 유용한지를 중점적으로 조사하였다. 소비자의 빈번한 구매 패턴을 알아내기 위해 Apriori 알고리즘과 제품 권장 사항을 제공하기 위한 Collaborative Filtering 알고리즘을 사용하였으며 제품 연관성과 제품 추천 간의 차이점과 유사점을 찾기 위해 두 알고리즘에 대한 비교 연구를 하였다. ARM은 거의 한 유형의 CF의 하위 집합이며 ARM과 CF의 주요 차이점은 연관규칙 마이닝에서 일반적으로 “세션”(제품이 동일한 세션에 함께 표시됨)이며 모든 사용자에 대해 계산된다는 것이다. 사용자 또는 항목 기반 CF에서 구성 단위는 사용자(동일한 사용자가 소비한 제품)이

며 모든 사용자 세션에서 함께 판매되었는지 여부에 관계없이 계산된다고 말하고 있다.

Ting외 4인 (2014)은 식품 공급망의 품질 및 안전에 대한 소비자들의 요구에 관심을 가지게 되어 QSDSS(quality sustainability decision support system)라는 새로운 시스템을 제안하고 있다. 이 시스템은 연관 규칙 마이닝과 Dempster의 조합 규칙을 이용하여 만들어졌다. QSDSS의 목표는 식품 제조 회사의 관리자가 식품의 품질과 안전을 유지하기 위하여 최적의 물류 프레임워크를 세우도록 지원하는 것이다. 이 시스템의 적용가능성과 효율성을 설명하기 위해 홍콩 레드와인 회사의 사례 연구를 수행했으며 결과적으로 전통적인 식품 품질 보증 방식보다 객관적이고 편향되지 않았으며 구체적이라고 말한다.

Moe Moe Hlaing (2019)은 Electronic Showroom의 장바구니 분석 시스템을 개발하기 위해 ECLAT(Equivale-nce Class Transformation) 알고리즘을 사용하였다. ECLAT 알고리즘은 처리 시간을 줄임으로써 연관규칙을 빠르게 생성하며 유용한 정보를 제공한다. Electr-onic Showroom에서 실제 데이터를 수집하여 구현하였으며 이 시스템은 관련 제품을 함께 배치하고 고객에게 최상의 가격과 최신 내용을 안내해줄 수 있는 전자 전시장 관리자를 지원한다고 말하고 있다.

A. Gupta 외 3인 (2016)은 더 효율적인 연관규칙 학습을 위해 Apriori, ECLAT 및 FP-Growth 알고리즘을 비교하였다. 이 세 가지 알고리즘의 성능은 처리 시간 효율성을 기반으로 비교되며 비교 결과는 Apriori 알고리즘이 큰 데이터 세트에 대해 처리 시간이 가장 빨랐으며 보FP-Growth 알고리즘이 작은 데이터 세트에서 가장 처리 시간이 적게 나왔다고 말하고 있다. 또한 ECLAT 알고리즘은 다른 두 가지의 즘에 비해 빈번한 항목집합을 생성하는 데에 시간이 적게 걸린다고 말하였다.

Ⅲ. 제안 연구 모형

1. 연구 방법론

본 연구는 몽골 내 미용용품 및 개인용품에 관련된 품목의 판매량과 판매수익 데이터를 기반으로 분석한다. 전체 데이터는 2014년도부터 2021년도까지 총 8년 간 오프라인 매장을 통한 판매량과 금액으로 구성되어 있다. 기존 FP-Growth를 사용한 Association Rule을 적용하여 연관도 분석을 수행하면, 전체 상품의 개수가 많아 현실적인 시간 내에 유의미한 연관규칙을 찾아낼 수 없다. 이를 해결하기 위해 실제 분석 전에 불필요한 데이터를 제거하기에는 명확한 기준을 정하기 어렵다는 문제점이 존재한다.

이에 본 연구에서는 연관성 분석을 수행하기 전 유틸리티 분석을 이용한 자체적인 데이터 필터링을 구현하여 처리 속도를 높이고 데이터의 품질을 높이는 순차적 레이어 방법론을 제시하고자 한다.

우선 연관상품들을 묶어 카테고리를 생성하고 그 위에 같은 방식으로 만들어진 상위 카테고리가 존재하는 피라미드식 구조를 형성한다. 피라미드의 한 층을 하나의의 레이어로 지정하면 점진적 구조를 가진 레이어층들이 생성된다. 이중 상위 계층의 레이어에 각 항목의 가중치와 발생빈도를 고려할 수 있는 높은 유틸리티 항목집합 마이닝(High Utility Itemset Mining) 기법을 적용하여 일정 임계치 이상의 유의미한 항목만을 추출한다. 이후 추출된 항목들의 하부 레이어에서 FP-Growth를 이용한 연관도 분석을 실시하여 도출된 결과로 분석을 실시하였다.

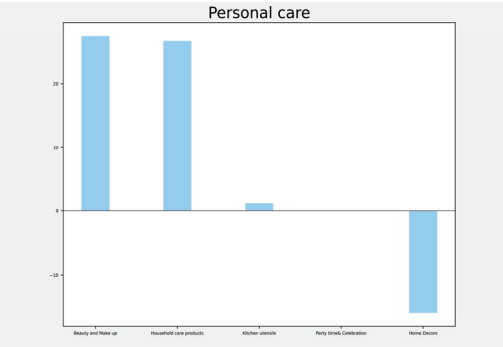
레이어는 데이터가 지나치게 복잡하게 되는 것을 피하고 제품의 용도만을 판단하기 위해 제품 각각의 Code 단위가 아닌 용도별 제품 묶음 관점에서 생성하였다.

본 실험에 사용된 데이터는 용도별 제품 Code 묶음의 가장 상위 항목인 Main Category로부터

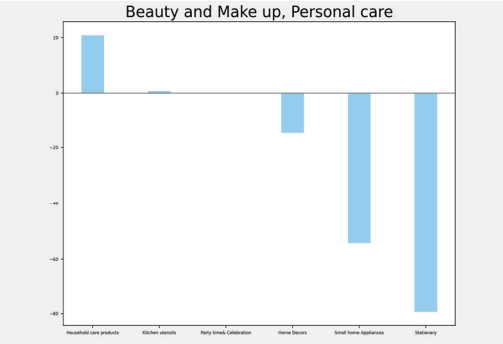
Category 및 Sub-Category로 구성되어 있으며, 최하위 Item은 브랜드별 상품 코드가 부여되어 있다. 우선 Main Category에서 유틸리티 분석을 실행하여 가장 높은 유틸리티 값을 지닌 묶음에 속한 요소만을 선택하여 하부 레이어를 구성한다. 최종적으로는 가장 하위 레이어인 Sub Category에서 일련의 과정을 거쳐 정제된 항목들에 대한 연관도 분석을 실시하여 제품의 판매 패턴을 탐색하는 것을 목표로 한다.

2. Utility Association Rule

<그림 1>은 요소 중 하나인 Personal Care와 타 요소들과의 유틸리티 증감 관계를 보여준다.



〈그림 1〉 집합에 대한 유틸리티 증감율



〈그림 2〉 {Beauty and make up, Personal care}에 대한 증감율

Personal care와 유틸리티 증가율이 높아 가장 연관도가 높다고 판단되는 Beauty and Make up이 묶여 같이 팔린 증감율을 <그림 2>에서 볼 수 있다.

<그림 1>과 <그림 2>가 보여주듯 높은 유틸리티 항목집합 마이닝은 유틸리티의 증감에 따라 연관도가 높은 집합을 탐색해나가는 알고리즘이다.

먼저 i는 각각 상품의 유틸리티이며 t는 기간의 유틸리티로 정의하였다. 집합 I는 각각의 상품의 유틸리티 i의 집합으로 정의했으며, D는 한 기간의 장바구니 t의 유틸리티의 집합으로 정의했다.

$$I=\{i_1,i_2,\cdots\ i_m\},\ D=\{T_1,T_2,\cdots\ T_m\} \qquad (1)$$

각각의 장바구니에서 연관되는 상품들의 빈도수와 가격을 곱해 얻어지는 이득을 유틸리티로 정의하였으며 항목집합 X의 유틸리티는 다음과 같이 정의된다.

$$u(X,T_q)=\sum_{i_p\in T_q}u(i_p,T_q) \qquad (2)$$

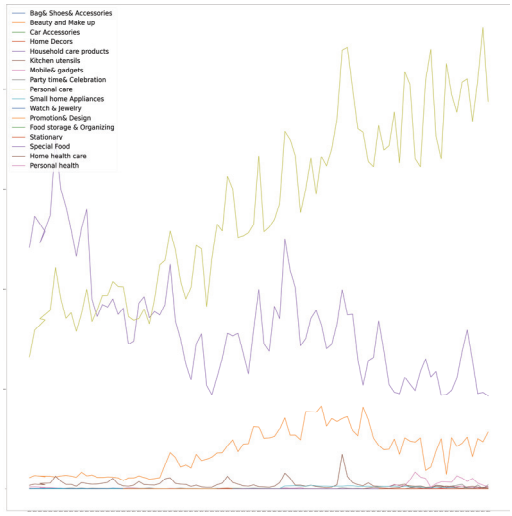
표 2에서 Personal care의 Utility는 총 판매금액/판매량*판매량은 다음과 같이 계산된다.

$$u(P,T_{2014})=216513960/38535\times38535=216513960$$

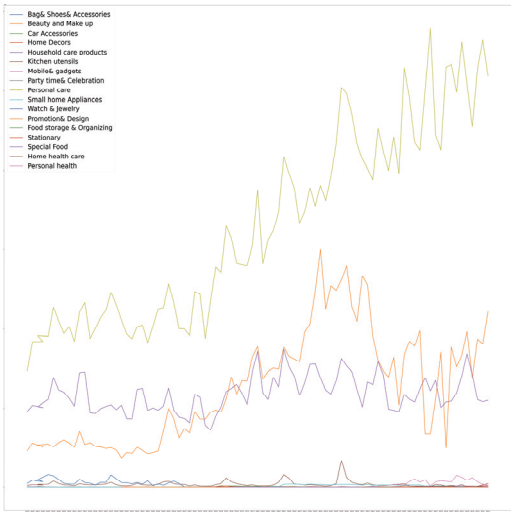
3. 원천 데이터 개요

연구에 사용된 데이터는 몽골 내 미용용품 및 개인용품에 관련된 품목들의 판매 수량과 판매총액을 매장별로 기입한 시계열 데이터이다. 품목들은 상위 카테고리에 묶여있으며 카테고리는 Main Category → Category → Sub Category 순으로 점진적으로 하강하며 세분화되는 형태를 취하고 있다. 데이터의 기초가 되는 가장 하부에는 상품코드와 그에 따른 브랜드가 기입되어있다.

데이터의 수집 기간은 2014년 1월부터 2021년 12월까지로 총 8년치를 확보했으며 각 연도마다



〈그림 3〉 Main Category – Qty



〈그림 4〉 Main Category – Amount

12개의 월별 데이터로 나누어진다. <그림 3>과 <그림 4>는 Main Category 내 항목들의 판매량과 수익을 월별로 정리한 시계열 형태의 그래프이다.

〈표 1〉 Main Category 별 정보

Main category	Amount	Qty
Bag&Shoes & Accessories	413997	6
Beauty and Makeup	11871926865	711120
Car Accessories	20763035	1123
Food storage & Organizing	1608430	407
Home Decors	18733399	1675
Home health care	18071568	8126
Household care products	11435085503	3191010
Kitchen utensils	367481263	86801
Mobile& gadgets	18951571	7119
Party time & Celebration	16565784	10443
Personal care	31289872899	5358157
Personal health	144724648	29146
Promotion& Design	3336619	105
Small home Appliances	101488344	15883
Special Food	199985	15
Stationary	4124030	353
Watch & Jewelry	230531096	1311

분석의 시작이 되는 Main Category의 개수는 총 17개이다. 이중 Personal care의 판매금액이 8년치를 통틀어 31,289,872,899₮[1]으로 가장 높았고 Special Food가 1,999,85₮로 가장 낮았다. 판매량은 역시 Personal care가 5,358,157건으로 가장 많았으며 Bag& Shoes& Accessories가 6건으로 가장 적었다. 이를 통해 데이터가 상당히 불균형한 것을 알 수 있다.

4. 데이터 전처리

전체 데이터 중 2013년도의 데이터는 4월 과 5월의 데이터가 중복되는 오류가 발견되어 제외시켰다. 또한 Personal Care의 Amount와 QTY가 너무 높아 분석시 정확도와 연관성의 저하를 초래할 것으로 우려되었으나, Sub Category 레이어로 전환할 때 연관된 하위 항목이 다양하기 때문에 별도로 제거하지 않았다.

점진적 분석을 시행할 시 가장 하위 항목은 Sub-Category로 지정하였다. 실제 데이터에서 최하위 수준은 Item은 상품의 브랜드별로 상품 코드가 정의되었기 때문에 동일한 상품임에도 불구

1) 몽골 투그릭

하고, 서로 다른 상품으로 인지될 가능성이 있으며 실제 현장에서 활용 가능한 연관규칙이 생성되기 어렵다.

Category에서는 레이어를 생성하지 않고 바로 아래 단계인 Sub Category에서 분석이 진행된다. 이처럼 Category에서 분석을 실시하지 않는 이유는 이미 Main Category 레이어에서 요소 간의 상관성을 분석해서 범위를 좁힌 상태이기에 Category 내에서의 분석을 진행하게되면 동일한 Category 내 상관성에만 집중하게 될 수 있기 때문이다.

최종적으로 각 월별 데이터 내의 상품 용도 파악에 필수적인 Main Category-Category-Sub Category를 남겼으며 브랜드와 상점명 등 불필요한 분류를 제거하였다.

최종 형태는 연도 및 월을 한 장바구니로 판단하여 시간에 따른 레이어 별 상품 판매 개수와 총 판매금액에 대한 14년도부터 21년도까지의 데이터를 추출했다.

〈표 2〉 전처리 결과 예시

Year month	Category	Amount	QTY
2014-01	Personal care	216513960	38535
2014-01	Beauty & Make up	56136729	2786
2014-01	Home Decors	31996	4
2014-02	Personal care	147523993	26496
2014-02	Beauty & Make up	46422705	2315
2014-02	Home Decors	43794	6

Ⅳ. 연관성 분석 결과

아래 <표 3>은 Main Category 수준에서 상위 10개의 Utility 분석 결과이다. 항목집합은 각각의 항목(상품)들이 묶여 만들어진 집합이며, Utility는 최상위 항목 집합의 유틸리티에 대한 비율을 나타내며 다음과 같이 표현하였다.

$$Utility_i / Utility_{\max} \quad (4)$$

항목집합 1, 2, 3, 4는 낮은 수치로 비율이 떨어지는 것을 확인할 수 있다. 항목집합 1, 2, 3은 가장 작은 항목을 가진 항목집합 4에서 파생된 것이며 다른 항목이 추가되어 항목집합 1에서 나타나는 Utility가 가장 높은 유틸리티를 갖는 것을

〈표 3〉 Main Category 수준에서의 상대적 Utility

No.	항목집합	Utility
1	'Beauty and Make up', 'Household care products', 'Kitchen utensils', 'Party time& Celebration', 'Personal care'	1.0000
2	'Beauty and Make up', 'Household care products', 'Kitchen utensils', 'Personal care'	0.9997
3	'Beauty and Make up', 'Household care products', 'Party time& Celebration', 'Personal care'	0.9933
4	'Beauty and Make up', 'Household care products', 'Personal care'	0.9930
5	'Beauty and Make up', 'Home Decors', 'Household care products', 'Kitchen utensils', 'Party time& Celebration', 'Personal care'	0.8606
6	'Beauty and Make up', 'Home Decors', 'Household care products', 'Kitchen utensils', 'Personal care'	0.8603
7	'Beauty and Make up', 'Home Decors', 'Household care products', 'Party time& Celebration', 'Personal care'	0.8553
8	'Beauty and Make up', 'Home Decors', 'Household care products', 'Personal care'	0.8550
9	'Beauty and Make up', 'Kitchen utensils', 'Party time& Celebration', 'Personal care'	0.7920
10	'Beauty and Make up', 'Kitchen utensils', 'Personal care'	0.7917

확인할 수 있다.

항목집합 5,6, 7, 8 역시 항목집합 4에서 파생되었다. 하지만 항목집합 4에 비해 상대적 Utility 비율값이 감소하는 것을 확인할 수 있다. Utility 값이 감소하는 원인으로 Home Decors 항목의 추가라는 점을 확인할 수 있으며, Home Decors는 판매되는 빈도수에 비해 가치와 Utility를 개선하는 데 도움이 되지 않는 것으로 판단할 수 있다.

항목집합 9, 10은 항목집합 9에서 파생된 집합이지만 항목집합 1에 포함되어 중복된다.

상위 항목 중 항목집합 4, 9와 같이 기준이 되는 항목에서 Home Decors등 Utility를 감소시키는 항목을 제거하고, Party time& Celebration, Kitchen utensils과 같이 Utility를 증가시키는 항목을 추가하여 연관도 분석에 사용될 데이터로 선정하였다.

<표 4>는 Utility 분석을 통해 선정된 데이터를 연관도 분석을 한 상위 10개의 결과이다. Confidence는 Confidence가 높을수록 신뢰성이 올라가 유용한 규칙임을 의미한다. Lift는 값이 1보다 크다면 우연적일 확률이 적다는 것을 의미한다. Conviction은 높을수록 독립성이 줄며 상호 연관성을 가짐을 의미한다.

<표 5> 는 Confidence, Lift, Conviction 값의 변동점을 표기하였다. 1번 규칙과 2번 규칙을 살펴보면 모든 수치들이 감소함을 알 수 있다. 세면용품 위주 구성인 1번 규칙과 다르게 2번 규칙에서는 Wet Wipes라는 청소 용품이 추가가 된 것이 요인으로 보인다. 하지만 이는 Lift값이 두 번째로 높은 연관규칙으로서 세면용품과 Wet Wipes의 새로운 연관성을 찾은 것으로 볼 수 있다.

3번 규칙과 4번 규칙을 살펴보면 Lift 값은 감소하였으나, Confidence와 Conviction은 증가함을 볼 수 있다. 데이터셋을 살펴보면 3번 규칙보다

〈표 4〉 Sub-Category 수준의 연관도 분석 결과

NO.	Antecedents	Consequents	support	confidence	lift	leverage	conviction
1	'Makeup remover', 'Ear waxing'	'Bar Soap', 'Facial cream'	0.6667	0.9143	1.2539	0.1350	3.1597
2	'Makeup remover', 'Ear waxing'	'Conditioner', 'Bar Soap', 'Facial cream'	0.6667	0.9143	1.2539	0.1350	3.1597
3	'Makeup remover', 'Ear waxing'	'Exfoliators & Masks', 'Bar Soap', 'Facial cream'	0.6667	0.9143	1.2539	0.1350	3.1597
4	'Makeup remover', 'Ear waxing'	'Toilet Paper', 'Bar Soap', 'Facial cream'	0.6667	0.9143	1.2539	0.1350	3.1597
5	'Makeup remover', 'Ear waxing'	'Cotton Balls', 'Bar Soap', 'Facial cream'	0.6667	0.9143	1.2539	0.1350	3.1597
6	'Makeup remover', 'Ear waxing'	'Bar Soap', 'Hair dye', 'Facial cream'	0.6667	0.9143	1.2539	0.1350	3.1597
7	'Makeup remover', 'Ear waxing'	'Liquid Hand Soap', 'Bar Soap', 'Facial cream'	0.6667	0.9143	1.2539	0.1350	3.1597
8	'Makeup remover', 'Ear waxing'	'Shampoo', 'Bar Soap', 'Facial cream'	0.6667	0.9143	1.2539	0.1350	3.1597
9	'Makeup remover', 'Ear waxing'	'Tissues', 'Bar Soap', 'Facial cream'	0.6667	0.9143	1.2539	0.1350	3.1597
10	'Bar Soap', 'Facial cream'	'Makeup remover', 'Tissues', 'Ear waxing'	0.6667	0.9143	1.2539	0.1350	3.1597

〈표 5〉 Sub-Category 내 유의미한 연관 규칙 목록

No.	antecedents consequents	confidence	lift	conviction
1	'Bar Soap', 'Facial cream', 'Tissues', 'Hair dye', 'Liquid Hand Soap' 'Shampoo', 'Makeup remover', 'Ear waxing'	0.9143	1.2539	3.1597
2	'Wet Wipes', 'Makeup remover', 'Ear waxing' 'Bar Soap', 'Facial cream'	0.9130	1.2522	3.1146
3	'Bar Soap', 'Facial cream', 'Shampoo', 'Tissues', 'Liquid Hand Soap' 'Wet Wipes', 'Makeup remover', 'Ear waxing'	0.9000	1.2522	2.8125
4	'Makeup remover', 'Face Powder', 'Ear waxing' 'Bar Soap', 'Facial cream'	0.9104	1.2486	3.0243
5	'Tissues', 'Shampoo', 'Bar Soap', 'Facial cream' 'Makeup remover', 'Face Powder', 'Ear waxing'	0.8714	1.2486	2.3495

4번 규칙의 빈도가 줄어들어 Lift 값은 낮지만 연관 비율이 높아 Confidence와 Conviciton 값이 높아졌기에 고려할 필요가 있다.

Ⅴ. 결 론

연관도 분석 특성상 시간이 오래걸리기 때문에 데이터 정재과정이 필요하나 그 기준을 잡기 어려웠다.

본 연구는 유틸리티의 증감을 척도로 피라미드 구조의 레이어와 유틸리티 분석을 이용하여 점진적으로 데이터를 정제하여 연관도 분석을 위한 새로운 기준을 세웠다. 이에 따라 가중치가 낮은 항목을 제외하는 규칙을 생성했다.

정재된 데이터를 이용하여 연관도 분석을 실행 시 Wet Wipes와 세면도구들 간의 연관성 등 의미있는 연관규칙을 발견하였으며 수치들의 증감을 통해 연관성을 고려해 볼 만한 항목 묶음을 얻을 수 있었다.

이를 활용해 주요 매장에 대해 상품 배치나 구성 등과 같이 판매를 촉진시키기 위한 다양한 방안을 제시할 수 있다. 또한, 단순 판매 여부가 아닌 판매량과 매출을 고려하여 도출된 상관성이 높은 상품 묶음에 대한 통합 수요예측을 수행하여 정확도를 높일 수 있으며, 유통 채널 별 재고관리에 활용될 수 있다.

최고 유틸리티 항목의 가중치가 높은 분석이기 때문에 레이어를 거쳐 제거되는 요소가 필연적으로 생기게 됨으로 특성에 잘 부합하는 레이어를 생성해야 하며 제거되는 데이터에 대해 고려할 필요가 있다.

참 고 문 헌

[1] Ambuj Gupta, Sudhakar Hannah, Kapoor Shivam and Anand Shivangi,, "Performance Comparison of Apriori, Eclat and FP-Growth Algorithm for Association Rule Learning", 2016.

[2] Bhargav Anshul, Robin Prakash Mathur and Munish Bhargav,, "Market basket analysis using artificial neural network", 1-6, 2014.

[3] Dubey Shish Kumar, Sonu Mittal, Seema Chattani and Vinod Kumar Shukla,, "Comparative Analysis of Market Basket Analysis through Data Mining Techniques", 239-243, 2021.

[4] Hidayat A. A., A. Rahman, R. M. Wangi, R. J. Abidin, R. S. Fuadi and W. Budiawan,,

"Implementation and comparison analysis of apriori and fp-growth algorithm performance to determine market basket analysis in Breiliant shop", 1402, 7, 2019.

[5] Hlaing Moe Moe., "ECLAT based market basket analysis for electronic showroom", 2019.

[6] Liu, Run-Qing, Young-Chan Lee, and Hong-Lei Mu. "Customer classification and market basket analysis using K-means clustering and association rules: evidence from distribution big data of korean retailing company." Knowledge Management Research, 19, 4, 56-76, 2018.

[7] Lin Chun-Wei, Tzung-Pei Hong and Wen-Hsiang Lu,, "An effective tree structure for mining high utility itemsets", Expert Syst.Appl., 38, 6, 7419-7424, 2011.

[8] Martinez Marcos, Belén Escobar, Maria E. Garcıa-Dıaz and Diego P. Pinto-Roa,, "Market basket analysis with association rules in the retail sector using Orange. Case Study: Appliances Sales Company", CLEI Electron, 24, 2, 2021.

[9] Nandintsatsral Amarsanaa, "코로나19가 가져온 몽골 화장품 시장의 새로운 트랜드", KOTRA&KOTRA 해외시장 뉴스, 2020.

[10] Pillai Jyothi and O. P. Vyas,, "User centric approach to itemset utility mining in Market Basket Analysis", International Journal on Computer Science and Engineering, 3, 1, 393-400, 2011.

[11] Prasad K. Rajendra., "Optimized high-utility itemsets mining for effective association mining paper", International Journal of Electrical and Computer Engineering (IJECE), 7, 5, 2911-2918, 2017.

[12] S. S. Khedkar and S. Kumari,, "Market Basket Analysis using A-Priori Algorithm and FP-Tree Algorithm", 2021 International Conference on Artificial Intelligence and Machine Vision (AIMV), 1-6, 2021.

[13] Ting S. L., Y. K. Tse, GTS Ho, Sai Ho Chung and Gu Pang,, "Mining logistics data to assure the quality in a sustainable food supply chain: A case in the red wine industry", Int J Prod Econ, 152, 200-209, 2014.

[14] Wang Le, Shui Wang, Haiyan Li and Chunliang Zhou,, "Improved Strategy for High-Utility Pattern Mining Algorithm", Mathematical Problems in Engineering, 2020, 2020.

[15] 김진형 and 김병욱, "데이터 카디널리티에 따른 FP-Growth 알고리즘의 효율성 분석", 한국정보처리학회 학술대회논문집, 26, 1, 33-35, 2019.

[16] 정병수, 아메드, 파한, 이인기 and 용환승,, "Prefix-Tree 를 이용한 높은 유틸리티 패턴 마이닝 기법", 정보과학회논문지: 데이타베이스, 36, 5, 341-351, 2009.

[17] 황정희., "유틸리티-리스트를 이용한 높은 유틸리티 항목집합 마이닝", 한국디지털콘텐츠학회논문지, 21, 3, 579-586, 2020.

저 자 소 개

방 선 호(Sun-Ho Bang)

·2021년 8월 : 인천대학교 전자공학과 (공학사)
·2021년 9월 : 인천대학교 동북아물류대학원 물류시스템학과 석사과정
·<관심분야> : 빅데이터, SCM

장 지 영(Ji-Young Jang)

·2021년 8월 : 우석대학교 토목환경공학과 (공학사)
·2021년 9월 : 인천대학교 동북아물류대학원 물류경영학과 석사과정
·<관심분야> : 스마트시티, 라스트마일

Tsatsral Telmentugs

·2007년 : Huree ICT University (InformationTechnology)
·2012년~2019년 : Nomin Trading CO.LTD (Import Manager)
·2018년~2020년 : Noming Trading Co.LTD (Senior Manager)
·2020년~2021년 : Nomin Holding LLC (Executive Director)
·2021년~현재 : 인천대학교 동북아물류대학원 물류경영학과 석사과정
·<관심분야> : 빅데이터, 수요예측

이 강 현(Kang-Hyun Lee)

·2022년 2월 : 청주대학교 전자공학과 (공학사)
·2022년 3월~현재 : 인천대학교 동북아 물류대학원 물류시스템학과 석사과정
·<관심분야> : 빅데이터, 머신러닝

신 광 섭(Kwnag-Sup Shin)

·2003년 2월 : 서울대학교 산업공학과 (공학사)
·2006년 2월 : 서울대학교 산업공학과 (공학석사)
·2012년 2월 : 서울대학교 산업공학과 (공학박사)
·2012년 2월~현재 : 인천대학교 동북아물류대학원 교수
·<관심분야> : 빅데이터 활용, 솔루션