Journal of The Korea Convergence Society
Vol. 13. No. 3, pp. 203-216, 2022
e-ISSN 2713-6353
DOI : https://doi.org/10.15207/JKCS.2022.13.03.203

전자상거래 이용시 연관성 분석을 통한 맞춤형 상품추천 모델 설계

MingFei Yang[1], 박기용[2], 최상현[3*]
[1]충북대학교 경영정보학과 석사, [2]충북대학교 빅데이터협동과정 연구교수, [3]충북대학교 빅데이터협동과정 교수

Design of customized product recommendation model on correlation analysis when using electronic commerce

MingFei Yang[1], Kiyong Park[2], Sang-Hyun Choi[3*]
[1]Master's Degree, Management Information System, Chungbuk National University
[2]Research Professor, Department of Big Data, Chungbuk National University
[3]Professor, Management Information System, Department of Big Data, Chungbuk National University

요 약 본 연구에서는 COVID-19의 영향과 온라인 시장을 중심으로 구매패턴이 변화하는 현 경영환경의 시대에서 온라인 배송업체의 구매정보와 상품정보를 기반으로 군집분석과 연관성 분석을 실시하였다. 고객군집, 상품군집, 그리고 교차결합을 통해 데이터를 세분화시켜 결합군집을 생성하여 학문적으로 새로운 방안의 군집분석을 시도하였으며, 각각의 군집분석 결과를 토대로 연관성 분석을 하였다. 연관성 분석 결과, 상대적으로 결합군집에서 더 많은 연관 규칙이 도출 되었으며, 중복률은 더 적은 것으로 분석되어 효율성이 매우 높은 것으로 나타났다. 이는 고객의 니즈에 맞게 상품을 추천하기 위해서는 결합군집이 가장 적합한 모델이라고 판단된다. 결합군집 모델은 소비자에겐 시간 절약과 유용한 정보를 제공하면서, 해당 업체에는 판매량을 증가시키는 등의 긍정적인 효과를 가져올 것으로 사료된다. 향후 연구과제로써, 다양한 특성을 갖고 있는 다수의 온라인 배송업체들을 대상으로 비교·분석한다면 좀 더 명확하고 유의미한 연구결과를 도출할 수 있을것으로 기대된다.

주제어 : 온라인, 전자상거래, 상품추천 모델, 군집 분석, 연관성 분석

Abstract In the recent business environment, purchase patterns are changing around the influence of COVID-19 and the online market. This study analyzed cluster and correlation analysis based on purchase and product information. The cluster analysis of new methods was attempted by creating customer, product, and cross-bonding clusters. The cross-bonding cluster analysis was performed based on the results of each cluster analysis. As a result of the correlation analysis, it was analyzed that more association rules were derived from a cross-bonding cluster, and the overlap rate was less. The cross-bonding cluster was found to be highly efficient. The cross-bonding cluster is the most suitable model for recommending products according to customer needs. The cross-bonding cluster model can save time and provide useful information to consumers. It is expected to bring positive effects such as increasing sales for the company.

Key Words : Online, Electronic commerce, Product recommendation model, Cluster analysis, Correlation analysis

*Corresponding Author : Sang-Hyun Choi(pky3489@chungbuk.ac.kr)
Received November 9, 2021
Revised March 7, 2022
Accepted March 20, 2022
Published March 28, 2022

1. 서론

급격하게 변화하고 있는 경영 환경의 시대에서 고객의 니즈를 파악하는 것은 과거부터 매우 중요한 과제로 인식되고 있었으며[1], 소비자가 제품 선택하고 서비스의 질을 요구하면서 상품 구매의 상호간의 연관성을 확인하는 문제는 학술적으로도 중요한 연구의 대상이 되고 있다[2]. 최근 정보통신 기술의 발전은 상품 구매 패턴에 상당한 영향을 주고 있으며, Television, 컴퓨터, 태블릿 PC, 스마트폰 등 다양한 기기를 통해 직접 매장을 방문(off-line)하지 않고 온라인(on-line) 채널을 통해 시간, 장소 등에 구애받지 않고 쇼핑을 간편하게 할 수 있게 되었다[3]. 쇼핑과 구매 패턴이 온라인 형태로 변화하면서 전자상거래 플랫폼 업체에도 많은 영향을 끼치고 있다. 쇼핑몰 유통업체, 판매업체 등 전자상거래(electronic commerce) 플랫폼 창업자의 급속한 증대[4]로 이어졌다. 또한, 데이터의 수집 및 활용이 용이해지면서 소비자의 선택과 구매의 연관성에 관한 분석이 가능해졌다. 더불어, 2020년에는 전 세계적으로 COVID-19 팬데믹이 발생하였으며, Fig. 1과 같이 대한민국에서도 마찬가지로 2020년 2월 말에 COVID-19가 시작되어 8월과 10월, 2차례에 걸쳐 확진자 수가 급격히 증가한 것을 확인할 수 있다.

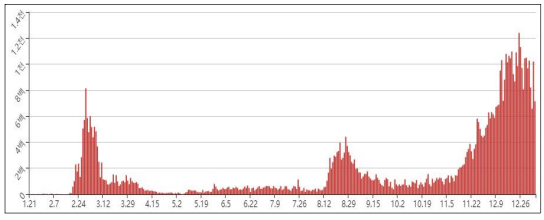


Fig. 1. COVID-19 confirmed cases in Korea(2020)
Source: https://coronaboard.kr/

2020년 COVID-19 팬데믹 이후, 소비자들은 직접 매장을 방문(off-line)하여 상품을 구매하는 방법보다는 전자상거래 방식의 온라인(on-line) 채널을 이용하여 상품을 구매하는 활동의 비중이 높아졌다[5]. SNS에서는 '1인', '무인', '비대면', '격리', '소규모', '소자본' 등의 키워드가 많이 등장하고 있다[6]. 이처럼 COVID-19의 영향으로 소비자의 소비패턴, 구매활동 그리고 생활환경까지도 큰 변화가 이루어졌으며, 전자상거래 업도 활성화를 이루게 되었다.

전자상거래(electronic commerce)는 다양한 고객의 특성에 맞춰 가장 적절한 유형의 제품이나 정보를 손쉽게 제공함으로써 개인화된 서비스가 가능하다[7]. 또한, 제품 선택과 결제, 배송까지 빠른 시간내에 이루어질 수 있어 쉽고 편리하게 구매 활동을 할 수 있다. 오프라인 마켓에서는 상품을 제한된 장소에 진열하고 판매하여 상품의 수가 제한적이고 운반의 어려움이 있어 소량의 상품을 구매하는 비중이 높은 편인 데에 반해[8], 온라인 구매는 제품은 운반이 쉽고 많은 상품 구매가 가능하여 한 번에 다량의 상품을 구매하는 경우가 많아지고 있다[9]. 이와 같은 이유로, 오프라인 시장 보다는 온라인 시장을 선호하는 추세로 변화가 이루어지고 있다.

따라서 소비자들의 요구를 정확하게 파악하고, 고객의 니즈에 맞게 상품을 전략적으로 추천하고 판매하게 되면, 구매 활동 시간을 절감시킬 수 있는 것은 물론, 소비자의 구매 의욕을 향상시킬 수 있다는 장점이 있다. 또한, 소비자가 다량의 상품을 구매할 경우 다양하고 복합적인 유형의 상품들 사이에서 파생되는 연관성을 분석하여 적용한다면 효과적인 판매 전략을 수립할 수 있을 것이다. 이에 본 연구는 최소한의 고객 정보를 기반으로 가장 효율적인 분석 방법을 통해 빠르고 정확하게 고객 맞춤형 상품을 추천하는 모델을 개발하고자 한다.

2. 이론적 배경

2.1 연관성 분석 기반의 연구 동향

최근 디지털 고도화가 더해진 빅데이터 시대로 변모하면서 대용량의 데이터를 기반으로 새로운 정보 도출이 가능한 데이터마이닝(data mining) 분석기법이 주목을 받고 있다. 데이터마이닝 분석은 데이터와 데이터 간의 상관 관계를 파악하고 다양한 패턴을 찾아내는 기법으로 이를 분석하고 유의미한 정보로 변환하는 과정까지를 포함한다. 데이터 정보는 이미 존재하고 있지만, 데이터베이스로부터 지식을 발견하여 새로운 규칙과 사실을 연구하고 찾아냄으로써 사용하는 대상에게 효율적인 의사결정을 할 수 있도록 도움을 줄 수 있다. 이러한 연관성 분석기법을 활용해 기존의 연구에서는 상관 관계를 입증하고 정확성을 향상 시키는 등 분석결과를 마케팅 전략으로 활용하고자 하였다.

자세히 살펴보면, Ryu는 투자자의 매매 동향을 분석하기 위하여 연관성 분석을 적용하는 방법을 제안하였다.

COVID-19 발병 전 후로 기간을 구분하여 다양한 상품별로 투자자 간의 매매패턴 분석을 통해 투자 주체가 갖고 있는 성향을 파악하였으며, 다양한 투자 주체와 투자 상품들의 연관성을 분석하여 객관적이고 정량적인 결과값을 도출하였다[10]. Cho는 연구기간 동안 수차례에 걸쳐 수집된 청년 패널 데이터(한국고용정보원)의 종단 자료를 기반으로 연관 규칙 마이닝(sequential pattern)을 적용하여 잠재성장모형(latent growth modeling)의 성장궤적 기울기를 추정함으로써 모형의 적합도와 정확도를 높였다[11]. Park는 5대 범죄에 해당하는 강도, 절도, 폭력에 대한 발생을 사전에 예측하여 범죄 발생을 예측하고자 하였다. 범죄 관련 키워드를 웹 검색을 통해 해당 범죄의 발생 빈도수를 분석하여 실제 범죄 발생 건수와의 연관성 분석을 실시함으로써 상관성이 있음을 입증하였다[12] Won은 농촌진흥청에서 제공하고 있는 농식품 소비자 패널조사에서 제공하고 있는 소비자의 농식품 구매내역 정보를 활용하여 연관성 분석을 실시하였다. 구매처는 대형마트, 기업형 슈퍼마켓, 전통시장으로 유형화하였으며, 시기는 봄, 여름, 가을, 겨울로 구분하고, 구매대상은 30대, 40대, 50대로 구분하여 분석하였다. 분석결과 연령대에 따라 구매하는 장소가 변화하였으며, 소매점포에서 묶음 판매와 상품 홍보 등을 활용했을 경우 매출액이 증가하는 것을 근거로 마케팅 전략을 수립하는 방법을 제시하였다[9].

2.2 군집 분석 기반의 상품추천 모델 연구 동향

군집 분석은 데이터 정보들 간의 유사성과 상호연관성에 근거하여 동질적인 군집으로 분류하고, 달리 분류된 군집과의 상이성을 분류하는 기법이다. 이러한 군집분석을 적용한 기존 연구에서는 군집화를 통해 데이터 정보의 패턴 분석과 예측분석을 통해 사용자의 만족도를 향상시키고자 하였다.

자세히 살펴보면, Shin은 국민 개개인의 특성을 반영하여 만성질환의 발생 가능성을 추론하기 위해 다중 상황의 지식 추론모델을 제안하였다. 생활습관을 통한 상황 정보는 국민영양조사에서 제공한 데이터를 활용하여 전처리 과정후 군집분석과 연관규칙을 각각 적용하여 사용자 개개인에게 헬스케어 방법을 제공하여 해당 질병에 대한 발병률을 낮추었다[13]. Yoon은 롯데 멤버스 고객의 상품 구매 이력과 고객의 이용업종 데이터를 활용하여 군집화 분석을 통해 집단별 예측모델을 구현하고자 하였다. 전체 데이터를 활용한 예측모델보다 군집화를 통해 예측모델을 분석하는 것이 정확도와 예측성능을 향상시킬 수 있음을 증명하였다[14]. Kim은 고객의 구매 이력 데이터(구매수량, 구대금액)를 활용하여 고객의 니즈를 반영한 상품을 탐색하고 분류하고자 k-means 군집분석 모형을 적용하였다. 통계적 기법과 기계학습 기법의 비교와 분석을 통해 고객 세분화를 위한 군집 분석을 실시하여 군집의 품질을 정량적으로 평가하였다. 상품군집의 유형을 토대로 고객의 구매 특성과 패턴을 예측할 수 있으며, 선호도 또한 예측이 가능해짐으로써 기존 고객의 만족도를 향상시키고 새로운 고객을 유입시킬 수 있는 마케팅 전략으로 활용하고자 하였다[15]. An은 농산물을 가공하여 만든 식품을 실용적인 판매 전략에 적용하여 판매량을 향상시키기 위한 목적으로 구매 패턴에 대한 연구를 진행하였다. 농촌진흥청에서 구득한 농식품 소비자 패널 데이터인 구매 이력 데이터를 기반으로 k-means 군집 분석을 실시하였으며, 분석결과 유통채널과 판매 시기, 그리고 소비자의 특성 등을 통해 분류된 군집별로 구매패턴의 차이점을 도출하여 판매 전략 개발에 활용하고자 하였다[16].

2.3 선행연구와의 차별성

기존 상품을 추천하기 위한 시스템을 목적으로 연관성 분석과 군집 분석을 진행한 연구에서는 사전에 구축되어 있는 대용량의 데이터를 기반으로 유의미한 규칙과 정확도를 향상시키는 데에만 집중하였다. 또한, 이러한 방법은 데이터 정보가 부족한 경우와 대상에 대한 세부적인 프로파일을 활용하지 못했을 경우 군집의 유형화는 물론, 예측 성능도 저하될 수 있는 한계점을 갖고 있다.

이에 본 연구에서는 온라인 플랫폼에서 구축한 최소한의 고객정보를 토대로 고객군집, 상품군집, 그리고 교차결합을 통해 적절한 군집으로 세분화시켜 새로운 결합군집을 도출하는 등 좀 더 구체적이고 다양한 시도를 하였다. 이를 토대로 연관성 분석을 실시함으로써 각 군집별 성능을 비교·평가함으로써 신속하고 정확도를 향상시킨 고객 맞춤형 상품 추천 모델을 설계했다는 점에서 차별성을 지닌다. 본 연구의 분석결과는 고객에게 유용한 정보를 제공함은 물론 구매 활동 시간을 절감시킬 수 있을 것으로 기대한다.

3. 연구 방법

3.1 연구대상 및 자료수집

본 연구에서는 온라인 유통업체 'M사'의 약 1년(2020년 1월 23일부터 2020년 12월 30일)의 기간 동안 고객정보 및 거래정보 데이터를 Table 1과 같이 획득하였다. 고객정보는 거래번호, ID, 성별, 나이, 구매유형 등 12가지 유형의 데이터를 획득하였으며, 거래정보는 상품 이름, 상품 무게, 상품 단가, 구매 수량, 지불 금액 등의 17가지 유형의 데이터를 획득하였다. 고객정보와 거래정보 데이터는 총 182,886건으로 집계되었다.

Table 1. Customer and transaction information items

Customer inf. items	Transaction info. items
ono	product_name
member_id	weight
sex	big
age	mid
mobile	small
join_date	small_4th
join_ref	sell_price
order_seq	last_order_date
order_date	buy_ea
reserve_time	pay_price
address	total_price
receipt_type	pay_type
	total_weight
	milage_price
	coupon
	dlv_price
	branch_no

3.2 데이터 전처리 과정 (Data Preprocessing)

온라인 유통업체 'M사'의 고객정보와 거래정보는 업무를 관리하고 사용하기 위해 해당 업체의 업무 특성에 맞게 기록하고 정리하였기 때문에 데이터마이닝(data mining) 기술을 적용하여 진행하기에 적절하지 않은 데이터가 다수 존재하고 있었다. 특히, 누락 되거나 결측된 값, 그리고 의미 없는 값 등이 포함되어 있는 경우 분석결과가 상이하게 나올 수 있을 뿐 아니라 그 질도 현저히 낮아질 수 있다. 따라서 데이터 기초분석과 연관성 분석을 진행하기에 앞서 부정확한 데이터를 제외하거나 올바르게 데이터를 가공하는 등의 데이터 전처리 과정(Data Preprocessing)이 필요하다. 본 연구에서는 182,896건의 온라인 유통업체 'M사'의 고객정보와 거래정보를 토대로 데이터 정리 → 결측값과 이상치 처리 → 데이터 가공 → 상품 정리 → 데이터 통합 순으로 데이터 전처리 과정(Data Preprocessing)을 진행하였다.

데이터 전처리 과정(Data Preprocessing)은 아래와 같다.

1) 데이터 정리: 고객정보에서 ono(거래 번호), mobile(구매 유형), join_date(가입 날짜), join_ref(가입유형)의 변수를, 거래정보 중에서는 big(대분류), branch_no(구매 지점)의 필요하지 않은 변수는 제외시켰다.
2) 결측값, 이상치 처리: 데이터 수집 시, 다양한 원인으로 인해 누락되거나, 잘못된 값이 입력되는 등의 오류가 발생할 수 있다. 따라서 데이터가 누락된 약 34,000건의 값과 age(연령), pay_price(지불금액)의 변수 값이 잘 못 입력된 것을 확인할 수 있었다. 따라서 결측값과 이상치로 확인 된 데이터를 제외시켰다.
3) 데이터 가공: 재범주화, 범주화, 변수의 신규생성 등의 과정을 통해 데이터를 보다 효율적으로 사용할 수 있도록 가공하였다. 주소 변수를 통해 지역(시, 군, 구)별로, 주문날짜 변수를 통해 시간(계절, 요일, 시간)별로 변수를 재범주화시켰다. 배송료와 관련된 변수는 배송비 여부로, 쿠폰할인 금액과 관련 된 변수는 쿠폰사용 여부 등의 범주형 변수로 변형시켜 범주화시켰다. 그리고 상품 이름 변수를 통해 저열량 식품[1], 즉석 식품[2], 신선 식품[3]으로 새로운 변수를 생성하였다.
4) 상품 정리: 연관성 분석 진행 시 다른 상품과 관련하여 지지도와 신뢰도에 영향을 미칠 수 있는 변수들이 해당되며, 상위 빈도 10% 상품 (과자·쿠키·파이, 봉지라면 외 48개 상품), 빈도수가 100개 이하인 상품 (복사용지, 곤약·도토리묵 외 85개 상품)들을 제외하였다.
5) 데이터 통합: 데이터 정리 〉 결측값, 이상치 〉 데이터가공 〉 상품정리의 과정을 통해 나온 변수와 데이터를 통합하는 과정이다. 통합된 데이터의 정보는 Table 2와 같으며 총 182,685건으로 집계되었다.

1) 한국 보건복지부에서 제공하고 있는 비만 예방 및 관리를 위한 바른 식생활 가이드인 '저열량 레시피'에 근거하여 변수를 생성하였음.

2) 식품 의약품 안전처'에서 제공하고 있는 가정간편식 나트륨 저감화 기술 가이드인 '즉석섭취 식품 편'을 근거로 변수를 생성하였음.

3) 야채, 과일, 해산물, 등 미가공된 식품을 기준으로 변수를 생성하였음.

Table 2. Integrated information items

member_id	receipt_type
sex	weight
age	mid
order_seq	small
order_date_Season	small_4th
order_date_Week	sell_price
order_date_Hour	buy_ea
reserve_time_Season	pay_price
reserve_time_Week	total_price
reserve_time_Hour	pay_type
product_name	total_weight
low_kcal_food	milage_price
fresh_food	coupon
fast_food	dlv_price
address	

3.3 데이터 기초분석

구매패턴을 확인하기 위해서 182,685건에 대한 상품의 중분류 변수 데이터를 기준으로 기초분석을 실시하였다. 중분류의 범주는 Table 3과 같이 16가지로 구분되어 있으며, 각 품목에 대하여 구매 빈도, 평균 가격, 평균 구매개수, 평균 제품무게 등을 분석하였다. 구매빈도는 과자·간식·시리얼 제품이 36,013개로 가장 많은 것으로 나타났고, 이어서 유제품·냉장·냉동 제품이 33,781개, 라면·면류·간편식·통조림 제품이 28,699개의 순으로 나타났다. 제품의 평균가격은 쌀·잡곡의 제품이 31,516원으로 가장 비싼 것으로 나타났으며, 이어서 화장지·생리대·기저귀 제품이 5,069원, 정육·달걀 제품이 4,183원의 순으로 나타났다. 구매개수는 세탁용품·청소용품이 3.16개로 가장 많은 것으로 나타났으며, 물·음료·커피·차·분유 제품이 2.71개, 정육·달걀 제품이 2.63개의 순으로 나타났다. 제품의 평균 무게는 쌀·잡곡이 11,403g으로 가장 무거운 것으로 나타났으며, 채소·과일·두부·견과류 제품이 3,019.5g, 물·음료·커피·차·분유 제품이 2,965.4g의 순으로 나타났다. 이는 데이터 기초분석을 통해서 상품의 종류에 따라서 유의미한 차이가 있는 것을 보여주며, 이를 통해 적절한 군집으로 분류가 가능하다는 것을 확인할 수 있었다.

Table 3. Transaction information according to classification

Classification \ Average	Transaction frequency	Price (won)	Purchase	Weight (g)
snacks	36,013	1,827	1.42	165.7
refrigerated·frozen food	33,781	2,932	1.65	581.7
remen, can	28,699	2,432	1.78	420.7
fruit, vegetable	21,520	2,393	1.18	3,019.5
water, coffee	17,805	2,513	2.71	2,965.4
flour, olive oil	10,729	2,942	1.33	760.4
meat, egg	8,067	4,183	2.63	814.1
washing·cleaning supplies	5,510	2,960	3.16	1,174.1
kitchen supplies	5,046	2,039	1.60	220.0
marine products	4,408	2,936	1.27	70.6
tissue	4,050	5,069	2.61	115.1
hair·body care	2,702	3,868	1.40	400.9
rice·cereals	1,518	31,516	1.06	11,403.0
stationery	1,230	1,406	1.44	16.9
daily necessities	1,108	2,946	1.38	42.8
pets	498	2,823	1.82	673.4

고객정보와 거래정보 중 거래 기준 데이터 12,737건을 추출하여 기초분석을 진행한 결과, Table 4와 같이 나타났다. 여성 고객은 10,776명(84.5%), 남성 고객은 1,971명(15.5%)로 나타났으며, 주요 고객 연령층은 30대(4,733명)와 40대(5,667명)가 가장 높은 비율을 차지하고 있었다. 또한, 연령대별로 평균 지불금액과 평균 거래개수, 그리고 평균 무게 역시 30대와 40대가 높은 비율을 차지하고 있는 것으로 나타났다.

Table 4. Transaction information by customer age group

Age \ Average	Customers	Price (won)	Transaction frequency	Weight(g)
0-20	139	23,758	12.0	13,543
20-29	1,064	24,193	12.0	15,945
30-39	4,733	25,373	12.9	16,720
40-49	5,667	24,967	12.7	16,736
50-60	1,021	22,981	11.3	14,823
Over 60	113	22,581	10.2	9,907
Total	12,737			

2020년 유통업체 'M사' 고객들의 월별 평균 이용량은 1,061건으로 Fig. 2와 같이 나타났으며, 특히 8월(1,522건)과 12월(1,504건)에 상대적으로 높은 이용량

을 보이고 있다. 이는 Fig. 1에서 나타났듯이 2020년 2월 말에 COVID-19가 발발하면서 유통업체 'M사'의 이용량에도 영향을 미친 것으로 판단된다. 2020년 대한민국 COVID-19 확진자 현황이 급증한 3월, 8월, 11월과 마찬가지로 유통업체 'M사'의 이용량 역시 급격하게 증가하는 것을 확인할 수 있었다. 이는 COVID-19가 고객들의 소비행태를 변화시켜 온라인 유통업체에도 영향을 주고 있음을 나타낸다.

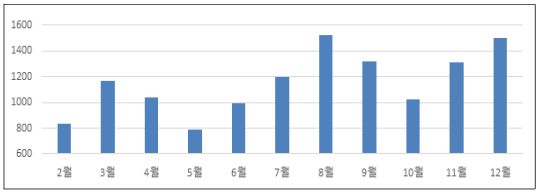


Fig. 2. Monthly usage of 'Company M' (2020)

3.4 방법론

군집 분석(cluster analysis)은 데이터의 상호연관성을 기반으로 동질적인 집단을 분류하는 방법으로[17], 계층적 군집화(Hierarchical Clustering) 방법과 비계층적 군집화(Non-Hierarchical Clustering) 방법으로 분류할 수 있다[18]. 계층적 군집화는 각각의 개체 간 거리로부터 최단 거리에 있는 개체를 시작으로 결합하는 과정을 반복함으로써 나무 모양의 계층구조를 형성해 나가는 기법이다[19]. 계층적 군집화 기법은 군집 형성 과정을 명확하게 파악할 수 있지만, 자료의 크기가 방대해질수록 모델의 계산량이 급격하게 증가하여 분석하는데 어려움을 겪을 수 있다[15]. 비계층적 군집화 기법은 군집의 수(K개)를 사전에 결정하고 각 레코드로부터 군집의 중심까지의 거리를 계산하여, 최단 거리의 군집을 할당하는 기법이다. 이러한 기법 중에 가장 많이 사용되는 기법은 K-평균 군집화(K-means Clustering) 기법이 있다[20]. K-평균 군집화(K-means Clustering) 기법은 K개의 군집 중심좌표를 설정한 후, 각 개체를 가까운 군집에 배정하는 알고리즘이다. 이 기법은 대량의 자료에서 군집을 발견하는데 효과적인 방법이다[21]. 최적의 군집인 K 개수를 결정하는데 대표적으로 엘보우(elbow) 방법과 실루엣(silhouette) 방법 2가지가 이용되고 있다. 엘보우 방법은 클러스터 개수를 증가시켰을 때 각 중심 간의 평균 거리가 더 이상 의미있는 감소가 발생하지 않는 경우의 군집의 수 K를 선택하는 방법이다. 실루엣 방법은 정량적으로 클러스터링의 품질을 계산하는 방법이다. 실루엣 분석 방법은 엘보우 방법보다 계산하는 데에 더 많은 시간이 소요된다는 단점이 있다[22].

연관규칙(AR; Association Rule)은 항목들(items) 간의 '조건-결과(if-then)' 식으로 나타나는 유용한 패턴을 의미하며[23], 조사하고자 하는 항목 X와 항목 Y 사이의 서로 연관성 있는 X→Y 형태의 규칙을 찾아내어 관계를 살펴보고, 유용한 규칙을 찾아내는 데이터마이닝 기법이다[24]. 연관성 분석은 상품 또는 서비스 간의 유용하고 의미있는 관계가 존재하는지 확인해보고자 할 때 가장 적합한 방법이라고 할 수 있다[25]. 또한 원리가 간단하여 누구나 쉽게 응용이 가능하며, 분석 되어진 결과가 명확하고 정확도가 높기 때문에 설명력이 높은 편이다[26]. 분석결과로 도출된 연관규칙은 지지도(support), 신뢰도(confidence), 향상도(lift) 등의 지표를 기반으로 평가할 수 있다[9]. 지지도(support)는 사용자가 동시에 항목 X와 항목 Y를 포함하는 확률로 정의할 수 있으며[27], 수식(1)과 같다.

$$\text{Support} = P(X \cap Y) \tag{1}$$

신뢰도(confidence)는 항목 X가 거래 항목 Y를 포함시키는 확률의 정도를 나태내는 조건부 확률로써 항목 X를 구매한 경우 항목 Y도 구매할 확률이며[28], 수식(2)과 같다.

$$\text{Confidence} = \frac{P(X \cap Y)}{P(X)} = P(Y \mid X) \tag{2}$$

향상도(lift)는 연관규칙의 강도(strength)를 측정하는 지표로써[24], X와 Y항목 간의 향상도 값이 1일 경우 상호독립적인 것을 나타내며, 1보다 큰 값은 양의 상관관계, 1보다 작은 값은 음의 상관관계를 나타내는 것을 의미한다[27]. 수식 (3)과 같다.

$$\text{Lift} = \frac{P(Y \mid X)}{P(Y)} \tag{3}$$

연관규칙은 크게 두 단계의 과정으로 진행된다. 첫 번째 단계에서는 최소한의 지지도가 사용자가 정의한 값을 넘어서는 규칙만을 선택하게 된다. 두 번째 단계에서

는 선택된 규칙 중에서 최소한의 신뢰도 값을 넘어서는 규칙을 선택하는 과정으로 진행되는데, 해당 단계에서는 신뢰도 값 대신 향상도 값을 기준으로 활용할 수 있다. 즉, 연관규칙에 관련 되어 있는 아이템들을 수 많은 고객들이 보유하고 있는 상태라면, 향상도 값을 기준으로 적용한다면 좀 더 유의미한 규칙과 패턴을 도출할 수 있는장점이 있다[29].

본 연구에서는 분석의 대상이 되는 고객과 거래 정보를 서로 유사하거나 연관성이 있는 개체와 같은 군집에 속할 수 있도록 그룹화하며, 다르거나 무관한 개체와는 다른 군집에 속하도록 그룹화할 수 있도록 비계층적 군집화(Non-Hierarchical Clustering) 방법을 활용하여 군집분석을 실시 한다. 군집분석을 토대로 생성된 군집은 해당 그룹 안의 데이터로부터 숨겨져 있는 유용한 패턴이나 의미있는 관계나 규칙을 찾아내고 탐색하기 위해 연관성 분석을 실시하고자 한다.

4. 실증 분석

4.1 군집분석의 결과

본 연구에서 군집분석의 목적은 최적의 군집 수를 결정하고, 생성된 군집들을 대상으로 연관성 있는 규칙과 유용한 패턴을 찾아냄으로써 군집마다 유의미하고 적절한 상품을 찾아내기 위함이다.

군집분석은 고객정보와 거래정보로 분류하여 3가지 방법으로 분석하였다. 첫 번째, 고객에 대한 군집을 형성할 때 가장 일반적으로 고객정보를 사용한다. 따라서 성별, 나이, 이용횟수 등에 대한 고객정보를 활용하여 고객군집 분석을 실시하였다. 두 번째, 동일한 고객이라고 하더라도 구매패턴이 변화하는 문제가 발생하기 때문에 거래기록인 총 무게, 거래개수, 즉석식품, 단품가격 등의 거래정보를 기반으로 군집분석을 실시하였다. 세 번째, 데이터를 세분화하기 위하여 고객군집과 상품군집의 분석결과를 교차결합(cross join)하여 결합군집이라 새로운 군집을 도출하여 분석하였다.

'M사'의 고객정보와 거래데이터를 활용하여 Python의 scikit-learn 패키지로 k-means Clustering 군집분석을 실시하였다. 고객 군집 수를 설정하기 위하여 Fig. 3과 같이, 상품 군집 수를 설정하기 위하여 Fig. 4와 같이 엘보우(Elbow) 기법을 적용하여 최적의 군집 수를 분석하였다. 분석결과, 군집의 수가 1개에서 2개로 변화될 때 변동성이 크게 감소하지만, 군집의 수가 3개 이후에는 변동성이 미비한 것으로 나타났다. 또한, 군집이 2개, 3개의 경우 소요시간이 짧은 것으로 나타났다. 이는 군집의 개수가 3개 일 때 최적임을 나타낸다고 할 수 있다.

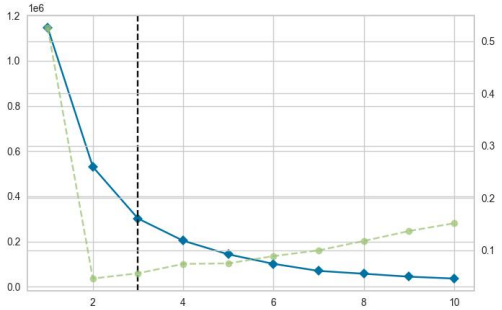


Fig. 3. Analysis number of customer clusters

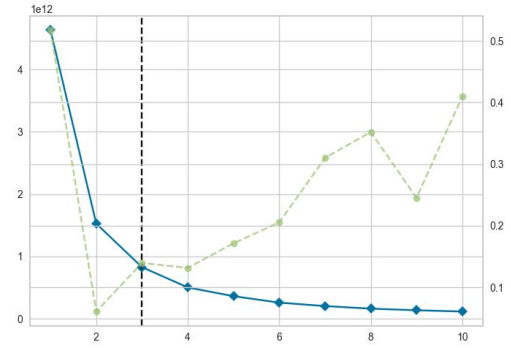


Fig. 4. Analysis number of product clusters

엘보우(Elbow) 기법 적용 결과 3개의 고객군집에 대한 기초 통계분석 결과는 Table 5와 같다. 고객군집 별 특성을 살펴보면, '고객군집 1'은 평균 나이가 가장 많으며, 평균 이용횟수도 다소 높은 것으로 나타났다. 또한 평균 지불금액은 상대적으로 가장 낮았으며, 구매하는 상품의 평균 무게도 가장 가벼운 것으로 나타났다. '고객군집 2'는 제품을 구매하는 평균 나이가 가장 젊으며, 평균적인 이용횟수도 가장 적게 이용하는 집단으로 나타났다. '고객군집 3'은 평균적으로 이용하는 횟수와 지불금액도 가장 높으며, 구매하는 상품의 무게 역시 상대적으로 무거운 상품을 구매하는 집단으로 나타났다. 군집간의 특성을 살펴본 결과, 평균적인 거래개수와 구매개수의 항목에서는 군집에 따라 차이가 없음을 확인할 수 있었다.

Table 5. Basic statistical analysis of customer clusters

Average	Total	Age	No. of use	Price	Weight	Transaction frequency	Purchase
Cluster 1	3,083	49.7	9.76	23,567	15,040	12	20.3
Cluster 2	2,377	28.0	7.03	24,938	16,235	12	21.9
Cluster 3	7,277	39.3	9.93	25,380	17,057	13	22.6

Table 6. Basic statistical analysis of product clusters

Average	Total	Age	No. of use	Price	Weight	Transaction frequency	Purchase	Fast food purchase (rate)
Cluster 1	2,644	40.0	10.31	20,853	8,430	9	15.0	0.00%
Cluster 2	5,005	39.9	9.61	22,474	5,934	10	18.0	100.00%
Cluster 3	5,088	39.5	8.60	29,286	30,875	17	29.4	93.81%

Table 7. Basic statistical analysis of cross-bonding clusters

Average	Total	Age	No. of use	Price	Weight	Transaction frequency	Purchase
Cluster1	690	50.0	10.16	19,567	8,169	8	14.7
Cluster2	1,288	50.0	9.77	21,498	5,783	10	17.0
Cluster3	1,105	49.3	9.48	28,477	30,120	16	27.6
Cluster4	464	27.5	7.08	21,507	8,723	9	14.9
Cluster5	966	27.7	6.71	23,147	5,914	10	18.0
Cluster6	947	28.6	7.37	28,448	30,444	17	29.3
Cluster7	1,490	39.2	11.38	21,243	8,460	9	15.1
Cluster8	2,751	39.4	10.57	22,693	6,012	10	18.4
Cluster9	3,036	39.3	8.66	29,843	31,284	18	30.2

엘보우(Elbow) 기법 적용 결과 3개의 상품군집에 대한 기초 통계분석 결과는 Table 6과 같다. 상품군집 1'은 평균 나이가 가장 많으며, 평균적인 지불금액, 거래개수, 그리고 구매개수가 가장 낮은 집단으로 나타났다. 즉석식품의 구매율은 0.00%로 구매를 전혀 하지 않는 것으로 나타났다. '상품군집 2'는 평균적인 지불금액과 평균 거래개수는 '상품군집 1'과 유사한 것으로 나타났지만, 무게 항목에서 가장 가벼운 상품을 구매하는 집단으로 나타났다. 즉석식품의 구매 항목에서는 구매 비율이 100.00%로 나타나 다른 군집과는 비교되는 특이점이라고 할 수 있겠다. '상품군집 3'은 평균적인 지불금액, 무게, 거래개수, 구매개수 등 가장 높은 수치를 나타내고 있는 군집으로 확인할 수 있었다. 특히, 평균 총무게 항목에서 다른 군집보다 평균 약 4~6배 이상 무거운 것으로 나타나 상당히 무게가 나가는 상품을 구매하는 군집임을 추측할 수가 있다. 군집 간의 특성을 살펴본 결과, 평균적인 이용횟수 항목에서는 군집에 따라 차이가 없음을 확인할 수 있었다.

고객군집과 상품군집의 교차결합을 통해 Table 7과 같이 더욱 세분화 된 9개의 군집으로 결합군집을 도출하였다. 결합군집은 고객군집과 상품군집의 분석결과보다 더욱 더 상세하고 차별성 있는 군집을 생성할 뿐만 아니라, 이를 활용하여 좀 더 유의미하고 유용한 패턴의 연관규칙 발견이 가능해졌다.

결합 군집에 대한 기초 통계분석 결과 9개의 군집 특성에 대하여 살펴보았다. '결합 군집 1, 4, 7' 그룹의 군집은 '상품군집 1'의 특성과 유사한 것으로 보이지만 고객군집의 특성으로 분류됨으로써 각 군집 간의 차이가 발생하고 있음을 확인할 수 있었다. 이를 토대로 연관성 분석 시 좀 더 세부적이고 숨겨져 있는 규칙을 발견할 수 있을 것으로 기대한다.

'결합 군집 1, 2, 3' 그룹의 군집은 모두 평균 나이는 50세, 이용횟수는 10회에 가깝게 나타났다. 이는 '고객군집 1'을 상품군집으로 세분화 한 군집으로 해석할 수 있음을 나타낸다. 따라서 상품군집의 특성을 반영하고 있는 '결합군집 1'은 평균적으로 지불하는 금액이 19,567원으로 가장 낮은 금액으로 나타났으며, '결합군집 2'는 평균 무게가 5,783g으로 가장 가볍게 나타났다.

'결합군집 3'은 평균 지불금액이 28,477원, 평균 무게는 30,120g으로 다른 군집들에 비해 상대적으로 높게 나타났다.

'결합 군집 4, 5, 6' 그룹의 군집은 평균 나이가 약 28세로 가장 어리며, 이용횟수 역시 평균 8회 미만으로 가장 적게 나타난 군집으로 나타났다. 이는 '고객군집 2'의 특성을 상품군집으로 세분화 한 군집으로 해석할 수 있음을 나타낸다. 해당 군집들의 거래정보를 분석해보면 '결합군집 4'는 평균 지불금액, 거래개수, 구매개수가 낮은 군집이며, '결합군집 5'는 평균 무게가 가벼운 제품들을 구매하는 군집이다. '결합군집 6'은 평균 거래개수, 구매개수가 많아짐으로써 평균 지불금액, 무게도 함께 높은 편에서 속하는 군집으로 분류된다.

'결합 군집 7, 8, 9' 그룹의 군집은 평균 나이가 약 39세로 나타났으며, 평균적인 이용횟수도 8회 이상을 보이고 있다. 이는 '고객군집 3'의 특성을 상품군집으로 세분화 한 군집으로 해석할 수 있음을 나타낸다. '결합군집 7'은 평균적인 지불금액이 낮고, 구매개수가 적은 고객들을 그룹화 한 군집이다. '결합군집 8'은 평균 무게가 가벼운 제품을 구매하는 군집으로 나타났다. '결합군집 9'는 평균 지불금액, 무게, 거래개수, 구매개수 등이 9개의 결합군집 중에서 가장 높은 것으로 나타났다. 이는 제품 구매 시 다량의 제품을 일괄적으로 구매하는 것으로 볼 수 있겠다.

4.2 연관성 분석의 결과

연관성 분석은 데이터 간에 유의미한 관계를 도출하고자 하는 탐색적 분석 방법으로, 지지도(support)를 기반으로 분석하기 때문에 데이터에서 빈발하는 속성이 높은 상품을 중심으로 연관성이 있는 규칙을 생성하게 된다. 본 연구에서는 고객군집, 상품군집 그리고 결합군집을 대상으로 연관성분석을 실행하였으며, 분석결과 도출 된 연관규칙을 비교·분석함으로써 결합군집을 중심으로 연관성 분석에 대한 상품추천의 유용성을 확인하고자 하였다. 각 분석에서 연관성 분석의 연관규칙이 10개 내외로 도출되도록 최소지지도와 최소신뢰도의 값을 설정하였다.

Table 8. Association rules by customer cluster

	no	Association rules (X → Y)		support	confidence	lift
Cluster 1	1	ice cream, frozen	cup·cone for ice	0.0097	0.341	9.55
	2	cup·cone for ice	ice cream, frozen	0.0097	0.273	9.55
	3	orange juice	grape juice	0.0094	0.192	7.05
	4	grape juice	orange juice	0.0094	0.345	7.05
	5	cucumber	green onion	0.0081	0.145	4.11
	6	green onion	cucumber	0.0081	0.229	4.11
	7	pumpkin	onion	0.0071	0.150	3.94
	8	onion	pumpkin	0.0071	0.188	3.94
	9	pumpkin	bean sprouts	0.0091	0.190	3.01
	10	bean sprouts	pumpkin	0.0091	0.144	3.01
Cluster 2	1	chicken·duck	pig	0.0084	0.290	4.89
	2	pig	chicken·duck	0.0084	0.142	4.89
	3	fried food	pork·fish cutlet	0.0084	0.202	2.89
	4	pork·fish cutlet	fried food	0.0084	0.120	2.89
	5	curry, jjajang	pig	0.0084	0.156	2.63
	6	pig	curry, jjajang	0.0084	0.142	2.63
	7	instant rice	tissue	0.0109	0.161	1.69
	8	tissue	instant rice	0.0109	0.115	1.69
	9	tissue	rice	0.0126	0.132	1.57
	10	rice	tissue	0.0126	0.150	1.57
Cluster 3	1	onion	potato	0.0122	0.149	2.81
	2	potato	onion	0.0122	0.231	2.81
	3	lettuce	pig	0.0124	0.333	2.61
	4	perilla leaf	pig	0.0111	0.297	2.32
	5	onion	green onion	0.0140	0.170	2.07
	6	green onion	onion	0.0140	0.170	2.07
	7	pig	chicken·duck	0.0133	0.104	1.93
	8	chicken·duck	pig	0.0133	0.247	1.93
	9	pig	green onion	0.0172	0.135	1.63
	10	green onion	pig	0.0172	0.208	1.63

Table 9. Association rules by product cluster

	no	Association rules (X → Y)		support	confidence	lift
Cluster 1	1	pumpkin	mushroom	0.0159	0.288	3.57
	2	mushroom	pumpkin	0.0159	0.197	3.57
	3	lettuce	pig	0.0087	0.261	2.88
	4	chicken·duck	pig	0.0083	0.259	2.85
	5	mushroom	chili (pepper)	0.0091	0.113	2.29
	6	chili (pepper)	mushroom	0.0091	0.185	2.29
	7	mushroom	onion	0.0083	0.103	1.86
	8	onion	mushroom	0.0083	0.150	1.86
	9	fish cake	bean sprouts	0.0091	0.178	1.77
	10	pumpkin	pig	0.0083	0.151	1.66
Cluster 2	1	pumpkin	bean sprouts	0.0112	0.199	2.38
	2	bean sprouts	pumpkin	0.0112	0.134	2.38
	3	pig	bean sprouts	0.0114	0.118	1.42
	4	bean sprouts	pig	0.0114	0.137	1.42
	5	curry, jjajang	pig	0.0124	0.133	1.38
	6	pig	curry, jjajang	0.0124	0.128	1.38
	7	curry, jjajang	dumpling	0.0130	0.139	1.18
	8	dumpling	curry, jjajang	0.0130	0.111	1.18
	9	tuna	coffee	0.0106	0.120	1.05
	10	pig	dumpling	0.0104	0.108	0.92
Cluster 3	1	potato	onion	0.0108	0.258	3.98
	2	onion	potato	0.0108	0.167	3.98
	3	lettuce	pig	0.0106	0.370	3.65
	4	potato	pig	0.0106	0.254	2.50
	5	onion	pig	0.0161	0.248	2.45
	6	pig	onion	0.0161	0.159	2.45
	7	chicken·duck	pig	0.0108	0.244	2.41
	8	carrot	green onion	0.0114	0.215	1.99
	9	onion	green onion	0.0130	0.200	1.86
	10	cucumber	green onion	0.0149	0.185	1.72

고객군집과 상품군집의 연관성 분석 결과는 Table 8, 9와 같이 나타났다. 구체적으로 고객 군집의 연관성 분석 결과, 군집 별 연관규칙을 탐색할 수 있었다. 다만, 음영처리 된 '파', '양파', '닭·오리고기', '돼지고기' 등의 제품은 모든 군집의 연관규칙에 나타나는 것으로 확인되었다. 특히, '고객군집 2, 3'에서는 같거나 유사한 연관규칙이 탐색 되었다. 상품군집의 연관성 분석 결과 역시 군집 별 연관규칙을 탐색할 수 있었다. 고객군집에서 연관성분석 결과와 동일하게 '애호박', '닭·오리고기', '상추', '돼지고기', '양파', '콩나물' 등의 많은 제품이 모든 군집에서 연관규칙이 나타나는 것을 확인할 수 있었다. 특히 '닭·오리고기 → 돼지고기', '상추 → 돼지고기' 등의 연관규칙은 여러 군집에서 동시에 같은 형태로 나타나는 것을 확인할 수 있었다. 이처럼 고객군집과 상품군집에서의 연관성 분석은 각각 생성 된 군집에서 다양한 종류의 상품과 관련하여 연관규칙이 생성되는 것과 함께 제품들의 다수가 중복(음영)되는 것을 확인할 수 있다. 이는 유사하거나 동일한 연관규칙이 반복해서 나타나고 있음을 의미하며, 효과적인 상품 추천을 위해서는 부적절한 모델이라고 할 수 있다.

효율적인 상품을 추천하기 위해서는 세분화 된 군집별로 상품의 종류가 중복되지 않고 다양한 연관 규칙이 많이 생성되는 것이 효과적이다. 하지만 고객군집과 상품군집의 연관성 분석의 결과는 다수의 상품이 2개 이상의 군집에서 나타났으며, 2개의 중복 된 연관 규칙도 도출 되었다. 따라서 본 연구에서는 고객군집과 상품군집을 교차결합함으로써 좀 더 데이터를 세분화하여 유용성 있은 군집을 생성함으로써 적합한 분석 결과를 도출하여 상품 추천을 하고자 하였다.

결합군집의 연관성 분석 결과는 Table 10과 같이 나타났으며, 평균 8.0개의 유일한 상품이 도출되었다. 앞서 분석된 고객군집의 5.3개와 상품군집의 3.3개 보다 상대적으로 유일한 상품의 개수가 더 많이 도출된 것을 확인할 수 있었다. 또한, 연관규칙의 중복률(개수)은 결합군집이 평균 3.3%(3개)로 나타나 고객군집 6.6%(2개)와 상품군집 6.6%(2개) 보다 1/2배 이상 더 적게 중복된 것으로 분석되었다. 이를 통해 교차결합하여 설계된 결합군집을 활용하여 맞춤형 상품을 추천하는 것이 더 효율적인 효과를 나타내는 것이 확인 되었다.

Table 10. Association rules by cross-bonding cluster

	no	Association rules (X → Y)		support	confidence	lift
Cluster 1	1	bag, trash basket	cleaning supplies	0.0130	0.231	4.30
	2	cleaning supplies	bag, trash basket	0.0130	0.243	4.30
	3	chili (pepper)	mushroom	0.0130	0.273	3.04
	4	pumpkin	mushroom	0.0203	0.259	2.89
	5	eggplant	stationery	0.0203	0.226	2.89
	6	rice	bath supplies	0.0116	0.157	2.35
	7	bath supplies	Daily necessities	0.0116	0.174	2.35
	8	onion	mushroom	0.0101	0.206	2.29
	9	food for baby	pumpkin	0.0101	0.167	2.13
	10	notebook, pencile	spinach	0.0130	0.167	2.02
Cluster 2	1	tissue	wet wipes	0.0142	0.211	2.85
	2	wet wipes	tissue	0.0142	0.192	2.85
	3	bath supplies	wet wipes	0.0112	0.172	2.33
	4	wet wipes	bath supplies	0.0112	0.152	2.33
	5	pumpkin	bean sprouts	0.0142	0.202	1.93
	6	red pepper sauce	salt	0.0142	0.136	1.93
	7	tuna	wet wipes	0.0135	0.129	1.74
	8	wet wipes	anchovy, kelp	0.0135	0.182	1.74
	9	Chinese cabba	salt	0.0105	0.149	1.42
	10	air freshener	Oral care	0.0120	0.120	1.15
Cluster 3	1	ice cream, frozen	cup·cone for ice	0.0180	0.465	11.01
	2	cup·cone for ice	ice cream, frozen	0.0180	0.426	11.01
	3	orange juice	grape juice	0.0153	0.230	5.44
	4	grape juice	orange juice	0.0153	0.362	5.44
	5	orange juiceb	bag, trash basket	0.0108	0.162	1.82
	6	bag, trash basket	green onion	0.0162	0.182	1.82
	7	energy drinks	sports drinks	0.0162	0.162	1.82
	8	old snacks	tissue	0.0144	0.143	1.49
	9	tissue	old snacks	0.0144	0.150	1.49
	10	soybean milk	carbonated water	0.0126	0.141	1.47
Cluster 4	1	milk	soup, sauce	0.0108	0.217	4.39
	2	soup, sauce	milk	0.0108	0.217	4.39
	3	candy bar	chocolate	0.0108	0.250	3.74
	4	ketchup	olive oil	0.0108	0.217	3.36
	5	rice cake	ketchup	0.0108	0.200	2.99
	6	wet wipes	olive oil	0.0108	0.192	2.97
	7	pizza	pig	0.0108	0.313	2.84
	8	chocolate	Instant soup, side dish	0.0108	0.200	2.21
	9	flour	bean sprouts	0.0129	0.240	2.06
	10	gravy	pie, cake	0.0216	0.185	1.68
Cluster 5	1	perilla leaf	cucumber	0.0105	0.297	5.75
	2	cucumber	perilla leaf	0.0105	0.204	5.75
	3	wafers	chocolate	0.0105	0.186	2.46
	4	chocolate	Instant soup, side dish	0.0105	0.139	2.46
	5	Instant soup, side dish	curry, jjajang	0.0105	0.145	1.54
	6	udon	dumpling	0.0105	0.200	1.54
	7	curry, jjajang	dumpling	0.0172	0.184	1.41
	8	dumpling	curry, jjajang	0.0172	0.132	1.41
	9	chocolate	udon	0.0105	0.151	1.16
	10	Pigs' Feet, sundae	kimchi	0.0144	0.138	1.06
Cluster 6	1	spaghetti	spaghetti sauce	0.0147	0.333	5.68
	2	spaghetti sauce	spaghetti	0.0147	0.250	5.68
	3	ketchup	mayonnaise	0.0126	0.194	3.77
	4	mayonnaise	ketchup	0.0126	0.245	3.77
	5	fried food	pork·fish cutlet	0.0126	0.245	2.92
	6	seasoning	spaghetti	0.0105	0.244	2.71
	7	low-fat milk	rice	0.0115	0.282	2.42
	8	ssamjang	instant rice	0.0115	0.216	2.29
	9	rice soup	tissue	0.0115	0.289	2.26
	10	banana	tissue	0.0136	0.271	2.12

Table 10. (Continued)

	no	Association rules (X → Y)		support	confidence	lift
Cluster 7	1	rice cake	pickled radis	0.0120	0.214	3.42
	2	pickled radis	rice cake	0.0120	0.191	3.42
	3	crab stick	pickled radis	0.0120	0.205	3.27
	4	pickled radis	crab stick	0.0120	0.191	3.27
	5	mushroom	green onion	0.0133	0.132	2.84
	6	green onion	mushroom	0.0133	0.286	2.84
	7	rice cake	ketchup	0.0133	0.227	2.53
	8	fish cake	crab stick	0.0133	0.148	2.53
	9	sugar	starch syrup	0.0107	0.190	2.12
	10	fish cake	garlic	0.0107	0.178	1.72
Cluster 8	1	chicken·duck	pig	0.0106	0.286	2.82
	2	onion	pig	0.0109	0.168	1.66
	3	tuna	coffee	0.0165	0.167	1.65
	4	pig	bean sprouts	0.0165	0.163	1.65
	5	시리얼	tuna	0.0127	0.156	1.45
	6	tissue	tuna	0.0106	0.136	1.26
	7	season the meat	frozen rice	0.0134	0.125	1.08
	8	pepper, spice	onion	0.0120	0.132	1.08
	9	oyster sauce	chicken·duck	0.0113	0.125	1.04
	10	season the pork	pig	0.0109	0.115	1.02
Cluster 9	1	carrot	potato	0.0128	0.234	3.35
	2	beef	pig	0.0125	0.396	2.34
	3	carrot	green onion	0.0138	0.251	2.24
	4	lettuce	pig	0.0177	0.370	2.19
	5	daikon	green onion	0.0102	0.214	1.91
	6	season the pork	pig	0.0128	0.273	1.61
	7	rice	sweet potato	0.0295	0.263	1.56
	8	potato	rice	0.0177	0.254	1.50
	9	beef	cheese	0.0115	0.211	1.25
	10	seaweed	pig	0.0108	0.202	1.20

5. 결론

정보통신 기술은 계속해서 발전하고 있으며, 고객이 상품을 쇼핑하고 구매하는 패턴에 상당부분 영향을 끼치고 있다. 과거 오프라인의 구매 형태에서 현재는 온라인 구매 형태의 패턴으로 간편하게 쇼핑을 할 수 있게 되었으며, 더욱이 COVID-19의 영향으로 외출 대신 온라인으로 상품을 주문하는 등 집안에서 모든 것을 해결하려는 형태로 생활 패턴이 빠르게 변화하고 있다. 소비자는 제품·서비스를 선택함에 있어 쉽고 빠르게 원하는 정보를 파악하고 구매하는 것이 중요한 요소가 된 것이다. 따라서 급변하고 있는 경영 환경 측면에서 소비자들의 구매 욕구를 향상시키면서, 니즈를 충족시킬 수 있는 판매 전략이 필요하다.

본 연구에서는 고객 맞춤형 상품을 추천하는 모델을 설계하기 위하여 온라인 배송업체 'M사'의 고객정보 및 거래정보 데이터 182,886건을 토대로 군집분석과 연관성분석을 하였다. 효과적인 고객 맞춤형 상품을 추천하기 위해서는 적절한 군집으로 세분화해야 한다. 고객군집, 상품군집, 그리고 교차결합을 통해 데이터를 세분화시켜 결합군집을 생성하여 새로운 방안의 군집분석을 시도하였다. 각각의 군집분석 결과를 토대로 연관성 분석을 하였으며, 연관성 분석 결과 도출된 연관규칙을 비교·분석함으로써 각 군집 별 성능을 평가하였다.

연관성 분석 결과, 고객군집과 상품군집은 3개의 군집에서 30가지의 연관 규칙을 도출하였으며, 고객군집 평균 5.3개, 상품군집 평균 3.3개의 유일한 상품을 구매하는 것으로 분석되었다. 또한, 중복률 확인결과를 보면 고객군집과 상품군집 모두 동일하게 평균 6.6%(2개)가 중복되는 것으로 나타났다. 반면, 결합군집에서는 9개의 군집에서 90가지의 연관 규칙을 도출하였으며, 고객군집 평균 8.0개의 유일한 상품을 구매하는 것으로 분석되었다. 또한 중복률은 평균 3.3%(3개)으로 분석되어 고객군집과 상품군집의 분석결과보다 효과적임을 알 수 있다. 따라서 고객의 니즈에 맞게 맞춤형 상품을 추천하기 위해서는 결합군집 모델이 가장 효과적인 군집으로 분석되었다.

전체적인 결과를 종합하면 고객정보와 거래정보를 교차결합하여 군집을 한 번 더 세분화한 결합군집이 더 많

은 상품의 연관규칙이 도출되었으며, 중복률 또한 적게 나타나 더욱 유의미한 결과를 나타냈다. 이는 고객의 니즈에 맞게 상품을 추천하기 위해서는 결합군집이 가장 적합한 모델이라고 판단된다. 온라인 시장에서 결합군집 모델을 적용한다면, 소비자에겐 유용한 정보를 제공하면서, 해당 업체에는 판매량을 증가시키는 등의 긍정적인 효과를 가져올 것으로 기대된다. 하지만, 본 연구에서는 온라인 배송업체 'M사'의 데이터만으로 분석을 실시하여 범위적인 부분에서 한정적이라는 한계가 있어, 향후 좀 더 다양한 특성을 갖고 있는 온라인 배송업체들을 대상으로 비교하고 분석한다면 좀 더 유의미한 연구가 될 것으로 사료된다.

REFERENCES

[1] J. M. Lee. (2020). *A Study of Diners' Purchase Association Rule Using Data Mining Methods*. Doctoral dissertation. KH University, Seoul.

[2] M. J. Shaw, C. Subramaniam, G. W. Tan & M. E. Welge. (2001). Knowledge management and data mining for marketing. *Decision support systems, 31 (1)*, 127-137.

[3] S. S. Oh. (2018). *A Study on the Customer Profile based Recommendation System using Association Rules Analysis for Online Duty Free Stores*. Master's Degree. IH University, Incheon.

[4] D. S. Jin & J. W. Lee. (2012). Impacts of Social Commerce in E-commerce: In perspective of Social Commerce Analysis Model. *Korea Association for International Commerce and Information, 14(1),* 369-390.

[5] Korea Information Society Development Institute. (2020). *Analysis of changes in e-commerce usage behavior due to COVID-19*. [Brochure]. Jincheon : Y. S. Oh.

[6] C. Lin. (2021). *Implementation of E-commerce Personalized Recommendation System Based on Web Data Mining*. Doctoral dissertation. HN University, Gwangju.

[7] B. C. Jang. (2002). *A study on multi-criteria individualized commodity recommendation in e-shopping mall*. Master's Degree. SKK University, Seoul.

[8] S. S. Kim. (2012). *An improved product recommender system based on association rules using extended information sources*. Master's Degree. KM University, Seoul.

[9] E. S. Won & S. Y. Kim. (2020). An Analysis of Consumers Purchasing Patterns for Fresh Food Products Using Association Rules. *Journal of Agriculture & Life Sciences, 54(4)*, 111-122.

[10] J. P. Ryu & H. J. Shin. (2021). Big Data Analysis of Financial Product Transaction Trends Using Associated Analysis. *Journal of the Korea Convergence Society, 12(12),* 49-57.

[11] Y. B. Cho, J. H. Jun & B. Choi. (2019). A Methodology for Improving fitness of the Latent Growth Modeling using Association Rule Mining. *Journal of the Korea Convergence Society, 10 (2),* 217-225.

[12] J. M. Park, K. R. Park & Y. S. Chung. (2018). Analysis of relationship between frequency of crime occurrence and frequency of web search. *Journal of the Korea Convergence Society, 9(5),* 15-20.

[13] D. H. Shin, M. J. Kim, S. Y. Oh & K. Chung. (2019). Knowledge Reasoning Model using Association Rules and Clustering Analysis of Multi-Context. *Journal of the Korea Convergence Society, 10(9),* 11-16.

[14] S. Y. Yoon. (2018). *A study on the importance of clustering in prediction model construction: Through consumer case analysis*. Master's Degree. HS University, Asan.

[15] D. H. Kim. (2018). A Study on the Customer Segmentation Using Purchase History Big Data for Target Marketing. Master's Degree. SK University, Seoul.

[16] M. U. An, E. S. Won, S. Y. Kim & D. H. Yoo. (2019). Development of Sales Strategies for Agricultural Products Using Lift-based Association Rules Network: A Focus on Large Supermarkets and Traditional Markets. *Korea Internet Electrornic Commerce Association, 19 (3),* 105-127.

[17] M. R. Anderberg. (1973). *Cluster analysis for applications*. New York : Academic Press.

[18] J. W. Jo. (2006). *Classification Analysis by Using the K-Means Clustering*. Master's Degree. JA University, Seoul.

[19] H. Y. Woo & C. H. Park. (2013). Active Learning based on Hierarchical Clustering. *Korea Information Processing Society, 2(10)*, 705-712.

[20] G. Shmueli, N. R. Patel & P. C. Bruce. (2011). *Data Mining for Business Intelligence: Concepts, Techniques, and Applications in Microsoft Office Excel with XLMiner*. John Wiley & Sons.

[21] J. Jeon. (2015). *Verification for patent data clustering based on text-mining.* Master's Degree. SKK University, Seoul.

[22] Y. S. Lee, P. They, J. H. Lee & J. M. Kil. (2018). A Study on Research Paper Classification Using Keyword Clustering. *Korea Information Processing Society, 7(12)*, 477-484.

[23] J. M. Ko, K. S. Jang & D. J. Hwang. (2005). *Understanding and Using Data Mining.* Ulsan : USU Press.

[24] J. Y. Park, T. W. Lee, C. L. Wang & T. H. Hong. (2012). Data Mining for Relationship Recommendation in Social Networks. *The Korea Society of Management Information Systems Conference, 2012(1)*, 508-512.

[25] K. S. Nam, H. J. Kim & J. H. Oh. (2002). Analysis for simultaneous activities using the data mining's association rule. *Korea Social Research Center, 17 (1)*, 37-156.

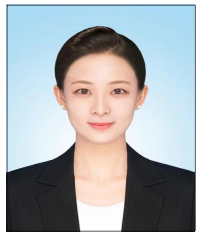
[26] J. S. Kim, Y. A. Do, J. W. Ryu & M. W. Kim. (2001). A Collaborative Recommendation System Using Neural Networks for Increment of Performance. *The Korean Brain Society, 1(2)*, 233-244.

[27] H. Bak, J. H. Kim & Y. J. Kim. (2013). An Analysis for Deriving New Convergent Service of Mobile Learning : The Case of Social Network Analysis and Association Rule. *The Korea Society of Management Information Systems, 15(3)*, 1-37.

[28] H. C. Ahn, I. K. Han & K. J. Kim. (2006). The Product Recommender System Combining Association Rules and Classification Models: The Case of G Internet Shopping Mall. *The Korea Society of Management Information Systems, 8(1)*, 181-201.

[29] D. Wielenga, B. Lucas & J. Georges. (1999). *Enterprise Miner: Applying Data Mining Techniques Course Note,* SAS Institute Inc, Cary, NC.

MingFei Yang [정회원]



· 2019년 2월 : 충북대학교 경영정보학과(경영학사)
· 2021년 8월 : 충북대학교 경영정보학과(경영학석사)
· 관심분야 : 정보보호, 빅데이터
· E-Mail : 506325026@qq.com

박 기 용(Kiyong Park) [정회원]

· 2018년 8월 : 충북대학교 도시공과(공학박사)
· 2020년 3월 ~ 2021년 8월 : Adjunct Professor, Urban and Regional Planning, Michigan State Univ.
· 2021년 3월 ~ 현재 : 충북대학교 빅데이터협동과정 연구교수
· 관심분야 : 빅데이터, 통계
· E-Mail : pky3489@chungbuk.ac.kr

최 상 현(Sang-Hyun Choi) [정회원]

· 1998년 2월 : 한국과학기술원 경영정보공학(공학박사)
· 1996년 9월 ~ 2002년 9월 : LG CNS 책임컨설턴트
· 2002년 10월 ~ 2011년 8월 : 국립경상대학교 산업시스템공학부
· 2011년 9월 ~ 현재 : 충북대학교 경영정보학과 교수
· 관심분야 : 빅데이터, 데이터마이닝, 스마트팩토리
· E-Mail : chois@cbnu.ac.kr