The Journal of The Institute of Internet, Broadcasting and Communication (IIBC)
Vol. 22, No. 6, pp.185-190, Dec. 31, 2022. pISSN 2289-0238, eISSN 2289-0246

https://doi.org/10.7236/JIIBC.2022.22.6.185
JIIBC 2022-6-27

오픈 데이터를 이용한 사용자 리뷰 분석 방법

Analysis Method of User Review using Open Data

최태호[*], 황만수[**], 김능회[***]

Taeho Choi[*], Mansoo Hwang[**], Neunghoe Kim[***]

요 약 오픈 데이터는 많은 경제적 가치를 지닌다. 우리나라 뿐만 아니라 여러 나라에서도 오픈 데이터를 확장하고 활용하기 위하여 갖은 정책과 노력을 기울이고 있다. 하지만 우리나라는 많은 데이터를 가지고 있음에도 불구하고 이를 잘 활용하지 못하여 효과를 보지 못하고 있어 여러 산업에서 다양한 시도가 필요하다. 그중에 패션 산업에서는 소비자의 단순 변심에 의해 교환, 환불 문제가 가장 많다. 이러한 문제를 해결하기 위해 서비스 제공자에게 더 나은 피드백이 필요하며 소비자의 요구사항이 반영된 사용자 리뷰와 함께 불만 사항이 개선된 이미지를 보여줌으로써 해결하고자 한다. 본 논문에서는 소비자의 요구사항을 파악하기 위해 온라인 쇼핑몰 사이트에서 사용자 리뷰 분석을 진행하고 K-fashion 데이터의 속성을 활용하여 제품에 대한 속성을 정의한다. 제품에 대한 사용자의 요구를 불만 속성으로 정의하고, 이에 해당하는 속성을 가진 라벨링 데이터를 검색하여 서비스 제공자에게 사용자의 요구사항을 텍스트 데이터나 속성 뿐만 아니라 이미지 또한 제공하여 제품 개선에 도움을 주고자 한다.

Abstract Open data has a lot of economic value. Not only Korea, but many other countries are doing their best to make various policies and efforts to expand and utilize open data. However, although Korea has a large amount of data, the data is not utilized effectively. Thus, attempts to utilize those data should be made in various industries. In particular, in the fashion industry, exchange and refund problems are the most common due to unpredictable consumers. Better feedback is necessary for service providers to solve this problem. We want to solve it by showing improved images of dissatisfactions along with user reviews including consumer needs. In this paper, user reviews are analyzed on online shopping mall websites to identify consumer needs, and product attributes are defined by utilizing the attributes of K-fashion data. The users' request is defined as a dissatisfaction attribute, and labeling data with the corresponding attribute is searched. The users' request is provided to the service provider in forms of text data or attributes, as well as an image to help improve the product.

Key Words : Open Data, Text Mining, User Review

[*]학생회원 군산대학교 소프트웨어학과
[**]정회원 신한대학교 IT융합공학부(교신저자)
[***]정회원 군산대학교 소프트웨어학과(교신저자)
접수일자 2022년 11월 10일, 수정완료 2022년 11월 30일
게재확정일자 2022년 12월 9일

Received: 10 November, 2022 / Revised: 30 November, 2022 / Accepted: 9 December, 2022
Corresponding Author: mshwang@shinhan.ac.kr, nunghoi@kunsan.ac.kr
School of IT Convergence Engineering, Department of Software Science & Engineering, Shinhan University, Korea, Department of Software Science & Engineering,Kunsan National University, Korea

Ⅰ. 서 론

오픈 데이터는 많은 경제적 가치를 가진다. 유럽 연합 내에서 오픈 데이터의 경제적 가치는 2020년까지 3,250억 유로의 가치와 매년 17억 유로의 정부예산을 절감시키고 25000개의 일자리 창출 등으로 추정하였다[1]. 이러한 가치를 지닌 오픈 데이터를 확장하기 위해 우리나라는 정부, 지자체 등에서 공공 데이터의 개방을 꾸준히 늘려가고 있다. 2022년 11월 기준 공공데이터 포털에 73,386건의 데이터가 등록되어 있다.

각국에서는 많은 경제적 가치를 지닌 오픈 데이터를 여러 방면에서 활용하고 있다. 스웨덴의 음악 스트리밍 서비스 업체 'Spotify'는 음반 데이터 사이트인 'MusicBrainz'의 음악에 관한 메타 데이터를 이용한다. 독일의 'Viomedo'는 임상 실험의 조건에 맞는 환자를 매칭하는 플랫폼으로 임상 시험 등록 장부 오픈 데이터를 활용하여 서비스로 승화시킨 사례이다. 이외에도 현재 위치 주변의 AED(Automated External Defibrillator)를 띄워서 맵을 열람할 수 있는 일본의 'Coaido 119', 지역의 범죄 발생률, 주택 가격, 학교의 질 등 다양한 데이터를 바탕으로 살기 좋은지 정보를 가시화해서 보여주는 서비스인 미국의 'NeighborhoodScout' 등이 있다[2].

우리나라는 여러 정책과 국가의 지원으로 많은 오픈 데이터의 개방이 이루어졌다. WWW 재단은 2013년 10월 전 세계 77개국 대상으로 오픈 데이터 이니셔티브 현황과 효과 분석을 목적으로 하는 오픈데이터 현황 보고서인 ODB(Open Data Barometer)를 발표하였다. ODB 보고서는 데이터 아젠다 별 분석, 국가 및 지역별 분석과 함께 각 국의 준비도, 실행력, 효과, 3개 항목에 대한 국가 별 순위를 제공한다. 한국은 ODB 평점 54.21로 12위에 위치하고 있으며 준비도(77.19)는 상대적으로 매우 높게 나타난 반면 효과(24.56)는 취약하였다[3]. 준비된 데이터에 비해 활용이 되고 있지 않거나 미흡하다는 것이다. 그만큼 다양한 분야의 활용에 대한 시도가 필요하다.

기존에는 사용자 리뷰를 분석하여 소비자의 제품에 대한 요구사항을 파악하기 위해 수많은 시도가 있었다. 소비자의 요구사항은 서비스 제공자에게 있어 제품에 대한 개선으로 이어지는 중요한 데이터이다. 서비스 제공자는 소비자의 요구사항에 대해 정확하게 파악하는 것이 중요하다. 본 논문에서는 사용자 리뷰를 분석해 나온 요구사항을 오픈 데이터의 라벨링 데이터 속성으로 변환시켜 오픈 데이터의 이미지와 함께 요구사항의 속성을 서비스 제공자에게 제공하는 방법을 제안하고자 한다.

본 논문의 구성은 2장에서 오픈 데이터와 사용자 리뷰에 대한 배경을 소개한다. 3장에서는 본 논문에서 제안하는 방법에 대한 프로세스와 시스템 구조도를 자세하게 설명한다. 4장에서는 이해를 돕기 위해 제안한 방법에 실제 데이터를 토대로 시뮬레이션을 보여준다. 마지막 5장은 결론과 향후 연구에 대해 서술하였다.

Ⅱ. 배 경

1. 오픈 데이터

오픈 데이터는 모든 사람이 공개적으로 액세스, 활용, 편집 및 공유할 수 있는 데이터이다. 데이터를 활용하는데 있어 저작권이나 규칙 등이 없는 자유로운 사용을 목적으로 하는 것이다. 다양한 문제를 해결하는데 오픈 데이터의 이용이 촉진되고, 다른 데이터와 조합한 새로운 가치 창출이 활성화되고, 공통적인 형식, 부가적 정보(메타 데이터)의 기술 형식 등이 설계되는 효과가 있다. 오픈 데이터는 민간 영역에서 제공하는 데이터와 공공 영역에서 제공하는 데이터가 있다[4-5].

2. 사용자 리뷰

사용자 리뷰는 온라인 쇼핑몰 사이트에서 이용자들의 견해가 담겨있는 데이터이다. 쇼핑몰 사이트에 따라 다양한 방법을 통해 제품에 대한 만족도, 평가, 불만사항 등을 수집한다. 사용자 리뷰는 빅 데이터 분석으로 전체적인 사용자의 견해를 데이터화 하여 표현할 수 있다. 대표적인 분석 방법 중 하나로 사용자 리뷰에서 제품의 속성들을 평가할 수 있는 오피니언 마이닝이 있으며 각 제품의 여러 속성에 대한 평가를 긍정 혹은 부정으로 구분하여 표현할 수 있다. 이러한 분석 기법으로 리뷰에 나타난 사용자의 긍정적인 또는 부정적인 감정과 제품에 대한 불만 사항 등을 추출하여 정리할 수 있다[6].

Ⅲ. 오픈 데이터를 이용한 사용자 리뷰 분석 방법

본 논문에서는 쇼핑몰 제품 페이지에서 사용자 리뷰를 분석하고 제품에 대해 소비자가 원하는 개선 사항을 반영하여 오픈 데이터에서 이미지와 함께 개선 사항을 출력하는 방법을 제안한다.

1. 프로세스 구성

가. 불만 사항 정의

불만 사항을 정의하기 위해서는 해당 제품에 대한 소비자의 불만 표현이 필요하다. 소비자의 불만 표현은 댓글에서 추출한다. 쇼핑몰 사이트에서 낮은 평점 위주의 제품 댓글을 크롤링하여 사용한다. 이렇게 추출한 데이터는 한국어 전처리기를 이용하여 단어 형태로 분리한다. 분리한 단어는 K-fashion 데이터의 속성과 비교하여 일치하는 단어 중 빈도수가 높은 단어를 제품의 불만 속성으로 정의한다. 정의된 불만 속성은 제품의 개선 사항에 필요한 데이터로 이용된다.

나. 제품 속성 정의

제품의 속성 정의 단계에서는 K-fashion 데이터에서 정의한 데이터 속성 분류 체계에 따라 제품에 해당하는 속성들을 추출하여 제품 속성으로 정의한다. 해당 쇼핑몰 홈페이지에서 제품의 정보를 제공받거나, 해당 제품에 대한 정보 페이지에서 제공되는 정보에 따라 속성에 해당하는 데이터들을 제품의 속성으로 K-fashion 데이터 가이드라인에 따라 라벨링을 진행한다. 정의된 속성은 불만 속성과 중복되는 부분을 교체하고 나머지 부분을 담당한다.

다. 데이터 검색 및 이미지 출력

불만 사항이 반영된 제품 속성 데이터를 만들기 위해서는 불만 속성과 제품 속성이 필요하다. 제품 속성에서 불만 속성과 중복되는 부분을 불만 속성에 해당하는 값으로 바꾸어준다. 불만 속성이 명확하게 정의되어 있지 않은 경우에는 해당 속성값 이외에 모든 값을 개선점으로 한다. 이러한 과정으로 나온 불만 사항이 반영된 제품 속성 데이터는 라벨링 데이터에서 검색을 위해 사용된다. 해당 속성을 가진 라벨링 데이터 검색 단계는 불만 사항이 반영된 제품 속성 데이터와 일치하는 이미지를 라벨링 데이터에서 검색하는 단계이다. K-fashion 데이터는 라벨링 데이터와 원천 데이터로 나뉘는데 라벨링 데이터는 원천 데이터인 이미지의 속성을 정의한 데이터를 말한다. 불만 사항이 반영된 제품 속성과 일치하는 라벨링 데이터를 찾으면 해당 라벨링 데이터에 맞는 이미지를 원천 데이터에서 검색한다. 원천 데이터는 라벨링 데이터에서 정의한 속성에 맞는 이미지 데이터이다.

2. 시스템 구조

가. 추출

추출 부분은 쇼핑몰 사이트를 크롤링하여 얻은 데이터로 제품과 제품에 대한 불만을 추출하여 K-fashion 데이터 라벨링 가이드라인에 따른 속성으로 정의하는 부분이다. 크롤러는 제품 페이지에서 제품 상세 데이터와 댓글 데이터들을 모은다. 제품의 댓글 데이터는 단어 형태로 분리되어 불만이 많은 단어들을 골라 불만 속성을 정의하기 위해 전처리기로 이동한다. 전처리기를 통해 분리된 데이터는 단어마다 언급된 횟수를 세어 가장 많이 언급이 된 단어를 불만 속성으로 정의하고 속성 저장소에 저장한다. 제품 페이지에서 추출된 제품 상세 데이터는 K-fashion 데이터 라벨링 가이드라인에 따라 속성이 정의된 후 속성 저장소에 저장된다.

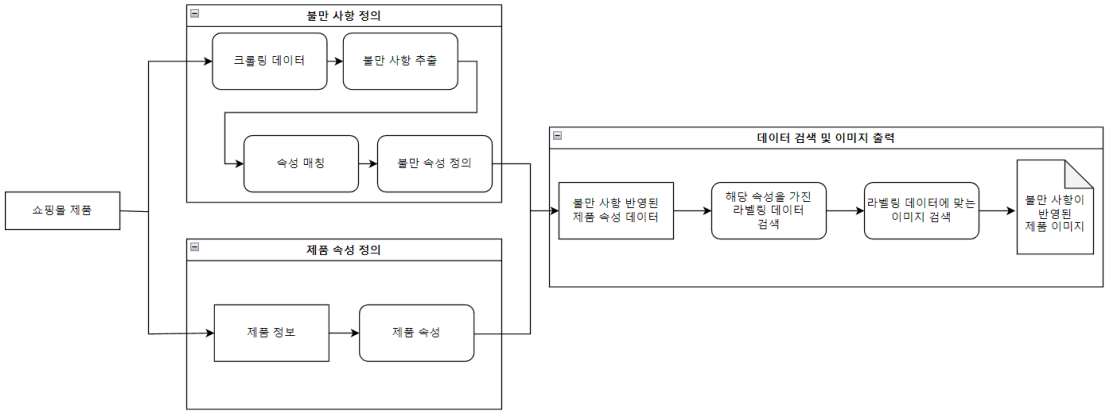

그림 1. 프로세스 구성도
Fig. 1. Process diagram

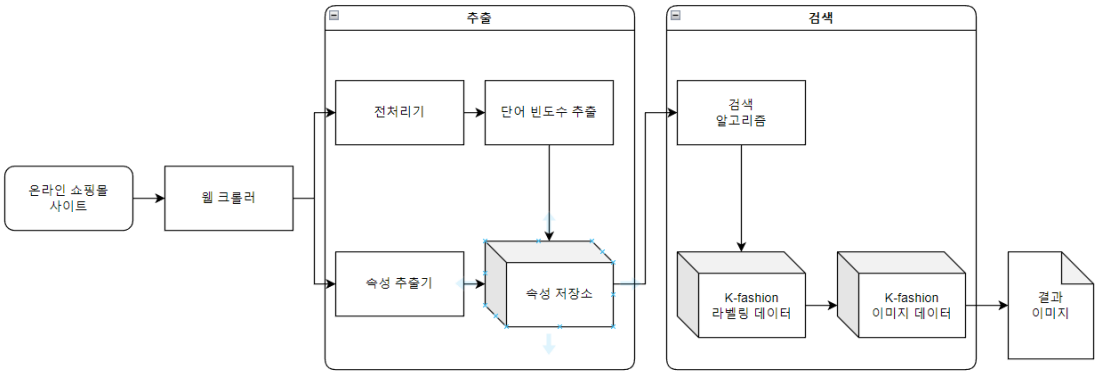

그림 2. 시스템 구조도
Fig. 2. System Architecture

나. 검색

검색 부분은 속성 저장소에서 추출된 속성 데이터와 K-fashion 라벨링 데이터를 비교한다. 속성 저장소에서 가져온 속성은 검색 알고리즘을 통해 K-fashion 라벨링 데이터에 가장 유사한 데이터를 검색한다. 검색 알고리즘이 해당 불만 사항이 반영된 속성을 가진 제품 데이터와 일치하는 데이터를 라벨링 데이터에서 찾는다. 검색된 라벨링 데이터는 해당하는 이미지를 검색 알고리즘이 찾아내어 출력한다. 위 방법을 거치면 사용자 리뷰를 분석하여 제시된 불만 속성과 함께 해당 불만 속성이 개선된 예시 이미지를 서비스 제공자에게 제공할 수 있다.

Ⅳ. 실 험

실제 쇼핑몰 사이트 댓글을 토대로 본 논문에서 제안하는 방법을 적용하였다. 실험 데이터로 글로벌 온라인 쇼핑몰에서 여성 하의 의류 중 임의의 제품을 선정하였다. 실험에 쓰일 데이터는 사이트의 제품 정보를 이용하였으며 K-fashion 데이터의 속성 가이드라인을 통해 해당 제품을 라벨링 하였다.

1. 추출

제품의 불만 속성을 정의하기 위해 해당 제품의 댓글을 낮은 평점순으로 정렬한 뒤 상위 50개 항목으로 실험을 진행하였다. 크롤링을 통해 얻은 50개 댓글 데이터에 전처리를 거쳐 나온 요구사항을 K-fashion 라벨링 데이터 가이드라인에 따른 속성에 맞추어 집계했다.

표 1. 불만 속성
Table 1. Dissatisfaction attributes

색상	28건
기장	13건
핏	17건

저장된 데이터는 한국어 전처리기를 통해 속성 형태로 분리되어 빈도수 단위로 저장한다. 낮은 평점순의 댓글은 해당 제품에 대한 불평을 주로 표현하기 때문에 주로 불만 속성에 대한 단어가 포함되어 있다. 표 1은 상위 3개의 K-fashion 데이터의 속성으로 집계된 데이터의 빈도수를 나타낸 표이다. 해당 제품에 대해 가장 불만이 많은 속성은 '색상' 속성으로 나타났다. '색상' 속성은 하위 속성 중 하나로 해당 제품의 '색상'의 속성인 '화이트' 에 대한 불만이 가장 많은 것을 확인할 수 있다.

제품에 대한 불만 속성을 추출하였으면 해당 제품에 대한 속성을 추출한다. 제품에 대한 정보는 제품의 상세 페이지를 통해 K-fashion 라벨링 데이터 가이드라인에 따라 제품 속성을 정의 할 수 있다.

표 2. 제품 속성
Table 2. Product attributes

분류	세부 속성
대분류	하의
기장	발목
색상	화이트
카테고리	팬츠
소재	우븐
프린트	무지
핏	노멀

정확한 불만 사항에 대해서는 소비자가 원하는 색상에 대한 데이터를 찾을 수 없기 때문에 제품 속성에 불만 속성과 겹치는 항목에 대해 해당 항목을 제외하여 제품 속성에 불만 속성을 적용시킨다.

표 3. 불만 사항이 적용된 제품 속성
Table 3. Product attributes subject to dissatisfactions

분류	세부 속성
대분류	하의
기장	발목
색상	'화이트'를 제외한 전부
카테고리	팬츠
소재	우븐
프린트	무지
핏	노멀

2. 검색

제품의 모든 속성에 대한 정의 절차 이후에는 K-fashion 라벨링 데이터에서 해당 제품과 유사한 속성의 라벨링 데이터를 검색한다. 가장 유사한 라벨링 데이터를 찾았다면 해당 라벨링 데이터가 가리키는 원본 데이터를 K-fashion 데이터에서 찾아내 표시한다.

표 4. 불만 사항이 적용된 제품 속성과 유사한 라벨링 데이터
Table 4. Labeling data similar to dissatisfaction product attributes

분류	세부 속성
대분류	하의
기장	발목
색상	그린
카테고리	팬츠
소재	우븐
프린트	무지
핏	노멀

가장 유사한 속성을 가진 라벨링 데이터는 21305번의 하의 카테고리이며 표 3에서 해당 속성을 확인할 수 있다. 해당 제품의 속성에서 '색상'만 그린이며 다른 속성은 모두 같다. 라벨링 데이터의 특정이 이루어졌다면 해당 라벨링 데이터가 가리키는 원본 데이터를 확인해서 불만 사항이 적용된 제품과 가장 유사한 이미지를 얻을 수 있다.

그림 3. 불만사항이 적용된 제품 이미지
Fig. 3. Product image with dissatisfactions applied

V. 결 론

본 논문에서는 오픈 데이터 중 하나인 K-fashion 데이터를 활용하여 사용자 리뷰를 분석하고 소비자가 제품에 대해 가지는 불만 사항을 속성으로 추출하여 해당 속성을 개선한 제품에 대한 이미지를 제공하는 시스템을 제안하였다.

서비스 제공자는 본 논문의 방법을 활용하여 제품에 대한 사용자의 요구사항을 속성과 이미지 정보로 같이 제공받기 때문에 제품에 대한 개선이 훨씬 수월해 질 것으로 기대된다.

향후 연구로는 본 논문의 방법을 실제 쇼핑몰 사이트에 적용하여 서비스 제공자와 함께 유용성을 검증하고 현업의 추가적인 요구사항을 도출하여 시스템을 보완해 나갈 계획이다.

References

[1] 2eConsulting, Seo Junhyeong, "Core Tasks for Spreading Open Data", March 2019, https://www.2e.co.kr/news/articleView.html?idxno=205783

[2] IRS Global, "Open Data Use Cases", September 2021, https://www.irsglobal.com/bbs/rwdboard/15175

[3] Cha, Young-Il, Noh, Keun-Bae, Kim, Jong-Bae, Han, Kyeong-Seok, "Trend of open data platform for public data opening", Journal of the Korea Institute of Information and Communication Engineering 15, 2 (2014) : 37-42.

[4] Wikipedia, https://en.wikipedia.org/wiki/Open_data

[5] Open Data Handbook, URL : https://opendatahandbook.org/guide/ko/what-is-open-data/

[6] Lim, Young-hee, Kim, Hong-bumm. "A study on the sentiment analysis using big data of hotels online review", Korean Journal of Hospitality & Tourism 28, 7 (2019) : 105-123.
DOI: https://doi.org/10.24992/KJHT.2019.10.28.07.105.

[7] Seokhyun Song, Samyoul Lee, Yeol Shin, Jaiyong Lee, "A Study on the Effectiveness of Korea's Open Government Data Policy: Ecosystem Perspective", Journal of Korean Associastion for Regional Information Society 20, 4 (2017) : 1-34.
DOI: https://doi.org/10.22896/karis.2017.20.4.001

[8] Hyunki Park, Jaekyoung Ahn, "Demand Forecasting for G2B E-commerce Using Public Data : A Case Study of Public Procurement Service", The Journal of Korean Institute of Information Technology 12, 10 (2014) : 113-121.
DOI: https://doi.org/10.14801/kitr.2014.12.10.113

[9] Yoon Eulyo, "A Study on the Big Data Utilization in Domestic Fashion Industry", Journal of Cultural Product & Design, 50 (2017) : 191-200.
DOI: https://doi.org/10.18555/kicpd.2017.50.17

[10] Keun-Hyung Kim, Sung-Ryoel Oh, "Methodology for Applying Text Mining Techniques to Analyzing Online Customer Reviews for Market Segmentation", Journal of the Korea Contents Association 9, 8 (2009) : 272-284.

[11] Shin, Eun-Ja, "An Analysis on Trends and Tasks of Open Data Policy in the Digital Era", Journal of the Korean Society for Information Management 32, 3 (2015) : 49-68.
DOI: https://doi.org/10.3743/KOSIM.2015.32.3.049

[12] An, Hyosun, Lee, Inseong, "An Investigation of a Sensibility Evaluation Method Using Big Data in the Field of Design -Focusing on Hanbok Related Design Factors, Sensibility Responses, and Evaluation Terms-", The Korean Society of Clothing and Textiles 40, 6 (2016) : 1034-1044.

저 자 소 개

최 태 호(학생회원)

- 2017년 3월 ~ 현재 : 군산대학교 소프트웨어학과 학사과정
- 주관심분야 : 데이터 마이닝, 데이터베이스

황 만 수(정회원)

- 1986년 8월 : 중앙대학교 전자계산학과 (이학석사)
- 2001년 8월 : 숭실대학교 컴퓨터학과 (공학박사)
- 1993년 3월 ~ 현재 : 신한대학교 IT융합공학부 정교수
- 주관심분야 : 요구공학 ,소프트웨어공학

김 능 회(정회원)

- 2008년 8월 : 고려대학교 컴퓨터학과 (이학석사)
- 2017년 8월 : 고려대학교 컴퓨터 ·전파통신공학과 (공학박사)
- 2021년 3월 ~ 현재 : 군산대학교 소프트웨어학과 조교수
- 주관심분야 : 요구공학, 소프트웨어공학

※ 이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. 2021R1H1A2095710).