석사학위 논문

온라인 소비자물가지수 작성을 위한 온라인 가격 데이터 정제에 관한 연구

A Study on Filtering of Online Price Data for Creating an Online Consumer Price Index

2017년 6월

숭실대학교 대학원

IT융합학과

박 원 배

석사학위 논문

온라인 소비자물가지수 작성을 위한 온라인 가격 데이터 정제에 관한 연구

A Study on Filtering of Online Price Data for Creating an Online Consumer Price Index

2017년 6월

숭실대학교 대학원

IT융합학과

박 원 배

석사학위 논문

온라인 소비자물가지수 작성을 위한 온라인 가격 데이터 정제에 관한 연구

지도교수 정 윤 원

이 논문을 석사학위 논문으로 제출함

2017년 6월

숭실대학교 대학원

IT융합학과

박 원 배

박 원 배의 석사학위 논문을 인준함

심 사 위 원 장 유 명 식 인

심 사 위 원 김 동 성 인

심 사 위 원 정 윤 원 인

2017년 6월

숭실대학교 대학원

목 차

표 목 차

그 림 목 차

국문초록

온라인 소비자물가지수 작성을 위한 온라인 가격 데이터 정제에 관한 연구

박 원 배

IT융합학과

숭실대학교 대학원

국가 통계로 활용되고 있는 소비자 물가지수(CPI)는 전국적으로 오프라인 대형 마켓 위주의 현장 조사이기 때문에 많은 시간과 비용을 소비하고 있다. 그리고 최근에는 많은 소비자들의 구매 패턴이 오프라인 마켓에서 온라인 마켓으로 이동하고 있어 온라인 소비자 물가지수의 개발이 필요한 시점이다. 그러나 아직까지 국내에서는 온라인 부분의 소비자 물가지수에 대한 연구가 더디게 진행되고 있다.

온라인 소비자 물가지수는 일반적으로 인지도가 높은 온라인 마켓에서 웹크롤링을 통해 대량의 데이터를 수집하여 안전한 저장소에 저장하고 물가지수의 정확성을 위해 수집된 데이터를 정제한다. 그리고 정제된 데이터를 계산식에 적용하여 물가지수를 작성하는 등 많은 복잡한 과정이 필요하다. 이러한 과정 중 특히 수집된 온라인 가격 정보를 정제하는 작업은 최종 물가지수의 정확성을 위해서 매우 중요하다. 따라서 본 연구에서는 온라인 가격의 특성을 파악하고 이에 따른 정확한 물가지수 산출을 위한 데이터 정제 방안을 제시하고자 한다.

ABSTRACT

A Study on Filtering of Online Price Data for Creating an Online Consumer Price Index

PARK, WONBAE

Department of IT Convergence

Graduate School of Soongsil University

The Consumer Price Index(CPI), used for official statistics, is costly and time consuming due to the fact that it is the result of field investigations. And nowadays, amounts of online consumption are growing more than those of offline. In spite of the need of developing the Online Consumer Price Index, it still progresses slowly within the country.

The Online Consumer Price Index is made through much complicated processes. First, we need to collect data from generally popular online market by web crawling. These data are saved in storage of safety which is available to store large amounts of data. Next, they need to be cleaned for accurate measures of the Online Consumer Price Index. Lastly, we need to apply those data to certain

formula to generate the Online Consumer Price Index. In particular, data cleaning is much important process to generate accurate numerical values of the Online Consumer Price Index. Therefore, this study figures out the peculiarities of online prices and suggests the way to clean data for the exactitude Online Consumer Price Index.

제 1 장 서 론

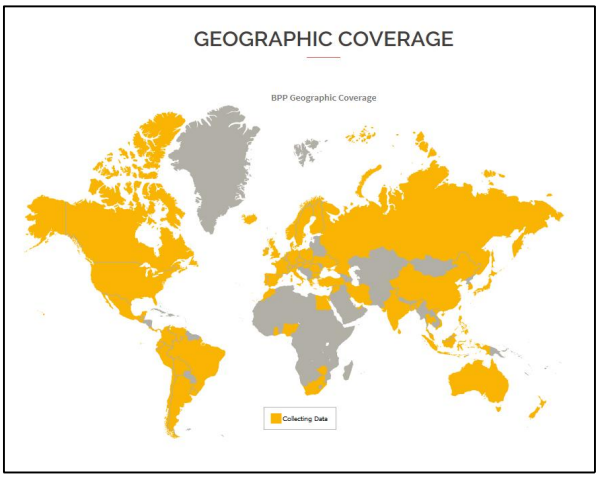
1.1 현 소비자물가지수의 한계

통계청의 통계 설명 자료에 의하면 소비자물가지수(CPI)의 조사 목적은 “상품과 서비스의 가격을 조사하여 소비자물가지수를 작성하고 그 결과를 정부 재정, 금융정책의 기초자료로의 이용하고, 가계수지, 국민소득 계정 등 다른 경제지표의 디플레이터로 사용”[1] 이라고 한다. 즉, 경제지표의 디플레이터로 사용하기 위해 매월 소비자물가지수를 작성하여 발표하고 있는 것이다. 2017년 초 현재 소비자물가지수는 소비자가 일상생활을 영위하기 위해 구입하는 상품과 서비스 중 총 460개의 지정된 품목별로 조사 대상 상품의 규격을 정하여 매월 지역별로 소매점 및 오프라인 대형 마트, 서비스업체에 조사원이 현장 방문을 통하여 가격 정보를 수집하고 소비자물가지수를 작성한다. 이러한 소비자물가지수는 월 단위로 발표하고 있으나, 월 단위 물가추이로는 인플레이션의 변화를 감지하는데 한계점이 있다. 또한 소비자물가지수를 작성함에 있어 오프라인 대형 마켓을 중심으로 현장 조사를 실시하므로 전국적으로 많은 인력과 시간, 비용이 소비되며, 여러 논문에서 소비자가 실제 체감하는 물가지수와는 괴리가 있음을 지적하고 있다[2]. 국내의 상품 및 서비스의 구매 패턴을 보면 여전히 오프라인 위주의 구매율이 온라인의 구매율 보다는 높긴 하지만 전자제품, 일반적인 가정용품, 서적 등의 구매율은 오프라인 구매율을 추격하고 있다. 2014년 논문 인터넷쇼핑의 성장이 소비자 물가에 미친 영향[7]에서 인터넷쇼핑 판매액이 연간 10% 증가하면 소비자물가는 연간 0.08% 감소하며, 인터넷쇼핑의 거래 비중이 높은 상품 군(여자 의류, 신발, 가정용 섬유제품, 가정용 기구, 항공 및 수상여객, 전화 및 팩스 장비, 서적, 단체 여행 등)의 물가는 인터넷쇼핑 판매액 증가에 음

(陰)의 영향을 받는 것으로 연구되었다. 이렇듯 온라인 전자상거래가 소비자물가지수에 영향을 주고 있으며, 소비자의 구매 패턴이 오프라인에서 온라인으로 변화하는 상황에서 현재의 소비자물가지수는 온라인의 물가지수 상황을 반영하지 못하고 있어 온라인 소비자물가지수의 개발이 필요한 시점이다.

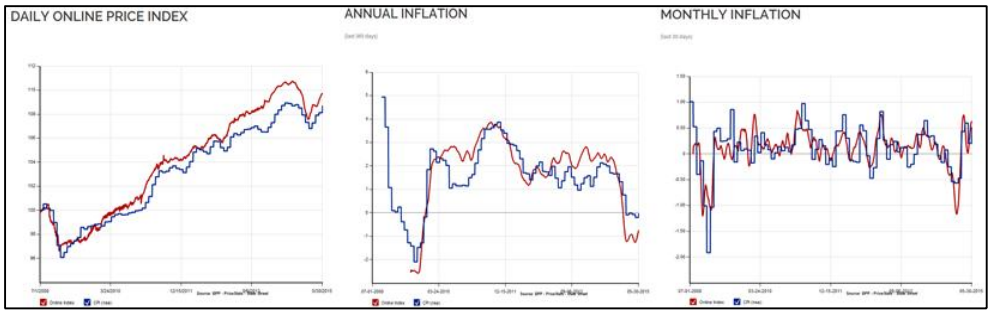
1.2 온라인 소비자물가지수의 현황

온라인 소비자물가지수는 2000년대 중반부터 미국의 매사추세츠 공과대학(MIT)에서 BPP(Billion Prices Project)라는 이름으로 전 세계 몇몇 국가의 대표 온라인 마켓으로부터 가격 정보를 수집하여 연구 중에 있다. 아래 그림은 BPP사이트에서 전 세계의 가격정보를 수집하고 있는 국가를 나타내고 있으며[3], 여기에는 우리나라도 포함되어 있다.



[그림 1-1] BPP 가격 수집 대상 국가

또한 BPP 웹사이트에서는 Daily online price index, Annual inflation, Monthly inflation 정보를 제공해 주고 있다[3]. 그러나 공개되는 자료는 Us Daily Index만을 대상으로 한다.



[그림 1-2] BPP Us Daily Index

MIT의 BPP에서는 데이터의 수집 대상을 선정할 때 아마존과 같은 온라인 전용 소매점은 제외하며, 월마트와 같은 온라인과 오프라인 판매를 함께 하는 다채널 판매 소매점을 중심으로 데이터를 수집하고 있다.

또한 제 3자의 가격 제공자에 의한 수집은 배제하고 각 소매점 사이트를 대상으로 웹 크롤러를 이용하여 가격 데이터를 직접 수집하는 방법을 사용하고 있다. 이렇게 각 소매점을 대상으로 데이터를 직접 수집하는 방법은 더 많은 노력이 필요하지만, 실제 거래와 연계된 가격을 수집할 수 있는 기회를 극대화하고 제 3자가 데이터를 가공하는 것을 막을 수 있기 때문이라고 한다. 수집된 데이터는 공용 데이터베이스로 표준화를 수행하고 각 상품의 품목 분류를 수행한다.

수집 대상 소매점의 상품을 온라인 물가지수에 포함시키기 전에 일반적으로 1년 넘게 해당 소매점에서 수집한 데이터를 분석하여 신뢰할 수 있는 가격 정보인지를 파악하고 있다. 그러나 현재 MIT의 BPP에서 공개하고 있는 정보는 데이터 수집 및 분류, 정제 등의 내용이 상세하지 않아, 우리나라에 맞는 온라인 소비자물가지수를 작성하기 위한 별도의 연구가 반드시 필요하다.

국내에서는 2013년에 행정자치부와 통계청이 함께 온라인 소비자물가지수 작성을 위한 시스템을 개발하였다. 국내에서 개발한 온라인 소비자물가지수 작성 시스템은 데이터 수집 단계에서 상품의 카테고리와 물가지수 품목을 맵핑하고, 몇 가지 데이터 정제 프로세스를 거쳐 온라인 소비자물가지수를 작성하고 있다.

또한 통계청은 2016년 11월 전자신문 기사에서 “통계청은 5년 주기로 그동안 변화를 반영해 소비자물가지수를 개편한다. 현재 사용하는 소비자물가지수는 2010년 기준이다. 종전에는 483개 품목 954개 제품 가운데 52개 제품의 온라인쇼핑 거래가격을 반영했다. 제품 수 기준으로 5.45%

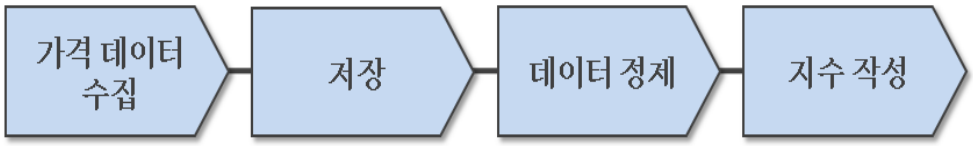
다. 통계청은 이 비중을 약 10%로 확대한다. 100개 전후 제품의 온라인 쇼핑 거래가격이 소비자물가지수에 반영되는 것이다."[4] 라고 밝혀 온라인 가격에 대한 물가지수 반영이 진행 중임을 확인 하였다.

특히 이 뉴스에서 "소비자물가지수 보조 성격의 '온라인 소비자물가지수'도 개발한다. 이마트 등 대형마트, 11번가 등 오픈마켓 제품 판매가격을 조사해 하루 단위로 물가 변화 정보를 제공할 것이다."[4] 라고 밝혔다. 이렇듯 온라인 소비자물가지수에 대한 개발이 국가 차원에서 진행되고 있다.

2015년에는 BPP를 기반으로 한 온라인 상품의 댓 글 정보 양에 따라 판매량을 추측하고, 수집된 상품에 추측된 판매량에 따른 가중치를 부여하여 온라인 소비자물가지수를 작성하는 것에 관한 연구가 석사학위 논문으로 작성된바가 있다[5]. 하지만 해당 논문에서는 온라인 소비자물가지수 작성에만 집중하고 있으며 데이터 품질을 보장할 수 있는 데이터 정제에 대해서는 연구가 진행되지 않았다.

한편 2010년대 들어 화두가 되었던 감성분석을 소비자물가지수에 적용하고자 하는 발표도 있었다. 한국데이터사이언스학회 2013년 학술대회에 자료 "온라인 물가지수 분석을 위한 빅데이터 융합분석 방법"[8]에 의하면 소비자가 체감하는 소비자물가지수를 소셜 감성분석을 활용하여 정량적으로 수치화하고, 이를 소비자물가지수와 함께 시각화하여 비교 분석함으로서 두 지수 간의 차이를 모니터링하는 방법이 제시되었다.

지금까지 설명한 온라인 소비자물가지수는 일반적으로 다음 [그림 1-1]과 같은 과정으로 작성될 것이다.



과정	설명
가격 데이터 수집	온라인 마켓에서 상품의 가격 데이터 수집
저장	수집된 데이터를 안전한 대용량 저장소에 저장
데이터 정제	온라인 소비자물가지수 작성을 위한 데이터의 품질 정비
지수 작성	품질이 정비된 데이터를 이용하여 온라인 소비자물가지수 작성

[그림 1-3] 온라인 소비자물가지수 작성 프로세스

온라인 마켓에서 수집 가능한 가격 데이터는 오프라인에서 수집하는 모든 품목을 대상으로 할 수 없다. 예를 들어 통계청의 소비자물가지수 품목에 포함되어 있는 상하수도료, 공동주택관리비, 김치찌개백반, 김밥 등과 같은 품목들이다. 가격 데이터의 수집은 국내 소비자들이 많이 이용하는 온라인과 오프라인 판매 채널을 모두 보유한 대표적인 온라인 마켓을 기준으로 하여야 한다. 이렇게 수집된 데이터는 유실되지 않도록 안전한 저장소에 저장하고 소비자물가지수 품목별로 정확하게 분류되고 오류 데이터를 정제하여야 작성된 온라인 소비자물가지수의 품질을 보장할 수 있다.

1.3 현 소비자물가지수(CPI)와 온라인 소비자물가지수

현재의 소비자물가지수와 온라인 소비자물가지수를 아래 [표 1-1]과 같이 간단히 비교해 보았다.

[표 1-1] 소비자물가지수와 온라인 소비자물가지수의 특성

항목	현 소비자물가지수	온라인 소비자물가지수
조사대상처	일반시장, 백화점, 할인점 등에서 조사권역의 가격 대표성이 높은 곳에 대하여 전국적으로 약 26,000곳을 선정하여 조사	시장점유율이 높거나 또는 특정 품목의 경우 가격을 대표할 수 있는 온라인 상품 판매 사이트(다양한 판매 채널을 보유한 소매점)
조사품목	소비자가 일상생활에서 구입하는 상품과 서비스 460개 품목(2016년부터 481개에서 460개 품목으로 조정됨)	온라인으로 가격정보 수집이 가능한 약 290여개 품목
조사규격	시장 점유율이 높고, 계속적으로 가격을 조사할 수 있는 상품과 서비스를 유형과 성격, 거래 조건 등을 종합적으로 고려하여 선정	수집 가능한 모든 상품을 대상으로 가격 정보를 수집하기 때문에 조사규격이 따로 없음
조사주기	월	일
조사방법	면접(방문)조사 원칙	온라인 가격정보 수집
물가지수 작성	기준시점고정 가중산술평균법(라스파이레스 산식)을 사용하여 계산	가중치 없는 기하평균을 사용하여 구한 개별 품목 가격 변동을 합산하여 가중산술평균법으로 계산

본 연구에서는 매월 발표되고 있는 소비자물가지수의 460개 품목 중 통계청에서 수집하여 공개하고 있는 온라인 상품 가격 데이터를 활용하여 온라인 소비자물가지수 작성을 위한 데이터 정제 방안을 제시하고, 각 데이터 정제 단계별 기술 통계 데이터를 비교한다. 또한 각 품목별로 데이터 정제 전과 후로 구분하여 온라인 소비자물가지수를 작성하고 소비자물가지수(CPI)와 비교를 수행 하였다.

제 2 장 본 론

2.1 온라인 수집 가능 가격 데이터

통계청에서는 식료품 및 비주류음료, 주류 및 담배, 의류 및 신발, 주택, 수도, 전기 및 연료, 가정용품 및 가사 서비스, 보건, 교통, 통신, 오락 및 문화, 교육, 음식 및 숙박, 기타상품 및 서비스의 총 12개 대분류로 이루어진 총 460개의 품목을 오프라인 가격 조사를 통하여 소비자물가지수로 매월 발표하고 있다(2016년 이전은 481개 품목). 통계청에서 발표하는 대분류별 물가지수 품목 개수는 아래 [표 2-1]과 같다.

[표 2-1] 소비자물가지수 품목 개수

대분류	품목개수
식료품 및 비주류음료	133개 품목
주류 및 담배	7개 품목
의류 및 신발	30개 품목
주택, 수도, 전기 및 연료	16개 품목
가정용품 및 가사 서비스	49개 품목
보건	32개 품목
교통	32개 품목
통신	6개 품목
오락 및 문화	55개 품목
교육	20개 품목
음식 및 숙박	44개 품목
기타상품 및 서비스	36개 품목

통계청에서는 오프라인으로 조사를 수행하여 460개의 모든 품목에 대하여 가격 데이터 수집이 가능하지만, 온라인에서의 가격 수집은 온라인의 특성상 모든 품목의 가격 정보를 수집하는 것은 사실상 불가능하다. 온라인에서 가격 정보의 수집이 가능하다고 판단한 품목은 약 290여개로 수집할 수 없는 가격 정보는 온라인 마켓에서 판매하지 않는 주류 및 담배, 주택, 수도, 연료 등의 재화와 서비스 등이다. 온라인 가격 정보의 수

집은 웹 크롤러를 활용하여 수집할 수 있으나, 본 연구에서는 통계청에서 수집하여 공개하고 있는 123개 소비자물가지수 품목의 가격 정보를 활용하였다. 통계청에서 수집하여 공개 중인 가격 데이터 품목은 아래 [표 2-2]와 같다.

[표 2-2] 온라인 수집 가능 소비자물가지수 품목

대분류	품목
식료품 및 비주류음료	쌀, 콩, 밀가루, 국수, 라면, 껌, 잼, 즉석식품, 케찹, 햄, 차음료, 마른멸치, 육류통조림, 전복, 사과, 커피, 당면, 카레, 소금, 참기름, 배, 김밥김, 토마토, 생수, 설탕 등 총 62개 품목
의류 및 신발	남자정장, 여자정장, 점퍼, 아동화, 여자하의, 스웨터, 야자상의, 운동복, 청바지, 실내화, 남자하의, 유아복 등 24개 품목
오락 및 문화	공책, 복사용지, 필기구, 스케치북, 포스트잇, 등 9개 품목
가정용품 및 가사서비스	세탁세제, 방향제, 살충제, 섬유유연제, 가정용 비닐용품, 전구, 헤어드라이어 등 10개 품목
기타상품 및 서비스	면도기, 칫솔, 치약, 샴푸, 구강세정제, 립스틱, 화장수, 염색약, 썬크림, 로션 등 18개 품목

2.2 기초 데이터 분석

통계청에서는 소비자물가지수 품목 중 123개 품목에 대해서 온라인 가격 정보를 수집하여 공개하고 있다. 본 연구에서는 공개되어 있는 온라인 가격 데이터에 대해서 2015.01 ~ 2016.12월까지의 건수, 최대 가격, 최소 가격, 평균 가격 및 표준 편차 등을 분석한 결과 총 건수는 약 6억 4천 8백만 건 이었고, 이중 티셔츠 품목이 약 1천 9백 17만 여건으로 가장 많았으며, 생강 품목이 약 51만 여건으로 가장 작았다. 최대 가격으로는 공책 품목에서 상식적으로 이해할 수 없는 수준인 약 900조원이 넘는 상품이 있었으며 해당 상품을 확인해본 결과 일반적인 초등학생용 노트로 확인 되었다. 최소 가격은 50원으로 전구, 회화용구, 헤어드라이어, 필기구, 점퍼, 여자하의 품목 등 여러 품목에서 나타났다. 대부분의 품목에서 평균과 표준편차의 차이가 크게 나타났으며, 특히 공책 품목에서 7조 1천 7백만 원 이상 이었고, 염색약 품목에서는 17,757원으로 대부분 데이터의 분포가 평균 중심에 모여 있음을 확인할 수 있었다. 그러나 많은 품목에서 평균과 표준편차에 매우 큰 차이가 있었다.

또한 거의 모든 품목에서 해당 품목으로는 분류할 수 없는 상품이 다수 포함되어 있었다. 예를 들어, 빵 품목에 이/미용 관련, 주방 관련 품목 등이 포함되어 있었고, 공채 품목에는 시계, 뱃지, 스탬프, 볼펜 등이 포함되어 있었다. 물가지수 작성 특성상 다른 품목의 상품이 포함되어 있다면 물가지수가 왜곡될 가능성이 있다. 이러한 현상은 데이터 수집 단계에서 온라인 마켓의 상품 카테고리와 소비자물가지수 품목을 연결하여 수집하는 과정에서 발생하는 문제점으로 생각된다. 온라인 마켓에서 수집된 많은 데이터가 수집 단계 혹은 원본 데이터 그 자체의 문제로 인하여 발생하므로 온라인 소비자물가지수를 작성하기 위해서는 데이터 정제가 반드시 필요하다. 이러한 문제점들은 어떤 온라인 마켓에서 어떤

방식으로 가격정보를 수집 하더라도 발생할 수밖에 없는 문제일 것이다.

아래 [표 2-3]은 통계청에서 수집하여 공개한 123개 품목의 가격 데이터 중 2015년 01월부터 2016년 12월까지 데이터를 대상으로 기초 분석을 수행한 자료이다.

[표 2-3] 기초 데이터 분석

품목명	최댓값	최솟값	평균값	표준편차	백분위(25%)	백분위(50%)	백분위(75%)	총 상품 개수
공책	908,558,700,000,000	90	64,863,313,526	7,017,282,358,038	2,400	4,560	9,200	5,944,885
빵	807,607,700,000,000	100	8,741,537,420	2,345,451,443,842	7,900	15,800	24,360	979,485
전구	790,007,490,000,000	50	1,391,233,047	561,732,455,771	6,960	17,600	53,350	9,992,961
헤어 드라이어	90,085,520,000,000	50	1,573,903,095	369,712,033,526	21,250	33,000	49,800	1,247,310
컴퓨터 소모품	70,066,480,000,000	50	159,921,470	96,653,781,277	17,130	40,700	102,920	10,300,379
남자외투	50,047,520,000,000	100	118,835,093	75,586,444,356	69,600	120,300	205,200	7,898,637
회화용구	90,085,520,000,000	50	156,834,128	70,271,423,570	5,760	13,810	35,040	7,644,608
필기구	3,002,850,000,000	50	47,430,317	8,495,508,147	2,830	7,800	32,970	12,208,967
남자내의	990,094,030,000	80	62,676,271	7,205,416,939	15,400	26,880	43,680	8,130,434
유아복	3,002,850,000,000	100	26,423,532	6,503,820,913	17,600	29,200	46,750	14,695,198
남자하의	990,094,030,000	100	43,553,556	5,928,701,417	30,710	42,440	58,000	7,120,425
복사용지	990,094,030,000	50	39,178,706	4,357,956,924	3,850	12,000	24,180	6,476,888
남자상의	990,094,030,000	90	18,024,990	4,015,606,346	26,620	40,210	62,300	11,821,993
남자구두	2,465,280,000,000	100	4,951,771	3,417,014,248	60,300	125,480	320,400	7,287,291
여자내의	3,002,850,000,000	70	10,750,334	3,325,580,146	11,040	18,000	34,000	15,209,542
실내화	3,464,010,000,000	100	1,060,805	1,902,023,272	7,020	11,840	19,860	9,950,566
여자외투	1,853,470,000,000	90	2,743,698	1,888,695,026	66,310	121,720	239,000	9,341,539
스케치북	950,903,900,000	50	2,006,296	880,639,065	3,350	6,800	15,260	3,371,329
기록매체	1,651,960,000,000	75	665,703	868,718,581	12,750	28,290	75,610	8,871,431
남자정장	1,661,050,000,000	100	505,147	715,859,965	97,920	154,860	217,550	5,384,055
점퍼	1,818,270,000,000	50	356,370	616,368,129	49,800	85,300	161,100	17,404,692
여자하의	1,020,600,000,000	50	246,504	301,775,067	22,230	36,670	63,000	13,510,400
운동복	183,850,000,000	90	144,931	109,751,929	34,400	52,580	77,760	14,494,285
원피스	119,950,000,000	60	109,760	62,832,191	32,000	53,580	89,100	10,933,457
여자상의	138,220,000,000	50	92,141	56,891,708	28,500	45,000	79,800	11,805,224
프린터	1,440,000,000	90	238,923	968,080	36,700	103,180	245,740	9,256,264
티셔츠	499,999,008	50	51,304	346,556	18,700	31,000	51,330	19,177,305
여자구두	63,999,980	100	139,066	166,165	39,650	62,000	178,170	9,654,015
스웨터	8,339,700	90	71,210	113,997	25,310	40,140	65,770	6,736,436
쌀	25,000,000	90	35,579	111,946	16,800	30,500	46,900	4,040,420
여자정장	9,365,640	90	85,625	106,014	32,800	55,000	100,880	5,664,031
청바지	5,796,200	90	75,376	96,441	26,600	42,000	80,640	9,680,778
고춧가루	520,310	100	69,114	87,926	13,000	30,180	80,800	1,224,179
운동화	19,593,500	100	94,477	84,332	46,920	74,800	118,150	11,388,468
등산복	1,321,210	90	87,155	80,436	35,197	60,800	108,280	7,274,616
페이스 파우더	1,077,420	100	53,594	70,138	18,000	33,520	60,690	3,880,619
치약	3,039,900	50	24,956	67,533	5,620	10,570	23,000	6,186,126
파운데이션	1,678,110	100	51,590	66,405	18,000	33,600	59,000	4,618,298
참기름	342,720	100	53,021	63,686	13,100	27,300	62,400	2,288,629
분유	859,000	60	64,138	59,768	20,420	42,000	92,700	1,034,976
아동복	8,623,000	70	47,994	56,716	18,500	34,120	59,200	12,270,945
혼합조미료	1,003,200	50	40,600	53,849	7,420	16,240	51,650	4,437,729
생강	463,190	100	41,396	53,420	7,500	18,200	51,200	513,494
커피	9,900,000	90	29,697	48,105	8,980	16,800	31,000	6,265,925
식용유	1,101,660	100	35,116	47,034	9,370	19,630	44,760	4,771,083

습기 제거제	466,500	90	28,178	46,601	3,600	9,900	29,500	3,500,340
전복	343,800	100	66,793	45,751	34,040	59,900	87,000	933,632
고사리	346,750	100	32,256	45,106	7,000	16,900	36,200	874,599
영양크림	1,266,600	100	46,336	44,963	18,910	33,670	58,780	3,535,362
포스트잇	1,036,800	50	19,498	44,856	2,800	6,740	17,260	5,027,129
면도기	1,213,770	50	30,778	43,782	10,130	17,900	29,050	5,559,976
미역	747,000	100	33,175	43,780	6,900	16,590	43,300	2,315,022
커피크림	346,750	100	34,904	42,786	8,000	17,300	43,600	1,387,456
살충제	394,020	100	25,329	42,414	5,500	11,100	24,000	5,643,598
햄	1,596,000	100	30,185	41,026	6,100	15,310	36,000	3,007,476
화장수	1,793,000	60	41,386	40,628	15,000	28,000	54,000	7,922,143
참치캔	650,000	100	31,277	37,612	5,520	16,800	43,730	4,791,307
소금	5,172,750	100	27,754	37,235	7,030	16,500	33,700	3,879,973
마른멸치	643,060	100	47,423	37,150	19,700	39,900	65,000	2,865,647
차음료	1,382,400	90	28,547	36,969	9,200	17,490	35,000	3,021,195
콩	1,300,000	100	27,033	36,686	7,500	13,800	30,900	1,707,539
칫솔	3,550,050	50	15,938	36,496	4,100	8,400	17,140	9,331,986
차	749,700	60	27,418	35,509	8,550	16,500	31,900	7,775,254
북어채	345,670	100	38,208	35,427	12,000	30,000	52,000	1,156,599
드레싱	1,488,780	50	26,326	35,380	5,450	12,500	33,310	2,341,386
청소용 세제	418,770	100	26,288	34,855	6,620	15,000	30,000	3,160,032
배	337,400	100	37,560	34,519	10,000	31,000	52,800	902,430
간장	1,479,600	90	31,542	33,611	9,000	19,040	43,700	3,209,965
땅콩	342,920	100	28,422	33,565	8,500	15,720	33,900	1,803,984
샴푸	1,113,480	50	32,355	33,423	12,100	20,600	40,560	6,452,972
당면	339,280	100	29,540	33,173	6,000	15,050	45,040	3,030,430
토마토	340,470	100	27,967	32,686	12,900	19,900	31,500	612,226
고등어	332,500	100	41,504	32,456	19,700	34,210	54,530	2,172,595
기능성 음료	1,200,000	90	32,127	31,929	14,900	24,500	39,310	1,801,099
욕류 통조림	558,870	100	26,834	31,178	5,890	16,350	37,440	1,127,825
가정용 비닐용품	782,800	50	23,910	30,951	4,500	9,350	33,960	2,598,618
설탕	342,470	50	24,594	30,584	5,750	14,000	31,400	1,743,362
사과	333,000	100	44,346	30,506	22,900	38,000	57,800	2,015,350
바디워시	3,672,100	50	28,615	30,072	11,090	18,050	35,900	5,819,451
카레	333,680	100	23,848	29,927	4,000	11,800	31,630	3,140,429
소시지	465,000	100	21,423	29,892	6,020	10,970	22,100	2,160,586
이유식	336,290	100	18,515	29,853	3,880	8,300	22,190	765,238
치즈	340,760	90	21,285	29,739	6,900	12,870	21,100	2,357,296
케찹	340,760	50	21,703	29,734	4,200	8,790	31,180	1,722,942
김밥김	1,153,820	90	29,316	29,710	10,500	21,300	38,000	3,461,797
찹쌀	340,760	90	25,412	29,510	7,630	16,000	32,900	1,172,279
씨리얼 식품	389,200	100	22,312	29,488	6,530	11,690	24,250	1,032,029
후르츠 칵테일	558,870	100	25,565	28,847	4,700	12,100	41,000	2,638,035
로션	2,120,970	50	32,914	28,773	14,440	24,500	41,180	8,416,778
단무지	993,430	100	21,330	28,684	4,480	11,400	25,000	1,548,785
화장지	663,960	50	30,278	28,538	9,900	19,850	43,000	9,527,360
고추장	340,760	100	33,438	28,485	11,600	27,700	46,020	3,029,660
스프	338,250	100	24,194	28,294	4,400	14,040	34,020	2,651,579
생수	346,750	90	21,737	28,238	6,190	13,610	27,390	1,323,643
식빵	346,380	100	18,224	28,052	3,400	9,280	20,500	1,769,858
밀가루	340,760	100	23,518	28,012	4,710	14,500	31,680	3,217,851
즉석식품	390,390	100	18,659	26,470	3,020	9,900	26,100	2,441,374
잼	338,250	80	20,344	26,457	6,000	10,540	22,000	2,075,238
고구마	335,600	100	25,748	26,257	11,900	18,900	30,520	1,197,684
화장비누	20,000,000	70	18,051	25,991	5,310	11,400	22,000	5,257,935
김치	1,000,000	100	31,180	25,916	15,700	26,170	39,900	1,624,810
국수	337,400	100	21,620	25,807	4,020	10,700	33,200	3,081,936

된장	339,280	100	30,936	25,793	12,000	26,000	42,340	2,700,292
썬크림	961,700	100	28,098	25,449	13,300	21,900	34,000	4,255,503
아이스 크림	312,230	100	19,520	25,301	5,000	10,300	22,250	1,633,952
혼합음료	431,660	100	22,147	25,213	11,010	16,900	27,000	1,025,242
과일주스	1,000,000	90	27,858	24,915	9,000	22,700	39,950	5,763,274
껌	340,760	80	17,673	24,729	3,500	9,670	19,400	2,308,495
아동화	1,533,010	100	38,562	24,663	25,270	33,310	45,570	14,876,501
키친타월	299,700	100	20,563	24,394	5,490	11,500	29,230	2,128,655
립스틱	10,487,200	100	23,717	24,277	9,670	17,800	34,490	6,333,366
구강 세정제	361,230	80	17,006	23,427	4,200	8,800	19,250	3,310,770
냉동식품	979,120	90	19,950	22,817	6,900	12,050	25,500	2,934,087
클린징 크림	5,225,000	50	21,408	21,150	8,800	15,000	26,600	6,303,945
세탁세제	950,000	90	22,561	20,485	9,900	16,740	29,500	10,183,743
비스킷	336,210	80	19,185	20,112	3,700	13,790	28,950	3,703,229
방향제	1,566,740	70	20,968	19,564	8,190	15,780	28,330	8,213,550
부엌용 세제	429,330	90	17,823	18,922	6,260	11,730	22,680	5,344,658
라면	339,800	90	23,124	18,624	8,250	21,250	33,300	5,296,128
섬유 유연제	325,900	100	19,383	18,292	7,990	13,900	23,900	5,711,698
스낵과자	737,350	80	19,482	18,097	4,900	17,970	28,000	4,497,751
탄산음료	435,400	90	23,683	18,001	13,500	21,000	30,840	4,675,482
염색약	1,009,000	80	14,975	17,757	6,520	10,100	15,810	5,870,792

2.3 온라인 가격 데이터의 문제점

수집된 가격 정보의 기초 분석을 통해서 확인한 결과 수집 데이터의 대부분 품목에서 [표 2-4]와 같은 공통된 문제점을 가지고 있다는 것을 확인할 수 있었다.

[표 2-4] 수집 데이터의 문제점

문제점		내용
품목 분류		수집된 상품이 해당되는 품목이 아닌 다른 품목과 맵핑되어 있는 문제점
가격	비정상적인 가격	수집된 상품의 가격이 일반적인 상식선에서 벗어나 매우 높거나 낮은 문제점 예) 공책의 가격이 900조원을 넘는 경우 쌀의 가격이 50원인 경우
	가격의 변화	수집된 상품의 가격 변화가 매우 급격하게 변화하는 문제점 예) 오늘의 상품가격/어제의 상품가격의 비율의 변화가 매우 높거나 낮은 경우
상품의 연속성 부재		특정 상품이 시간의 흐름에 따라 연속적으로 존재하지 않는 문제점 ※ 이 문제점은 데이터 수집 단계의 문제점으로 본 연구에서는 다루지 않음

서론에서도 거론 하였듯이 2013년에 구축된 온라인 물가지수 작성 시스템은 상품의 가격정보를 수집 시 온라인 소매점의 상품 카테고리와 소비자물가지수 품목을 맵핑하는 방법을 사용하고 있다. 온라인 소매점의 상품 카테고리와 소비자물가지수 품목을 100% 1:1로 맵핑하는 것은 사실상 불가능 하며 100% 정확히 1:1 맵핑이 된다고 해도 온라인 마켓에서 상품의 카테고리를 잘못 분류한 경우에는 동일한 문제가 발생할 수 있다. 또한 수집된 가격 정보에서 명확한 이유는 알 수 없지만 공책 품목의 가격이 900조원을 넘는다거나 헤어드라이어 품목의 가격이 50원 밖에 되지 않는 등 비정상적인 가격에 대한 문제점도 확인할 수 있었다.

그리고 상품 가격의 변화를 확인해본 결과 하루 차이로 상품의 가격이 수백 퍼센트를 오르거나 내리는 문제점도 확인 되었다. 이러한 현상은 온라인 마켓에서 상품을 판매하는 과정에서 상품에 대한 속성의 잘못된 변경으로 인한 가격 변경 현상으로 예상된다. 예를 들어 오늘까지는 특정상품ID에 대해서 남자내의로 판매하던 상품이 다음날에는 남자구두로 판매되는 경우이다. 이 경우 오늘과 다음날의 상품 가격이 큰 폭으로 변경될 수 있고 소비자물가지수 작성 시 동일한 상품으로 인식하게 된다. 만일 온라인 소비자물가지수 작성 시 이러한 상품을 포함한다면 잘 못된 소비자물가지수가 작성될 것이다. 상품의 연속성 부재와 관련된 문제점은 데이터를 수집하는 단계에서 발생하는 문제점으로 수집 대상 사이트에서 웹크롤러의 차단 또는 웹크롤러의 오류 등으로 인하여 수집을 하지 못했거나 또는 상품의 판매 중단 등으로 발생하는 문제점으로 생각된다. 상품의 판매 중단인 경우는 어쩔 수 없는 상황이지만 다른 문제로 인하여 상품의 연속성을 보장하지 못한 경우라면 정상적인 온라인 소비자물가지수를 작성하는데 지장을 초래하게 된다. 본 연구에서 활용한 통계청 공개 온라인 가격 데이터는 상품의 연속성 문제가 존재하고 있지만, 본 연구에서 대응할 수 없는 문제이므로 상품의 연속성 문제에 대해서는 다루지 않을 것이다. 이렇듯 온라인 마켓에서 수집한 가격 정보는 많은 문제점을 가지고 있다. MIT의 BPP에서 특정 소매점의 수집 가격 데이터를 1년 이상 모니터링 하는 이유가 여기에 있는 것이다.

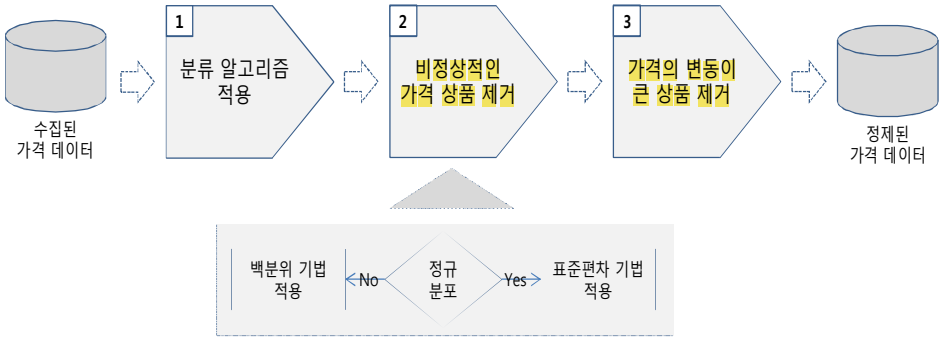
2.4 온라인 가격 데이터 정제 방안

앞서 제시한 온라인 마켓에서 수집된 가격 데이터의 문제점은 정확한 온라인 소비자물가지수를 작성하기 위해 반드시 해결되어야 한다. 본 연구에서는 [표 2-5]와 같은 방법으로 온라인 마켓에서 수집된 가격 데이터를 정제하는 방안을 제시하고자 한다.

[표 2-5] 데이터 정제 방안

문제점		데이터 정제 방안
품목 분류		기계학습의 일종인 텍스트(문서) 분류 알고리즘을 활용하여 수집된 상품을 올바른 품목으로 분류 - 상품의 판매제목에 대해 텍스트 분류 알고리즘을 적용
가격	비정상적인 가격	소비자물가지수 품목별 표준편차를 활용하여 평균값으로부터 많이 벗어나 있는 가격 데이터 정제 * 데이터가 정규분포를 따르는 경우 - |당일가격-당일품목평균| >= (2 * 당일표준편차) 인 경우 정제 * 데이터가 정규분포를 따르지 않는 경우 - 데이터의 백분위수를 5%단위로 산정하여 최적의 백분위수 범위 결정
	가격의 변화	가격비(오늘가격/어제가격)가 품목별로 최소와 최대 비율의 범위를 벗어나는 경우 데이터 정제 - 가격비율 변동이 ±200%이상인 경우 정제

위에서 제시한 데이터 정제 방안을 온라인 마켓에서 수집한 가격 데이터에 적용하는 절차는 [그림 2-1]과 같다.



[그림 2-1] 온라인 가격 데이터 정제 절차

본 연구에서는 통계청에서 공개한 123개 물가지수 품목의 가격 데이터 중 표준편차가 가장 큰 상위 5개의 품목인 공책, 빵, 전구, 헤어드라이어, 컴퓨터소모품만을 대상으로 데이터 정제 알고리즘을 적용하였다. 그러나 통계청에서는 품목별 정해진 규격에 해당하는 상품만을 조사하여 소비자물가지수를 작성하지만 본 연구에서는 온라인 가격 정보에서 규격 제한 절차는 생략하였다. 다만 통계청의 조사 규격의 대략적인 가격대를 고려하였다. 아래 [표 2-6]은 5개의 데이터 정제 알고리즘 적용 대상 품목에 대한 통계청 소비자물가지수 기준 상품 규격이다.

[표 2-6] 통계청 품목별 기준 상품 규격

품목명	기준 상품 규격
빵(Bread)	1) 단팥빵, 제과점 판매, 비닐포장 2) 모카빵, 제과점 판매, 비닐포장
전구(Light bulb)	일반조명용 형광램프
컴퓨터소모품 (Computer supplies)	프린터용잉크, 검정색, 정품
공책(Notebook)	1) 중.고등학생용, 내지 24매 내외 2) 초등학생용, 내지 24매 내외
헤어드라이어(Electric dryer)	전기식드라이어

2.4.1 분류 알고리즘의 적용

분류 알고리즘은 기계학습 알고리즘 중 하나로 데이터의 분류를 위해서 학습 데이터가 필요한 알고리즘이다. 온라인 마켓에서 수집한 상품을 올바른 품목으로 분류하기 위해서는 상품의 품목을 특정할 만한 텍스트(문장)가 필요하다. 온라인 마켓에서 웹크롤링을 통해 수집 가능한 상품의 정보 중 품목 분류를 위해 활용할 수 있는 항목은 상품명, 상품의 판매 제목 글 그리고 상품의 상세 설명 글이다. 그러나 상품의 상세 설명의 경우 문자 형식의 데이터가 아닌 그림이나 동영상 등의 데이터인 경우 많아 적용이 매우 어렵다. 그러므로 상품의 품목 분류를 위해 적용할 만한 정보는 상품의 판매 제목 글이 가장 유력하다. 또한 통계청에서 수집하여 공개한 가격 데이터에는 상품의 품목 분류를 위해 활용할 수 있는 항목으로 상품의 제목 글 밖에 없어 본 연구에서는 상품의 제목 글을 분류 알고리즘의 적용 데이터로 활용하였으며, 분류 알고리즘은 오픈 소스 기계학습 도구인 머하웃(Mahout)의 나이브베이즈(NaiveBayes) 알고리즘을 사용하여 분류 모델을 만들었다. 위키백과 사전에 따르면 나이브베이즈 분류 알고리즘은 “텍스트 분류에 사용됨으로써 문서를 여러 범주(예:스팸, 스포츠, 정치)중 하나로 판단하는 문제에 대한 대중적인 방법”[6] 으로 알려져 있다.

분류 모델 적용을 위한 학습 데이터는 2016년 08월 01일 ~ 2016년 08월 15일까지의 데이터를 활용하여 각 품목별로 T와 F 두 개의 분류로 구분 하였다(T = 품목에 해당하는 상품, F=품목에 해당하지 않는 상품). 학습 데이터 중 30%는 시험용 데이터로 활용하였다. 컴퓨터 소모품 품목의 경우 매우 다양한 종류 및 가격대의 상품이 있어 통계청의 품목 규격에 맞추어 프린터 잉크와 토너에 해당하는 상품으로 제한하였다. 각 품목별로 분류 모델의 신뢰도를 보면 전구, 공책, 컴퓨터 소모품, 헤어드

라이어는 60%이상을 보였으나, 빵 품목의 경우 T로 분류된 상품과 F로 분류된 상품간의 단어에 대한 충돌이 매우 많이 발생하여 35%로 매우 낮은 신뢰도를 보였다. 주기적인 학습과정을 거치면 정교한 모델을 만들 수 있을 것으로 판단된다. 품목별 분류 모델의 평가 결과는 아래 [표 2-7]과 같다.

[표 2-7] 분류 알고리즘 평가

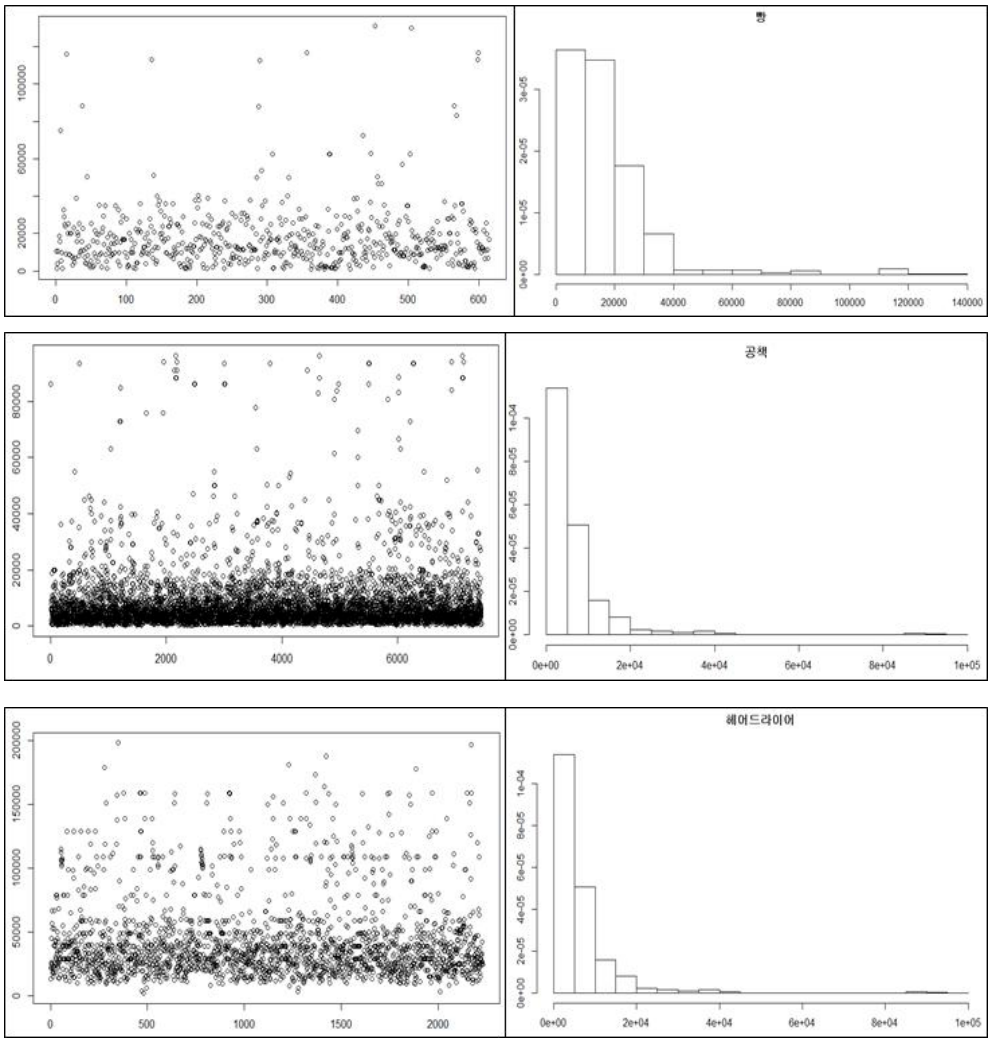
품목명	Accuracy	Reliability	Reliability (standard deviation)
빵	52.27%	35.1843%	0.424
전구	99.11%	64.9574%	0.563
공책	96.91%	63.6344%	0.553
컴퓨터소모품	99.31%	65.3498%	0.566
헤어드라이어	99.61%	65.5603%	0.568

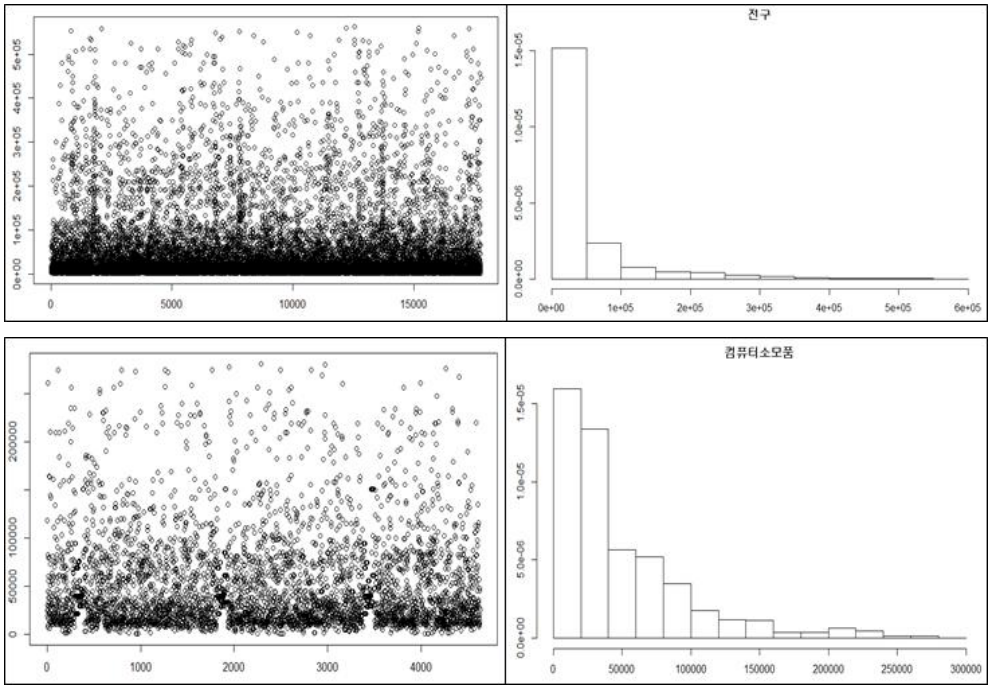
분류 모델을 적용하여 2015년 01월 01일 ~ 2016년 12월 31일까지의 데이터에 대해서 기본 분석을 재 수행하였다. 기분 분석을 재 수행 결과 공책과 헤어드라이어 품목에서 최댓값과 표준편차가 매우 좋아진 것을 확인할 수 있었다. 그러나 빵, 전구, 컴퓨터 소모품 품목은 약간 낮아지기는 했지만 여전히 표준편차가 매우 높았다. 아래 [표 2-7]은 분류 모델을 적용한 후 재 수행한 기본 분석 결과이다.

[표 2-8] 분류 모델 적용 후 기본 분석

품목명	최댓값	최솟값	평균값	표준편차	백분위(25%)	백분위(50%)	백분위(75%)	총 상품 개수
공책	10,000,000	90	9,380	38,939	2,400	4,260	7,800	3,939,343
빵	415,039,422,000	120	39,958,807	3,518,185,186	6,580	12,400	21,000	306,405
전구	150,001,431,000,000	50	320,523,635	208,505,473,203	6,500	15,500	46,700	7,165,208
헤어 드라이어	5,059,890	50	44,531	47,928	22,900	33,500	49,000	894,745
컴퓨터 소모품	1,901,930,000,000	50	7,158,510	3,314,618,111	25,000	57,600	144,500	4,510,109

다음으로 이렇게 분류된 데이터의 가격이 정규 분포를 따르는지를 2016년 08월 01일자 데이터를 활용하여 확인해 보았다. 대부분의 품목에서의 가격 데이터는 정규 분포를 따르지 않고 대부분 낮은 가격대에 위치하였다. 아래 [그림 2-2]는 분류 모델 적용 후 데이터 분포를 확인한 자료이다. 데이터 분포의 확인은 오픈소스 통계 분석 툴인 R을 활용하였다.





[그림 2-2] 분류 모델 적용 후 데이터 분포

모든 품목의 데이터가 정규 분포를 따르지 않고, 낮은 가격대(시작 가격대)에 존재함을 알 수 있다. 이러한 현상은 대부분의 일자별 데이터에서 공통적임을 확인 하였다.

2.4.2 비정상적인 가격의 상품 정제

앞서 기초 데이터 분석에서 제시한 것처럼 수집된 가격 데이터에는 상식적으로 이해할 수 없이 매우 높거나 낮은 가격을 가진 상품 데이터가 존재한다. 이러한 상품을 정제하기 위해서 데이터가 정규 분포를 따르는 경우 표준편차법을 활용하는 방안과 정규 분포를 따르지 않는 경우 백분위수법을 활용하는 경우를 제시 하였다. 먼저 데이터가 정규 분포를 따르는 경우에는 경험법칙(정규분포를 보이는 데이터의 95%는 표준편차의 2배 이내 범위에 있다)을 적용하여 전체 데이터 중 약 5%의 비정상적인 데이터를 제거하게 된다. 상품 가격 데이터가 정규 분포를 따르는 경우 다음과 같은 공식을 적용하여 비정상적인 가격의 상품을 정제하며, 표준편차의 2배 밖에 존재하는 비정상적인 데이터를 제외한 95%의 데이터가 포함되므로 경험상수(w)는 일반적으로 2를 적용하면 된다. 만일 정제 대상 데이터를 조절 하고자할 때는 w 값을 적절히 조정하면 된다.

$$P = |Price(t) - \overline{Price}(t)| <= (w \times Ps(t))$$

* P = 적용대상상품, Price = 상품의 가격, t = 데이터수집일자
* w = 경험상수, Ps = 상품가격의 표준편차

그러나 본 연구에서 분석 대상으로 선정된 데이터 중 분류 알고리즘이 적용된 결과 데이터는 정규 분포의 형태를 보이지 않는 것으로 확인 되었으므로 백분위수법을 적용하여 데이터를 정제하여야 한다. 백분위수법을 적용하기 위하여 각 품목별 상품 가격의 백분위수를 5분위단위로 확인하였다. 아래 [표 2-9]은 분류 모델이 적용된 상품의 품목별 백분위수를 5분위단위로 조사한 결과이다.

[표 2-9] 분류 모델 적용 후 품목별 백분위수

백분위수	공책	빵	전구	헤어 드라이어	컴퓨터 소모품
최솟값	90	120	50	50	50
5%	950	1,980	1,800	11,900	9,000
10%	1,400	2,900	2,900	15,160	12,000
15%	1,780	3,800	3,900	18,500	15,620
20%	2,050	5,020	5,160	20,340	20,000
25%	2,400	6,580	6,500	22,900	25,000
30%	2,670	7,900	7,890	25,000	29,900
35%	3,000	8,990	9,220	26,990	35,000
40%	3,400	9,990	10,930	29,000	40,680
45%	3,800	11,400	13,000	30,500	48,210
50%	4,260	12,400	15,500	33,500	57,600
55%	4,800	13,500	19,000	36,000	68,440
60%	5,320	15,000	23,000	38,660	80,000
65%	6,010	16,900	29,000	40,100	95,800
70%	6,800	18,900	36,380	45,000	116,090
75%	7,800	21,000	46,700	49,000	144,500
80%	9,030	24,000	60,600	55,800	180,000
85%	10,890	27,900	80,270	65,000	225,900
90%	14,500	34,800	116,600	88,000	290,890
95%	24,320	50,910	192,060	118,000	428,840
최댓값	10,000,000	415,039,422,000	150,001,431,000,000	5,059,890	1,901,930,000,000

백분위수를 보면 모든 품목에서 상식선에서 생각할 수 있는 금액대가 5% ~ 95% 사이에 들어와 있다. 그러나 [표 2-6] 품목별 기준 상품 규격을 보면 5% ~ 95% 범위에 있는 가격을 그대로 적용할 수 없다. 명확한 상품의 모델을 알 수는 없지만 기준 상품 규격을 확인하여 대략적인 상품의 가격대를 적용하였다. 빵은 최솟값 ~ 10%, 전구는 5% ~ 30%, 컴퓨터소모품은 5% ~ 30%, 공책은 25% ~ 45%, 헤어드라이어는 10% ~ 30%의 백분위수를 적용하여 데이터를 정제하였다. 본 연구에서 확인해보지는 않았지만 백분위수의 범위별로 데이터를 정제하여 확인하는 방법도 고려해 볼 수 있을 것으로 생각한다. 아래 [표 2-10]은 품목별로 위에서

제시한 백분위수 기준으로 데이터를 정제한 후 기본 분석을 재 수행한 결과이다.

[표 2-10] 비정상 가격 제거 적용 후 기본 분석

품목명	최댓값	최솟값	평균값	표준편차	백분위 (25%)	백분위 (50%)	백분위 (75%)	총 상품 개수
공책	16,880	1,100	3,103	737	3,000	3,410	1,800	803,549
빵	5,150	120	2,122	807	1,480	2,000	2,800	31,736
전구	13,720	1,040	4,720	1,902	3,160	4,500	6,210	1,803,032
헤어 드라이어	30,800	6,290	20,439	3,714	18,500	20,700	23,100	181,467
컴퓨터 소모품	106,490	3,500	22,030	12,029	12,880	17,800	29,690	1,138,570

비정상적인 가격 정제 후 총 분석대상 상품 개수를 분류 알고리즘 적용 후와 비교해보면 공책은 20.4%, 빵은 10.4%, 전구는 25.2%, 헤어드라이어는 20.3%, 컴퓨터소모품은 25.2% 밖에 남지 않은 것을 확인할 수 있다. 본 연구에서는 최종 품목별 소비자물가지수를 작성하여 통계청 소비자물가지수와 비교하기 위하여 통계청의 품목별 규격을 최대한 만족하는 가격대를 적용하여 분석대상 상품 개수가 많이 줄었지만 표준편차는 많이 완화되었음을 확인할 수 있다. 그리고 [표 2-10] 비정상 가격 제거 적용 후 기본 분석에서 제시한 정보와 [표 2-9] 분류 모델 적용 후 품목별 백분위수를 비교해봤을 때 이해할 수 없는 수치 정보가 있을 것이다. 예를 들어 헤어드라이어 품목의 비정상 가격 제거 시 백분위수 10% ~ 30%를 적용하였고 [표 2-9] 분류 모델 적용 후 품목별 백분위수의 헤어드라이어 품목의 10%는 15,160원이고 30%는 25,000원 이었다. 그러나 [표 2-10] 비정상 가격 제거 적용 후 기본 분석의 최솟값이 6,290원이 도출되었다. 이런 현상이 발생한 이유는 [표 2-9] 분류 모델 적용 후 품목별 백분위수의 모수는 일자와 상관없이 전체를 대상으로 하였지만, 데이터를 정제할 때는 대상 일자별 백분위수를 적용하였기 때문이다.

2.4.3 가격 변동이 큰 상품 정제

일반적으로 상품의 가격은 일단위로 변동이 크지 않을 것이다. 다만, 특별 할인 등으로 인하여 일시적으로 가격이 급락하거나 또는 급등할 수는 있겠지만 이런 경우도 일반적으로 2배 이내일 것이다. 가격이 급등 또는 급락한 가격정보를 사용하여 소비자물가지수를 작성하게 되면 왜곡이 발생할 것이다. 가격의 급락 및 급등에 대해서는 오늘가격과 어제가격의 비율로 정제를 하였다. 본 연구에서는 다음과 같은 공식을 적용하여 2단계(비정상적인 가격의 상품 제거)까지 정제가 진행된 데이터에 대해서 가격 변동이 큰 상품을 정제하였으며, 가격 변동 비율은 전일가격의 2배를 적용 하였다. 즉 MAX(금일가격/전일가격, 전일가격/금일가격) > 2 이면 분석 대상 데이터에서 제외되는 것이다. 예를 들어 전일가격이 100이고 금일가격이 210인 경우 MAX(210/100, 100/210) = MAX(2.1, 0.48) = 2.1 이므로 분석대상 데이터에서 제외된다. 변동의 비율은 상황에 따라 적절하게 조정이 가능할 것이다.

P = Ratio >= MAX(Price(t)/Price(t-1), Price(t-1)/Price(t))

* P = 적용대상상품, Price = 상품 가격, t = 데이터수집일자

* Ratio = 가격 변동 비율

아래 [표 2-11]은 2단계(비정상적인 가격의 상품 제거)까지 진행된 데이터에 대해서 위의 공식을 적용하여 데이터를 생성하고 기본 분석을 수행한 결과이다.

[표 2-11] 가격비율 정제 적용 후 기본 분석

품목명	최댓값	최솟값	평균값	표준편차	백분위 (25%)	백분위 (50%)	백분위 (75%)	총 상품 개수
공책	16,880	1,100	3,103	737	2,640	3,000	3,410	803,091
빵	5,800	120	2,122	807	1,480	2,000	2,800	31,675
전구	13,720	1,040	4,720	1,902	3,150	4,500	6,210	1,800,658
헤어 드라이어	30,800	6,290	20,438	3,714	18,500	20,700	23,100	181,259
컴퓨터 소모품	106,490	3,500	22,034	12,031	12,890	17,800	29,700	1,137,789

[표 2-11] 가격비율 정제 적용 후 기본 분석은 [표 2-10] 비정상 가격 제거 적용 후 기본 분석과 비교하여 큰 변동이 없음을 알 수 있다.
그러나 가격 변동이 비상식적으로 큰 데이터가 전체 데이터에 영향을 줄 수 있으므로 반드시 제거 되어야 한다.

2.5 데이터 정제 결과

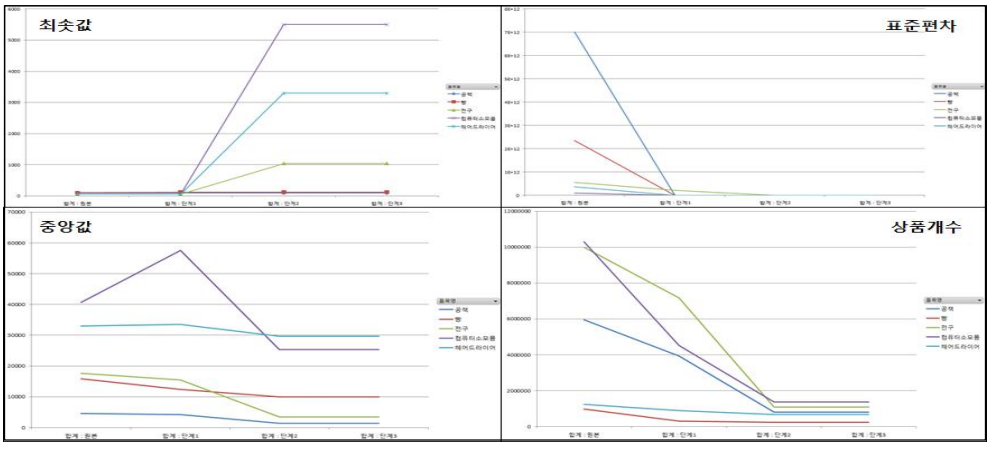
지금까지 원본 데이터로부터 총 세 가지의 데이터 정제 방안을 적용하여 데이터의 변화를 관찰 하였다. 첫 번째로 데이터 분류 알고리즘을 적용하여 해당 품목에 포함되지 않아야할 상품을 정제하였고, 두 번째로 백분위수를 적용하여 비정상적인 가격 범위에 있는 상품을 정제하였다. 세 번째로 가격 변화의 비율을 이용하여 2배 이상의 급격한 가격 변동이 존재한 상품을 정제 하였다. 아래 [표 2-12]는 원본 데이터부터 최종 분석용 데이터가 생성되기까지의 데이터 변화를 비교하여 보았다.

[표 2-12] 단계별 기본 분석 비교

품목명	항목	원본	단계1	단계2	단계3
최댓값	공책	908,558,700,000,000	10,000,000	16,880	16,880
	빵	807,607,700,000,000	415,039,422,000	5,150	5,800
	전구	790,007,490,000,000	150,001,431,000,000	13,720	13,720
	헤어드라이어	90,085,520,000,000	5,059,890	30,800	30,800
	컴퓨터소모품	70,066,480,000,000	1,901,930,000,000	106,490	106,490
최솟값	공책	90	90	1,100	1,100
	빵	100	120	120	120
	전구	50	50	1,040	1,040
	헤어드라이어	50	50	6,290	6,290
	컴퓨터소모품	50	50	3,500	3,500
평균값	공책	64,863,313,526	9,380	3,103	3,103
	빵	8,741,537,420	39,958,807	2,122	2,122
	전구	1,391,233,047	320,523,635	4,720	4,720
	헤어드라이어	1,573,903,095	44,531	20,439	20,438
	컴퓨터소모품	159,921,470	7,158,510	22,030	22,034
표준편차	공책	7,017,282,358,038	38,940	737	737
	빵	2,345,451,443,842	3,518,185,186	807	807
	전구	561,732,455,771	208,505,000,000	1,902	1,902
	헤어드라이어	369,712,033,526	47928.50929	3,714	3,714
	컴퓨터소모품	96,653,781,277	3,314,618,111	12,029	12,031
백분위 (25%)	공책	2,400	2,400	3,000	2,640
	빵	7,900	6,580	1,480	1,480
	전구	6,960	6,500	3,160	3,150
	헤어드라이어	21,250	22,900	18,500	18,500
	컴퓨터소모품	17,130	25,000	12,880	12,890
백분위 (50%)	공책	4,560	4,260	1,800	3,000
	빵	15,800	12,400	2,800	2,000
	전구	17,600	15,500	6,210	4,500
	헤어드라이어	33,000	33,500	20,700	20,700
	컴퓨터소모품	40,700	57,600	17,800	17,800
백분위	공책	9,200	7,800	1,800	3,410

(75%)	빵	24,360	21,000	2,800	2,800
	전구	53,350	46,700	6,210	6,210
	헤어드라이어	49,800	49,000	23,100	23,100
	컴퓨터소모품	102,920	144,500	29,690	29,700
총 상품 개수	공책	5,944,885	3,939,343	803,549	803,091
	빵	979,485	306,405	31,736	31,675
	전구	9,992,961	7,165,208	1,803,032	1,800,658
	헤어드라이어	1,247,310	894,745	181,467	181,259
	컴퓨터소모품	10,300,379	4,510,109	1,138,570	1,137,789

위의 표에서 세 가지 데이터 정제과정을 거치면서 총 상품개수가 급격히 줄어 들은 것을 확인할 수 있다. 상품의 개수가 줄어든 가장 큰 원인은 통계청의 품목별 규격에 어느 정도 접근 위하여 백분위수를 이용하여 가격대의 조정을 수행하였기 때문이다. 빵의 경우 초코파이, 쿠키 등의 과자류 데이터가 많이 정제 되었다. 또한 단계 2를 거치면서 표준편차와 최대/최솟값이 많이 안정화 된 것을 확인할 수 있다. 온라인 소비자물가지수를 산출함에 있어 분석 표본 데이터를 많이 확보하는 것도 중요하지만 해당 품목에 부합되는 상품을 대상으로 상식적인 가격과 특히 급등/급락하는 데이터를 정제하여 올바른 데이터를 확보하는 것이 중요하다. 본 연구에서 제시한 데이터 정제 결과를 비교하기 위해서 각 품목의 정제전과 정제후의 데이터를 [그림 2-3]과 같이 그래프로 비교하였다.



[그림 2-3] 데이터 정제 단계별 지표 추이

2.6 데이터 정제 전/후의 온라인 소비자물가지수 비교

원본 데이터로부터 총 3단계로 데이터 정제를 통하여 생성된 최종 분석 대상 데이터와 정제 전 데이터를 이용하여 개별품목지수를 산출하여 CPI(소비자물가지수) 비교하였다. 단, 정제 전 데이터로 개별품목지수를 산출시 백분위수를 이용하여 95%이하의 데이터만을 활용하였다. 이렇게 한 이유는 최소한의 데이터 정제를 수행하지 않은 경우 편차가 너무 심하여 온라인 소비자물가지수 산출 후 비교 자체를 할 수 없었다. 온라인 소비자물가지수의 개별품목지수는 Jevons Index를 적용하여 품목 내 가중치가 없는 기하평균으로 산출하였다.

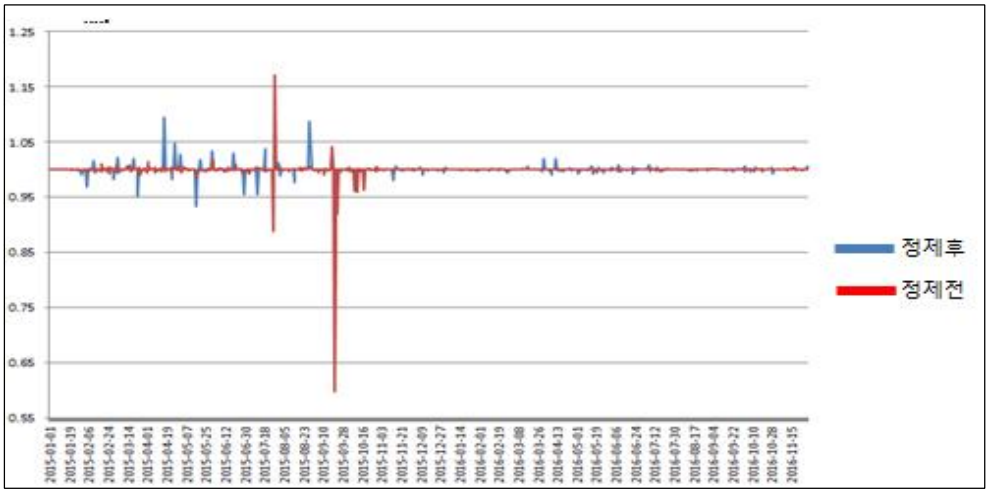
$$R_{t,t-1}^{j} = \prod_{i} \left(\frac{p_t^i}{p_{t-1}^i} \right)^{\frac{1}{nj,t}}$$

$R_{t,t-1}^{j}$ = 개별품목지수, p_t^i = 시점t에서의 상품 I 에 대한 가격,

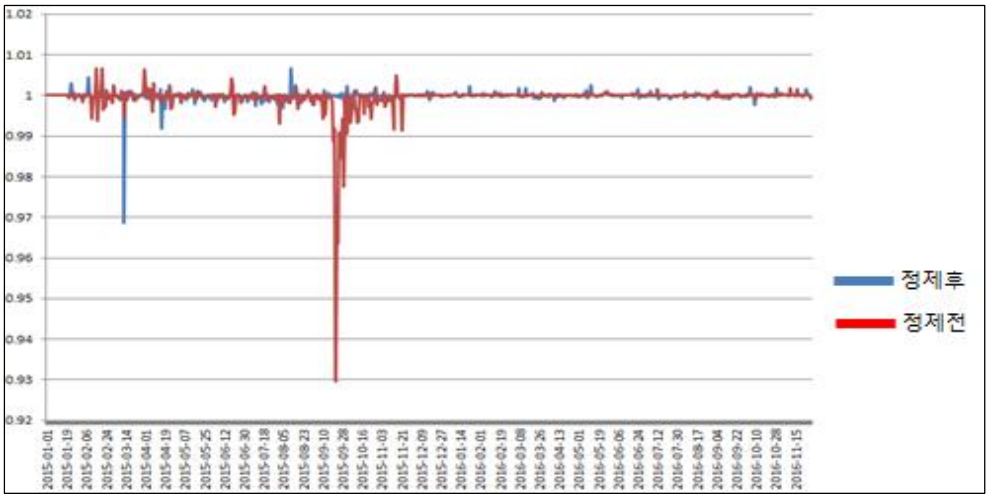
nj,t = 당일 표본에서 추출된 품목 j 에서의 상품의 개수

아래 [그림 2-4] ~ [그림 2-8]까지는 2015년 01월 01일 ~ 2016년 11월 30일까지 정제 전과 정제 후 데이터를 이용하여 개별품목지수를 산출하고 일자별 변동내역을 비교한 차트이다. 차트에서 알 수 있듯이 정제 전 데이터로 산출한 개별품목지수의 경우 매우 불안정하게 변동되고 있다. 특히 2015년도 데이터에서 그 현상이 매우 두드러지게 나타나고 있으며, 이는 2015년도의 원본 가격 데이터의 불안정 요소로 판단된다. 컴퓨터소모품 품목의 2015년 05월 ~ 12월 데이터에서 변동이 매우 심하다. 그러나 정제된 이후에는 많이 안정되었음을 확인할 수 있다. 좀 특이한 점은

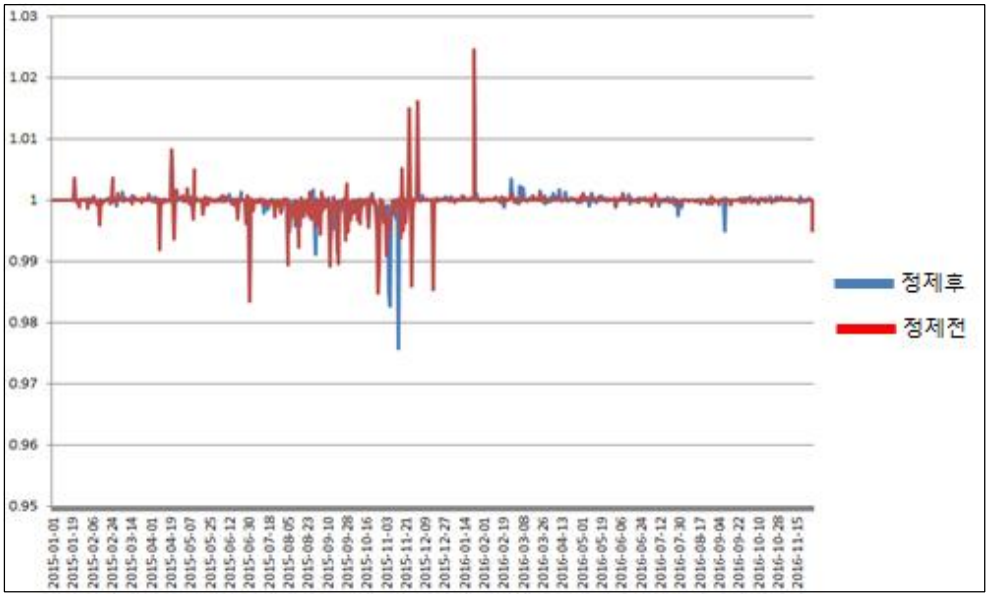
전구 품목의 정제 후 데이터에서 2015년 03월에 아래로 길게 하락한 점이다. 또한 헤어드라이어 품목도 2015년 8월 ~ 12월 데이터에서 오히려 정제 전 데이터보다 더 불안정한 모습을 보인다. 2015년도 몇몇 구간에서 정제 후 데이터도 변동이 심한 모습을 보이지만 대체로 정제 전 데이터로 산출한 개별품목지수보다는 안정적인 모습을 보이고 있다.



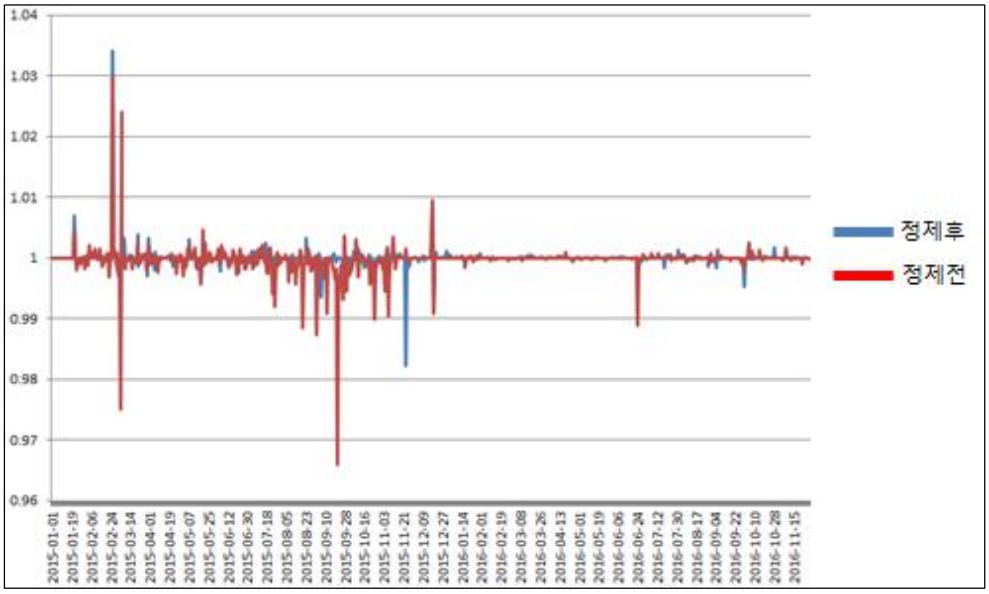
[그림 2-4] 개별품목지수(빵)



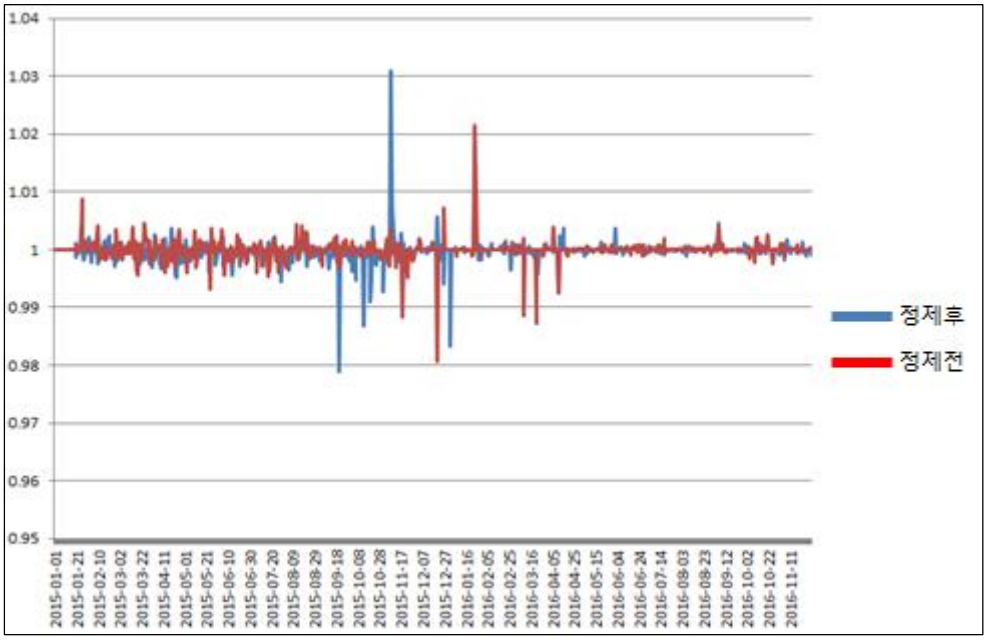
[그림 2-5] 개별품목지수(전구)



[그림 2-6] 개별품목지수(컴퓨터소모품)



[그림 2-7] 개별품목지수(공책)



[그림 2-8] 개별품목지수(헤어드라이어)

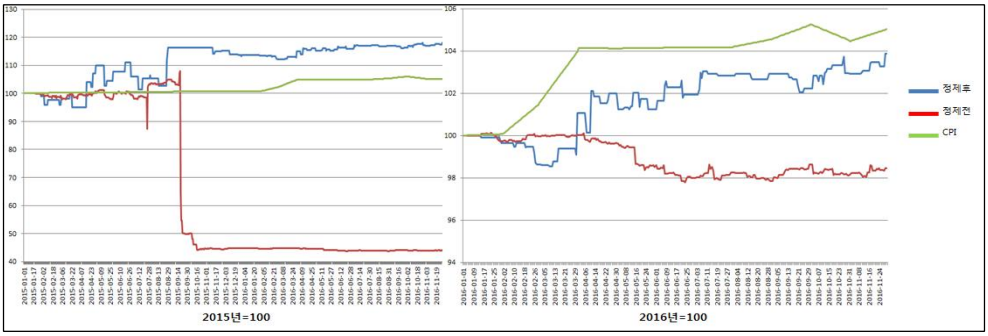
그리고 이렇게 작성된 개별품목지수를 일자별 누적 곱으로 품목 소비자물가지수를 산출하고 CPI와 추이 비교를 수행 하였다. 본 연구에서 산출한 온라인 소비자물가지수와 CPI는 기본이 되는 데이터가 상이하여 단순 비교를 수행하는 것이 의미가 없을 수도 있다. CPI와의 비교에 의미를 부여하기 위해서는 MIT의 BPP와 같이 CPI 산출시 조사 대상이 된 상품의 상세 규격에 해당되는 상품만을 대상으로 온라인 소비자물가지수를 작성하여야 한다. 다만, 본 연구에서 CPI와 비교를 수행하는 이유는 정제후의 데이터로 산출한 온라인 소비자물가지수가 정제전의 데이터로 산출한 온라인 소비자물가지수보다 CPI에 좀 더 가까이 있는지, CPI의 추이를 반영하고 있는지 확인하기 위함이다. 일자별 품목 소비자물가지수는 아래와 같이 Chain Index를 적용하여 기준시점부터 특정시점까지 계산된 개발 품목지수의 누적 곱으로 산출 하였다.

$$I_t^i = R_{1,0}^j \cdot R_{2,1}^j \cdot \cdot \cdot R_{t,t-1}^j$$

※ 시점 t 에서의 품목 소비자물가지수

아래 [그림 2-9] ~ [그림 2-13]은 2015년 01월 01일 ~ 2016년 11월 30일까지 정제 전과 정제 후 데이터를 이용하여 품목별 소비자물가지수를 2015년 01월을 100으로 하는 경우와 2016년 01월을 100으로 하는 2가지 경우를 산출하고 CPI와 함께 비교한 차트이다. CPI도 2015년 01월을 100으로 환산한 경우와 2016년 01월을 100으로 환산한 경우로 비교하였다. CPI는 월 단위이고 온라인 소비자물가지수는 일 단위이기 때문에 CPI의 금월과 전월의 차이를 해당 월의 일수에 균등하게 배분하여 선형적으로 증가하도록 CPI를 일단위로 변경하였다. 따라서 해당 월의 CPI는 시작지점, 추이선이 꺽이는 지점, 끝나는 지점임을 미리 밝혀둔다. 지수의 기준 시점을 2015년과 2016년 두 가지로 측정한 이유는 공개된 원본 데이터에서 2015년 데이터보다 2016년의 데이터가 좀 더 신뢰성을 가지고 있다고 판단하였기 때문이다.

품목 빵의 경우 정제 전 온라인 소비자물가지수는 하락하는 모습이지만, 정제 후 온라인 소비자물가지수와 CPI는 상승하는 모습이다. 정제 후 지수는 2015년을 100으로 한 경우 2016년 지수가 2015년의 영향을 받아 CPI보다 매우 높게 상승한 것으로 보이나 2016년을 100으로 하여보면 CPI 추이와 매우 유사하게 이동하는 것을 확인할 수 있다. 또한 CPI보다 선행하는 모습도 보이고 있다.

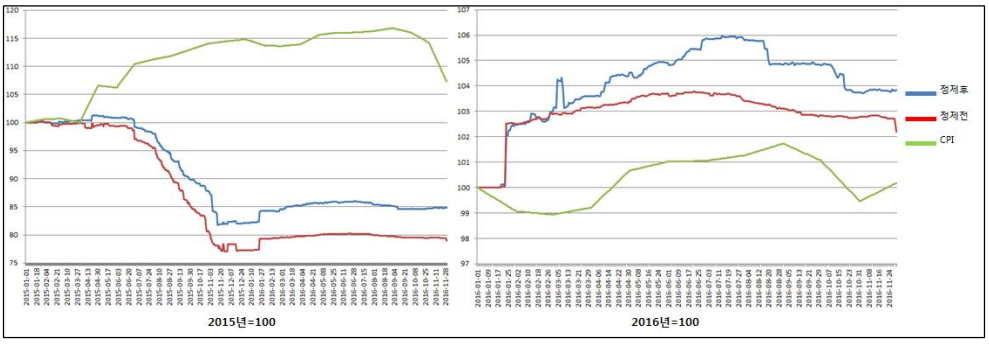


[그림 2-9] 물가지수의 비교(빵)

전구 품목의 경우 정제 전 지수는 2015년을 100으로 한 경우 급격한 하락을 보이고 있다. 정제 후 지수의 경우도 하락의 모습을 보이나 급격한 모습은 아니다. 2016년을 100으로 한 경우를 보면 정제 전/후의 지수가 유사한 추이를 보이고 있다. 그러나 정제 후의 지수가 CPI와 더 유사한 모습을 보이고 있으며, 역시 CPI보다 선행하는 모습을 보이고 있다.

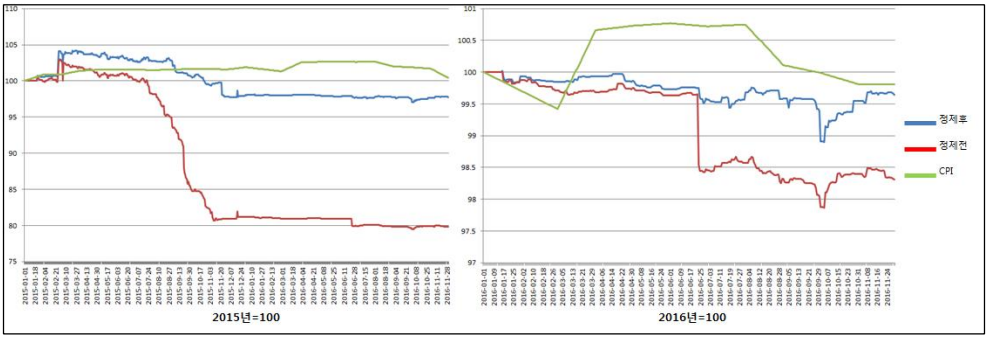
[그림 2-10] 물가지수의 비교(전구)

컴퓨터소모품 품목 경우도 정제 전의 지수보다 정제 후의 지수가 CPI와 유사한 모습을 보인다. 2015년을 100으로 한 경우 전혀 다른 추이를 보이지만, 2016년을 100으로 한 경우 정제 후 지수와 CPI와 크기의 차이는 있지만, 추이는 CPI와 거의 유사한 모습이다. 또한 CPI보다 선행하는 모습도 볼 수 있다.



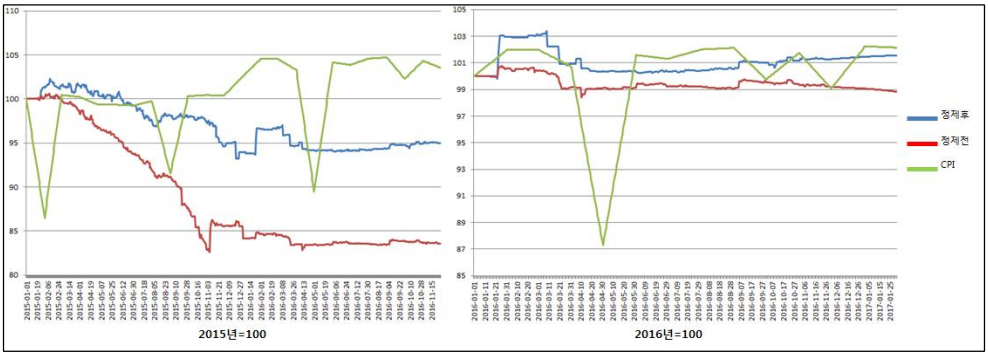
[그림 2-11] 물가지수의 비교(컴퓨터소모품)

공책 품목의 경우에는 전제 전과 정제 후 지수 모두 CPI의 추이를 반영하지 못하는 모습이다. 다만 정제 전의 지수는 급격한 하락의 모습을 보이고 있으나, 정제 후의 지수는 안정적인 모습을 보인다.



[그림 2-12] 물가지수의 비교(공책)

헤어드라이어 품목의 경우 CPI가 매우 불규칙한 모습을 보이고 있다. 정제 전 지수는 2015년을 100으로 한 경우나 2016년을 100으로 한 경우나 모두 하락의 모습을 보이고 있다. 정제 후의 지수는 2016년을 100으로 한 경우 CPI의 추이를 잘 반영하고 있다. 또한 CPI에 선행하는 모습도 보인다.



[그림 2-13] 물가지수의 비교(헤어드라이어)

지금까지 다섯 가지 품목에 대해서 온라인 소비자물가지수의 정제 전과 정제 후 그리고 CPI와의 비교를 수행하였다. 모든 품목에서 정제 전 지수보다 정제 후 지수가 더 안정적이고 CPI를 잘 반영한 것을 확인 하였다. 또한 CPI보다 선행하는 모습도 확인할 수 있었다. 다만 공책 품목의 경우에는 정제 후의 지수도 CPI의 추이를 잘 따라가지 못하는 문제점이 있었으며, 컴퓨터 품목의 경우 CPI의 추세는 잘 따라가지만 지수의 차이가 3이상으로 큰 모습이 보였다. 이는 원천 데이터와 데이터 정제 과정을 확인할 필요성이 있다.

제 3 장 결 론

지금까지 온라인 소비자물가지수에 대한 연구 상황과 온라인 소비자물가지수를 작성하는 일반적인 절차, 그리고 온라인 소비자물가지수를 정확히 작성하기 위해 수집한 가격 데이터를 정제하기 위한 방안을 제시하였고, 실제 데이터를 정제하고 온라인 소비자물가지수를 작성하여 CPI와 비교를 수행하였다.

정확한 온라인 소비자물가지수를 작성하기 위하여 본 연구에서 제시한 온라인 가격 데이터를 정제하는 방안은 다음과 같다.

첫째, 수집된 상품을 해당되는 소비자물가지수 품목에 맵핑하기 위해 기계학습 알고리즘의 한 종류인 분류 알고리즘을 적용하였다.

둘째, 이렇게 분류된 상품들 중 비정상적인 가격 데이터를 정제하기위해 표준편차 또는 백분위수를 활용 하였다.

셋째, 비정상적인 가격의 변화를 보이는 상품을 정제하기위해 금일가격/어제가격의 비율을 활용하였다.

이렇게 제시한 세 가지 데이터 정제 방안으로 가격 데이터를 정제하여 품목 소비자물가지수를 작성한 결과 데이터 정제 후 온라인 소비자물가지수가 정제 전 지수보다 좀 더 안정적이고 CPI를 잘 반영하는 것을 직접 확인 하였다. 본 연구에서는 품목별 소비자물가지수를 작성하는 것으로 하였지만, 만일 온라인에서 수집 가능한 전체 품목을 대상으로 품목별 가중치 없이 온라인 소비자물가지수를 작성하고자 한다면, 데이터 정제를 위한 분류 알고리즘의 적용은 의미가 없을 수도 있다.

이번 연구에서 제시한 정제방안의 구조와 절차 그리고 알고리즘 등에 대해서 논란의 여지가 있을 수 있다. 그러나 이번 연구를 통해서 정확한 온라인 소비자물가지수를 작성하기 위해서는 최소한의 데이터 정제가 반드시 필요하다는 것은 확인할 수 있었다.

본 연구에서는 온라인 가격 데이터의 기본 분석을 통하여 표준편차가 매우 높은 5개의 품목을 선정하여 데이터 정제를 수행 하였지만, 향후 국가 통계로서 활용되기 위해서는 더 정교한 데이터 정제 방안으로 다른 품목으로 확대 적용하는 연구가 필요할 것으로 생각된다. 통계청에서 2013년을 시작으로 온라인 소비자물가지수와 관련하여 연구를 진행하고는 있지만, 아직까지 국내에서는 연구가 활발하게 이루어지지 않고 있는 게 현실이다. 현재의 소비자물가지수가 갖고 있는 현실과의 괴리 해결, 소비자가 물건을 구매하는 채널의 변화(오프라인에서 온라인으로)를 소비자물가지수에 빠르고 정확하게 적용할 수 있는 방안 등 여러 사안들을 해결하기 위해서는 다양한 연구가 필요하다. 또한 소비자의 상품에 대한 구매빈도, 개별 상품의 판매량 등도 소비자물가지수에 많은 영향을 주는 변수이므로 이를 적용하기 위한 연구도 진행될 필요성이 있다.

참고문헌

[1] 통계청, 통계청 통계 설명 자료, (04,18,2017) “http://meta.narastat.kr/metasvc/index.do”

[2] 조지성,김관수,안동환, “가공식품의 소비자체감물가지수 개발을 위한 연구”, 소비문화연구 제17권 제4호, 2014

[3] MIT, BPP Web Site, (04,27,2017) “http://bpp.mit.edu”

[4] 전자신문, 물가지수에 반영하는 온라인쇼핑 2배 높인다, (04,28,2017) “http://www.etnews.com/20161129000215”

[5] 강유진, “온라인 마켓 상품 댓글 정보를 이용한 온라인 물가지수”, 고려대학교 석사학위논문, 2015

[6] 온라인 위키백과, 나이브 베이즈 분류, (04,28,2017) “https://ko.wikipedia.org/wiki”

[7] 황성혁, 이정희, “인터넷쇼핑의 성장이 소비자 물가에 미친 영향”, 한국중소기업학회, 기업가정신과 벤처연구(JSBI)(구 벤처경영연구), 17권, 1호, pp.19-30, 2014

[8] 이상호, “온라인 물가지수 분석을 위한 빅데이터 융합분석 방법”, 한국데이터사이언스학회, 11, 2013