논문 17-42-02-22 The Journal of Korean Institute of Communications and Information Sciences '17-02 Vol.42 No.02
https://doi.org/10.7840/kics.2017.42.2.453

공간 태그된 트윗을 사용한 밀도 기반 관심지점 경계선 추정

신 원 용◆°, 둥 부 도*

Density-Based Estimation of POI Boundaries Using Geo-Tagged Tweets

Won-Yong Shin◆°, Dung D. Vu*

요 약

사용자들은 그들의 관심이 관심지점 (POI: Point-of-Interest)과 관련이 있다는 사실을 언급하기 위해 위치 기반 소셜 네트워크에 체크인하거나 그들의 상태를 올리는 경향이 있다. 관심지역 (AOI: Area-of-Interest)을 찾는 기존 연구는 대부분 위치 기반 소셜 네트워크로부터 수집된 공간 태그된 사진과 함께 밀도 기반 군집화 기법을 사용하여 수행되었다. 반면, 본 연구에서는 POI 중심을 포함한 하나의 군집에 해당하는 POI 경계선을 추정하는 데에 초점을 맞춘다. 트위터 사용자들로부터의 공간 태그된 트윗을 사용하여 POI 중심으로부터 도달할 수 있는 적절한 반경을 찾음으로써 POI 경계선을 추정하는 밀도 기반 저복잡도 두 단계 방법을 소개한다. 두 단계 밀도 기반 추정을 통해 선택된 공간 태그의 convex hull로써 POI 경계선을 추정하는데, 각 단계에서 다른 크기의 반경 증가를 가정하여 진행한다. 제안한 방법은 기본 밀도 기반 군집화 방법보다 계산 복잡도 측면에서 우수한 성능을 가짐을 보인다.

Key Words : AOI, Density, Geographic Distance, Geo-Tagged Tweet, POI Boundary

ABSTRACT

Users tend to check in and post their statuses in location-based social networks (LBSNs) to describe that their interests are related to a point-of-interest (POI). While previous studies on discovering area-of-interests (AOIs) were conducted mostly on the basis of density-based clustering methods with the collection of geo-tagged photos from LBSNs, we focus on estimating a POI boundary, which corresponds to only one cluster containing its POI center. Using geo-tagged tweets recorded from Twitter users, this paper introduces a density-based low-complexity two-phase method to estimate a POI boundary by finding a suitable radius reachable from the POI center. We estimate a boundary of the POI as the convex hull of selected geo-tags through our two-phase density-based estimation, where each phase proceeds with different sizes of radius increment. It is shown that our method outperforms the conventional density-based clustering method in terms of computational complexity.

※ 본 연구는 2017년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업입니다(2014R1A1A2054577).
※ 본 연구는 미래창조과학부 및 정보통신기술연구진흥센터의 정보통신·방송 연구개발사업의 일환으로 수행하였습니다 [B0101-16-1361, 국가 공공안전서비스를 위한 LTE기반 재난통신 시스템·단말 개발].
◆° First and Corresponding Author : Dankook University, Department of Computer Science and Engineering, wyshin@dankook.ac.kr, 정회원
* Dankook University, Department of Computer Science and Engineering, vudodung85@gmail.com, 학생회원
논문번호 : KICS2017-01-005, Received January 8, 2017; Revised February 7, 2017; Accepted February 7, 2017

Ⅰ. 서 론

빅 데이터의 빠른 수집 및 처리가 가능해짐과 동시에 빅 데이터 분석 및 데이터 마이닝 연구가 큰 관심을 받고 있다. 대표적인 빅 데이터의 활용 예는 인터넷 망의 효율적 운영을 위한 빅 데이터 트래픽 분석[1], 빅 데이터 기반 개인화된 추천 시스템 개발[2], 딥러닝 모델 중 하나인 순환신경망을 사용한 언어 문장 생성[3] 등이 있다.

이 중에서도 Foursquare와 Flickr와 같은 위치 기반 소셜 네트워크는 최근에 급속도로 성장하고 있다. 이들은 수백만의 사용자들이 사진, 비디오, 음악, 텍스트 등과 같은 공간 태그된 미디어 콘텐츠를 공유할 플랫폼을 제공해준다. 이러한 공간 태그들로부터 수집된 위치 정보 덕분에 위치 기반 소셜 네트워크를 통한 다양한 관심지점 (POI: Point-of-Interest) 관련 연구[4-13]가 수행되어 왔다. 일반적으로 사용자들이 POI를 방문할 때, 사용자들은 그들의 관심이 POI와 관련이 있다는 사실을 언급하기 위해 온라인 체크인을 하거나 그들의 상태를 업로드하는 경향이 있다.

1.1 관련 연구

크게는 POI 관련하여 1) POI 추천[4-7], 2) 관심영역 (AOI: Area-of-Interest) 추정[8-13], 이렇게 두 가지 형태의 연구 방향이 있다. 먼저, 사용자들이 과거 체크인 기록으로부터 다른 시간에 다른 장소를 방문하는 경향이 있음을 관찰함으로써 시간 인지 기반 POI 추천 기술[4,5]이 소개되었다. 위치 인지 기반 추천 시스템[6,7]도 소개되었는데, POI 추천을 위해 아이템 기반 협력 필터링 기술[6]과 의미론적 분석 기술[7]을 사용하였다. 한편, AOI에 대한 데이터의 관련성이 POI와 데이터가 생성된 위치 사이의 지리적 거리에 따라 변화하기 때문에, AOI를 추정[8-13]하는 것은 근본적으로 중요하다. AOI를 추정하는 기존 연구는 대부분 위치 기반 소셜 네트워크로부터 수집된 공간 태그된 사진들과 함께 밀도 기반 군집화 방법을 사용함으로써 수행되었다. DBSCAN (Density-Based Spatial Clustering of Application with Noise)[9,10]은 원래 AOI 추정을 위해 고안된 기술은 아니지만, 가장 일반적으로 사용되고 있는 밀도 기반 군집화 알고리즘이다. 입력 데이터의 총 개수를 n_{in}이라 할 때, DBSCAN은 $O(n_{in}\log n_{in})$의 평균 실행시간 계산 복잡도와 함께 다수의 군집을 찾을 수 있다 (최악의 계산 복잡도는 $O(n_{in}^2)$로 주어짐). 공간 태그된 사진을 사용하는 대신에, AOI 추정을 위해 공간 태그된 텍스트 데이터가 활용되었는데, 텍스트 내의 묘사 (즉, POI 이름)와 위치 사이의 상관관계가 사용되었다[11]. 또한, 주어진 POI 이름에 대해 관심지역의 quality 측정 기반으로 또 다른 군집화 방법이 제안되었다[12]. 최근에 단일 군집으로 구성된 사회적 POI 경계선을 새롭게 정의하고, precision과 recall의 조화평균을 나타내는 F1 score 측정을 통해 정확도 측면에서 우수한 경계선 추정 방법도 소개되었다[13].

그 밖에도 POI와 관련되어 수행된 여러 연구가 있다. 수동, 지도 및 자동 분류 방법을 사용하여 각각의 근처 POI feature class에 대한 공간 태그된 트윗과 그들의 위치 사이의 관계를 조사하는 연구가 수행되었다[14]. 애매한 영역의 경계선을 유도하는 연구도 다방면으로 수행되었는데, 이를 위해 웹 페이지 검색[15], 최단거리 경로 그래프[16] 등이 사용되었다.

1.2 주요 기여도

본 논문에서는 POI에 대한 지리적 근접성을 반영함으로써, 대응하는 POI 중심으로 포함하는 convex hull 형태를 가진 단일 군집으로 새로운 POI 경계선을 먼저 정의한다. 이 POI 경계선은 좀 더 작은 스케일에서 추정된 AOI 안에 있는 하나의 고밀도 군집 (아마도 가장 높은 밀도를 가진 군집)을 나타낼 것이다. 그리고, 밀도 기반으로 사용자들이 가장 관심을 나타내는 군집에 대응하는 경계선을 추정하는 새로운 알고리즘 제안에 초점을 맞춘다. 제안한 알고리즘이 입력 데이터의 총 개수인 n_{in}에 따라 선형 스케일로 증가하는 계산 복잡도를 가짐을 보인다.

POI 경계선을 추정함으로써 다양한 위치 광고 서비스에 응용될 수 있다. 보다 구체적으로, POI (예: 쇼핑 몰) 홍보에 목적이 있는 회사의 마케팅 전략으로써, 온라인 브로셔가 해당 지역을 방문하는 사용자들에게 전달될 수 있다. 이에 회사 운영자들은 추정된 POI 경계선을 통해 구체적인 마케팅 존을 파악할 수 있을 뿐만 아니라 마케팅 비용을 절감할 수 있는 효과도 얻을 수 있다.

위치 기반 소셜 네트워크로부터 수집된 공간 태그된 사진 대신에, 본 논문에서는 트위터에서 수집된 공간 태그된 트윗[17-20]을 사용하는데, 이는 위치 기반 소셜 네트워크로부터 수집된 것보다 훨씬 더 방대한 양의 사용자 계정 및 데이터 기록을 포함한다. 보다 구체적으로, POI 경계선이 공간상에서 convex hull 형태로 어떻게 생성되는지를 결정하기 위해, 영국과 미국에 있는 사용자들로부터 수집한 수많은 공간 태그

된 트윗을 포함한 데이터셋을 분석한다. 이 두 개의 지역 집합들은 인구통계학적으로 비교될 수 있기에 선택되었는데, 이는 해당 지역에서 비교를 통한 의미 있는 분석을 가능하게 하는 충분한 트위터 데이터를 보유하고 있기 때문이다.

데이터셋 수집 이후에, 밀도 기반으로 POI 경계선 추정을 위한 새로운 두 단계 방법을 제안한다. 그리고, 제안 방법의 계산 복잡도가 알고리즘에서 사용된 입력 데이터의 수에 선형적으로 증가하고 (즉, $O(n_{in})$), 이는 기존 밀도 기반 군집화 방법인 DBSCAN[9,10]의 계산 복잡도보다 낮음을 보인다. 제안하는 추정 알고리즘은 다음의 과정을 따른다.

- 먼저 Wikipedia 개념과 함께 POI를 수집한다.
- 위치 기반 소셜 네트워크에서 체크인 데이터와 함께 주제의 유사성을 조사하는 방식과는 달리, 본 연구에서는 트위터 사용자들이 140자 이내로 트윗한다는 사실로부터 쿼리 (query) 처리를 통해 모든 관련된 공간 태그된 트윗을 수집한다. 구체적으로, POI 이름 (예: POI 전체 이름 또는 약어)을 포함한 텍스트를 찾는 쿼리 과정을 수행한다.
- 그 후에 POI 중심과 각 트윗이 포스팅된 위치 사이의 거리를 계산한다.
- POI 중심으로부터 도달할 수 있는 적절한 반경을 간단히 찾음으로써, 두 단계 접근 방식을 통해 선택된 공간 태그의 convex hull (즉, 최 가장자리 영역)로 POI에 대한 다각형 형태의 경계선을 추정한다. 더욱 정확한 결과를 제공하기 위해 각 단계에서는 다른 크기의 반경 증가를 가정하여 진행한다.

방대한 양의 공간 태그된 트위터 데이터 기반으로, 분석 결과는 선형 스케일의 실행시간 계산 복잡도와 함께 정밀한 POI 경계선을 제공함을 확인할 수 있다.

본 논문의 구성은 다음과 같다. Ⅱ장에서는 데이터셋에 대해 설명한다. Ⅲ장에서는 용어 정의 및 추정 알고리즘을 포함한 방법론을 설명한다. Ⅳ장에서는 실험결과를 보인다. Ⅴ장에서는 논문을 요약한다.

Ⅱ. 데이터셋

본 장에서는 트위터 데이터 및 POI 수집에 대해 설명한다.

2.1 트위터 데이터 수집

본 논문에서는 트위터 Streaming API를 통해 수집한 데이터셋을 사용한다. 데이터셋은 2015년 7월 29일부터 2015년 8월 29일까지 대략 한 달 간 영국과 미국에서의 트위터 사용자들로부터 기록된 방대한 양의 공간 태그된 트윗으로 구성된다. 한 달 간의 단기간 수집된 데이터셋은 널리 알려진 POI들에 대한 경계선을 추정하는데 충분하다. 영국에서 수집된 데이터셋은 629,881명의 다른 사용자들로부터 포스팅된 18,682,819개의 공간 태그된 트윗으로 구성되는 반면, 미국에서 수집된 또 다른 데이터셋은 2,139,483명의 다른 사용자들로부터 포스팅된 58,118,361개의 공간 태그된 트윗으로 구성된다. 각 트윗은 자신의 필드 (field) 이름에 의해 구별되어지는 수많은 요소를 포함한다. 데이터 분석을 위해 본 연구에서는 트윗으로부터 다음 네 가지 중요한 필드를 채택하도록 한다.

- user_id_str: 특정 사용자에 대한 스트링으로 표현되는 유일한 ID
- text: POI 이름을 포함한 실제 UTF-8 텍스트
- lat: 트윗의 위치에 대한 위도
- lon: 트윗의 위치에 대한 경도

2.2 POI 수집

본 절에서는 Wikipedia 개념과 함께 POI를 수집하는 방법을 소개한다. 쿼리 처리를 통해 텍스트 필드가 관련된 POI 이름을 포함하는 그러한 공간 태그된 트윗을 얻는다. 본 연구에서는 분석을 위해 London과 Los Angeles에 위치한 두 가지 POI를 사용한다. 여기서, POI는 우리가 고려하는 위에서 언급한 것뿐만 아니라 관광객이나 사용자들이 자주 체크인하는 경향이 있는 인기 있는 장소를 포함한 잠재적으로 관심이 있는 다양한 곳을 의미한다. 경계선 크기가 서로 다를 것으로 예상되는 그러한 POI들을 선택하도록 한다.

네 개의 POI에 대한 대표적인 속성은 표 1에서 정

표 1. 네 가지 관심지점
Table 1. Four POIs

POI name	(latitude, longitude)	The number of geo-tagged tweets	Initial radius Δr_1 (m)
London Eye	(51.503300°, -0.119700°)	2,178	80
Victoria and Albert Museum (V&A)	(51.496667°, -0.171944°)	1,098	120
Dodger Stadium	(34.072686°, -118.240603°)	3,666	120
Los Angeles International Airport (LAX)	(34.053718°, -118.242642°)	4,100	2,000

리하였다. 표 1 안의 두 번째와 세 번째 열은 각각 POI 중심 좌표 및 텍스트가 POI 이름을 포함하는 공간 태그된 트윗의 총 수를 나타낸다. 어림 감정으로 인해 POI의 지리적 영역을 포함하는 POI 중심으로부터의 대략적인 최소 거리는 구글 맵 (Google Maps)으로부터 얻어지고 초기 반경 Δr_1으로 표기되는데, 이는 표 1의 마지막 열에서 확인할 수 있다. 표에서 볼 수 있듯이 본 연구에서는 다른 종류의 POI들을 선택하였기 때문에, 관련된 공간 태그된 트윗의 수와 초기 반경은 POI에 따라 크게 차이가 나는 것을 알 수 있다.

Ⅲ. 방법론

본 장에서는 용어 정의 및 추정 알고리즘을 포함한 전체적인 연구 방법론을 설명한다.

3.1 정의

먼저 "POI 경계선"에 대해 다음과 같이 정의한다.

정의 1. POI 경계선은 대응하는 POI 중심을 포함하는 하나의 convex hull 형태의 고밀도 군집을 나타낸다. 경계선 안에 있는 POI 중심으로부터 사전에 결정된 반경 증가와 함께 생성된 모든 환형 (annulus)은 최소 하나 이상의 공간 태그를 포함해야 한다.

본 연구에서의 정의는 보다 큰 스케일에서 두 개 이상의 고밀도 군집으로 구성된 AOI에 대한 기존 정의[8-10]와는 다르다. 이러한 POI 경계선은 일반적으로 모든 군집 중에서 가장 높은 밀도를 가진 군집을 표출할 가능성이 높다.

3.2 밀도 기반 POI 경계선 추정 알고리즘

이제 정밀한 POI 경계선에 대한 밀도 기반 두 단계 추정 알고리즘을 설명한다. 앞서 언급한 바와 같이, 본 연구에서는 각 POI에 대해 적절한 반경을 찾는 것에 초점을 맞춘다. 먼저 알고리즘 전개 시 필요한 변수부터 소개한다. 사용자 u의 공간 태그된 텍스트 데이터를 t_u, 사용자 u의 좌표를 l_u, POI 중심 좌표를 c라고 하자. D_{in}과 D_{out}을 각각 텍스트가 POI 이름을 포함하는 공간 태그된 트윗의 집합 및 추정된 POI 경계선 내 모든 공간 태그된 트윗의 집합이라고 하자. 또한, 중심 c와 반경 r을 가진 원 안에 존재하는 공간 태그된 데이터 집합을 $D_{(c,r)}$로 표기한다. POI 중심 c와 사용자 u의 위치 사이의 지리적 거리인 $d(c,l_u)$는 구면 코사인 법칙 (spherical law of cosines)을 사용하여 계산될 수 있는데, 이는 1 미터 이내의 거리로 추정된 거리의 well-conditioned된 결과를 제공해준다.

첫 번째 추정 단계에서는, 주어진 POI에 대해 비연속적으로 증가하는 변수인 아래와 같이 주어지는 반경 r_i와 함께 중심 c를 갖는 원을 사용한다.

$$r_i = \triangle r_1 > 0 \tag{1}$$

여기서, 특정 조건을 만족하면 반경 r_i는 각 단계마다 $\triangle r_1$만큼 증가하고, 업데이트된 값은 저장된다. $d(c,l_u)$가 r_i보다 작을 때, $(t_u,l_u)\in D_{(c,r_i)}$가 성립한다. 이제, 첫 번째 단계에서의 업데이트 조건에 대해 설명한다. 특정 환형에서의 공간 태그된 트윗의 수를 아래와 같이 정의하자.

$$dn_1 = |D_{(c,r_i)} \setminus D_{(c,r_{i-1})}| \tag{2}$$

만약 dn_1이 주어진 임계값 $\triangle \eta > 0$보다 크다면, POI에 대한 세 개의 변수 r_i, $D_{(c,r_i)}$, dn_1은 반복적으로 업데이트된다. 그렇지 않으면 이 과정은 종료된다. 여기서, 임계값 $\triangle \eta$는 반경 $\triangle r_1$에 따라 적응적으로 결정될 수 있는데, 이에 대한 설명은 나중에 다루도록 한다.

이제, 두 번째 추정 단계를 설명하도록 하자. 이 단계는 첫 번째 단계와 비교하여 더 정확한 결과를 얻기 위해 수행된다. 다음과 같이 주어지는 반경 $r_{i,j}$와 함께 중심 c를 갖는 원을 사용한다.

$$r_{i,j} = r_{i-1} > 0 \tag{3}$$

이 단계에서의 증가 반경 구간을 $\triangle r_2$라 정의하되, 이 값은 GPS 오차로 인한 거리보다는 큰 값으로 설정하도록 한다. 보다 구체적으로 아래와 같이 가정한다.

$$\triangle r_2 = \frac{\triangle r_1}{Q} \tag{4}$$

여기서, Q는 구간 분할 정도를 나타내는 변수이다. 마찬가지로, $r_{i,j}$는 또 다른 조건 하에서 매 단계마다 $\triangle r_2$만큼 증가하고, 업데이트된 값은 저장된다. 이 단계에서, 특정 환형에서의 공간 태그된 트윗의 수를 아래와 같이 정의하자.

$$dn_2 = |D_{(c,r_{i,j})} \setminus D_{(c,r_{i,j-1})}| \tag{5}$$

만약 dn_2가 1보다 크거나 같다면, 세 개의 변수 $r_{i,j}$, $D_{(c,r_{i,j})}$, dn_2는 반복적으로 업데이트된다. 그렇지 않으면 이 과정은 종료되고, 집합 $D_{out}=D_{(c,r_{i,j})}$가 최종적으로 얻어진다. 위에서 설명한 추정 알고리즘에 대한 전반적인 절차는 표 2에서 요약된다.

그 후에, quick-hull과 같은 알고리즘을 사용하여 convex hull 문제를 해결함으로써, POI 경계선에 대응하는 POI 중심 및 D_{out} 안의 점들을 포함한 가장 작은 convex 다각형을 찾을 수 있게 된다.

표 2. 밀도 기반 두 단계 추정 알고리즘
Table 2. Density-based two-phase estimation algorithm

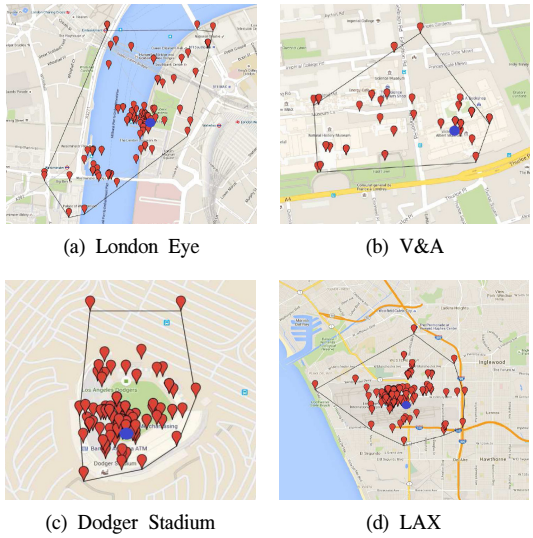
Input: D_{in}, c, $\triangle r_1$, and $\triangle r_2$
Output: D_{out}
Initialization: $i \leftarrow 1$; $j \leftarrow 0$; $r_i \leftarrow 0$; $r_{i,j} \leftarrow 0$, $dn_1 \leftarrow 0$; $dn_2 \leftarrow 0$; $D_{(c,r_i)} \leftarrow \{(t_u,l_u) | d(c,l_u) < \triangle r_1,\ (t_u,l_u) \in D_{in}\}$
1: do
2: $i \leftarrow i+1$
3: $r_i \leftarrow r_i + \triangle r_1$
4: $D_{(c,r_i)} \leftarrow \{(t_u,l_u) | d(c,l_u) < r_i,\ (t_u,l_u) \in D_{in}\}$
5: $dn_1 \leftarrow |D_{(c,r_i)} \setminus D_{(c,r_{i-1})}|$
6: while $dn_1 > \triangle \eta$
7: $i \leftarrow i-1$
8: $r_{i,j} \leftarrow r_i$
9: do
10: $j \leftarrow j+1$
11: $r_{i,j} \leftarrow r_{i,j} + \triangle r_2$
12: $D_{(c,r_{i,j})} \leftarrow \{(t_u,l_u) | d(c,l_u) < r_{i,j},\ (t_u,l_u) \in D_{in}\}$
13: $dn_2 \leftarrow |D_{(c,r_{i,j})} \setminus D_{(c,r_{i,j-1})}|$
14: while $dn_2 \geq 1$
15: $D_{out} \leftarrow D_{(c,r_{i,j})}$
16: return D_{out}

Ⅳ. 실험 결과

본 장에서는 실험을 통하여 추정된 POI 경계선들을 도시화하고, 제안 알고리즘의 계산 복잡도를 보인다.

4.1 POI 경계선 추정 결과

Ⅲ장에서 보인 제안한 추정 알고리즘을 사용하여 실험 결과를 먼저 도시한다. 본 연구에서는 실험의 간



(a) London Eye (b) V&A

(c) Dodger Stadium (d) LAX

그림 1. POI 경계선 추정
Fig. 1. Estimation of POI boundaries

소화를 위해 $Q=10$을 가정한다. 또한, $\triangle\eta$는 $\triangle r_1$과 $\triangle r_2$ 사이의 관계에 따라 100으로 적절히 주어질 수 있다. 구글 맵 위에서 추정된 POI 경계선을 그림 1에서 보인다. 하나의 파란색 원은 POI 중심을 나타내고, 집합 D_{out}에서 선택된 공간 태그된 트윗들은 빨간색 핀으로 표기된다. 이 때 몇 개의 핀들은 거의 중첩됨을 확인할 수 있다.

4.2 계산 복잡도

다음으로, 기존 DBSCAN 알고리즘과 제안하는 두 단계 추정 알고리즘을 계산 복잡도 측면에서 비교 분석한다. Ⅱ장에서 언급한 데이터셋을 사용하여 LAX 경계선을 추정하는데 있어서, calling process 시스템의 instruction 실행을 위해 요구되는 CPU 시간으로 전체적인 평균 실행시간을 측정한다.

DBSCAN 알고리즘의 두 가지 중요한 변수인 반경 및 인접 점들의 최소 개수는 각각 2km 및 5로 주어지는데, 이 때 알고리즘을 통해 총 37개의 군집이 찾아진다. POI 경계선을 얻기 위해 37개 군집 중 POI 중심을 포함하는 고 밀도 군집을 하나 선택한다. 즉, 나머지 36개의 군집은 필터링된다. 이 경우 경계선 내에 있는 POI 중심으로부터 최대 거리는 3.782km로 주어진다. 반면, 제안한 알고리즘을 사용할 경우 즉시 하나의 군집을 출력으로 반환하게 된다. 제안한 기법에서 $\triangle r_1=2$km, $Q=10$, $\triangle\eta=100$으로 가정함으로써 POI 중심으로부터 도달할 수 있는 최대 거리는 3.872km로 주어지게 되는데, 이는 DBSCAN의 경우

와 상담히 유사함을 확인할 수 있다. 그림 2에서는 DBSCAN 및 제안한 추정 알고리즘에 대해 초 단위로 실행시간을 집합 D_{in} 안에 있는 공간 태그된 트윗의 수에 따라 도시하였다. $|D_{out}|$이 n_{in}보다 상대적으로 느리게 스케일할 때 (즉, $|D_{out}|=o(n_{in})$), 제안하는 알고리즘의 계산 복잡도는 $O(n_{in})$으로 주어짐을 알 수 있다. 반면에, DBSCAN 알고리즘의 복잡도는 $n_{in}\log n_{in}$으로 스케일하는 것으로 알려져 있다. n_{in}이 클 경우, 두 알고리즘 사이의 성능 격차는 점점 더 벌어지게 된다. 그림 2에서 비교를 위해 점근적인 곡선도 도시되었다. 두 개의 곡선 모두 실험 결과와 매우 유사한 경향을 보임을 확인할 수 있다.

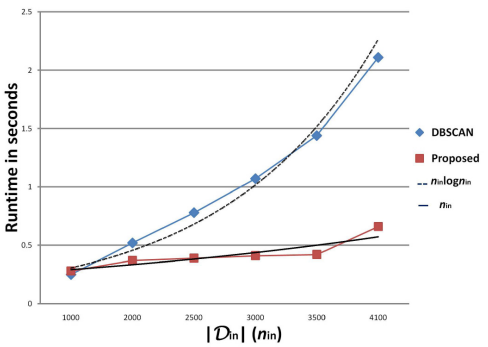


그림 2. 실행시간 계산 복잡도
Fig. 2. Runtime computational complexity

V. 결 론

본 논문에서는 POI 경계선을 POI 중심을 포함한 convex 다각형 형태로 새롭게 정의하고 이를 찾을 수 있는 밀도 기반 저 복잡도 POI 경계선 추정 알고리즘을 소개하였다. 구체적으로, 트위터 사용자들로부터 공간 태그된 트윗을 사용하여 POI 중심으로부터 도달할 수 있는 적절한 반경을 밀도 기반 두 단계로 찾는 방법을 제안하였는데, 각 단계에서 다른 크기의 반경 증가를 가정하여 진행하는 추정 알고리즘을 제안하였다. 영국 및 미국에서의 데이터셋을 사용하여 네 가지 POI에 대해 추정된 경계선 결과를 보였다. 또한, 제안한 방법이 POI 이름을 포함하는 공간 태그된 트윗 입력의 개수에 따라 선형적으로 증가함을 보임으로써, 기존 밀도 기반 군집화 방법인 DBSCAN보다 계산 복잡도 측면에서 우수함을 입증하였다.

References

[1] H.-M. An, S.-K. Lee, K.-S. Sim, I.-H. Kim, S.-H. Jin, and M.-S. Kim, "Big-data traffic analysis for the campus network resource efficiency," *J. KICS*, vol. 40, no. 3, pp. 541-550, Mar. 2015.

[2] J. W. Kim and K.-H. Park, "Personalized group recommendation using collaborative filtering and frequent pattern," *J. KICS*, vol. 41, no. 7 pp. 768-774, Jul. 2016.

[3] Y.-H. Kim, Y.-K. Hwang, T.-G. Kang, and K.-M. Jung, "LSTM language model based Korean sentence generation," *J. KICS*, vol. 41, no. 5, pp. 592-601, May 2016.

[4] Q. Yuan, G. Cong, Z. Ma, A. Sun, and N. Magnenat-Thalmann, "Time-aware point-of-interest recommendation," in *36th Proc. Int. ACM SIGIR'13*, pp. 363-372, Dublin, Ireland, Jul. 2013.

[5] M. Ye, P. Yin, W.-C. Lee, and D.-L. Lee, "Exploiting geographical influence for collaborative point-of-interest recommendation," in *34th Proc. Int. ACM SIGIR'11*, pp. 325-334, Beijing, China, Jul. 2011.

[6] J. J. Levandoski, Sarwat, A. Eldawy, and M. F. Mokbel, "LARS: A location-aware recommender system," in *Proc. IEEE ICDE2012*, pp. 450-461, Washington DC, USA, Apr. 2012.

[7] J.-W. Son, A.-Y. Kim, and S.-B. Park, "A location-based news article recommendation with explicit localized semantic analysis," in *Proc. ACM SIGIR'13*, pp. 293-302, Dublin, Ireland, Jul. 2013.

[8] J. Liu, Z. Huang, L. Chen, H.-T. Shen, and Z. Yan, "Discovering areas of interest with geo-tagged images and check-ins," in *Proc. ACM MM'12*, pp. 589-598, Nara, Japan, Oct. 2012.

[9] S. Kisilevich, F. Mansmann, and D. Keim, "P-DBSCAN: A density based clustering algorithm for exploration and analysis of attractive areas using collections of geo-tagged photos," in *Proc. COM.Geo2010*, Bethesda,

MD, USA, Jun. 2010.

[10] M. Ester, H.-P. Keriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," *Data Mining Knowledge Discovery*, vol. 96, no. 34, pp. 226-231, 1996.

[11] S. V. Canneyt, S. Schockaert, O. V. Laere, and B. Dhoedt, "Detecting places of interest using social media," in *Proc. 2012 IEEE/WIC/ACM WI-IAT'12*, pp. 447-451, Macau, SAR of the People's Republic of China, Dec. 2012.

[12] A. Skovsgaard, D. Sidlauskas, and C. S. Jensen, "A clustering approach to the discovery of points of interest from geo-tagged microblog posts," in *Proc. IEEE MDM'14*, pp. 178-188, Brisbane, Australia, Jul. 2014.

[13] D. D. Vu, H. To, W.-Y. Shin, and C. Shahabi, "GeoSocialBound: An efficient framework for estimating social POI boundaries using spatio-textual information," in *Proc. Int. ACM SIGMOD GeoRich2016*, pp. 1-6, Sanfrancisco, CA, USA, Jun. 2016.

[14] S. Hahmann, R. S. Purves, and D. Burghardt, "Twitter location (sometimes) matters: Exploiting the relationship between georeferenced tweet content and nearby feature class," *J. Spatial Inf. Sci.*, no. 9, pp. 1-36, Sept. 2014.

[15] A. Arampatzis, M. van Kreveld, I. Reinbacher, C. B. Jones, S. Vaid, P. Clough, H. Joho, and M. Sanderson, "Web-based delineation of imprecise regions," *Comp. Environ. Urban Syst.*, vol. 30, no. 4, pp. 436-459, Jul. 2016.

[16] M. Berg, W. Meulemans, and B. Speckman, "Delineating imprecise regions via shortest-path graphs," in *Proc. 19th ACM SIGSPATIAL'11*, pp. 271-280, Chicago, Illinois, USA, Nov. 2011.

[17] Y. Takhteyev, A. Gruzd, and B. Wellman, "Geography of Twitter networks," *Social Netw.*, vol. 34, no. 1, pp. 73-81, Jan. 2012.

[18] J. Julshrestha, F. Kooti, A. Nikravesh, and K. P. Gummadi, "Geographic dissection of the twitter network," in *Proc. ICWSM-12*, pp. 202-209, Dublin, Ireland, Jun. 2012.

[19] W.-Y. Shin, B. C. Singh, J. Cho, and A. M. Everett, "A new understanding of friendships in space: Complex networks meet Twitter," *J. Inf. Sci.*, vol. 41, no. 6, pp. 751-764, Dec. 2015.

[20] A.-C. Lee, G.-E. Seo, W.-Y. Shin, D. Kim, and J. Cho, "Tweet bot detection using geo-location information," in *Proc. KICS Winter Conf.*, pp. 1-2, Jeju Island, Korea, Jun. 2015.

신 원 용 (Won-Yong Shin)

2002년 : 연세대학교 기계전자공학부 학사
2004년 : KAIST 전자전산학과 석사
2008년 : KAIST 전자전산학과 박사
2009년 5월~2011년 10월 : Harvard University Postdoctoral Fellow
2011년 10월~2012년 2월 : Harvard University Research Associate
2012년 3월~현재 : 단국대학교 컴퓨터학과 조교수
<관심분야> 정보이론, 통신이론, 신호처리, 빅데이터분석, 온라인소셜네트워크분석

둥 부 도 (Dung D. Vu)

2009년 : Hanoi University of Science and Technology, Telecommunications & Electronics 학사
2014년 : Hanoi University of Science and Technology, Telecommunications Engineering 석사
2015년 3월~현재 : 단국대학교 컴퓨터학과 박사과정
<관심분야> 빅데이터분석, 데이터마이닝, 온라인소셜네트워크분석