SNS 데이터 기반 지역 트렌드 분석

조재현*, 문남미*
*호서대학교 컴퓨터정보공학과
e-mail:jaehyeon99@naver.com

SNS data based regional trend analysis

Jae-hyeon Cho*, Nam-me Moon*
*Division of Computer and Information Engineering, Hoseo University

요 약

도시 속 상업 공간에서는 공간적 위치에 따른 지리적 이점이나 판매하는 상품뿐만 아니라, 해당 공간 속에서 소비자가 느낄 수 있는 문화와 감성이 소비자가 소비를 유하게 하는 중요한 요소가 되기도 한다. ICT 서비스 환경이 자리를 잡아 감과 동시에 제4차 산업 혁명이 도래하고 있는 현대 정보화 환경 속에서 소비자들은 자신의 심리나 감성, 정서에 들어맞는 공간에 방문하며 소비하고 SNS를 통해 공유한다. SNS는 Social Network Service의 줄임말로 너무나 일반적으로 우리 일상에 들어와 있는 개념이다. SNS의 시작은 마케팅의 한 분야로 시작된 것으로 판단된다. SNS를 이용한 홍보마케팅은 21세기에 접어들면서 고객들의 주관적인 개개인의 욕구 충족과 감성을 중시하게 됨으로써 예전보다 더 복합적이며 정교해졌다. 본 연구는 SNS 데이터를 블로그, 카페, 페이스북, 인스타그램에서 지역 명칭을 키워드로 1년간 콘텐츠를 크롤링하며, 형태소 분석기를 통해 학습할 수 있도록 데이터 전처리 작업을 한다. 마지막으로 딥러닝 알고리즘인 RNN 중 LSTM을 사용하여 감성 분석 학습 모델을 만들어서 지역별 콘텐츠의 주요분야, 긍/부정을 판별한다. 이렇게 분석한 데이터를 이용해 각 지역만의 특색과 인기 분야, 비인기 분야, 더 나아가 유망한 분야를 알아본다.

1. 서론

도시 속 상업 공간에서는 공간적 위치에 따른 지리적 이점이나 판매하는 상품뿐만 아니라, 해당 공간 속에서 소비자가 느낄 수 있는 문화와 감성이 소비자가 소비를 유하게 하는 중요한 요소가 되기도 한다. 특히 상업 공간이 보유한 문화와 감성이 방문자에 불과하던 개인을 소비자가 되도록 유도한다는 것은 구매의 전제조건이 상업 공간으로의 진입이기 때문이다. ICT 서비스 환경이 자리를 잡아 감과 동시에 제4차 산업 혁명이 도래하고 있는 현대 정보화 환경 속에서 소비자들은 더 방대한 전문 지식에 접근할 수 있고, 나아가 자신의 심리나 감성, 정서에 들어맞는 공간에 방문하며 소비하고 SNS에 공유한다[1].

SNS는 Social Network Service의 줄임말로 너무나 일반적으로 우리 일상에 들어와 있는 개념이다. 온라인과 같은 인터넷상에서 개개인의 인간관계를 강하하거나 새로운 인적 네트워크를 형성함으로써 폭넓은 네트워크를 형성할 수 있게 해주는 서비스로서 두루 이용되고 있으며, 정보화 시대에 모든 분야에서 다양하게 적용되고 있다 [2].

본 연구는 SNS 데이터를 블로그, 카페, 페이스북, 인스타그램에서 지역 명칭을 키워드로 1년간 콘텐츠를 크롤링하며, 형태소 분석기를 통해 학습할 수 있도록 데이터 전처리 작업을 한다. 마지막으로 딥러닝 알고리즘인 RNN 중 LSTM을 사용하여 지역별 콘텐츠의 주요분야, 긍/부정을 판별한다.

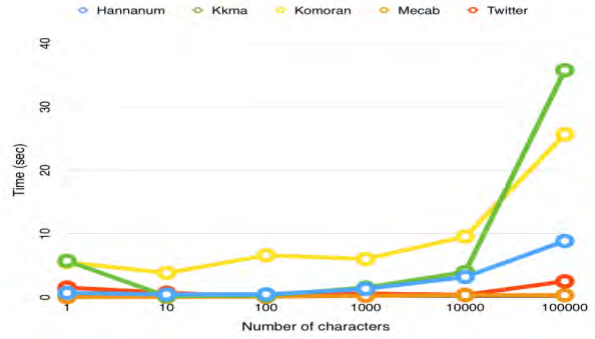
2. 주요기술

네이버 데이터 랩은 검색어로 알아보는 대한민국, 지역 통계, 데이터 융합분석, 공공데이터의 메뉴를 제공한다. 지역 통계는 네이버 검색데이터와 다른 기관/기업 데이터를 통해 만들어진 정보로 지역별, 업종별 추이를 확인할 수 있으며 각 변인의 상대적인 검색량을 100을 최고치로 설정하여 제공한다. 다시 말하면 입력한 단어의 추이를 하나로 합산하여 해당 변인이 네이버에서 얼마나 검색되는지 관련 데이터를 그래프로 제공하는 것이다[3].

Selenium은 웹앱을 테스트하는데 이용하는 프레임워크다. webdriver라는 API를 통해 운영체제에 설치된 Chrome 등의 브라우저를 직접 제어하게 된다. 직접 브라우저를 직접 작동시킨다는 것은 JavaScript를 이용해 비동기적으로 혹은 뒤늦게 나타나는 컨텐츠를 가져올 수 있다는 것이다. 또한, 요소에 쉽게 접근하기 위해서 웹 브라우저의 개발자 도구에서 각 요소의 xpath 경로를 지원해주

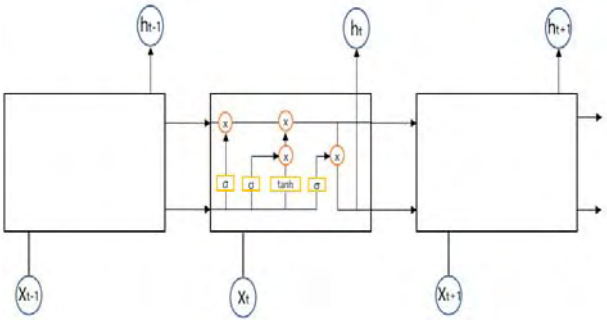
기 때문에 개발할 때 편리하다.

KoNLPy(Korean NLP in python)는 한국어 정보처리를 위한 파이썬 패키지이며 자연어처리를 갓 배우기 시작한 학생이나 자연어처리를 연구 목적으로 사용하려는 연구자에게 적합하다. KoNLPy 형태소 분석기로는 꼬꼬마(Kkma), 한나눔(Hannanum), Mecab, Okt, 트위터(Twitter) 등이 있으며 로딩시간, 실행시간, 띄어쓰기 알고리즘, 단어의 의미와 주변부 관계, 사전에 포함되지 않는 단어처리 등 사용자가 중점으로 두는 상황에 맞춰 사용하면 된다[4].



(그림 1) KoNLPy 형태소 분석기 속도 비교

본 연구에선 KoNLPy 형태소 분석기로 한나눔을 사용했다. (그림 1)과 같이 한나눔은 다량의 단어를 분석할 때 다른 분석기에 비교해 로딩시간이나 실행시간이 비교적 짧고, 사전에 포함되지 않는 단어처리를 세밀하게 해주는 강점이 있다. 한나눔보다 빠른 트위터는 현재 KoNLPy에서 업데이트가 멈췄고, Mecab의 경우에는 Window 운영체제를 지원하지 않는다.

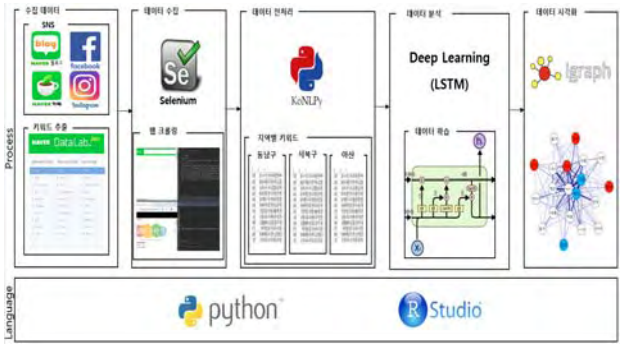


(그림 2) LSTM 구조

LSTM(Long Short Term Memory network)은 RNN의 한 종류로 관련 정보와 그 정보를 사용하는 지점의 거리가 멀어질 때 장기 의존성 문제가 발생하는 RNN의 문제를 해결하는 알고리즘이다. 전체 체인을 지나가는 셀 스테이트를 사용해 이전 학습 가중치를 큰 변동 없이 전달해 장기 의존성 문제를 해결한다. 또한, sigmoid 함수를 이용해 삭제할 정보를 결정한 뒤 또 다른 sigmoid 함수와 tanh 함수를 이용해 새로운 정보가 셀 스테이트에 저장될지 결정한다. LSTM 구조는 (그림 2)와 같다[5].

igraph는 효율성, 휴대성 및 사용의 용이성에 중점을 둔 네트워크 분석 도구 모음이며 무료 오픈소스이다. igraph는 네트워크 과학 및 관련 분야의 학술연구에 널리 이용된다.

3. 본론



(그림 3) SNS 데이터 분석 흐름도

본 연구는 지역 발전을 위한 SNS 데이터 분석을 위해 대한민국에서 많이 사용하는 SNS 4곳을 선정했고, 키워드는 천안과 아산의 동/면/읍으로 기간은 2017년 9월 1일부터 2018년 9일 1일까지 수집한다. 데이터 수집을 위한 웹 크롤러로 Selenium을 사용한다. 데이터 전처리는 한국어 형태소 분석기인 KoNLPy를 이용해서 특수문자와 URL을 제거하고 품사로 분류한다. 전처리된 데이터는 딥러닝 알고리즘 중에서 RNN의 장기의존성 문제를 해결한 LSTM을 사용한다. 수집한 데이터의 주요 단어와 컨텐츠의 긍/부정을 추출하고 주요 단어에 긍/부정에 따라서 점수를 부여해 감성을 판별한다. 이렇게 나온 주요 단어에 순위를 매겨서 지역마다 점수가 높은 단어와 낮은 단어로 그 지역의 특성을 알아낸다. SNS 데이터 분석을 위한 흐름도는 (그림 3)와 같다.

SNS 데이터는 사람들이 많이 이용하는 네이버 블로그, 네이버 카페, 페이스북, 인스타그램에서 앞서 선정한 키워드로 수집한다.

데이터 수집은 강력한 웹 크롤러인 Selenium을 사용한다. Chromedriver를 이용해서 Chrome을 제어한다. 네이버에서 블로그나 카페 리스트 검색은 세션에서 지역으로 keyword를 지역으로 하고, startDate는 '2017-09-01', endDate는 '2018-09-01'으로 옵션을 설정한다. 네이버 세션 페이지는 pageNo에 -1이나 마지막을 넘어가는 페이지를 넣어도 페이지가 끝나지 않기 때문에 마지막 페이지를 찾아야 한다. 이를 위해 검색된 총 컨텐츠 수를 수집하고 (총 컨텐츠 수) / (한 페이지에 보이는 컨텐츠 수)로 페이지 수를 구한다. 블로그의 한 페이지에 보이는 컨텐츠 수는 7개이며, 카페는 10개다. 네이버 블로그 크롤링의 특징은 switch_to_frame 함수를 이용하여 screenFrame과 mainFrame이라는 이름을 갖는 프레임을 고정해야 블로그 본문에 접근할 수 있다.

네이버 카페는 로그인 등 사용자 권한 관련한 예외처리에 중점을 두었다. 로그인이 필요한 경우는 크롤링을 시작할 때 네이버 로그인을 해야 하는데 CAPTCHA에서 막히게 된다. 그래서 크롤러를 디버그 모드로 실행시키고 로그인 화면에 멈춰서 수동으로 CAPTCHA를 실행해야 한다. 카페 가입이 필요한 컨텐츠는 생략한다.

페이스북과 인스타그램은 스크롤을 내려서 타임라인을 업데이트하는 특징이 있다. Selenium 기능 중에 자바스크립트를 실행하는 기능을 이용하여 스크롤을 제어할 수 있고, 스크롤 한 번에 나오는 타임라인은 9개다. 스크롤 횟수는 제한하지 않고 비교 사전을 만들어서 한번 스크롤하고 업데이트된 타임라인을 사전과 비교한다. 사전에 없으면 사전에 등록하고 스크롤을 반복하며, 사전에 있다면 크롤링을 멈추는 방법을 사용한다.

키워드

생활	음식	취미	뷰티	레저	관광
생활편의	주점	반려동물	뷰티	생활체육	숙박
휴대폰	유흥	강아지	헤어	생활체육	호텔
가전제품	Bar	고양이	피부	레저용품	모텔
인테리어	주점	동물병원	메이크업	골프용품	여관
생활용품	호프	동물용품	패션	레저공간	리조트
스튜디오	분식	취미	캐주얼웨어	볼링장	콘도
교육	분식	취미용품	유아옷	축구장	숙박
학교	중식	문화	남성패션	농구장	관광교통
학습장소	중식	영화관	패션잡화	당구장	렌터카
학원	서양식	전시	신발	축구장	그린카
유아교육	양식	예술쇼핑	여성패션		쏘카
보육	한식	뮤지컬			
직업학원	한식				
어학원	한정식				
건강	카페				
건강시설	카페				
병원	어류				
의료용품	회집				
의료시설	해물				
약국	아시아				
교통	아시아				
자당 구매					
주유소					
대중교통					
정비소					

(그림 4) 분류를 위한 키워드

분류 키워드 선정은 네이버 데이터 랩의 천안시 동남구, 천안시 서북구, 아산시 지역별 관심도와 카드사용통계를 이용한다. 키워드 분류는 (그림 4)처럼 크게 생활, 음식, 취미, 뷰티, 레저, 관광으로 나누고, 중분류, 소분류로 더 세분화하여 분류의 정확도를 높인다.

데이터 전처리는 수집된 데이터를 한국어 형태소 분석 라이브러리인 KoNLPy의 5가지 종류의 형태소 분석기 중 Mecab 형태소 분석기를 이용하여 분석한다. 이때 분석 옵션으로 정규화와 원형 찾기를 추가한다. 정규화란 “그래욬ㅋㅋㅋ”처럼 작성된 글을 “그래요”로 변환해주는 옵션이고, 원형 찾기 옵션은 “그래요”의 원형인 “그렇다”로 찾아주는 옵션이다. 전처리는 먼저 불필요한 특수문자와 URL을 제거하고 남은 데이터에서 핵심 단어는 분류 키워드를 통해 분류하고 감성을 나타내는 단어는 핵심 단어의 긍/부정을 판별한다[6].

데이터 학습은 LSTM을 이용하여 진행한다. 단어 벡터로 변환되는 Word Embedding 층으로 1000 노드, 중간층 128 노드로 하여 3층, 출력층 1 노드로 구성하고, 활성화 함수로 softmax를 사용하며 최적화 알고리즘으로 오차 감소 속도가 빠른 Adam Optimizer 알고리즘을 구성한다.

마지막으로 데이터의 수치화는 데이터의 가시성이 좋은 Rstudio를 사용하며, 여러 수치를 비교할 때 사용하는 igraph 라이브러리를 이용해서 수치화한다.

4. 결론

본 연구는 지역 발전에 도움을 주고자 SNS 분석 시스템을 제안한다. 10대부터 30대까지 연령층이 많이 사용하는 4곳의 SNS에서 해당 지역을 키워드로 한국어 형태소 분석기를 이용해 더 세부적으로 데이터를 처리하고, 딥러닝 알고리즘인 RNN 중 LSTM을 사용하여 감성 분석 학습 모델을 만들어서 지역 사용자들의 심리나 감성, 정서를 알아보고, 사용자들이 어떤 감정으로 글을 작성했는지까지 찾아낸다[7]. 더 나아가 SNS 데이터와 해당 지역의 공공데이터를 융합한다면 데이터 수집 측면에서 더 가치 있는 데이터를 얻을 수 있고, 해당 지역의 지역 행사나 학교축제 등 다양한 경우의 수를 머신러닝을 도입해서 보다 다양한 상황들에 맞는 분야를 제안할 수 있을 것이다.

본 연구는 문화체육관광부 및 한국콘텐츠진흥원의 2018년도 문화기술 연구개발 지원사업(R2018020083)으로 수행되었음.

참고문헌

[1] 유인진, 서봉군, 박도형. (2018). Smart Store in Smart City: 소비자 감성기반 상권분석 시스템 개발. 지능정보연구, 24(1), 25-52.

[2] 이상은, 이상미, 임상택. 2016. SNS 커뮤니케이션이 축제참가자 만족과 지역축제의 브랜드이미지에 미치는 영향-부산의 지역축제 참가자를 대상으로-. 한국관광학회 국제학술발표대회집, 80(0): 333-341

[3] 김종성. (2017). 스포츠관광 활성화와 빅데이터 활용에 관한 연구. 한국엔터테인먼트산업학회논문지, 11(3), 99-109.

[4] 박은정, 조성준. (2014). KoNLPy: 쉽고 간결한 한국어 정보처리 파이썬 패키지. 제26회 한글 및 한국어 정보처리 학술대회 논문집, 133-136.

[5] 손진광. (2017). RNN LSTM과 ACO를 이용한 감성분석을 통한 컨텐츠 추천 시스템에 관한 연구. 한국정보과학회 학술발표논문집, , 1033-1035.

[6] 이현아, 한경식. (2018). SNS와 온라인 커뮤니티 내의 스트레스 측정 모델 및 시각화 개발을 위한 연구. 한국통신학회 학술대회논문집, , 896-897.

[7] 김소희, 이병기, 조국애, 박동근, 신민아, 윤재영. (2018). SNS 유형과 게시물의 성격에 따른 ‘좋아요’ 의 표현방법이 사용자의 만족도에 미치는 영향. 한국HCI학회 학술대회, , 476-481.