K-POP 연관 해시태그 크롤링을 이용한 맛집 추천 시스템 개발

김화선[1], 이채연[1], 조서윤[1], 나정은[2]
[1]연세대학교 전기전자공학과
[2]연세대학교 학부대학
ajdandk12@yonsei.ac.kr, chaeyeon1013@yonsei.ac.kr, ynexito@yonsei.ac.kr,
jenah@yonsei.ac.kr

Development of Restaurant Recommendation System Using K-Pop Hashtag Crawling

Hwa-Seon Kim[1], Chae-Yeon Lee[1], Seo-Yun Cho[1],
Jeong-Eun Nah[2]
[1]Dept. of Electrical and Electronic Engineering, Yonsei University
[2]University College, Yonsei University

요 약

COVID-19 상황 속에서도 전 세계 Twitter K-POP 콘텐츠 관련 트윗 양은 78억 건 이상으로 매년 성장세를 보인다. Twitter 내 K-POP 팬들은 아티스트 관련 해시태그를 포함한 트윗을 작성하여 같은 팬덤끼리 실시간으로 정보를 전달하고 생산한다. 이러한 맛집 트윗들은 K-POP 팬들이 Twitter 내에서 신뢰도 있는 맛집 정보를 얻는 용도로 사용된다. 하지만 팬들이 정보를 얻기 위해서는 여러 맛집 해시태그로 검색하고 리트윗 수가 많은 트윗을 직접 찾아야 한다. 기존의 맛집 추천 시스템은 서비스 제공자 중심의 구조를 띤다. 서비스 제공자가 일방적으로 정보를 전달하거나, 사용자 리뷰 갱신 간격이 길다는 한계가 존재한다. 본 논문에서는 Twitter 내 K-POP 맛집 해시태그가 포함된 트윗을 Twitter API와 Tweepy를 사용하여 크롤링하였다. 수집한 데이터의 좋아요 수와 리트윗 수를 바탕으로 데이터 필터링을 진행하여 bot user와 광고 계정이 제외된 맛집 관련 트윗을 추출한다. 최종적으로는 추출한 트윗의 정보를 마커로 표시하여 웹 사이트를 제작하였다. K-POP 팬들은 맛집 해시태그를 검색하여 일일이 찾을 필요 없이 웹 사이트에 방문하여 맛집 위치를 확인할 수 있다. 웹 사이트 사용자의 위치가 지도상에 표시되어 가까운 맛집을 찾기도 편리하다. 본 논문에서는 맛집의 위치를 서대문구로 한정하여 진행했다.

1. 서론

Twitter Internal data에 따르면 최근 전 세계 Twitter 내 K 콘텐츠 관련 트윗 양은 K-POP 규모의 성장세를 따라가며 지난 10년간 546% 증가하였다[1]. 2021년 관련 트윗 수는 78억 건 이상 생성되었다[2]. 특히 COVID-19 이후, 4세대 K-POP 아티스트들은 이전 세대 아티스트들보다 2배 많이 트윗을 작성하며 온라인에서 팬들과 활발히 소통한다[1]. 국내에서는 COVID-19 방역 지침에 따라 K-POP 관련 오프라인 행사가 취소되거나 연기되며 Twitter를 이용한 온라인 소통은 팬과 아티스트 모두에게 중요해졌다.

Twitter에서 이용하는 공간인 트윗은 140자 단문 소통, 다자간 대화, 실시간성이라는 특징을 가진다. 사용자들은 작성된 트윗을 대상으로 리트윗(글 전달), 좋아요, 인용(댓글 작성) 기능을 이용하여 자신의 타임라인에 공유한다. 사용자가 짧은 글만 적으면 간편하게 자신의 의견을 전파하기 쉽기 때문에 Twitter는 인스타그램이나 페이스북 등의 다른 소셜 네트워크 서비스보다 간결하고 빠르다[3].

Twitter 내 K-POP 팬들은 실시간으로 콘텐츠와 의견을 공유한다. 팬들은 자신이 좋아하는 아티스트 관련 해시태그를 포함한 트윗을 작성하여 같은 팬덤끼리 맛집을 공유하기도 한다. 이러한 문화는 자신이 좋아하는 K-POP 아티스트에게 맛집을 추천하기 위한 의도로 시작되었으나, 현재는 팬들이 Twitter 내에서 신뢰도 있는 맛집 정보를 얻는 용도로 사용된다. 그 이유는 사용자가 트윗의 리트윗 수나 좋아요 수, 인용을 참고하여 음식점의 정보나 후기를 실시간으로 확인할 수 있기 때문이다.

Twitter 내의 트윗을 분석하여 다양한 분야에 활용하는 연구도 활발하다. 이종호는 Python을 이용한 트윗 크롤링 시스템을 제시하였다. Twitter에서 사

용자가 팔로우하는 대상의 트윗을 크롤링하여 수집하고, 수집된 트윗의 개수를 시간대에 따른 자료 개수의 빈도로 나타내었다[4].

김한나·정영섭은 특정 키워드를 중심으로 트윗과 리트윗 수를 수집하여 Twitter 사용자의 감성을 분석하고 사회적 이슈와 연결하였다. 데이터의 질적 분석과 양적 분석을 함께 수행하여 데이터의 신뢰성을 높이고 사용자의 의사 결정 과정을 확인하였다. 감성별 트윗 분류 결과, 긍정 트윗이 리트윗 수가 높고 부정 트윗은 좋아요 수가 높다는 결론을 얻었다[5].

본 연구는 위 기간 Twitter 내 K-POP 커뮤니티에서 사용되는 해시태그가 포함된 한국어 트윗을 크롤링하고 데이터 필터링 과정을 진행하여 Twitter 상의 인기 있는 맛집을 추천하는 시스템을 제안한다.

2. 기존 추천 시스템의 현황 및 한계

기존 맛집 어플리케이션 Foodinko은 카테고리 분류가 세밀하지 않아 사용자가 원하는 음식을 찾기 불편할 수 있다. 비공개 운영진이 맛집을 선정하는 기준이 불분명하여 데이터의 신뢰도가 낮다. 기존 맛집 웹 및 어플리케이션 Startip는 네이버 블로그와 후기를 확인한다. 후기가 갱신되는 간격이 길고, 후기를 다는 사용자들이 적다. 이러한 두 가지 특징은 정보의 다양성을 저하한다. 또한 음식 카테고리의 분류가 모호해 사용자가 원하는 음식을 파는 곳을 찾기 어렵다.

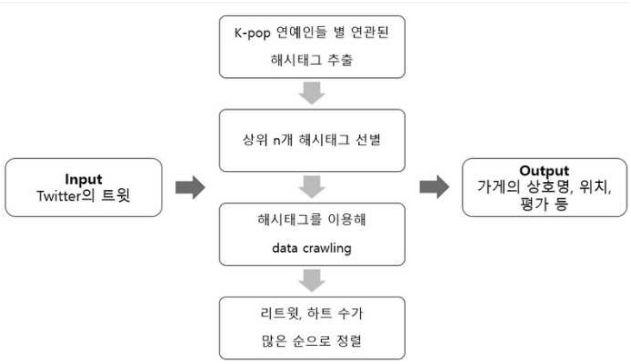
Foodinko와 Startrip 모두 정보 전달 구조가 운영진에서 사용자 방향으로 일방적이다. 이러한 구조에서 서비스 사용자는 제공되는 맛집 정보의 출처를 알기 어렵다.

<표 1> 기존 맛집 어플리케이션의 인터페이스

서비스 이름	Startrip	Foodinko
운영방식	- 비공개 운영진이 한국 연예인들이 방문한 맛집과 장소를 모아서 공유하는 방식 - Startrip 어플리케이션과 웹 사이트 운영	- 비공개 운영진이 인스타그램 이용자들의 추천 맛집을 모아서 공유하는 방식 - 인스타그램과 Foodinko 어플리케이션을 기반으로 페이스북 계정 또한 운영
장점	- 맛집 외에도 연예인이 방문한 장소 정보 제공 - K-POP 관련 이벤트 확인 편리	- 인스타그램 기반의 맛집 정보를 오랫동안 축적 - 광고 없는 서비스 제공
단점	- 리뷰 갱신 간격이 김 - 장소의 위치 정보를 텍스트로만 표기	- 맛집 정보의 출처 불분명 - 음식점의 카테고리 분류 모호

3. 트윗 크롤링을 이용한 맛집 추천 시스템 설계 및 구현

본 논문에서 제안하는 크롤링의 방식은 아래 그림 1과 같다. Twitter 내 K-POP 맛집 해시태그가 포함된 트윗을 Twitter API와 Tweepy를 사용하여 크롤링하였다. 이후 수집한 데이터의 좋아요 수와 리트윗 수를 바탕으로 데이터 필터링을 진행하여 bot user와 광고 계정을 제외한다. 최종적으로 얻은 트윗과 관련된 맛집의 정보를 저장한다.



(그림 1) 트윗 크롤링 및 필터링 시스템 구성

1) K-POP 연관 상위 해시태그 선별

Twitter 이용자들이 작성한 최근 100개의 게시글에서 특정 해시태그와 함께 쓰인 다른 해시태그를 추출하였다. 위 과정을 통해 추출하여 동시 언급된 순으로 정렬한 해시태그 중 상위 5개를 선별하여 사용자 맞춤 데이터의 수집을 준비한다.

```
[('몬베베가_몬베베에게_추천하는_맛집', 60),
 ('우석아_맛집으로_우따다따', 9),
 ('제노의_맛그당어', 2),
 ('원우야_여기_테이스티', 2),
 ('백현이를_위한_맛집투어', 2),
 ('채형원_너는세입도못먹겠지', 2),
 ('로하_이거먹으러_왔따나', 1),
 ('맛집two어', 1),
```

(그림 2) 추출된 연관 맛집 해시태그

2) 해시태그를 이용한 트윗 크롤링

이후, 특정 해시태그를 추적하여 그 해시태그가 포함된 트윗을 실시간으로 수집하기 위해 Python 내장 함수 GetStreamFilter()를 사용하였다. Tweepy의 Cursor 함수를 이용하여 일정 날짜 이후 작성된 트윗들의 작성자 이름, 팔로잉 및 팔로워 수, 작성 트윗의 전체 개수, 리트윗 수, 사용된 해시태그를 database 형태로 저장하였다. 완성된 database를 이용해 해시태그마다 csv 파일을 생성한 뒤 음식별 분류를 진행하였다.

(그림 5) Twitter 크롤링 결과 csv 파일

3) 크롤링 결과 리트윗 및 하트 순 필터링

크롤링 과정 이후 리트윗이 적고, 트윗 수나 팔로워가 0인 bot user이나 광고성 계정을 제거하였다. 저장된 csv 파일을 리트윗 및 하트 순으로 정렬하고 서대문구에 위치한 상위 30개의 맛집 정보를 추출하였다. 좋아요 수와 리트윗 수를 모두 수집하되, 정렬 과정에서 두 가지 지표의 합산 값이 같은 경우 리트윗 수를 우선하여 맛집 순위를 선정하였다.

4. 서대문구 맛집 추천 시스템 결과 활용

앞서 설계한 Twitter 크롤링 시스템을 이용하여 맛집 웹 사이트를 제작하였다. 필터링한 결과를 바탕으로 상위 30개 서대문구 맛집의 위치를 웹 사이트에 마커로 표시하였다. 웹 사이트 사용자의 위치는 지도에 표시되어 사용자의 주변 맛집 확인이 간편하다. (그림 6)은 앞서 설계한 Twitter 크롤링을 이용하여 얻은 맛집의 정보를 반영한 웹 사이트의 초기 화면에서 맛집 마커를 클릭한 화면이다.

(그림 6) 맛집 시스템 적용 웹 사이트 시작 화면

5. 결론

본 연구에서는 Tweepy와 Python을 이용하여 K-POP 맛집 관련 상위 5개 해시태그를 도출하였다. 이후 해시태그를 포함하는 트윗을 크롤링하고, 리트윗, 좋아요 수를 기준으로 인기도 순 필터링을 진행하여 얻은 상위 30개의 서대문구 맛집을 웹 사이트에 마커로 표시하였다.

본 연구의 시스템은 기존의 맛집 추천 시스템과 다르게 맛집 정보의 출처가 명확하다는 장점이 있다. 또한 맛집 리뷰 인용 트윗의 갱신 간격이 짧아 사용자 간 정보 공유가 활발하다. K-POP 팬들은 Twitter에 해시태그를 검색할 필요 없이 해당 시스템을 이용한 웹 사이트에 방문하여 주변 맛집의 위치를 확인할 수 있다. 기존의 방식대로 Twitter에 맛집 해시태그를 검색하면 뜨는 광고성 트윗과 계정이 필터링되었기 때문에 사용자는 더욱 간편하게 맛집 정보를 확인할 수 있다.

향후 연구에서는 맛집 웹 사이트에 출처 트윗의 url을 표시하여, 사용자가 맛집의 정보 공유를 간편하게 진행할 수 있도록 할 계획이다. 크롤링 기간을 연장하여 맛집 데이터의 정확도를 높이고, K-POP 팬덤 별 선호 맛집의 관계나 메뉴 선호도의 차이를 분석할 계획이다. 본 연구는 서대문구를 중심으로 진행되었지만 다른 지역에 대해서도 같은 알고리즘을 적용해 활용 범위를 확장할 수 있다.

본 논문은 과학기술정보통신부
정보통신창의인재양성사업의 지원을 통해 수행한
ICT멘토링 프로젝트 결과물입니다.

참고문헌

[1] Twitter Internal data, Global Tweets, July 2011 - June 2022 | Tweets including keywords related to Korean dramas, movies, and actors.

[2] Twitter internal data, Global Tweets, Jan 1 - Dec 31, 2021.

[3] S.-H. Hur, K.-S. Choi, "A Study on characteristics and types of tweet in twitter". Hanminjok Emunhak, vol. 61, pp. 455-494, 2012.

[4] J.-H. Lee, "Building an SNS Crawling System Using Python", Journal of the Korea Industrial Information Systems Research, vol. 23, no. 5, pp. 61-76, 2018.

[5] H. Kim and Y.-S. Jeong, "Social Issue Analysis Based on Sentiment of Twitter Users," Journal of Convergence for Information Technology, vol. 9, no. 11, pp. 81 - 91, 2019.