한국지형공간정보학회
(Journal of the Korean Society for Geospatial Information Science)
Vol.23 No.2 June 2015 pp.89-96

ISSN: 1598-2955 (Print)
ISSN: 2287-6693(Online)
http://dx.doi.org/10.7319/kogsis.2015.23.2.089

연구논문

위치기반 소셜 미디어 데이터의 텍스트 마이닝 기반 공간적 클러스터링 분석 연구

Spatial Clustering Analysis based on Text Mining of Location-Based Social Media Data

박우진* · 유기윤**
Park, Woo Jin · Yu, Ki Yun

要 旨

위치기반 소셜 미디어 데이터는 빅데이터, 위치기반서비스 등 다양한 분야에서 활용가능성이 매우 큰 데이터이다. 본 연구에서는 위치기반 소셜 미디어 데이터의 텍스트 정보를 분석하여 주요한 키워드들이 공간적으로 어떻게 분포하고 있는지를 파악할 수 있는 일련의 분석방법론을 적용해보았다. 이를 위해, 위치태그를 지닌 트윗 데이터를 서울시 강남지역과 그 주변지역에 대하여 2013년 8월 한달 간 수집하였으며, 이 데이터를 대상으로 하여 텍스트 마이닝을 통해 주요 키워드들을 도출하였다. 이러한 키워드들 중 음식, 엔터테인먼트, 업무 및 공부의 세 카테고리에 해당하는 키워드들만 추출, 분류하였으며 각 카테고리에 해당하는 트윗 데이터들에 대해서 공간적 클러스터링을 실시하였다. 도출된 각 카테고리별 클러스터들을 실제 그 지역의 건물 또는 벤치마크 POI들과 비교한 결과, 음식 카테고리 클러스터는 대규모 상업지역들과 일치도가 높았고 엔터테인먼트 카테고리의 클러스터는 공연장, 극장, 잠실운동장 등과 일치하였다. 업무 및 공부 카테고리 클러스터들은 학원 밀집지역 및 사무용 빌딩 밀집지역과 높은 일치도를 나타내었다.

핵심용어 : 소셜 미디어, 위치태그, 텍스트 마이닝, 공간적 분포, 클러스터링 분석

Abstract

Location-based social media data have high potential to be used in various area such as big data, location based services and so on. In this study, we applied a series of analysis methodology to figure out how the important keywords in location-based social media are spatially distributed by analyzing text information. For this purpose, we collected tweet data with geo-tag in Gangnam district and its environs in Seoul for a month of August 2013. From this tweet data, principle keywords are extracted. Among these, keywords of three categories such as food, entertainment and work and study are selected and classified by category. The spatial clustering is conducted to the tweet data which contains keywords in each category. Clusters of each category are compared with buildings and benchmark POIs in the same position. As a result of comparison, clusters of food category showed high consistency with commercial areas of large scale. Clusters of entertainment category corresponded with theaters and sports complex. Clusters of work and study showed high consistency with areas where private institutes and office buildings are concentrated.

Keywords : Social Media, Geo-tag, Text Mining, Spatial Distribution, Clustering Analysis

1. 서 론

1.1 연구배경 및 목적

스마트폰, 태블릿과 같은 개인 모바일 장비의 광범위한 보급으로 인해 트위터, 페이스북, 인스타그램과 같은 마이크로 블로그 서비스의 인기가 높아지면서 엄청난 양의 소셜 미디어 데이터가 쏟아지고 있다. 이에 따라 소셜 미디어 데이터는 도시계획, 마케팅, 재난재해

Received: 2015.05.21, revised: 2015.06.09, accepted: 2015.06.11

* 서울대학교 환경정화기술 및 위해성평가 연구센터 연수연구원(Postdoctoral Researcher, Center of Environmental Remediation and Risk Assessment, Seoul National University, woojin1@snu.ac.kr)

** 교신저자 · 정회원 · 서울대학교 건설환경공학부 정교수(Corresponding author, Member, Professor, Department of Civil & Environmental Engineering, Seoul National University, kiyun@snu.ac.kr)

등 광범위한 분야에서 연구의 대상이 되고 있으며, 특히 최근에는 개인적 관심의 분석 및 추천, 감정 분석, 네트워크 분석, 이벤트 탐지 및 추적 등 다양한 분석 방법론과 응용분야가 개발되고 있다(Sakaki et al., 2010; Java et al., 2007; Qu and Liu, 2011; Chae et al., 2012; Kouloumpis et al., 2011). 이 중에서도 위치정보를 포함하는 소셜 미디어 데이터는 기존의 공간 데이터, 통계 데이터와 접목되어 공간분석 방법론, 텍스트 마이닝[1] 기법 등을 통해 다양한 분야에서 활용될 수 있다(Park et al., 2015; Kim and Park, 2014; Choi and Yom, 2014).

소셜 미디어 데이터의 공간적 분포 패턴에 대해 분석한 사례로 Mei et al.(2006)의 연구에서는 웹 블로그들을 분석하여 태풍이나 아이팟 나노 출시 등의 이벤트가 발생하였을 때의 사회적 트렌드와 어떤 연관성이 있는지에 대한 연구를 시공간적 패턴 분석 기법을 통해 실시한 바 있다. Sakaki et al.(2010)의 연구에서는 일본 지역 내에서의 지진의 발생과 트위터 데이터의 공간적 분포 패턴 간의 유사성 및 시간적 추이를 분석하여 지진이 발생할 위치를 예측하는 모델을 개발한 바 있다. 그러나 이러한 사례들은 단순한 키워드로 관련 데이터를 필터링하여 발생빈도를 시공간적으로 분석하는 데에만 그치고 있어 다각도의 텍스트 마이닝 기법을 적용하지는 않은 한계가 있다.

소셜 미디어 데이터에 텍스트 마이닝 기법을 적용한 연구사례로, Ghosh and Guha(2013)의 연구에서는 미국 내의 트윗 데이터로부터 비만과 관련된 키워드들을 탐색하고, 이들 키워드를 포함하는 트윗 데이터의 공간적 분포를 분석한 사례를 들 수 있다. 또한, Widener and Li(2014)의 연구에서는 트윗 데이터의 텍스트에서 건강식품과 불량식품에 대한 감정에 관련된 내용을 추출하여 이들 트윗 데이터의 공간적 분포와 인구학적 통계 데이터와의 관련성을 분석한 바 있다. Gerber(2014)의 연구에서는 트윗 데이터의 텍스트로부터 범죄 관련 내용을 추출, 분석하고 이들 데이터의 시공간적 패턴을 파악함으로써 범죄 발생 예측 모델을 개발한 바 있다.

LBSNS 관련 서비스 중 Trendsmap[2]은 미국 내에서 지역별로 이용자들 사이에서 많이 언급되고 있는 키워드를 보여주는 기능을 제공하고 있으나 카운티 단위의 넓은 지역에 대한 시각화를 제공하여 세부적인 지역에 대해서는 파악하기 어렵다는 한계점을 가지고 있다. 유사한 사례로 San Diego State University의 Center for Human Dynamics in the Mobile Age에서 개발한 Geoviewer[3] 서비스를 들 수 있는데 이 서비스는 실시간으로 위치태그를 가지고 있는 트윗 데이터를 공간적으로 지도화하고 핫스팟과 그에 해당하는 주요 키워드를 3개씩 뽑아서 시각화하고 있으나 불용어(stopwords) 제거를 하지 않았기 때문에 추출된 키워드들은 큰 의미를 담지 못하는 일반적인 단어(예를 들어, “we”, “I’m”, “can” 등)가 대부분이다. 이러한 사례들은 공통적으로 각 지역에서 주요한 키워드들을 추출하는 데에만 그치고 키워드들이 그 지역에서 가지는 지역적 의미에 대해서는 살펴보지 않은 한계점이 있다.

이에 본 연구에서는 위치기반 소셜미디어 데이터로부터 중요한 키워드들을 텍스트 마이닝 기법을 통해 추출하고 이를 카테고리화 함으로써 대상지역에 대한 키워드들이 가지는 지역적 의미를 살펴보고자 하였다. 또한, 이들 키워드들의 공간적 분포 패턴을 살펴보고 이를 통해 특정 지역에서 어떤 키워드들이 주로 소셜 미디어 데이터 상에 나타나는지를 파악하고 이러한 분석 패턴이 그 지역의 주요한 벤치마크와 어떠한 연관성이 있는지 비교할 수 있는 일련의 분석방법론을 적용해보고자 하였다.

1.2 연구의 범위 및 방법

본 연구에서는 위치태그를 포함하는 트위터 데이터를 대상으로 하여 서울시 강남지역(강남구, 서초구 등)에 대한 주요 키워드를 추출하고 키워드들을 카테고리화한 후, 각 카테고리에 해당하는 트윗 데이터의 공간적인 분포를 분석하고, 이러한 분포와 그 지역의 주요한 벤치마크 POI(Point of Interest)와 어떠한 연관이 있는지 살펴보고자 하였다.

이를 위하여 본 연구에서는 먼저, 위치태그를 포함한 트윗 데이터들에 대해 텍스트 마이닝을 통해 트윗 데이터에 전체적으로 포함되어 있는 주요 키워드들을 도출하고 이를 ‘음식’, ‘엔터테인먼트’, ‘업무 및 공부’의 세 개 카테고리로 분류하였으며, 각 카테고리의 키워드들을 포함하는 트윗 데이터들을 그룹화한 후, 각 카테고리별 트윗 데이터 소그룹의 공간적 분포 패턴에 대해 클러스터링을 실시하였다. 클러스터링된 지역을 지도상에 표시한 후, 기존의 건물 및 주요 POI들과 중첩하여 각 클러스터의 지역적 특성과 비교해 보았다. Fig. 1은 본 연구에서 적용한 상세 연구흐름도를 나타낸다.

1) 텍스트 마이닝(Text Mining): 텍스트 데이터에 대하여 자연어 처리, 문서 분석 기법 등을 통해 유용한 정보를 추출, 가공하는 기술 (Wikipedia, http://en.wikipedia.org/wiki/Text_mining, last visited: 2015/6/8).

2) http://trendsmap.com (last visited: 2015/5/18)

3) http://vision.sdsu.edu/hdma/geoviewer/ (last visited: 2015/5/18)

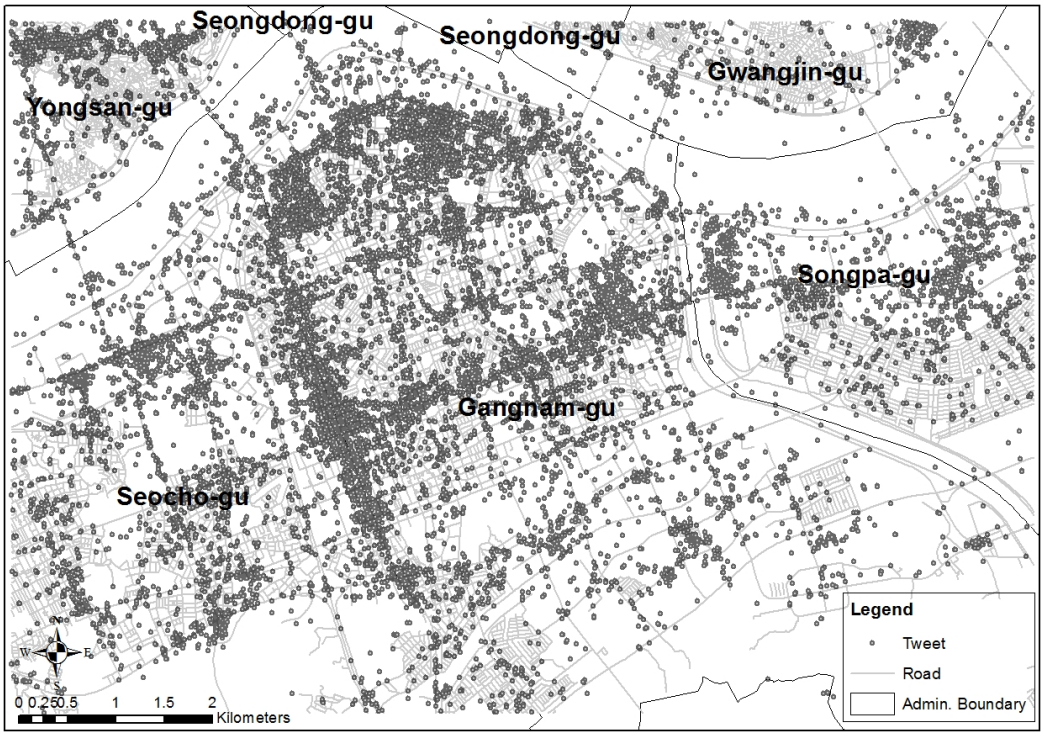

Figure 1. Tweet data in study area and administrative boundaries of district

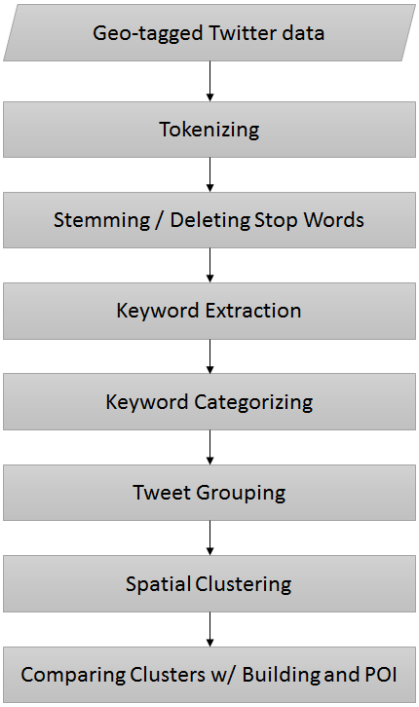

Figure 2. Workflow of this study

2. 대상데이터

본 연구에서는 위치기반 소셜미디어 데이터에 대해 경위도 속성정보를 지닌 트윗 데이터를 트위터 서비스에서 자체적으로 제공하는 인터페이스[4]를 통해 2013년 8월 한달 간 수집한 데이터[5]를 대상으로 하였다. 대상 지역의 공간적 범위는 좌상단 TM좌표와 우하단 TM 좌표를 각각 (198950.9, 448231.7), (209530.1, 441079.0)으로 하는 사각형 범위이다. 이 지역은 서울시 강남구를 중심으로 서초구, 송파구, 용산구, 성동구, 광진구 일부를 포함하는 지역이며, 이 지역은 주거지역, 상업지역, 업무지역 등이 모두 존재하는 지역이다. 대상 데이터인 트윗 데이터의 개수는 총 35,533개이며, 전체적인 분포는 Fig. 2와 같이 번화가 지역(강남역, 신

4) 트위터 공개 API: 소셜 네트워크 서비스 업체인 '트위터(Twitter)'에서 자사의 트윗 데이터의 활용을 지원하기 위해 개발자들에게 공개한 데이터 제공 인터페이스. 텍스트 메시지, 사용자, 위치정보 등의 정보를 수집할 수 있음. https://dev.twitter.com/overview/api (2014년 12월 28일 방문)

5) 트위터 API는 일주일 단위로 데이터를 수집할 수 있도록 제한이 되어 있기 때문에 본 연구에서는 한달 간의 데이터를 수집하기 위해서 매주 반복적으로 데이터를 수집하는 작업을 실시하였음.

사역, 압구정역, 코엑스, 이태원, 건대입구역, 신천 등)을 중심으로 몰려있는 현상을 발견할 수 있다.

3. 키워드 도출 및 카테고리화

본 절에서는 트윗 데이터의 텍스트 정보들에 대하여 텍스트 마이닝 기법을 적용함으로써 주요한 키워드를 도출하고 카테고리화하는 과정을 서술하였다. 본 연구에서 적용한 텍스트 마이닝 기법은 연구자가 직접 Java와 Matlab을 통해 구현한 프로그램 코드에 의해 실행되었다.

3.1 키워드 추출

본 연구에서 적용한 키워드 추출 프로세스는 다음과 같다. 먼저, 위치 태그된 트윗 데이터들에 대해 코멘트에 해당하는 텍스트 부분만을 모두 추출한 다음, 텍스트에 포함된 모든 단어들을 자연어 처리(Natural Language Processing) 기술을 통해 수집한다. 수집된 단어들에 대해 조사나 동사의 활용형 등을 제거하고 단어의 기본 원형으로 변환한다. 여기서 문장부호, 숫자, 대명사, 일상에서 흔히 사용되는 일반적인 단어들을 제거한다. 이러한 과정을 통해 추출된 단어들에 대해 전체 텍스트 내에서의 출현 빈도를 계산하였다. 아래의 Table 1은 단어들 중 가장 출현빈도가 높은 단어순으로 나열한 결과이다.

Table 1. Extracted terms and frequency in tweet text data

Term	Frequ-ency	Term	Frequ-ency
lunch	77	late	35
seoul	66	start	35
work	62	long	34
eat	58	tomorrow	34
sun	58	cafe	33
korea	57	car	32
hot	57	lot	31
star	56	music	31
park	52	mom	31
gangnam	48	sleep	30
meat	47	drink	30
flight	46	japan	30
coffee	43	cat	29
dinner	42	men	29
real	41	post	28
shop	38	arrive	27
cool	36	...	...

Table 2. Category and keywords used in this study

Category	Keywords
Food	lunch, coffee, dinner, cafe, drink, delicious, brunch, breakfast, beer, chichen, taste, kimchi, rice, lemon, choco, cheese, pasta, pizza, starbucks, dessert, butter, chef, bingsu, meat, soup, juice, restaurant, ramen
Entertainment	park, shop, music, watch, bulgeum, dance, studio, store, travel, concert, design, megabox, cgv, cinema, theater, sports, swim, game, hall, ball
Work and Study	work, study, job, business, office, school, alba, report, student, educate, book, university, teach

3.2 키워드 카테고리화

위의 3.1절에서 추출된 키워드들은 여러 가지의 카테고리로 분류할 수 있다. 가장 주요한 카테고리로는 '음식', '엔터테인먼트(유흥, 스포츠, 문화예술, 취미활동 등)', '업무 및 공부', '사람', '동물', '날씨 및 기후', '지명', '시간', '이동', '기타 일상행위' 등이 있었다. 이러한 카테고리들 중, '음식', '엔터테인먼트', '업무 및 공부'의 카테고리는 단어의 빈도수가 높은 동시에 지역적 특성을 가장 잘 표현할 수 있는 주요한 카테고리이며, 나머지 카테고리들은 지역적 의미와는 관련성이 높지 않은 일상적인 단어들로 판단된다. 따라서 본 연구에서는 '음식', '엔터테인먼트', '업무 및 공부' 이 세 가지 카테고리만을 분석의 대상으로 선택하였으며, 이들 카테고리에 해당하는 단어들을 재분류하였다. 이러한 카테고리의 분류와 주요 카테고리의 선정은 연구의 목적에 따라 연구자가 달리 적용할 수 있을 것으로 판단된다. Table 2는 본 연구에서 선택한 세 가지 카테고리와 각 카테고리에 포함된 단어들을 나열한 것이다.

4. 공간적 클러스터링 분석

본 절에서는 위에서 도출된 세 가지 카테고리 별 키워드들을 포함하는 트윗 데이터들을 재분류하고, 공간적 클러스터링 기법을 적용함으로써 각 카테고리 별 주요 클러스터를 탐지하는 과정을 서술하였다. 트윗 데이터 재분류 및 공간 클러스터링 기법의 구현 및 시각화는 ArcGIS SW (Ver. 10.3)에서 제공하는 분석기능들과 지도 시각화 기능들을 활용하였다.

4.1 트윗 데이터 카테고리별 재분류

본 절에서는 위의 3장에서 추출된 키워드들을 포함하는 트윗 데이터들을 탐색하여 카테고리별로 트윗 데이터를 재분류하였다. 분류 결과, '음식' 카테고리에 포함된 트윗 데이터는 총 3,781개로 전체 트윗 데이터의 약 10.6%를 차지하였으며, '엔터테인먼트' 카테고리에 포함된 트윗 데이터는 4,425개로 전체의 12.5% 정도를 차지하였다. 또한, '업무 및 공부' 카테고리에 포함된 트윗 데이터는 2,356개로, 전체의 6.6%를 차지하는 것으로 나타났다. Fig. 3, 4, 5는 각각 세 가지 카테고리별로 재분류된 트윗 데이터들을 나타낸다.

카테고리별로 재분류된 트윗 데이터들의 공간적 분포를 살펴보면 전체적으로 약간의 차이가 있음을 발견할 수 있다. '음식' 카테고리의 트윗 데이터들은 주로 이태원, 가로수길, 압구정 로데오거리, 강남역 지역에 집중적으로 분포하고 있으며, '엔터테인먼트' 카테고리의 트윗 데이터들은 주로 코엑스몰, 잠실 종합운동장, 예술의 전당 지역에 집중적으로 분포하고 있다. '업무 및 공부' 카테고리의 트윗 데이터들은 반포 지역에 약간 모여 있는 것 외에는 전체적으로 골고루 분포하고 있었다.

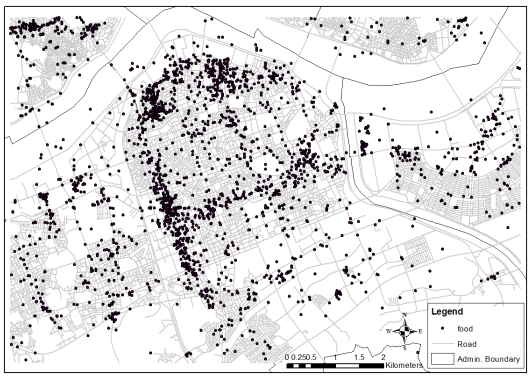

Figure 3. Tweet data classified as 'food' category

Figure 4. Tweet data classified as 'entertainment' category

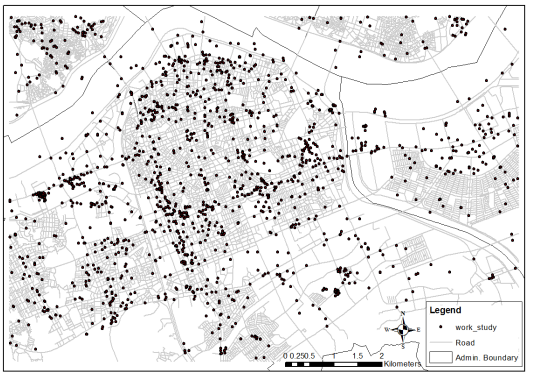

Figure 5. Tweet data classified as 'work and study' category

4.2 공간적 클러스터링

위의 절에서 재분류된 트윗 데이터들에 대하여 공간적 클러스터링을 실시하였다. 공간적 클러스터링 기법에는 계층적 클러스터링, 비계층적 클러스터링, 밀도기반의 클러스터링, 격자기반 클러스터링 등 다양한 기법들이 개발되어 왔다(Kang et al., 2004). 본 연구에서는 격자기반 클러스터링 기법을 적용하였는데 클러스터링 과정은 다음과 같다. 첫째, 대상지역을 격자로 나누고 각 격자 내에 포함되는 카테고리별 트윗 데이터의 개수를 구한다. 이때 격자의 크기는 대상지역의 범위(120.5㎢)와 트윗 데이터의 개수(35,533개)를 고려하여 50m로 결정하였다[6]. 둘째, 격자 내 트윗 데이터 개수에 대한 핫스팟 분석을 실시한다. 셋째, 핫스팟 분석결과로부터 신뢰수준 90% 이상의 핫스팟 지역으로 선정된 지역을 잘라낸다. 넷째, 잘라낸 지역에 대해 버퍼를 적용하여 핫스팟 지역에 대한 폴리곤 데이터를 형성한다. 이때 버퍼의 크기는 실험적으로 40m를 적용하였다. 이 40m 수치는 인접한 핫스팟들을 합쳐주는 동시에 두 격자 이상 떨어진 핫스팟은 분리시키는 수준으로 결정되었다. 다섯째, 일정면적 이하의 핫스팟 폴리곤은 제거하였다. 이때 핫스팟 폴리곤의 최소면적 기준은 35,000㎡을 적용하였는데 이 수치는 두개 이하의 격자 그룹으로 구성된 핫스팟을 제거할 수 있는 수준으로 결정되었다. 마지막으로, 격자형태의 핫스팟을 부드러운

6) 격자크기 $= \sqrt{\frac{(\text{대상지역 면적})}{(\text{점데이터 개수})}}$ (Yu, 1998)

곡선으로 변환하기 위해 선형 단순화 기법을 적용하였다. 이때 적용한 선형 단순화 기법은 곡선의 형태보존에 우수한 Bend Simplify 기법(Wang and Muller, 1998)이며 임계치는 전체적인 형상을 변형시키지 않는 범위 내에서 핫스팟 폴리곤을 곡선화 할 수 있도록 실험적으로 300m를 적용하였다.

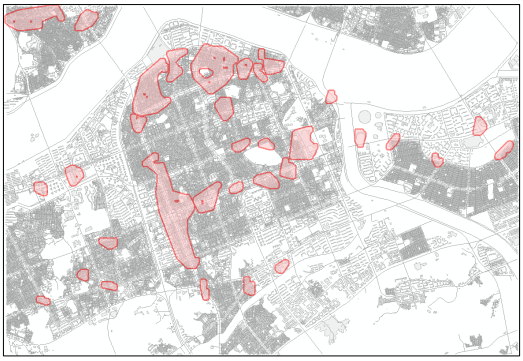

Figure 6. Clustering result of tweets in ‘food’ category

Figure 7. Clustering result of tweets in ‘entertainment’ category

Figure 8. Clustering result of tweets in ‘work and study’ category

Fig. 6, 7, 8은 각각 세 가지 카테고리에 대하여 클러스터링을 적용한 결과이다. 클러스터링의 위치를 나타내기 위해 건물 및 도로 중심선과 중첩해서 시각화하였다.

4.3 건물 및 벤치마크 POI와 클러스터링 결과 비교

위의 절에서 도출된 각 카테고리별 클러스터링 결과와 해당지역의 건물 및 주요 벤치마크 POI를 비교해본 결과는 아래와 같다.

첫째, ‘음식’ 카테고리에 해당하는 트윗 클러스터는 총 33개이며, 방배역 주변지역, 양재역 주변지역, 매봉역 4번 출구 지역, 도곡동 타워펠리스 인근, 교대역 주변지역, 신반포역 4번 출구 지역, 역삼역 주변, 선릉역, 선정릉, 포스코 사거리, 삼성동 코엑스몰을 연결하는 대규모 음식점 지역, 강남역, 교보타워 사거리, 우성아파트 사거리를 연결하는 대규모 상가밀집지역, 잠실운동장 신천역 지역, 잠실 롯데월드, 석촌호수 남쪽, 강남구청역 주변, 신사역과 가로수길, 압구정동, 청담동을 아우르는 대규모 상가지역, 이태원 지역 등의 상가지역과 일치하였다. 그러나 한남 오거리, 르네상스호텔 사거리, 서초3동 사거리, 잠실학원 사거리, 동부간선도로와 올림픽대로 교차지점 등 음식점이 많지 않은 지역과 일치하는 클러스터도 있었다.

둘째, ‘엔터테인먼트’ 카테고리에 해당하는 트윗 클러스터는 총 7개이며, 예술의 전당, 강남역 극장가, 코엑스몰, 잠실운동장과 일치하였으나 문화시설이 없는 신반포역 4번출구 지역과 일치하는 클러스터도 있었다.

셋째, ‘업무 및 공부’ 카테고리에 해당하는 트윗 클러스터는 총 29개로, 양재역 인근 지역, 교대역 주변 지역, 강남구청역 인근지역, 선정릉역 인근 지역, 테헤란로 인근의 고층빌딩 지역, 신사역과 가로수길 인근 지역, 청담동 패션거리, 잠실역 교차로 인근 지역과 같은 업무관련 건물 밀집지역들과 일치하였다. 또한, 방배동 학원거리, 신반포역 인근 학원지역, 예술의 전당 인근 지역, 고속터미널 내 대형서점, 개포동 경기여고 주변 지역, 압구정 현대아파트 주변지역, 잠실학원 사거리 주변지역 등과 같이 공부에 관련된 지역들과도 일치하였다. 특히 강남역과 교보타워 사거리를 아우르는 클러스터는 업무와 공부 관련 건물이 혼재하는 지역이라고 할 수 있다. 그러나 잠실운동장 주변지역과 같이 업무 또는 공부와 관련이 없는 지역과 일치한 클러스터도 있었다.

5. 결 론

본 연구에서는 위치기반 소셜 미디어 데이터로부터

텍스트 정보를 분석함으로써, 주요한 키워드를 도출하고, 이러한 키워드들을 담고 있는 메시지들이 공간적으로 어떻게 분포하고 있는지 살펴보고자 하였다. 이에 대한 결론은 다음과 같다.

1. '음식' 카테고리의 트윗 데이터들로부터 도출한 클러스터는 음식점 밀집지역을 포함하는 대규모 상가지역들과 일치하였다.
2. '엔터테인먼트' 카테고리의 클러스터들은 극장, 공연장, 경기장 등 주요한 문화 및 여가 관련 POI와 일치하였다.
3. '업무 및 공부' 카테고리의 클러스터들은 학원 밀집지역, 사무용 빌딩 밀집지역들과 대부분 일치하는 결과를 나타내었다. 그러나 몇몇의 클러스터들은 실세계의 건물 또는 POI와 일치하지 않는 경우도 있었다.

본 연구는 소셜 미디어 데이터에 대한 텍스트 마이닝과 공간분석 기법을 접목하고 이러한 분석결과와 실제 건물 및 POI들과의 관련성을 살펴보았다는 측면에서 의의가 있다고 할 수 있다. 이러한 노력들은 향후 소셜 미디어 데이터를 이용한 다양한 공간 빅데이터 분석 모델을 개발하는 데 활용될 수 있으며, 정책 의사결정 지원시스템 또는 위치기반 마케팅 분야에 있어서도 유용하게 활용될 수 있을 것으로 예상된다.

그러나 본 연구에서 텍스트의 대상 카테고리를 선정하고 카테고리 별 키워드를 분류하는 부분, 클러스터링 과정에서의 격자크기와 버퍼크기, 최소 클러스터 크기 기준 등은 본 연구의 대상 데이터와 연구의 목적에 맞게 연구자의 주관적인 판단이 개입되어 있다고 볼 수 있다. 따라서 이러한 과정을 보다 정량적, 합리적으로 접근하는 시도가 필요하며, 텍스트 마이닝 및 공간적 클러스터링 방법론 역시, 보다 고도화할 수 있는 추가 연구가 필요할 것으로 판단된다.

감사의 글

본 연구는 국토교통부 국토공간정보연구사업의 연구비지원(14CHUD-C061156-04)에 의해 수행되었습니다.

References

1. Chae, J., Thom, D., Bosch, H., Jang, Y., Maciejewski, R., Ebert, D. and Ertl, T., 2012, Spatiotemporal social media analytics for abnormal event detection using seasonal-trend decomposition, Proceedings of IEEE Conference on Visual Analytics Science and Technology, IEEE, pp. 143-152.
2. Choi, H. and Yom, J., 2014, Implementation of webGIS for integration of GIS spatial Analysis and social network analysis, Journal of the Korean Society of Surveying, Geodesy, Photogrammetry and Cartography, Vol. 32, No. 2, pp. 95-107.
3. Gerber, S., 2014, Predicting crime using Twitter and kernel density estimation, Decision Support Systems, Vol. 61, pp. 115-125.
4. Ghosh, D. and Guha, R., 2013, What are we 'tweeting' about obesity? Mapping tweets with topic modeling and Geographic Information System, Cartography and Geographic Information Science, Vol. 40, No. 2, pp. 90-102.
5. Java, A., Song, X., Finin, T. and Tseng, B., 2007, Why we Twitter: understanding microblogging usage and communities, Proceedings of WebKDD/ SNA-KDD 2007, ACM, pp. 56-65.
6. Kang, N., Kang, J and Yong, H., 2004, Performance comparison of clustering techniques for spatio-temporal data, Journal of Intelligence and Information Systems, Vol. 10, No. 2, pp. 15-37.
7. Kim, M. and Park, S., 2014, Construction and application of POI database with spatial relations using SNS, Journal of Korea Spatial Information Society, Vol. 22, No. 4, pp. 21-38.
8. Kouloumpis, E., Wilson, T. and Moore, J., 2011, Twitter sentiment analysis: The good the bad and the OMG! Proceedings of ICWSM 2011, AAAI, pp. 538-541.
9. Mardia, K. and Kent, J., 1979, Multivariate Analysis, Academic Press.
10. Mei, Q., Liu, C., Su, H. and Zhai, C., 2006, A probabilistic approach to spatiotemporal theme pattern mining on weblogs, Proceedings of the 15th international conference on World Wide Web, ACM, pp. 533-542.
11. Park, W., Eo, S. and Yu, K., 2015, Analyzing spatial correlation between location-based social media data and real estates price index through rasterization, Journal of the Korean Society for Geo-Spatial Information System, Vol. 23, No. 1, pp. 23-29.
12. Qu, Z. and Liu, Y., 2011, Interactive group suggesting for Twitter, Proceedings of HLT 2011, ACL, pp. 519-523.
13. Sakaki, T., Okazaki, M. and Matsuo, Y., 2010,

Earthquake shakes Twitter users: real-time event detection by social sensors, Proceedings of the 19th International Conference on World Wide Web, ACM.
14. San Diego State University, Center for Human Dynamics in the Mobile Age, 2015, GeoViewer, http://vision.sdsu.edu/hdma/geoviewer
15. Shin, J., 2004, Research on areal interpolation methods and error measurement techniques for reorganizing incompatible regional data units, Journal of the Korean Association of Regional Geographers, Vol. 10, No. 2, pp. 389-406.
16. Trendsmap solutions, 2009, Trendsmap, http://trendsmap.com
17. Wang, Z. and Muller, J., 1998, Line generalization based on analysis of shape characteristics, Cartography and Geographic Information Systems, Vol. 25, No. 1, pp. 3-15.
18. Widener, J. and Li, W., 2014, Using geolocated Twitter data to monitor the prevalence of healthy and unhealthy food references across the US, Applied Geography, Vol. 54, pp. 189-197.
19. Yu, K., 1998, Generalization of point feature in digital map through point pattern analysis, Journal of GIS Association of Korea, Vol. 6, No. 1, pp. 11-23.