반려동물 대규모 빅데이터 분석을 위한 시각화

Visualization for Large-Scale Big Data Analysis of Pets

황선우[1]· 오하영[2*]

Sun-Woo Hwang[1] · Ha-Young Oh[2*]

[1]Undergraduate Student, Department of Aplied Artificial Intelligence, Sungkyunkwan University, Seoul, 03063 Korea
[2]*Associate professor Professor, College of Computing & Informatics, Sungkyunkwan University, Seoul, 03063 Korea

요 약

오늘날 반려동물을 양육하는 사람들은 과거와는 반려동물을 다르게 취급한다. 그들에게 반려동물은 애완동물이 아닌 가족의 구성원이며, 그러므로 복지가 이루어져야할 대상이다. 이는 다양한 반려동물 상품의 출시와 반려동물 관련 산업의 가파른 성장을 불러왔으며, 그 시장의 가치는 매우 크다. 확대되는 시장에서 소비자를 유치하기 위해서는 그들의 선택을 단순화 해줄 수 있는 반려동물 맞춤형 추천 시스템이 필요하다. 본 논문에서는 추천 시스템을 구축하기 위한 데이터를 워드 클라우드, 파이차트, 꺾은선 그래프 등으로 시각화하고, 상품에 대한 소비자의 의견을 분석한다.

ABSTRACT

Today, people who raise pets treat pets differently from the past. For them, pets are family members, not pets, and therefore well-being of them is the subject to be achieved. This has led to the launch of various pet products and the exponential growth of the pet-related industry, and the value of the market is very high. In order to attract consumers in the expanding market, a customized recommendation system for pets is needed to simplify their owner's choices. In this paper, data for building a recommendation system are visualized with word clouds, pie charts, and line graphs, and consumer opinions on products are analyzed.

키워드 : 후기 시각화, 평점 시각화, 감정 분석

Keywords : Review Visualization, Rate Visualization, Sentiment Analysis

Received , Revised , Accepted
(출판사에서작성)

* Corresponding Author: Hayoung Oh (E-mail: hyoh79@gmail.com Tel:+82-2-583-8585)
Global Convergence, Sungkyunkwan University, Seoul, 03063 Korea

Ⅰ. 서 론

최근 반려동물을 양육하는 가구 및 양육되는 반려동물 수가 증가하고 있다. 농림축산식품부의 「2020년 동물보호에 대한 국민의식조사」에 의하면, 2015년부터 2020년까지 전국 반려동물 양육 가구는 457만 가구에서 638만 가구로 약 179만 가구 많아졌다. 또한 반려동물에 대한 인식 역시 변화하고 있다. 반려동물은 과거에는 애완동물이라는 인식이 강하여 소비자들이 반려동물의 사료 및 용품에 많은 비용을 투자하지 않았다. 그러나 현재는 반려동물이라는 명칭에서도 알 수 있듯이 반려동물 양육 가구에서 반려동물은 가족의 구성원으로 인식되는 경향이 두드러지게 나타난다. 그에 따라 인간처럼 반려동물의 복지를 고려하는 다양한 상품들이 등장하고 있다. 이는 반려동물 연관산업 시장 규모가 빠르게 커지고 있는 추세를 통해 확인할 수 있다. 한국농촌경제연구원은 2018년부터 2020년까지 1조에 가까운 성장이 있었음을 보고하며, 국내 반려동물 연관 산업의 시장 규모가 2027년에 이르러서는 6조 이상이 될 것으로 추산했다.

앞으로 크게 팽창할 것으로 예상되는 반려동물 시장에서 소비자를 확보하기 위해서는 반려동물에게 개인화된 추천 시스템이 필수적으로 요구될 것이다. 시장에는 반려동물의 복지를 추구하는 여러 소비자의 요구를 충족시키기 위해 다양한 상품 및 서비스가 출시되고, 각 동물들의 특수한 성질을 반영한 상품이 등장하고 있다. 그러므로 앞으로 시장 규모가 확대됨에 따라 소비자가 고려해야할 요소들이 복잡하게 증가하는 것은 예견된 미래이다. 반려동물 맞춤형으로 개인화된 추천 시스템은 복잡한 반려동물 시장에서 서비스 이용자의 반려동물에게 필요한 상품 및 서비스를 걸러 제안하여 사용자가 더 효율적인 소비를 할 수 있게 도울 수 있다.

반려동물 맞춤형 추천 시스템을 개발하기 위해서는 먼저 반려동물 상품과 그 상품들에 대한 소비자들의 반응 및 소비자들의 특성을 분석해야 한다. 그리고 이를 분석하려면 반려동물 상품 및 후기 데이터를 분석해야 한다. 본 논문에서는 상품과 상품에 대한 사용자들의 평가인 후기를 분석하여 시각화하고 추천 시스템에 사용될 데이터의 기반을 마련하였다.

2장에서는 추후 추천 시스템 개발에 참고할 관련 연구를 소개하였으며, 3장에서는 추천 시스템을 구축하는데 사용될 데이터를 설명하고 사용 방식을 안내하였다. 4장에서는 3장의 데이터를 분석하기 위해 이루어진 시각화와 시각화 결과가 시사하는 바를 도출했으며, 5장에서는 데이터 분석 결과와 한계 및 향후 연구를 제시한다.

Ⅱ. 관련 연구

2.1 영양성분 프로파일링 기반 사료추천 알고리듬

최근, 추천시스템은 다양한 분야에 적용되어 보편적으로 사용되고 있다. 그러나 반려견의 상태에 맞춰 필요한 영양성분을 가장 적절하게 포함하고 있는 사료를 제안하는 추천시스템은 부재중인 실황이다. 반려견 사료 추천시스템 알고리즘은 다음과 같은 세 가지 이유로 기존 추천시스템만을 이용하여 구축하기에는 한계가 있다. 첫째, 사용자의 선호도를 활용하여 제품 또는 서비스를 제안하는 기존 추천 시스템과는 달리 사료 추천시스템은 사용자로부터 입력 받은 반려견의 상태 및 특성에 맞춰 적합한 사료를 추천해주어야 한다. 둘째, 반려견에게 적합한 사료뿐만 아니라 부적합한 사료도 고려한 추천시스템을 구축해야 한다. 셋째, 사료를 구매한 사용자의 후기와 더불어 사료에 대한 전문가의 평가를 결합하여 보다 반려견의 상태에 알맞은 제품을 추천해줄 수 있어야한다. 위와 같은 한계점을 보완할 수 있는 사료추천 알고리즘을 설계하기 위해 재료, 영양소, 요리방법 등의 다양한 요소들을 고려해야 하는 조리법 추천에 사용되는 알고리즘을 활용하여 영양 성분 프로파일링 기반의 사료추천 알고리즘(NRA; Nutrient profiling-based Recommendation Algorithm)을 고안했다. 이 시스템의 성능을 평가하기 위해 제품을 구매한 사용자들의 만족 또는 불만족 후기만을 고려하여 사료의 적합성을 예측하는 베이스라인 모델의 성능 또한 측정해본 결과, 고안한 알고리즘의 AUC와 F1값이 베이스라인 모델보다 높음을 확인할 수 있었다. 그러나 질병견의 경우 알고리즘의 성능이 건강견의 성능보다 낮게 나타나는 문제가 발생했는데, 이는 어떤 질병견과 완벽하게 같은 질병들을 앓고 있는 질병견의 급여후기가 많지 않기 때문에 발생한 것이다.

2.2 빅 데이터를 활용한 애완동물 상품 추천 시스템 구현

키우는 반려동물의 수가 급증하면서 최근 반려동물 개인화 추천 시스템의 필요성이 대두되고 있다. 이 시스템은 사용자의 SNS, 쇼핑몰 내에서 클릭한 아이템 및 검색정보, IoT센서, 직접입력한 반려견의 정보 등을 이용하여 사용자의 정보를 수집하는 것을 시작으로 군집화, 메타데이터 등의 기능을 수행하는 데이터 정제 및 통합 모듈을 거쳐 의미 있는 데이터를 추출한다. 첫번째 전처리과정인 데이터 정제는 오류를 보정하고 분포를 정규화하는 과정이다. 두번째 과정인 데이터 통합은 정제된 데이터들을 상관성 분석 등의 작업을 거쳐 데이터를 통합하는 과정이다. 전처리를 거친 데이터는 데이터 분석기에 다시 사용되는데, 데이터 분석기는 성능을 측정할 수 있는 데이터 분류 모델을 사용한다. 개인 맞춤 서비스를 추천하기 위해 추천 분석기는 코사인 유사도를 이용하여 상품간 유사도를 분석하여 군집분석을 수행한다. 군집분석이 완료되면 이를 바탕으로 사용자에게 맞춤형 개인화서비스를 제공할 수 있게 된다.

2.3 레시피 데이터 기반의 식재료 궁합 분석을 이용한 레시피 추천 시스템 구현

본 논문에서는 두 포털에 공개된 레시피 데이터를 수집하고, 레시피 제목과 식재료 정보만을 파싱하여 레시피 데이터를 획득한다. 획득한 데이터에서 결측값 수식어를 삭제하는 방법으로 데이터를 정제하고, 전처리 및 후처리를 위한 대체가능한 식재료의 데이터 리스트를 작성한다. 데이터 추출 모듈을 통해 불필요한 수식어 등을 정제하는 과정을 거쳐 대체가능한 식재료로 전처리된 레시피는 데이터 관리 모듈에 의해 레시피 데이터로 저장된다. 레시피 식재료 데이터는 Word2Vec를 통해 벡터화가 이루어져 사용자가 입력한 식재료와 저장된 레시피 식재료 데이터와의 유사도를 계산할 수 있게 한다. 유사도 계산을 통하여 코사인 유사도가 일정 수준이상인 재료들을 궁합이 맞는 식재료로 추천하고, 추천된 식재료 및 입력한 식재료를 사용하는 레시피를 제안한다. 사용자가 레시피를 선택하면 그 레시피를 화면에 띄우는 단계를 끝으로, 식재료 궁합에 따른 추천 레시피 제공이 완료된다.

2.4 YOLO 기반 개체 검출과 Node.js 서버를 이용한 반려견 행동 분류 시스템 구현

본 논문은 반려견이 직접 착용해야 하는 IoT기기의 한계를 언급하며, 카메라로 반려견의 행동을 분석하고 정해진 카테고리로 분류하여 실시간으로 사용자에게 보고하는 시스템을 개발하였다. 시스템을 구축하는데 요구되는 기술은 크게 3가지로, 반려견을 인식하고 위치를 파악하는 기술, 반려견의 행동을 분류하는 기술, 분류 결과를 사용자에게 전송하는데 사용될 웹서버 기술이 있다. 먼저 반려견을 인식하고 위치를 파악하는 기술인 object detection은 YOLO를 이용한 학습으로 구현하였고, 웹 기반 이미지 학습 도구인 Teachable Machine의 모델을 통해 반려견의 행동을 주어진 9가지 분류체계에 맞춰 분류하는 학습을 하였다. 그 결과 객체 검출 성능은 인간의 인지능력과 비슷한 수준까지 높일 수 있었고, 행동의 종류는 98.9%의 정확도로 분류하는 모델을 구현할 수 있었다.

2.5 반려동물 사료 추천시스템을 위한 유사성 측정 알고리즘에 대한 연구

본 논문은 콘텐츠 기반 추천시스템을 사용한 반려동물 사료 추천시스템을 구현하기 위해 7대 주요 영양소에 해당하는 조지방, 조섬유, 조단백질, 조회분, 인, 수분, 칼슘을 이용하여 사료를 군집화했다. 국내 유통 중인 260여 개의 사료 중 실험을 위해 임의로 100개의 사료를 추출한 후, 각 영양소의 성분량을 정규화하는 과정을 거쳐 데이터를 정제했다. 이후 성분량이 정제된 사료 데이터를 계층적 클러스터링하기 위해 유클리드 거리, 쟈카드 거리, 코사인 거리, 맨하튼 거리를 이용하여 사료간의 유사도를 계산했다. 이와 같은 방법으로 개발된 추천 알고리즘은 영양성분이 유사한 사료를 추천할 수 있지만, 보다 높은 군집 정확도를 위해서는 원재료 데이터에 대한 분석 역시 요구된다. 또한 사용자와 사용자의 반려동물에 대한 빅데이터를 수집하여 분석한 후 이를 추천 시스템에 반영하면 더 개인화된 추천 서비스를 사용자에게 제공할 수 있을 것이다.

2.6 개체명 인식을 이용한 반려동물질병 질의응답시스템

이 논문에서는 한글로 반려동물의 증상을 입력했을 때 예상되는 질병명 및 동물병원 방문 여부를 판단해주는 반려동물 질병 개체명 모델을 제안한다. 모델은 네이버 지식 IN의 수의사 전문가가 답변을 한 질문과

답변 쌍을 데이터로 사용한다. 이 데이터에서 질병 개체명을 추출하기 위해 질병명 사전에 있는 단어들을 음절 단위로 인식하게 하여 학습 데이터의 형태소를 분석하고 개체명 인식 모델을 학습시킨다. 형태소가 분석된 후 일정 빈도 이상 언급된 질병명을 추출하여 사전을 만들었다. 학습된 질병 개체명 인식 모델은 사용자의 입력 내용과 유사한 상위 3개의 문서와 예상 질병을 함께 제시하고, 동물병원 찾기 서비스도 함께 제공한다.

Ⅲ. 연구 데이터

3.1 시각화에 사용한 데이터

본 논문은 주식회사 funNC로부터 제공받은 dataset을 사용하여 시각화를 진행했다. 제공받은 dataset은 "강아지대통령"에서 수집된 상품 정보, 상품 카테고리 정보, 주문 정보, 구매 후기 등의 정보를 포함한다. 본 논문은 카테고리와 상품 정보, 구매 후기, 평점만을 사용하여 데이터를 시각화를 하고 사용자의 경향성을 분석하는 것을 목표로 설정했다. 그러므로 '카테고리 코드(category)'를 기본키로 하며 '카테고리 이름(catnm)'을 속성으로 갖는 gd_category.csv와 '상품 고유 번호(goodsno)'를 기본키로 하고, '카테고리 코드(category)'를 외래키로 갖는 gd_goods_link.csv, 그리고 '상품 고유 번호(goodsno)'를 외래키로 갖고 '후기 제목(subject)', '후기 본문(contents)', '평점(point)'를 속성으로 갖는 gd_goods_review.csv 파일만을 사용하였다.

gd_category.csv의 category는 상품들이 분류되는 카테고리들을 정수형으로 표현한 속성이다. 카테고리는 계층적인 성격을 갖고 있는데, 상위 카테고리는 2자리 정수로 표현되고 하위 카테고리는 상위 카테고리의 두 자리 정수를 앞의 두자리로 갖는 5자리 정수로 표현된다. 상위 카테고리는 '사료', '치아', '건강관리', '위생/배변', '미용/목욕', '급식기/급수기', '하우스/울타리', '이동장', '의류/액세서리', '목줄/인식표/리드줄', '장난감', '훈련', '간식'그리고 속성값을 문자열로 갖는 catnm을 통하여 해당 카테고리 코드가 어떤 카테고리를 가리키는지 알 수 있다. 예를 들어 catnm이 '사료'인 카테고리의 category가 36일 때, category가 36,004인 카테고리는 '사료'를 catnm으로 갖는 카테고리의 하위 카테고리에 속하는 것이다. 실제로 category가 36004인 카테고리는 catnm이 '건식사료'인 것을 확인 할 수 있다. gd_goods_review.csv는 외래키인 goodsno를 사용하여 gd_goods_link.csv의 category를 가져 올 수 있으므로, 모든 후기의 카테고리를 분류할 수 있게 된다.

step1: 상위카테고리딕셔너리 생성

1. 파일 열기
   category_df <-
   openRead('dogpre.gd_category.csv')
   goods_link_df <-
   openRead('dogpre.gd_goods_link.csv')
   good_review_df <-
   openRead('dogpre.gd_goods_review.csv')

2. 상위카테고리('category' 값이100이하인 경우 상위카테고리) 추출 후'category'의 값을key로 하고 이 외의 열의 값들을value로 하는 딕셔너리 생성
   super_category <- **if** row of category_df['category']<100 **then** row of category_df
   supcate_dic <- {value of super_category['category']: value of other column of super_category}

step2: 전체 상품의 하위 카테고리를 상위 카테고리로 재분류

1. 중복 상품번호 제거
   goods_link_df <- **if** row of goods_link_df['goodsno'] is exist **then** remove the row

2. 상품의 상위카테고리 열의 값을0으로 초기화하고 하위카테고리와 상위카테고리 딕셔너리에 따라 값을 수정

```
super_category <-
if row of category_df 'category'<100
then row of category_df
for i ∈supcate_dic do
    if goods_link_df['category']= i
      then goods_link_df['supcate']= i
    elseif (goods_link_df['category']>=i*1000)

&(goods_link_df['category']<(i+1)*1000)
      then goods_link_df['supcate']= i
```

3.2 시각화 방식

Python의 여러 모듈을 사용하여 시각화 작업을 하였다. matplotlib, plotly 등의 라이브러리를 사용하여 워드 클라우드, 선 그래프, 파이 차트를 생성하였다.

Ⅳ. 시각화 결과

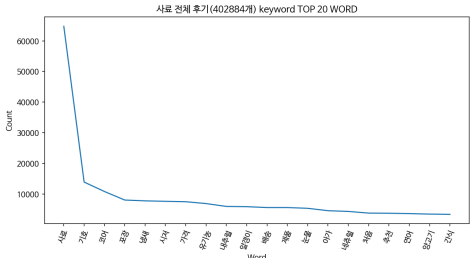

그림1. '사료' 카테고리 전체 후기 중 빈도수 상위 20개 단어

'사료' 카테고리 전체 후기 중 빈도수 상위 20개 단어

1. '사료' 카테고리 후기에서 명사만을 추출하는 전처리를 위한 라이브러리 불러오기
   m <- use library Mecab

2. '사료' 카테고리 후기에서 한글만을 추출하는 전처리를 위한 함수 정의 및 사용
   def kextract(s):
     hangul <-
     extract except of Hangul from review
     result <- **if** hangul **then**' '
     kextract(review_cate['content'])
     kextract(review_cate['subject'])
     if row of review_cate['content'] is NULL
      then remove the row

3. 꺾은선 그래프 생성을 위한 모듈 및 한글 폰트 설치
   install wordcloud
   download Korean font

4. 꺾은선 그래프 생성('사료' 카테고리 전체 후기 중 빈도수 상위 20개 단어)
   for r ∈ length of current_point **do**
     tmp <-subject_list[r]&content_list[r]
     tokens <- use m to tokenize tmp
     tokens <- **if**(tokens is not in stopword)
               and (length of token is not 1)

               then tokens
     insert tokens in dataset

   extract the top 20 words of frequency
   set x axis
   set y axis

   set axis name

   create line graph

반려동물 시장에서 가장 큰 비율을 차지하는 상품은 사료이므로 전체 후기 중 '사료' 카테고리의 후기를 대상으로 등장하는 단어들의 빈도수를 측정해보았다. 그 결과 카테고리 이름에 해당하는 '사료'가 다른 단어들보다 압도적으로 많은 횟수 언급되었으며, 뒤이어 반려견의 사료에 대한 반응을 나타내는 단어인 '기호', 보관기간 및 방식을 결정하는데 영향을 미치는 '포장', 그리고 사료를 섭취한 후 보는 변의 '냄새', '눈물' 증상 호전에 대한 후기들이 많아 위의 단어들이 많이 언급되었다. 이 외에는 상품명에 자주 언급되거나 배송 등의 단어가 순위 내에 기록되어 있다.

데이터의 하위 카테고리를 상위 카테고리로 분류하는 작업을 통하여 모든 상품 및 후기의 카테고리 분류를 시행하였다. 카테고리는 카테고리 이름과 카테고리 코드가 함께 저장되어 있는 dogpre.gd_category.csv 파일에서 추출할 수 있었다. 상위 카테고리는 10 이상 100 이하의 자연수로, 사료, 간식, 건강관리 등의 포괄적인 카테고리이다. 하위 카테고리는 상위 카테고리 내의 카테고리로, 사료라는 상위 카테고리에는 습식사료, 건식사료 등의 하위 카테고리가 있다. dogpre.gd_goods_link 파일은 모든 상품에 대한 데이터를 저장하고 있는 파일로, 상품 고유의 key에 해당하는 상품번호와 카테고리 정보를 포함하고 있다. 카테고리는 하위 카테고리로 분류된 경우가 많았는데, 모든 상품의 카테고리 분포를 보기 위하여 하위 카테고리를 상위 카테고리로 변환하여 상품의 상위 카테고리를 저장하는 작업을 수행하였다. 즉, 모든 상품을 분류 안됨, 의류/액세서리, 간식, 목줄/인식표/리드줄, 사료, 하우스/울타리, 장난감, 미용/목욕, 이동장, 급식기/급수기, 사은품, 위생/배변, 건강관리, 생활/가전, 훈련, 치아 순으로 총 16개 상위 카테고리로 재분류했다. 그 결과 다음과 같은 분포를 확인할 수 있었다.

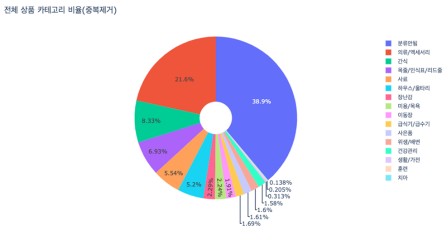

그림2. 전체 상품 카테고리 비율

전체 상품의 카테고리 비율을 나타낸 파이 차트 생성

1. 각 상위 카테고리 당 상품 개수를 pandas series 또는 딕셔너리 자료형으로 나타냄
   cate_counting <-
   {goods_link_df['supcate']: # of value}

2. 파이차트생성(전체 상품의 상위 카테고리 비율)
   labels <- super category's name
   values <- # of super category
   create pie chart

모든 상품의 후기에 대한 상위 카테고리 분류를 시행한 이후에는 모든 후기에 대한 상위 카테고리 분류 또한 시행했는데, 이는 상품번호와 후기 정보를 포함하는 dogpre.gd_goods_review.csv 파일과 이전에 모든 상품에 대하여 상위 카테고리를 분류한 작업을 통하여 할 수 있었다. 상위 카테고리가 분류된 상품 번호와 dogpre.gd_goods_review.csv 파일의 상품 번호가 같은 경우, 해당 후기의 상위 카테고리를 이전에 변환한 상위 카테고리로 설정하였다. 위와 같은 방법으로 후기에 대한 상위카테고리 분류를 시행한 결과, 간식 카테고리의 후기가 가장 많은 것을 확인할 수 있었다.

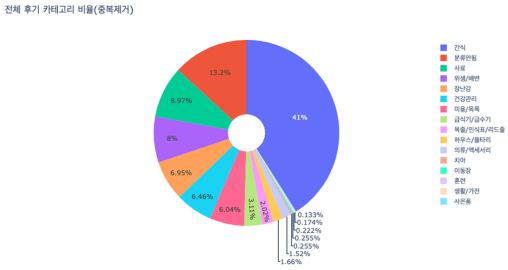

그림3. 전체 후기 카테고리 비율(중복 제거)

전체 후기의 카테고리 비율을 나타낸 파이 차트 생성

1. 상품번호를key로 왼쪽 외부 조인하여 후기 데이터프레임에 카테고리 정보 추가하고 카테고리 정보를 추가할 수 없을 경우 상위카테고리를0으로 설정
   review_cate <-goods_review_df ⋈goods_link_df
   if review_cate['supcate'] is NULL
   then review_cate['supcate'] <- 0

2. 각 상위 카테고리 당 후기 개수를pandas series 또는 딕셔너리 자료형으로 나타냄
   rcate_counting <-
   {review_cate['supcate']: # of value}
3. 파이차트 생성(전체 후기의 상위 카테고리 비율)
   labels <- super category's name
   values <- # of super category
   create pie chart

모든 상품에 대하여 상위 카테고리를 분류한 경우와 모든 후기에 대하여 상위 카테고리를 분류한 경우 모두 분류되지 않은 경우가 상당히 많은데, 이 원인은 dogpre.gd_goods_link 파일에

dogpre.gd_category.csv 파일에는 존재하지 않는 카테고리 코드가 있기 때문인 것으로 추측된다. 파일에 카테고리 코드가 존재하지 않으면 해당 상품의 상위 카테고리를 분류할 수 없고, 어떤 상품의 상위 카테고리를 분류할 수 없으면 해당 상품의 후기의 상위 카테고리를 분류할 수 없기 때문이다. dogpre.gd_category.csv 파일에는 존재하지 않으며 dogpre.gd_goods_link 파일에서만 등장하는 카테고리 코드는 현재는 사용되지 않는 과거에 사용하던 카테고리 코드로 추측된다. 보다 많은 상품과 후기의 카테고리를 분류하여 각 카테고리의 데이터로 사용하기 위해서는 과거의 카테고리 코드 데이터를 담고 있는 파일이 필수적일 것으로 예상한다.

구매자들의 상품에 대한 의견은 구매 후기를 통해 드러난다. 후기 데이터에는 구매자가 상품을 평가한 지표인 평점이 함께 제공되는데, 이를 이용하여 카테고리별로 긍정 및 부정 데이터 분류를 수행했다. 먼저 긍정과 부정 각각으로 분류되는 평점의 지표를 설정하기 위해 전체 후기들의 평점 분포를 파이차트로 시각화하여 확인하고, 각 평점에서 빈도수가 높은 단어들을 워드클라우드로 확인했다. 모든 후기의 평점의 분포를 시각화하기 전에 평점 값을 확인해본 결과 1점에서 5점까지의 값이 아닌 0점이 존재하는 경우도 전체 4,490,421건의 후기 중 7건 확인해볼 수 있었다. 0점인 후기 데이터가 어떤 내용을 포함하는지 출력해 본 결과, '좋아요 강아지가 좋아합니다.', '굿'과 같은 긍정적인 어휘들을 사용한 것을 확인 할 수 있었다. 이 데이터들은 유효하지 않은 평점을 포함하고 있으므로 전체 후기 데이터 평점 시각화를 할 때는 제외하여 총 4,490,414개의 후기 데이터에 대하여만 시각화를 진행했다.

표1. 이상치 제거 후 후기(4,490,414개) 평점 비율

평점	5	4	3	2	1
개수	3139856	885561	327637	83278	54082
비율(%)	69.9	19.7	7.3	1.85	1.2

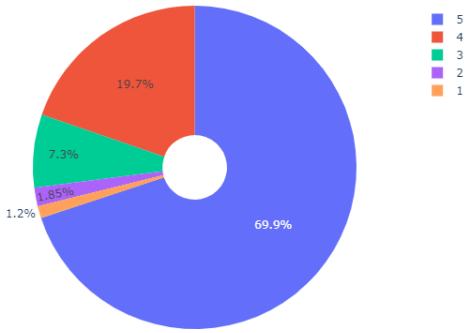

그림4. 이상치 제거 후 후기 평점 비율

전체 후기의 평점 비율 파이차트 생성

1. 평점 당 후기 개수를 pandas series 또는 딕셔너리 자료형으로 나타냄
   point_counting <- {review_cate['point']: # of value}
3. 파이차트 생성(전체 후기의 상위 카테고리 비율)
   labels <- points
   values <- review # of points
   create pie chart

[표1]에서 5점이 70%에 가까운 비율을 차지하고 있는 것을 확인할 수 있었다. 긍정 데이터와 부정 데이터 평점의 지표를 선정하기 위하여 각 평점을 기준으로 빈도수가 높은 100개 단어를 선정하여 워드클라우드를 생성해본 결과는 아래와 같다.

평점별 워드클라우드 생성

1. 모든후기(제목, 내용)에서 명사만을 추출하는 전처리를 위한 라이브러리 불러오기
   m <- use library Mecab

2. 모든후기(제목, 내용)에서 한글만을 추출하는 전처리를 위한 함수 정의 및 사용
   def kextract(s):
     hangul <-
     extract except of Hangul from review
     result <- **if**hangul**then**' '
     kextract(review_cate['content'])
     kextract(review_cate['subject'])
     if row of review_cate['content'] is NULL
      then remove the row

3. 워드클라우드생성을 위한 모듈 및 한글 폰트 설치
   install wordcloud
   download Korean font

4. 워드클라우드 생성(5~1점까지 평점별로 워드클라우드 생성)
   for i ∈ [5,4,3,2,1] **do**
     current_point <- review_cate['point'] = i
     subject_list <- current_point['subject'] to list
     content_list <- current_point['contents'] to list

     define stopwords

     for r ∈ length of current_point **do**
       tmp <-subject_list[r]&content_list[r]
       tokens <- use m to tokenize tmp
       tokens <- **if**(tokens is not in stopword)
                 and (length of token is not 1)

                 then tokens
       insert tokens in dataset

       extract the top 100 words of frequency

       create wordcrowd

그림5. 전체 후기 중 평점이 5점인 후기로 생성한 워드클라우드

[그림5]에는 '조아', '만족', '최고', '강추', '추천' 등의 긍정적인 단어들의 크기가 크다.

그림6. 전체 후기 중 평점이 4점인 후기로 생성한 워드클라우드

[그림6]은 '조아', '향기' 등의 긍정적인 단어와 '기호성'에서 분해된 것으로 모이는 '기호'라는 단어가 빈도수가 높은 것으로 확인된다.

그림7. 전체 후기 중 평점이 3점인 후기로 생성한 워드클라우드

[그림7]에는 [그림5], [그림6]에서도 공통적으로 나타나는 '조아', '간식'이 많이 등장한다. '간식'에 대한 후기가 1,842,271개로 가장 많아 여러 평점의 워드클라우드에 등장하는 것으로 추측된다.

그림8. 전체 후기 중 평점이 2점인 후기로 생성한 워드클라우드

[그림8]에는 부정적인 단어들이 두드러지게 등장하기 시작한다. '실망', '불편', '냄새' 등의 단어 크기가 커졌다.

그림9. 전체 후기 중 평점이 1점인 후기로 생성한 워드클라우드

평점 별로 워드클라우드를 생성하여 빈도수가 높은 단어들을 확인해본 결과, 평점이 5점 및 4점인 경우에는 '조아, 만족, 최고, 강추, 감사, 추천' 등의 긍정적인 단어들이 있고 부정적인 단어를 찾기 어려웠다. 반면 평점이 1,2점인 경우에는 '실망, 불편, 실패, 비추, 별루' 등의 단어들이 공통적으로 두드러지게 등장하는 것을 확인할 수 있었다. '냄새'는 [그림5]에서 [그림9]까지의 모든 워드클라우드에서 공통적으로 빈도수가 높은 단어이다. '냄새'는 4,5점에서는 '향기'와 함께 등장하는 반면 1,2점에서는 '향기'를 찾을 수 없었다. 평점이 중간값인 3점인 경우에는 4,5점에서 등장하는 '조아, 만족'같은 긍정적인 단어들과 함께 '보통, 불편'과 같은 다소 부정적인 어감이 드러나는 단어 또한 무시하기 어려운 빈도수로 등장했다. 그러므로 평점이 4,5점인 경우에는 긍정 데이터로, 1,2점인 경우에는 부정 데이터로 분류하기로 결정했다. 3점인 경우 시각화 시에는 표시했지만, 감성분류를 시행할 때는 사용하지 않을 예정이다.

13개 카테고리에 대하여 각각 평점 분류를 시행한 결과, 다음과 같은 비율이 나왔다.

각 카테고리의 평점 비율 파이차트 생성

1. 모든 카테고리의 후기 개수를 그래프에 나타내기 위해 카테고리별 후기 개수를 딕셔너리로 변환
   rcate_counting <-
   {review_cate['supcate']: # of value}
3. 파이차트 생성(각 카테고리의 평점 비율)
   for c ∈ rcate_countingdo
   c_dataframe <-
   if row of review_cate[‘supcate’]==c
   then the row
   c_coounting <-
   {c_dataframe['point']: # of value}
   labels <- points
   values <- review # of points
   create pie chart

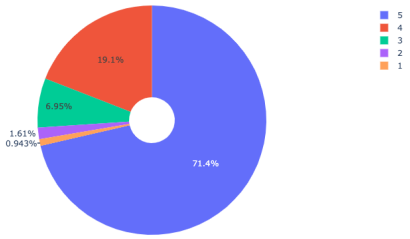

그림10. '간식' 카테고리 후기(1,842,271개)의 평점 비율

'간식' 카테고리는 후기가 가장 많은 카테고리로, 4, 5점 후기가 가장 많다.

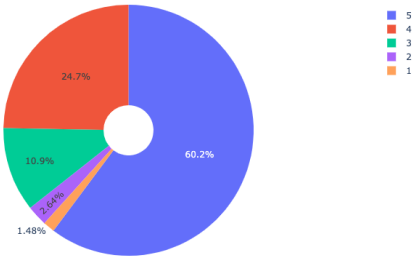

그림11. '건강관리' 카테고리 후기(289,902개)의 평점 비율

'건강관리' 카테고리는 3점 후기 비율이 10.9%로 높은 편이다.

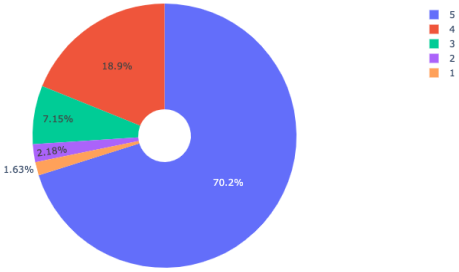

그림12. '급식기/급수기' 카테고리 후기(139831개)의 평점 비율

5점 후기는 980,097개로 70.2%를 차지하고, 4점 후기는 18.9%를 차지하므로 긍정적인 평점을 준 후기의 비율은 약 89.1%이다. 1점과 2점 후기는 각각 1.63%와 2.18%로 카테고리 내에서 두 평점이 차지하는 비율은 3.81%이다.

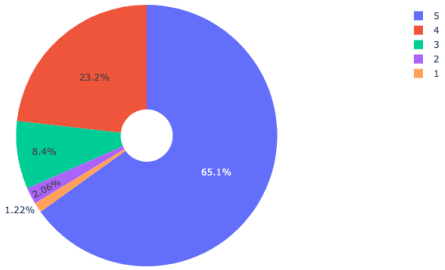

그림13. '미용/목욕' 카테고리 후기(271,309개)의 평점 비율

[그림5]와 [그림6]의 워드클라우드에 언급된 '향기'라는 단어가 '미용/목욕' 카테고리에서 파생되었을 것으로 추측했다. 그러나 4,5점 후기의 비율이 더 높음에서 불구하고 두 그림에서 나타나는 단어 '향기'의 크기에 비해 [그림8], [그림9]에서 등장하는 '냄새'의 크기가 더 크게 나타난다. 이는 '냄새'라는 단어가 다른 카테고리에서도 반복적으로 언급되는 것을 시사하는 것일 수도 있고, 단지 부정적인 후기가 적어 긍정적인 후기에서 보다 비슷한 빈도, 또는 더 적은 빈도로 언급되어도 단어의 크기가 더 크게 나타난 현상일 수 있다.

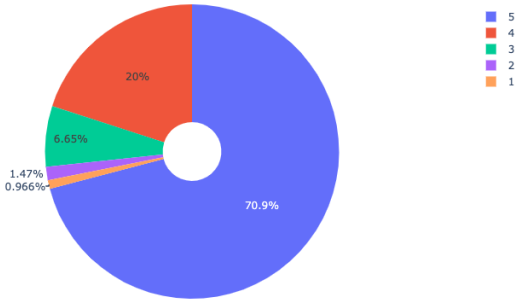

그림14. '사료' 카테고리 후기(402,884개)의 평점 비율

'사료' 카테고리는 '간식' 카테고리에 이어 후기 개수가 두 번째로 많은 카테고리이다. 3점 후기의 비율이 6.65%로 모든 카테고리를 통틀어 가장 낮은 비율을 보인다.

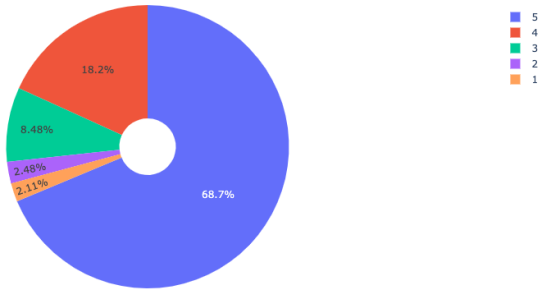

그림15. '사은품' 카테고리 후기(5,958개)의 평점 비율

'사은품' 카테고리는 상품을 분류하는 기준으로 적합하지 않으므로 고려하지 않는다.

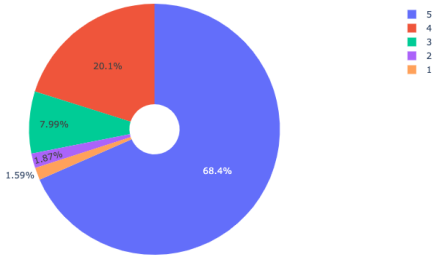

그림16. '생활/가전' 카테고리 후기(7,802개)의 평점 비율

[그림16]에서 확인할 수 있듯이 평점 5, 4, 3, 2, 1점의 비율은 각각 68.4%, 20.1%, 7.99%, 1.87%, 1..59%이다.

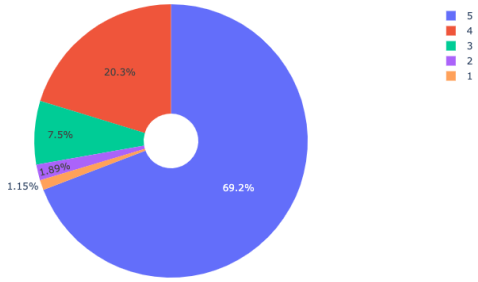

그림17. '위생/배변' 카테고리 후기(359,292개)의 평점 비율

'위생/배변' 카테고리의 긍정적 후기로 분류되는 4,5점 후기는 전체에서 89.5%를 차지하며 총 321,426개이다. 한편 3점은 7.5%의 비율을 차지하고 실제 후기 개수는 26,945개이다. 부정적인 후기로 분류되는 1,2점 후기는 약 3.04%이다.

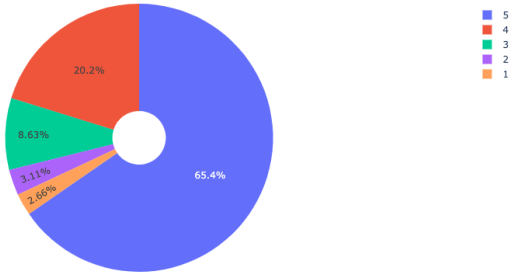

그림18. '의류/액세서리' 카테고리 후기(68,198개)의 평점 비율

'의류/액세서리' 카테고리의 5점 후기의 개수는 445,812개, 4점 후기의 개수는 13,795개, 3점 후기는 588개, 2점 후기는 2,120개, 1점 후기는 1,814개이다. 평점의 비율은 순서대로 65.4%, 20.2%, 8.63%, 3.11%, 2.66%이다.

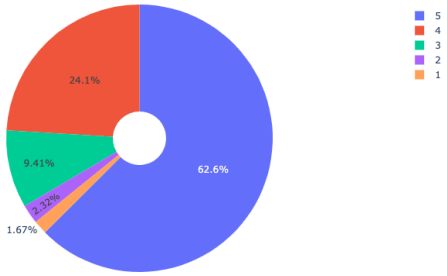

그림19. '이동장' 카테고리 후기(11,445개)의 평점 비율

'이동장' 카테고리의 후기 개수는 11,445개로, '사은품' 카테고리를 제외하고 후기 개수가 가장 적은 카테고리이다. 그러나 5점 후기 비율이 가장 낮은 '훈련' 카테고리(5점 후기가 4,740개)보다 5점 후기의 절대적인 양이 7,159개로 더 많다.

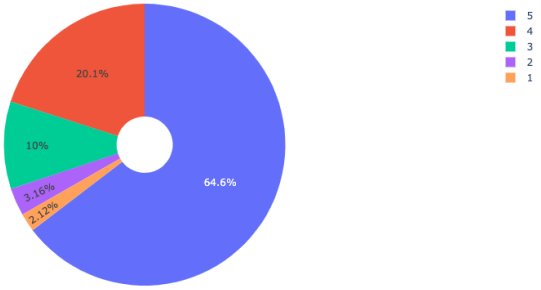

그림20. '장난감' 카테고리 후기(312,004개)의 평점 비율

5점 후기가 201,663개, 4점 후기가 62,564개, 3점 후기가 31,331개, 2점 후기가 9,870개, 그리고 1점 후기는 6,606개이다.

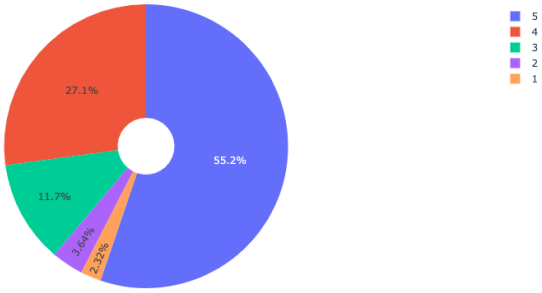

그림21. '치아' 카테고리 후기(11,459개)의 평점 비율

'치아' 카테고리는 '훈련' 카테고리 다음으로 1,2점 후기가 많은 카테고리이다. 1점 후기가 2.32%, 2점 후기가 3.64%로, 두 평점이 카테고리 내 전체 평점에서 5.96%을 차지한다. 또한 4점 후기 비율이 27%로 가장 높은 카테고리이기도 하다.

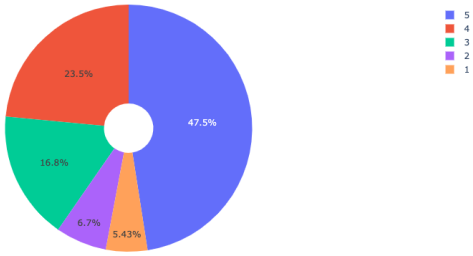

그림22. '훈련' 카테고리 후기(9,970개)의 평점 비율

[그림22]에서 확인할 수 있듯이 '훈련' 카테고리에 1,2점 후기 비율이 다른 카테고리의 후기들에 비해 많은 것을 확인할 수 있다. 이는 평점 5점 후기에서 등장하는 단어들의 빈도수를 시각화한 워드클라우드[그림5]에서 단어 '훈련'의 빈도수가 높은 것과는 대비되는 결과이다.

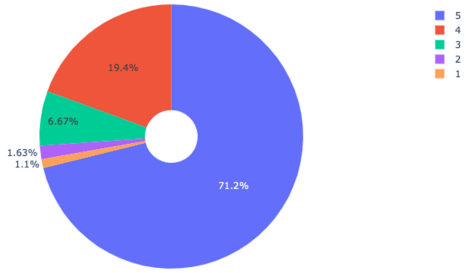

그림23. '하우스/울타리' 카테고리 후기(74,694개)의 평점 비율

가장 긍정 데이터 비율이 높은 카테고리이자 가장 부정 데이터 비율이 낮은 카테고리는 '사료'이다. [그림14]에서 확인할 수 있듯이 사료 카테고리의 후기 중 평점이 4, 5점인 후기는 전체 후기의 약 90.9%를 차지하고, 1, 2점인 후기는 약 2.436%를 차지한다. 한편 가장 부정 데이터 비율이 높고, 긍정 데이터 비율이 낮은 카테고리는 '훈련'이다. [그림22]에서 훈련 카테고리의 전체 후기 대비 4, 5점 후기 비율은 약 71%로, 사료 카테고리의 긍정 데이터 비율보다 약 20%p 낮다. 한편 부정 데이터는 약 12.13%로 10%p 가까이 차이가 난다. 또한 훈련 카테고리는 다른 카테고리에 비해 3점 후기의 비율이 눈에 띄게 높다. 평균적으로 3점 후기는 카테고리 내 전체 후기 중 10% 이하의 비율을 차지한다. 그러나 훈련 카

테고리의 경우 전체 후기 중 약 16.8%가 평점이 3점인 후기이다.

V. 결론 및 향후 과제

5.1 결론

반려동물 연관 시장은 빠르게 확장 중이고, 소비자를 확보하기 위해서는 상품 및 소비자에 대한 분석이 이루어져야한다. 본 논문에서는 반려동물 시장에서 가장 큰 비율을 차지하는 '사료' 카테고리의 전체 후기에서 빈도수 상위 20개 단어를 추출하여 소비자들이 사료를 구매할 때 어떤 부분에 주목하여 구매를 하는 지를 파악하였다. 그 결과 반려견의 사료에 대한 기호성 및 보관에 영향을 미치는 포장, 그리고 섭취 후 변의 냄새가 사료를 구매할 때 소비자가 고려하는 요소임을 알 수 있었다. 또한, 사료뿐만 아니라 각 카테고리에서 소비자가 신경쓰는 요소들을 파악하기 위해 전체 상품 및 후기에 대하여 카테고리 분류를 수행했다. 그러나 카테고리 코드가 유효하지 않은 상품들이 많아 전체 상품 중 38.9%가 카테고리가 분류되지 못했으며, 전체 후기 중 13.2%의 카테고리가 분류되지 않았다. 이후 전체 후기를 평점의 관점에서 분석하기위해 평점 별로 워드클라우드를 생성했는데, 5, 4점은 긍정적인 어휘가 대다수인데 반해, 1, 2점은 부정적인 어휘가 주를 이루었고, 3점은 긍정적인 어휘와 다소 부정적인 어휘가 함께 등장했다. 카테고리 별로 후기를 분류한 후 각 카테고리 마다 파이차트를 생성하여 평점 비율을 확인해 본 결과 '사료' 카테고리에 평점이 높은 후기 비율이 높은 것을 확인할 수 있었고, '훈련' 카테고리에 평점이 낮은 후기들의 비율이 높은 경향을 확인해 볼 수 있었다.

5.2 향후 연구 제시

견종 및 반려견의 특성에 따라 적합한 사료를 추천하기 위해서는 상품의 특징과 소비자의 반려견의 특징 및 소비자가 중요하게 여기는 요소 등을 추출하여야 한다. 과거의 카테고리 코드 또는 현재 사용하는 코드지만 사용한 파일에 저장되어 있지 않던 카테고리 코드에 대한 추가적인 데이터를 얻는다면 각 카테고리에 대한 더 많은 데이터를 확보할 수 있을 것이다. 더 많은 데이터를 사용하여 워드클라우드, 히트맵 등을 사용하여 각 카테고리에서 소비자들이 주요하게 생각하는 요소들을 추출할 수 있을 것이다. 또한 다른 평점들에 비해 소비자의 의견이 모호하게 드러나는 평점인 3점을 받은 상품과 후기를 분석하는 것도 소비자에게 맞춤형 상품을 추천하는 데 도움이 될 것이다.

REFERENCES

[1] H. Song, Y, Kim, "Nutrient Profiling-based Pet Food Recommendation Algorithm", *Journal of Information Technology Applications & Management*, Vol.25, No.4, pp. 145-156, 2018.

[2] S. Kim, "Implementation of a pet product recommendation system using big data", *Journal of the Korea Convergence Society*, Vol. 11. No. 11, pp. 19-24, 2020.

[3] S. Min, Y. Oh, "Implementation of Recipe Recommendation System Using Ingredients Combination Analysis based on Recipe Data", *Journal of Korea Multimedia Society*, Vol. 24, No. 8, pp. 1114-1121, 2021.

[4] Y. Jo, H. Lee, Y. Kim, "Implementation of a Classification System for Dog Behaviors using YOLI-based Object Detection and a Node.js Server", *The Journal of Korea Institute of Convergence Signal Processing*, Vol.21, no.1, pp. 29-37, 2020.

[5] S. Kim, "A Study of Similarity Measure Algorithms for Recomendation System about the PET Food", *Journal of the Korea Convergence Society*, Vol. 10. No. 11, pp. 159-164, 2019.

[6] H. Oh, J. Seo, H. Kim, H. Kim, S. Kim, H. Lee, "Pet Disease Q&A System Using Named Entity Recognition", *The Journal of Korean Institute of Information Technology*, Vol. 25, pp. 512-514, 2021.

황선우(Sun-Woo Hwang)

성균관대학교 인공지능융합전공, 소프트웨어학과 학사 재학 (2021~)
※관심분야 : 추천 시스템, 심리학

사진

오하영(Ha-Young Oh)

Sungkyunkwan University Professor (2019~)
Ajou University Professor (2016~2019)
Soongsil University Professor (2013~2016)
Ph.D. in computer engineering at Seoul National University (2013)