한국지형공간정보학회지
(Journal of the Korean Society for Geospatial Information System)
Vol.22 No.4 December 2014 pp.21-30
ISSN: 1598-2955 (Print)
ISSN: 2287-6693(Online)
http://dx.doi.org/10.7319/kogsis.2014.22.4.021

연구논문

공간통계분석기법을 이용한 소셜 네트워크 유력지역 탐색기법 연구

A Study on Detection Methodology for Influential Areas in Social Network using Spatial Statistical Analysis Methods

이영민* · 박우진** · 유기윤***
Lee, Young Min · Park, Woo Jin · Yu, Ki Yun

要 旨

최근 소셜 미디어의 활성화로 인해 소셜 네트워크상에서 수많은 자발적 추종자들을 확보한 새로운 형태의 유력자가 대두되고 있다. 소셜 네트워크상에서의 유력자를 탐색하는 연구들이 진행되어 왔고, 관련 서비스가 제공 중에 있으나 이들은 유력자 규명에 있어 위치기반 소셜 네트워크 서비스(LBSNS)가 가지고 있는 위치 정보에 대한 반영이 부족하다는 한계점을 가지고 있었다. 이에 본 연구에서는 공간통계분석기법을 이용하여 LBSNS 데이터를 대상으로 다양한 사회문화적 이슈에 대한 발언에 영향력을 가지는 유력자를 공간적으로 탐색하고, 이를 활용하는 방안을 제시하고자 하였다. 이를 위해 트위터의 지오태깅된 메시지를 분석 데이터로 사용하였으며, 서울시를 공간적 범위로 하여 한 달 동안 총 168,040건의 메시지를 수집하였다. 또한 '정치', '경제', 'IT'를 연구 대상 범주로 설정하고, 데이터 수집 기간 동안 이슈가 되었던 키워드들을 주어진 범주별로 분류하였다. 이를 바탕으로 키워드에 대한 유력자를 파악하기 위한 노출도를 도출하고, 이에 대해 서울시의 행정동을 기준으로 공간결합연산을 실시함으로써 각 키워드에 대한 행정동별 노출도를 산출하였다. 그리고 행정동별로 산출된 노출도의 공간적 의존성을 고려하여 유력지수를 도출하였으며, 키워드별로 상위의 유력지수를 보이는 지역을 유력지역으로 추출하여 이들의 공간적인 분포 특성과 키워드들 간의 공간적 상관성을 분석하였다. 실험 결과, 동일 범주 내에서 키워드 간의 공간적 상관계수는 0.3 이상으로 높은 상관성을 보였으며, 정치범주와 경제범주의 키워드 간 상관계수 역시 평균 0.3으로 비교적 높은 상관성을 보인 반면, 정치범주와 IT범주, 경제범주와 IT범주 키워드 간의 상관계수는 각각 0.18, 0.15로 낮은 상관성을 보였다. 본 연구는 유력자에 대한 연구를 공간 정보의 관점에서 구체화시켰다는 점에서 의의를 가지며, 향후에 gCRM(geographic Customer Relationship Management) 등의 분야에 유용하게 활용될 수 있을 것이다.

핵심용어 : 유력지역, 유력자, 노출도, 유력지수, 공간통계분석기법, LBSNS

Abstract

Lately, new influentials have secured a large number of volunteers on social networks due to vitalization of various social media. There has been considerable research on these influential people in social networks but the research has limitations on location information of Location Based Social Network Service(LBSNS). Therefore, the purpose of this study is to propose a spatial detection methodology and application plan for influentials who make comments about diverse social and cultural issues in LBSNS using spatial statistical analysis methods. Twitter was used to collect analysis object data and 168,040 Twitter messages were collected in Seoul over a month-long period. In addition, 'politics,' 'economy,' and 'IT' were set as categories and hot issue keywords as given categories. Therefore, it was possible to come up with an exposure index for searching influentials in respect to hot issue keywords, and exposure index by administrative units of Seoul was calculated through a spatial joint operation. Moreover, an influential index that considers the spatial dependence of the exposure index was drawn to extract information on the influential areas at the top 5% of the influential index and analyze the spatial distribution characteristics and spatial correlation. The experimental results demonstrated that spatial correlation coefficient was

Received: 2014.08.08, accepted: 2014.10.07

* 정회원 · 서울대학교 대학원 건설환경공학부 박사과정(Member, Department of Civil & Environmental Engineering, Seoul National University, daldanka@snu.ac.kr)
** 서울대학교 환경정화기술 및 위해성평가 연구센터 연수연구원(Center of Environmental Remediation and Risk Assessment, Seoul National University, woojin1@snu.ac.kr)
*** 교신저자 · 정회원 · 서울대학교 건설환경공학부 정교수(Corresponding author, Member, Department of Civil & Environmental Engineering, Seoul National University, kiyun@snu.ac.kr)

relatively high at more than 0.3 in same categories, and correlation coefficient between politics category and economy category was also more than 0.3. On the other hand, correlation coefficient between politics category and IT category was very low at 0.18, and between economy category and IT category was also very weak at 0.15. This study has a significance for materialization of influentials from spatial information perspective, and can be usefully utilized in the field of gCRM in the future.

Keywords : Influential Areas, Influentials, Exposure Index, Influential Index, Spatial Statistical Analysis Methods, LBSNS

1. 서 론

1.1 연구배경 및 목적

최근 각종 스마트 기기의 대중화와 함께 소셜 미디어가 활성화되면서 소셜 네트워크를 바탕으로 개인들이 영향력을 행사할 수 있는 공간이 더욱 다양해지고 있다. 또한 다양한 소셜 미디어를 통해 즉각적이고 직접적인 교류가 가능해지면서 기존의 전통적 대중매체를 통해서는 평판과 명성을 획득하기 어려웠던 개인들도 온라인 공간에서 수많은 자발적 지지자들을 확보함으로써 새로운 형태의 유력자(influential)[1]를 대두시키고 있다(Lee et al., 2010). 이를 증명하는 사례로, 최근 페이스북과 같은 소셜 미디어가 구글, 야후, MSN과 같은 검색 엔진보다 더 많은 트래픽을 발생시키는 현상이 나타났는데(Korea Internet and Security Agency, 2010), 이는 인터넷 이용자들이 검색 엔진에서 정보를 얻는 것보다 친구나 지인들의 추천이나 평판을 통해 신뢰성 있는 정보를 획득하려는 경향을 보여준다고 할 수 있다.

한편 GPS를 내장한 모바일 기기의 사용이 보편화되면서 다양한 소셜 미디어 중에서도 특히 LBSNS (Location Based Social Network Service, 이하 LBSNS)에 대한 이용이 증가하고 있으며, 기존의 SNS (Social Network Service, 이하 SNS)도 위치 정보를 포함하는 서비스를 추가하여 제공함으로써 LBSNS로 저변을 확장하고 있는 추세이다. 이렇게 기존 SNS가 확장된 형태의 LBSNS는 지오태깅(geotagging)[2] 기능을 활용함으로써 사용자가 작성하는 개인의 일상, 사회적 이슈, 그리고 사건사고와 같은 내용과 함께 자신의 위치 정보를 선택적으로 포함시킬 수 있다.

그동안 소셜 네트워크상에서 유력자를 탐색하는 연구(Java et al., 2007; Weng et al., 2010; Park et al., 2010; Cha et al., 2010; Yoo and Kim, 2013)가 진행되어 왔으나 이러한 연구들은 유력자 규명에 있어서 LBSNS가 가지고 있는 위치 정보에 대한 반영이 부족하다는 한계점을 보였다.

LBSNS 관련 서비스 중 Trendsmap[3]은 구글맵과 연동하여 지역별로 이용자들 사이에서 많이 언급되고 있는 키워드를 보여주는 기능을 제공하며, 이를 통해 어느 지역에 어떤 키워드와 사용자가 이슈가 되고 있는지를 파악할 수 있다. 그러나 Trendsmap 서비스는 소축척 레벨에서 넓은 지역에 대해서는 파악할 수 있지만 대축척 레벨에서 세부적인 지역에 대해서는 파악하기 어렵다는 단점을 가지고 있다. 또한 지도상에서 키워드에 대한 검색은 가능하지만 이에 대한 결과로서 지도를 기반으로 한 키워드 시각화 화면만을 제공할 뿐 정량적이고 수치적인 분석 기능을 제공하지는 못한다는 한계가 있었다.

따라서 본 연구는 위에서 지적한 연구 및 서비스가 가지고 있는 한계점을 극복하기 위한 방안으로, 공간통계분석기법을 이용하여 LBSNS 데이터를 대상으로 정치, 경제, IT 분야에서 이슈가 되고 있는 사안에 대한 발언에 영향력을 가지는 유력자를 공간적으로 탐색, 활용하는 방안을 제시하는 것을 목적으로 한다.

1.2 연구 범위 및 방법

본 연구의 대상이 되는 공간적 범위는 서울특별시 전역으로, 분석 단위는 서울시의 423개 행정동이며, 시간적 범위는 2013년 8월 5일부터 9월 5일까지이다. 현재 상용되고 있는 다양한 LBSNS 중 트위터(Twitter)[4]를 분석 대상 데이터로 사용하였으며, 해당 기간 동안 총 168,040건의 지오태깅된 트위터 메시지(tweet, 트윗[5])를 수집하였다. 또한 기존의 뉴스 분류 체계를 차용하여 '정치', '경제', 'IT'를 연구 대상 범주로 설정하고, 데이터 수집 기간 동안 사회적으로 이슈가 되었던 키워드들을 각 범주별로 분류하였다.

1) 유력자에 대해 다양한 정의가 존재하지만 본 논문에서는 소셜 네트워크상에서 메시지의 생성 및 확산 등에 영향력을 행사함으로써 사람들의 인지, 태도, 행동 등을 변화시키는 힘을 가지는 개인으로 정의하였음.

2) 신조어로, 사진이나 동영상 등 디지털 매체 내에 최신 위치 정보를 삽입시키는 것을 말함.

3) http://trendsmap.com/

4) https://twitter.com/

5) 트위터상에서 사용자가 게시한 메시지를 일컬으며, 140자의 글자 수 제한이 있음.

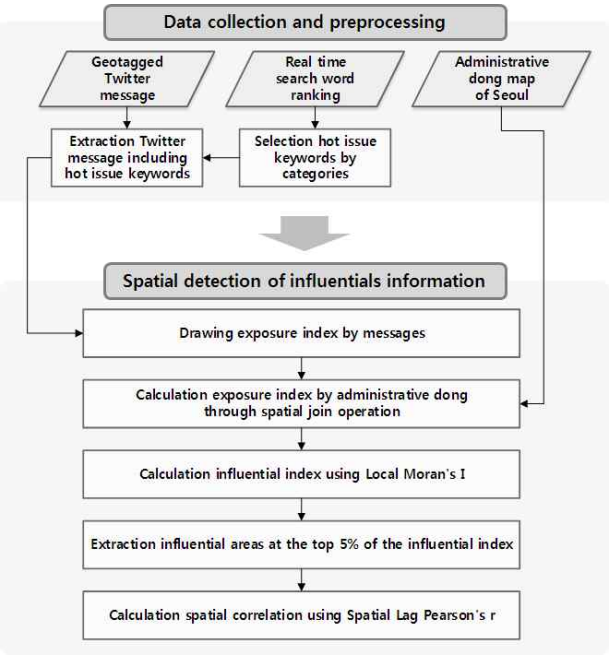


Figure 1. Flow Chart

유력자 정보의 공간적 탐색 과정은 크게 4단계로 구성되며 내용은 다음과 같다. 첫째, 특정 키워드에 대한 유력자를 파악하기 위한 노출도(exposure index)를 도출하고, 이에 대해 서울시 행정동을 기준으로 공간결합(spatial join)연산을 실시함으로써 각 키워드에 대한 행정동별 노출도를 산출하였다. 둘째, 유력자 파악에 있어 행정동별로 산출된 노출도의 공간적 의존성(spatial dependence)을 가중치로 고려하기 위해 Local Moran's I를 이용하여 유력지수(influential index)를 도출하였다. 셋째, 상위의 유력지수를 보이는 지역을 유력지역(influential area)으로 추출하여 이들의 공간적인 분포 특성을 분석하였다. 넷째, Spatial Lag Pearson's r을 이용하여 키워드별로 다양하게 추출된 유력지역들의 공간적인 분포 패턴 간의 상관성(correlation)을 분석하였다(Fig. 1).

본 논문의 구성은 다음과 같다. 2장에서는 유력자 정보의 공간적 탐색을 위한 분석기법을 설계하였으며, 3장에서는 설계한 분석기법을 실제 데이터에 적용한 결과를 작성하였다. 4장에서는 본 연구에 대한 결론과 향후 연구 과제를 도출하였다.

2. 유력자 정보의 공간적 탐색기법

2.1 데이터 수집

2.1.1 지오태깅된 트윗

트위터에서 제공하는 OpenAPI(Open Application Programming Interface)를 이용하여 서울시 전역을 대상으로 2013년 8월 5일부터 9월 5일까지 한 달 동안 168,040건의 지오태깅된 트위터 메시지를 수집하였으며, 분석 단위는 서울시의 423개 행정동으로 설정하였다. OpenAPI에서 제공하고 있는 속성 정보 중 본 연구에서 사용한 필드는 id, contributors, created_at, text, coordinates, favorite_count, retweet_count, followers_count이다(Table 1).

Table 1. Field Description Collected through OpenAPI of Twitter

Field	Type	Description
id	integer	The integer representation of the unique identifier for the Tweet
contributors	integer	An collection of brief user objects indicating users who contributed to the authorship of the tweet
created_at	string	UTC time when the Tweet was created
text	string	The actual UTF-8 text of the status update
coordinates	coordinates	Represents the geographic location of the Tweet as reported by the user(longitude first, then latitude)
favorite_count	integer	Indicates approximately how many times the Tweet has been favorited by Twitter users
retweet_count	integer	Number of times the Tweet has been retweeted
followers_count	integer	The number of followers the account currently has

2.1.2 범주별 핫이슈 키워드

기존 언론사들이 뉴스 게재 시 사용하고 있는 보편적인 분류 체계를 참고로 하여, 그중에서 '정치', '경제', 'IT'를 실험 대상 범주로 설정하였다. 그리고 국내 주요 포털사 중 네이버와 네이트에서 순위별로 제공하는 뉴스 기사를 기준으로 데이터 수집 기간 동안 이슈가 되었던 키워드들을 범주별로 각각 10개 이상씩 추출하였다. 그리고 이를 수집한 데이터에 대입하여 추출된 키워드를 포함하는 트윗의 개수가 30개 이상인 것만을 추려내었다. 결과적으로 정치범주에서 '국정원', '민주당', '박근혜', 경제범주에서 '세금', '전월세', '4대강' IT 분야에서 '페이스북', '아이폰', '블로그'를 핫이슈 키워드로 선정하였다(Table 2).

Table 2. Hot Issue Keywords Selected by Categories

Category	Keyword	Number of Tweets
Politics	National Intelligence Service(NIS)	281
	Democratic Party	167
	Park Geun-hye	119
Economy	Tax	79
	Jeonse and Monthly Rent	40
	Four Major River	38
IT	Facebook	102
	iPhone	91
	Blog	87

2.2 공간결합연산을 이용한 행정동별 노출도 설정

2.2.1 노출도

노출도(Exposure Index, EI)란 특정 단일 키워드를 포함하고 있는 메시지의 영향력을 나타내는 수치로 정의할 수 있다. Yoo and Kim(2013)의 트위터 검색 네트워크 분석에 대한 연구에서는 트윗 작성자의 팔로어 수와 그 트윗이 리트윗될 때 리트윗하는 사용자들의 팔로어 수의 합을 이용하여 노출도를 산정하고 있으며, 이는 팔로어 유력자의 개념과 리트윗 유력자의 개념을 모두 고려한 것이다. 본 연구에서는 이러한 기존 노출도에 해당 트윗이 즐겨찾기로 지정된 횟수를 추가적으로 더해줌으로써 영향력 있는 트윗의 노출도를 가중하는 효과와 함께 노출도의 장기적인 영향까지 고려하고자 하였다. 노출도는 식 (1)과 같이 계산할 수 있으며, 0 이상의 정숫값을 가진다.

$$EI = \begin{array}{l} Followers'\ number\ of\ writer\ + \\ Followers'\ number\ of\ retweet\ users\ + \\ Bookmark\ number\ of\ Tweet \end{array} \quad (1)$$

위의 식에 따라 각 범주별로 선정된 키워드를 포함하는 트윗에 대해 각각의 노출도를 계산하여 이를 해당 트윗 객체의 속성으로 부여하였다.

2.2.2 공간결합연산을 이용한 행정동별 노출도

특정 키워드를 포함하는 발언에 대한 유력자의 공간적 분포를 파악하기 위해 노출도가 부여된 트윗을 서울시의 423개 행정동에 대해 공간결합연산을 실시하였다. 이때 하나의 행정동에 여러 개의 트윗이 포함되는 경우, 해당되는 노출도 간의 단순 합을 통해 이를 속성으로 부여하였다. 결과적으로 노출도의 값이 높게 산출된 지역은 해당 키워드에 대한 발언에 영향력을 가지는 유력자가 다수 분포하고 있음을 의미한다.

2.3 Local Moran's I를 이용한 유력지수 설정

Lee(2001)에 따르면 위치 정보가 포함된 데이터들은 공간적 의존성(spatial dependence) 때문에 서로 독립적으로 존재할 수 없다. 따라서 유력자의 공간적 탐색에 있어 이에 대한 고려가 필요하다고 보았다. 이에 본 연구에서는 유력지수(Influential Index, II)라는 개념을 제안하였다. 유력지수란 각 키워드에 대해 행정동별로 산출된 노출도와 해당 노출도의 공간적 의존성의 정도를 가중치로 하여 나타낸 값을 곱하여 구해지는 수치로, 키워드가 2개 이상일 경우 이들의 가중 합이 유력지수가 된다. 이때 노출도의 공간적 의존성을 고려하기 위해 국지적 공간 의존성 지수 중 하나인 Local Moran's I를 사용하였다. Local Moran's I는 여러 연구에서 도시공간구조를 분석하는데 유용하게 사용될 뿐 아니라(Kim et al., 2009) 유력지역을 판단할 때 총 4가지의 경우(high-high, low-low, high-low, low-high)에 대해 판단할 수 있기 때문에 노출도 합산에 있어 다른 지수들보다 합리적으로 접근할 수 있다는 장점이 있다. 예컨대 또 다른 공간적 의존성 지수인 Getis-Ord의 G_i^*의 경우에는 hot spot(high-high) 또는 cold spot (low-low)의 오직 두 가지 경우에 대해서만 판단 가능하며, spatial outlier(high-low, low-high)는 탐지할 수 없다는 단점이 존재한다.

키워드에 대한 행정동별 유력지수는 식 (2)와 같이 계산하며, 결과적으로 표준화를 통해 0과 1사이의 실숫값을 가진다. 유력지수가 1에 가까울수록 해당 키워드에 대한 발언에 영향력을 가지는 유력자가 다수 분포하고 있음을 나타낸다.

$$II = \frac{\sum_{i}^{n}(Local\ Moran's\ I_i \times EI_i)}{(\sum_{i}^{n}(Local\ Moran's\ I_i \times EI_i)_{\max}} \quad (2)$$

여기서 i는 각 키워드, n은 키워드의 개수, $Local\ Moran's\ I_i$는 i번째 키워드에 대한 $Local\ Moran's\ I$, EI_i는 i번째 키워드에 대한 노출도이다.

2.4 Spatial Lag Pearson's r을 이용한 공간적 상관성 분석

2.4.1 Spatial Lag Pearson's r

이변량(bivariate) 간의 상관관계를 측정하는 지표인 Pearson's r은 데이터 집합의 공간적 분포를 반영하지 못한다. 이는 Figure 2를 통해 확인할 수 있는데, 공간적인 분포 패턴은 계속 변화하나 이를 반영하지 못한 r의 값은 그대로인 것을 볼 수 있다. 반면 Spatial Lag Pearson's r이란 Pearson's r에 spatial lag를 적용한 것으로, 이를 이용하면 이변량에 대해서 공간적 상관성을 계산할 수 있다.

spatial lag란 공간가중행렬에 의해 정의된 이웃들의 가중평균합으로서 다음과 같이 계산할 수 있다.

$$\widetilde{x_i} = \sum_j w_{ij} x_j \tag{3}$$

여기서 w_{ij}는 i와 j가 인접하면 1, 그렇지 않으면 0의 값을 갖는 이진행렬로 행표준화가 행해진 값이며, x_j는 j번째 변수를 의미한다.

또한 Spatial Lag Pearson's r은 다음 식과 같이 spatial lag가 적용된 변수 X와 Y의 공분산을 각 표준편차의 곱으로 나눔으로써 계산할 수 있다.

$$r_{\widetilde{X},\widetilde{Y}} = \frac{\sum_i (\widetilde{x_i} - \bar{\widetilde{x}})(\widetilde{y_i} - \bar{\widetilde{y}})}{\sqrt{\sum_i (\widetilde{x_i} - \bar{\widetilde{x}})^2} \sqrt{\sum_i (\widetilde{y_i} - \bar{\widetilde{y}})^2}} \tag{4}$$

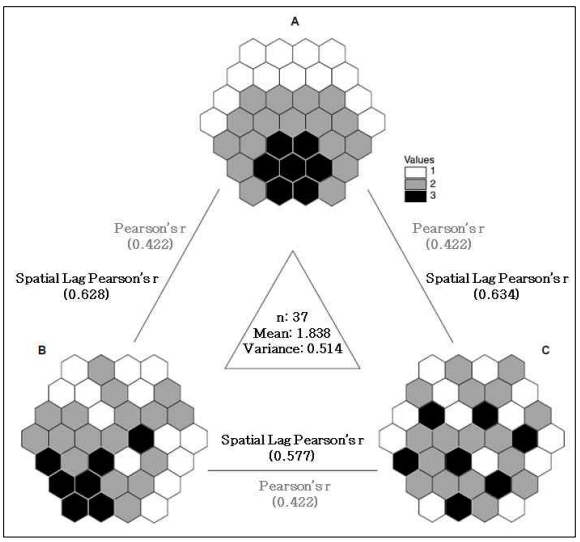


Figure 2. Difference of Spatial Distribution(Lee, 2001)

Table 3. Interpretation of Correlation Coefficient by Range(Taylor, 1990)

Range	Interpretation
-1.0 ~ -0.7	Strong negative correlation
-0.7 ~ -0.3	Moderate negative correlation
-0.3 ~ -0.1	Weak negative correlation
-0.1 ~ +0.1	No or negligible correlation
+0.1 ~ +0.3	Weak positive correlation
+0.3 ~ +0.7	Moderate positive correlation
+0.7 ~ +1.0	Strong positive correlation

여기서 $\widetilde{x_i}$, $\widetilde{y_i}$는 i 위치에서 두 변수 X와 Y의 spatial lag에 의한 값이며, $\bar{\widetilde{x}}$, $\bar{\widetilde{y}}$는 X와 Y의 spatial lag에 의한 이웃 값들의 평균이다. 결괏값은 -1에서 1 사이의 값을 가지며, 값의 범위에 따른 상관계수의 일반적인 의미는 Table 3과 같다.

2.4.2 공간적 상관성 분석

공간적 상관성 분석은 서로 다른 키워드에 대한 유력자의 공간적인 분포 패턴이 어느 정도의 유사도, 즉 상관성을 가지는지를 수치적으로 파악하기 위한 과정이다. 이를 위해 이변량 공간 상관성 측도인 Spatial Lag Pearson's r을 이용하였으며, 각 키워드에 대한 유력지수 값을 변량으로 활용하여 식 (5)와 같이 공간적 상관계수(Spatial Correlation Coefficient, SCC)를 구할 수 있다.

$$SCC_{AB} = \frac{\sum_i^n (\widetilde{\Pi(A)}_i - \overline{\widetilde{\Pi(A)}})(\widetilde{\Pi(B)}_i - \overline{\widetilde{\Pi(B)}})}{\sqrt{\sum_i^n (\widetilde{\Pi(A)}_i - \overline{\widetilde{\Pi(A)}})^2} \sqrt{\sum_i^n (\widetilde{\Pi(B)}_i - \overline{\widetilde{\Pi(B)}})^2}} \tag{5}$$

여기서 i는 각 행정동, n은 행정동의 개수, $\widetilde{\Pi(A)}_i$는 i번째 행정동에서 키워드 A에 대한 유력지수의 spatial lag에 의한 값, $\widetilde{\Pi(B)}_i$는 i번째 행정동에서 키워드 B에 대한 유력지수의 spatial lag에 의한 값이다. 이에 대한 결괏값은 0과 1사이의 범위를 갖게 되며, 값의 해석은 일반적으로 0.3 이상이면 뚜렷한 공간적 상관성이 있다고 보고, 0.7 이상이면 강한 공간적 상관성을 보인다고 할 수 있다(Table 3 참조).

3. 실험 적용 및 결과

3.1 키워드별 유력지수 산출 및 유력지역 추출

각 범주별 키워드에 대해 행정동별 유력지수를 산출하고, 이를 바탕으로 상위 5%의 값을 가지는 유력지역을 추출하여 공간적인 분포 특성을 살펴보았다.

먼저, 정치범주 키워드 중 '국정원'에 대해 유력지수가 가장 높은 지역, 즉 가장 영향력 있는 지역을 추출한 결과, 중구 명동이 가장 높은 값을 보였다. 이에 대한 공간적 분포를 살펴보면, 유력지역은 주로 한강 이남 지역에 분포하고 있으며, 그중에서도 강남구에 속하는 행정동(대치2동, 개포1동, 개포2동, 압구정동)이 4개로 가장 많았다. 또한 강남구에 속하면서 상위 유력지수 값을 가지는 지역들과 인접하고 있는 성동구 성수1가1동과 서초구 내곡동 및 양재2동의 유력지수 역시 높게 나타났다(Table 4, Fig. 3(a)).

정치범주의 두 번째 키워드인 '민주당'의 경우 중구 회현동의 유력지수가 가장 높게 나타났으며, 이를 공간적으로 살펴보면, 서울의 경제적 중심지에 해당하는 중구, 용산구, 영등포구에 높은 유력지수를 보이는 행정동이 주로 분포하고 있는 것을 확인할 수 있다. 이 외에도 성북구(길음1동, 정릉1동, 정릉4동)가 3개의 행정동을 포함하고 있었다(Table 5, Fig. 3(b)).

정치범주의 세 번째 키워드인 '박근혜'에 대한 유력지역들을 추출한 결과, 가장 영향력 있는 지역은 서초구 서초3동인 것으로 나타났다. 이에 대한 공간적 분포를 보면 서울시의 도심에 해당하는 종로구(숭인1동, 청운효자동, 종로1·2·3·4가동) 및 중구(명동, 소공동, 을지로동)와 서초구(양재1동, 서초3동, 서초4동) 지역에 높은 값들이 다소 밀집되어 분포하고 있는 것을 확인할 수 있었다(Table 6, Fig. 3(c)).

경제범주의 첫 번째 키워드인 '세금'에 대한 유력지역을 추출한 결과, 가장 영향력 있는 지역은 서초구 서초3동인 것으로 나타났다. 공간적 분포를 자치구별로 살펴보면, 서초구에 해당하는 행정동(서초3동, 양재1동, 방배1동, 방배3동)이 4개로 가장 많았으며, 이들 지역은 모두 지리적으로 인접하고 있다는 특징이 있었다(Table 7, Fig. 3(d)).

경제범주의 두 번째 키워드인 '전월세'의 경우 '세금'과 마찬가지로 서초구 서초3동의 유력지수가 가장 높게 나타났으며 공간적으로 살펴본 결과, 영등포구에 해당하는 여의동, 영등포동, 당산2동이 높은 유력지수를 보이면서 지리적으로 서로 인접하고 있는 것을 확인할 수 있었다. 반면 서초구의 경우에는 역시 3개의 행정동(서초3동, 반포1동, 방배4동)이 영향력 있는 유력지역으로 추출되었으나 반포1동과 방배4동의 일부분만이 인접하고 있을 뿐 서로 독립적으로 존재하고 있다는 점에서 영등포구의 분포 특성과는 차이를 보였다(Table 8, Fig. 3(e)).

Table 4. Top Five Influential Index(II) of The Keyword 'NIS'

Gu	Administrative dong	II
Jung-gu	Myeong-dong	1.000
Seocho-gu	Naegok-dong	0.755
Yeongdeungpo-gu	Yeouido-dong	0.444
Gangnam-gu	Gaepo 1-dong	0.387
Gangnam-gu	Daechi 2-dong	0.235

Table 5. Top Five Influential Index(II) of The Keyword 'Democratic Party'

Gu	Administrative dong	II
Jung-gu	Hoehyeon-dong	1.000
Jung-gu	Myeong-dong	0.670
Yeongdeungpo-gu	Dangsan 2-dong	0.605
Yongsan-gu	Hannam-dong	0.379
Seongbuk-gu	Gireum 1-dong	0.301

Table 6. Top Five Influential Index(II) of The Keyword 'Park Geun-hye'

Gu	Administrative dong	II
Seocho-gu	Seocho 3-dong	1.000
Seocho-gu	Yangjae 1-dong	0.892
Eunpyeong-gu	Eungam 1-dong	0.834
Dongjak-gu	Heukseok-dong	0.678
Gangnam-gu	Apgujeong-dong	0.613

Table 7. Top Five Influential Index(II) of The Keyword 'Tax'

Gu	Administrative dong	II
Seocho-gu	Seocho 3-dong	1.000
Gangbuk-gu	Samgaksan-dong	0.613
Yeongdeungpo-gu	Yeoui-dong	0.575
Nowon-gu	Gongneung 2-dong	0.305
Jung-gu	Sogong-dong	0.293

Table 8. Top Five Influential Index(II) of The Keyword 'Jeonse and Monthly Rent'

Gu	Administrative dong	II
Seocho-gu	Seocho 3-dong	1.000
Dongjak-gu	Heukseok-dong	0.922
Songpa-gu	Jangji-dong	0.791
Eunpyeong-gu	Eungam 1-dong	0.491
Dongjak-gu	Sangdo 1-dong	0.238

Table 9. Top Five Influential Index(II) of The Keyword ‘Four Major River’

Gu	Administrative dong	II
Mapo-gu	Hapjeong-dong	1.000
Seongdong-gu	Seongsu1-ga1-dong	0.610
Seocho-gu	Seocho 3-dong	0.509
Gangnam-gu	Apgujeong-dong	0.142
Gangseo-gu	Yeomchang-dong	0.077

경제범주의 세 번째 키워드인 ‘4대강’에 대한 유력지역은 마포구 합정동인 것으로 나타났다. 위의 내용에 대한 공간적인 분포를 살펴본 결과, 강남구(압구정동, 청담동)와 성동구(성수1가1동, 옥수동) 지역에 유력 행정동이 지리적으로 인접하며 집중적으로 분포하고 있었고, 서초구(서초3동, 서초4동, 방배4동) 역시 3개의 행정동이 서로 인접한 상태로 밀집 분포하고 있었다(Table 9, Fig. 3(f)).

IT범주의 첫 번째 키워드인 ‘페이스북’에 대해 가장 영향력 있는 지역은 용산구 이태원1동인 것으로 나타났다. 위의 값을 공간적으로 살펴본 결과, 키워드 ‘페이스북’에 대해 높은 유력지수를 보이는 행정동들이 마포구(용강동, 서교동, 서강동, 상암동, 성산2동) 및 송파구(오륜동, 오금동, 삼전동, 문정1동) 지역에 주로 밀집되어 분포하고 있는 것을 알 수 있었으며, 앞서 분석한 정치 및 경제범주에 속하는 키워드들과는 다소 상이한 분포 패턴을 보이고 있음을 확인할 수 있었다(Table 10, Fig. 3(g)).

Table 10. Top Five Influential Index(II) of The Keyword ‘Facebook’

Gu	Administrative dong	II
Yongsan-gu	Itaewon 1-dong	1.000
Gwanak-gu	Sinsa-dong	0.947
Yangcheon-gu	Sinwol 2-dong	0.805
Eunpyeong-gu	Galhyeon 2-dong	0.697
Mapo-gu	Yonggang-dong	0.580

Table 11. Top Five Influential Index(II) of The Keyword ‘iPhone’

Gu	Administrative dong	II
Jung-gu	Myeong-dong	1.000
Jung-gu	Sogong-dong	0.566
Gangnam-gu	Daechi 4-dong	0.441
Yeongdeungpo-gu	Yeoui-dong	0.255
Gangnam-gu	Apgujeong-dong	0.212

Table 12. Top Five Influential Index(II) of The Keyword ‘Blog’

Gu	Administrative dong	II
Seodaemun-gu	Sinchon-dong	1.000
Gangseo-gu	Hwagok 2-dong	0.032
Seocho-gu	Seocho 3-dong	0.013
Yeongdeungpo-gu	Singil 5-dong	0.008
Seodaemun-gu	Yeonhui-dong	0.006

IT범주의 두 번째 키워드인 ‘아이폰’의 경우, 중구 명동의 유력지수가 가장 높게 나타났다. 위의 내용을 공간적으로 살펴본 결과, 키워드 ‘아이폰’에 대한 유력자는 서울의 도심지에 해당하는 중구(명동, 소공동) 및 종로구(종로1·2·3·4가동, 사직동)와 강남구(압구정동, 신사동, 논현 1동, 삼성 1동, 대치4동) 지역에 주로 분포하고 있는 것을 확인할 수 있었다(Table 11, Fig. 3(h)).

IT범주의 세 번째 키워드인 ‘블로그’의 유력지역은 서대문구 신촌동인 것으로 나타났다. 이를 공간적 분포를 통해 살펴 본 결과, 지리적으로 인접 관계에 위치한 서대문구 연희동 및 신촌동과 마포구 연남동 지역의 유력지수가 높게 나타났으며, 전체적으로는 서울시 전역에 다소 산발적으로 분포하고 있는 것을 확인할 수 있었다(Table 12, Fig. 3(i)).

본 장에서는 범주별 키워드에 대한 행정 동별 유력지수를 산출하고, 이들 중 상위 5%에 해당하는 지역들의 공간적인 분포 특성을 탐색해 보고자 하였다. 그 결과, 서울시의 423개 행정동 중에서 정치범주에서 총 50개, 경제범주에서 총 46개, IT범주에서는 총 54개의 행정동이 각각 유력지역으로 추출되었다. 또한 각 범주별로 세 개의 키워드를 중첩 분석하여 교차되는 지역을 탐색한 결과, 정치범주는 3개 지역, 경제범주는 4개 지역, IT범주는 0개 지역이 도출되었다. 따라서 범주별 키워드에 대한 공간적 분포 패턴은 경제범주가 가장 유사하고, IT범주가 가장 상이하다는 것을 짐작해 볼 수 있다. 이러한 공간적 상관성에 대한 정량적인 분석은 3.2장에서 수행하였다.

3.2 유력지역 간 공간적 상관성 측정

범주별 키워드에 대해 추출한 유력지역들의 공간적 분포 패턴을 대상으로 공간 상관성 분석을 실시하였다. 그 결과, 동일 범주 내의 키워드 간 공간적 상관계수는 모두 0.3 이상의 값을 가지므로 범주 내에서는 키워드 간의 공간적인 분포 패턴이 유사하다는 결론을 내릴 수 있다. 특히 정치 및 경제범주에서 대체적으로 공간적 상관계수가 높게 나왔는데, 그중에서도 경제범주의 키

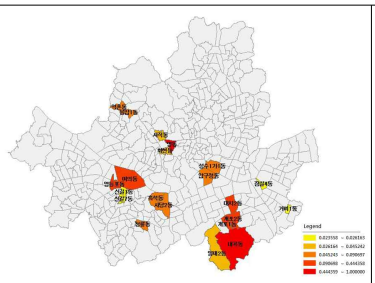
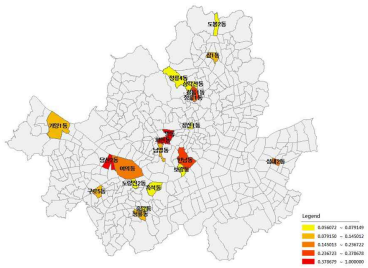
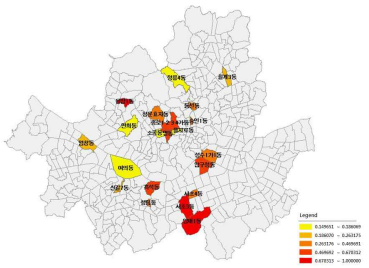
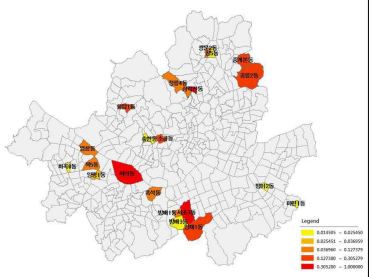
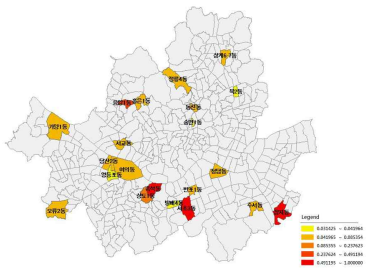
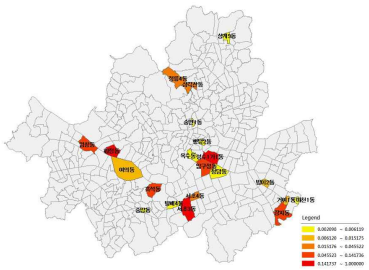
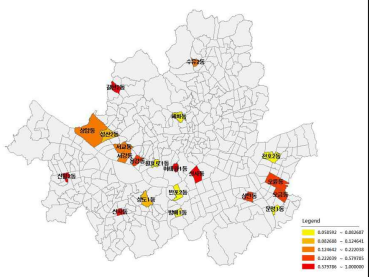
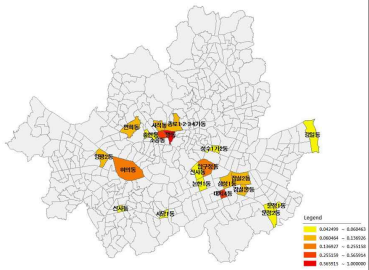
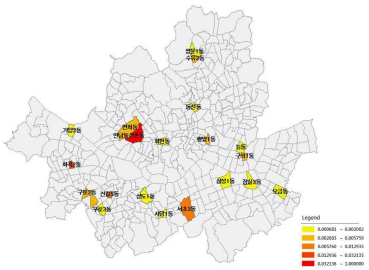
Category	Keyword		
Politics	 (a) 'NIS'	 (b) 'Democratic Party'	 (c) 'Park Geun-hye'
Economy	 (d) 'Tax'	 (e) 'Jeonse and Monthly Rent'	 (f) 'Four Major River'
IT	 (g) 'Facebook'	 (h) 'iPhone'	 (i) 'Blog'

Figure 3. Spatial Distribution of Top 5% of Influential Index by Keyword

Table 13. Spatial Correlation Coefficient between Influential Areas by Keyword

Division		Politics			Economy			IT		
		NIS	Democratic Party	Park Geun-hye	Tax	Jeonse and Monthly Rent	Four Major River	Facebook	iPhone	Blog
Politics	NIS	1	0.3840	0.4332	0.2674	0.1484	0.2274	0.0714	0.3529	0.0896
	Democratic Party	0.3840	1	0.3691	0.3926	0.3239	0.1668	0.0682	0.1983	0.0340
	Park Geun-hye	0.4332	0.3691	1	0.4274	0.3960	0.3943	0.2100	0.4202	0.2244
Economy	Tax	0.2674	0.3926	0.4274	1	0.4244	0.4671	0.0884	0.1475	0.2059
	Chonsei and Monthly Rent	0.1484	0.3239	0.3960	0.4244	1	0.3489	0.1903	0.1675	0.0944
	Four Major River	0.2274	0.1668	0.3943	0.4671	0.3489	1	0.1175	0.1281	0.1262
IT	Facebook	0.0714	0.0682	0.2100	0.0884	0.1903	0.1175	1	0.3097	0.3289
	iPhone	0.3529	0.1983	0.4202	0.1475	0.1675	0.1281	0.3097	1	0.3377
	Blog	0.0896	0.0340	0.2244	0.2059	0.0944	0.1262	0.3289	0.3377	1

워드 '세금'과 '4대강'의 상관계수가 0.4671로, 이들의 공간적 분포 패턴의 상관성이 가장 높은 것을 알 수 있다(Table 13).

한편 서로 다른 범주 간의 상관계수를 살펴보면, 정치범주와 경제범주의 키워드 간 상관계수는 평균 약 0.3으로 비교적 높은 상관성을 보였다. 특히, 정치범주의 '박근혜'와 경제범주의 '세금'의 상관계수가 0.4274로, 다른 범주의 키워드 간 상관계수 중 가장 높은 수치를 보였다. 반면, 정치범주와 IT범주의 키워드 간 상관계수는 평균 0.18, 경제범주와 IT범주의 키워드 간 상관도 지수는 평균 0.15로, 이들 범주 간 키워드들은 대체적으로 낮은 공간적 상관성을 갖는다는 것을 알 수 있다. 특히 정치범주의 '민주당'과 IT범주의 '블로그' 간의 상관계수는 0.0340으로 최하위를 기록했다(Table 13).

4. 결 론

본 연구에서는 공간통계분석기법을 이용하여 LBSNS 데이터를 대상으로 사회문화적으로 이슈가 되고 있는 키워드에 대한 발언에 영향력을 가지는 유력자를 공간적으로 탐색, 활용하는 방안을 제시하고자 하였다. 이를 위해 서울시를 대상으로 한 달 간 168,040건의 LBSNS 데이터를 수집하였으며, 기존의 뉴스 분류 체계를 차용하여 '정치', '경제', 'IT'를 연구 대상 범주로 설정하고, 데이터 수집 기간 동안 이슈가 되었던 키워드들을 범주별로 분류하였다. 또한 유력자 정보의 공간적 탐색 과정으로써 각 키워드에 대한 행정동별 노출도를 산출하고, 이를 바탕으로 노출도의 공간적 의존성을 가중치로 고려한 유력지수의 개념을 도출하여 실험에 적용하였다. 그리고 상위의 유력지수를 보이는 지역을 유력지역으로 추출하여 이들의 공간적인 분포 특성을 분석하였으며, 키워드별로 추출된 유력지역들의 공간적인 분포 패턴 간의 상관성을 분석하였다.

본 연구는 LBSNS를 포함하는 다양한 소셜 미디어의 사회문화적 영향력에 대한 막연하고 추상적인 연구의 한계를 극복하고, 나날이 발전하는 소셜 미디어 환경 하에서 새로운 방식으로 등장한 유력자에 대한 연구를 공간 정보의 관점에서 구체화시켰다는 점에서 의의를 가진다.

본 연구의 한계점으로는 이동성이 강한 LBSNS의 특성상 사용자의 주 활동 반경을 특정하기 어렵다는 점과 LBSNS 데이터의 부족으로 인해 특정 개인, 또는 기관에 의해 해당 지역이 유력지역으로 추출되는 문제가 발생한다는 것이다. 특히, 지오태깅된 트윗의 양은 트위터를 통해 생성되는 전체 메시지의 약 1 ~ 2% 정도를 차지하고 있기 때문에 데이터의 절대적인 수치 부족으로 인한 한계를 갖는다. 그러나 이러한 문제점들은 사용자의 위치 이력을 바탕으로 주 활동 지역을 규정하는 연구와 앞으로 LBSNS의 시장 확대로 인한 지속적인 데이터의 증가로 극복할 수 있을 것으로 보인다.

향후에는 LBSNS상에서 특정 주제에 대한 유력자를 탐지할 때 데이터의 이동성을 고려함으로써 보다 다양한 지수를 개발하는 연구가 필요할 것으로 보인다. 또한 회귀분석을 통해 본 연구에서 도출한 유력지수와 이에 영향을 미치는 세분화된 사회적 요인들에 대한 상관관계를 파악하고, 이에 대한 회귀모델을 도출함으로써 유력자의 공간적 분포를 예측하는 연구도 추가로 진행되어야 할 것으로 예상된다.

감사의 글

본 연구는 국토교통부 국토공간정보연구사업 연구비 지원(11첨단도시G10)에 의해 수행되었습니다.

References

1. Cha, M. Y., Haddadi, H., Benevenuto, F. and Gummadi, P. K., 2010, Measuring user influence in Twitter: The million follower fallacy, Proceedings of the 4th International AAAI Conference on Weblogs and Social Media, pp. 10-17.
2. Java, A., Song, X., Finin, T. and Tseng, B., 2007, Why we twitter: understanding microblogging usage and communities, Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis, pp. 56-65.
3. Kim, H. T., Kim, S. S. and An, S. H., 2009, The pattern of urban growth and measurement of spatial structural change in Daejeon metropolitan city, Journal of the Korean Society for GeoSpatial Information System, Vol. 17, No. 3. pp. 41-48.
4. Korea Internet and Security Agency, Internet & Security Issue, Vol. 2010, No. 3, p. 39.
5. Lee, S. I., 2001, Developing a bivariate spatial association measure: an integration of Pearson's r and Moran's I, Journal of Geographical Systems, Vol. 3, No. 4, pp. 369-385.
6. Lee, W. T., Cha, M. Y. and Park, H. Y., 2010, Influentials' role in mobile social media, Digital Convergence- based Prospective Study(Ⅱ), Korea

Information Society Development Institute, Vol. 2010, No. 27, pp. 1−123.
7. Park, H. S., Cha, M. Y. and Moon, S. B., 2010, Influentials ranking in social networks, Journal of Korean Institute of Information Scientists and Engineers, Vol. 28, No. 3, pp. 24−30.
8. Taylor, R., 1990, Interpretation of the correlation coefficient: a basic review, Journal of Diagnostic Medical Sonography, Vol. 6, No. 1, pp. 35−39.
9. Weng, J., Lim, E. P., Jiang, J. and He, Q., 2010, Twitterrank: finding topic−sensitive influential twitterers, In Proceedings of the third ACM International Conference on Web Search and Data Mining, pp. 261−270.
10. Yoo, B. K. and Kim S. H., 2013, Marketing strategies using social network analysis : Twitter's search network, The Journal of The Korea Contents Association, Vol. 13, No. 5, pp. 396−407.