논문 17-42-10-26 The Journal of Korean Institute of Communications and Information Sciences '17-10 Vol.42 No.10
https://doi.org/10.7840/kics.2017.42.10.2046

동시 최적화를 통한 개선된 사회적 관심지점 경계선 추정

신 원 용◆, 최 상 일°

Improved Estimation of Social POI Boundaries through Joint Optimization

Won-Yong Shin◆, Sang-Il Choi°

요 약

최근에 공간-텍스트 정보를 사용하여 사회적 관심지점 (POI: point-of-interest) 경계선을 추정하는 효율적인 알고리즘인 GeoSocialBound가 소개되었는데, 이는 지리적 마케팅에 유용하게 쓰일 수 있다. 본 레터에서는 추정 성능을 더 개선하기 위해 POI 중심을 업데이트하는 것을 허용함으로써 POI 중심과 원의 반지름을 찾는 동시 최적화 문제를 제안한다. 또한, 이를 해결하는 개선된 알고리즘을 보이고 F-measure와 이의 변화된 메트릭 측점에서 우수함을 보인다.

Key Words : F-measure, Geo-tagged tweet, Point-of-interest (POI), Social POI boundary, Spatio-textual information

ABSTRACT

Recently, an efficient algorithm, named GeoSocialBound, for estimating a social point-of-interest (POI) boundary using spatio-textual information was introduced, which is very useful in geomarketing. In this letter, to further improve the estimation performance, we propose a joint optimization problem of finding both the POI center and the radius of a circle by allowing to update the POI center. An improved algorithm to solve this problem is shown to outperform the baseline in terms of F-measure and its variant.

Ⅰ. 서 론

사용자들이 관심지점 (POI: point-of-interest)을 방문할 때, 그들의 위치가 POI와 관계되어 있다는 점을 표현하기 위해 온라인 체크인을 하거나 공간 기반 소셜 네트워크 (LBSN: location-based social network)를 통해 방문지의 사진을 올리는 경향이 있다. POI에 대한 데이터의 연관성은 POI와 데이터가 생성된 위치 사이의 공간 거리에 따라 변화하기 때문에, 관심 영역을 찾아서 다양한 상업적 광고에 위치 정보를 활용하는 연구는 중요하다고 할 수 있다. LBSN으로부터 수집된 공간 태그된 사진 대신에, 본 연구에서는 방대한 양의 사용자 계정과 레코드를 보유한 가장 인기 있는 마이크로블로그 중 하나인 트위터에서의 공간 태그된 트윗[1-3]을 활용한다. POI 경계선 (또는 관심영역)을 찾기 위해서는 가장 일반적으로 사용되어 온 밀도 기반 군집화 알고리즘인 DBSCAN (density-based spatial clustering of application with noise)[4]을 적용할 수 있지만, 이는 공간 좌표만을 이용할 뿐 소셜 네트워크에서의 텍스트 정보를 활용하지는 못한다. 반면, 최근에 트위터에서의 공간-텍스트 정보에 기반하여 사회적 POI 경계선을 추정하는 연구[5]가 수행되었다. 기존 연구[5]에서는 주어진 POI 중심이 원점인 원의 최적 반지름 r을 찾음으로써 POI 경계선을 추정하는 알고리즘인 GeoSocialBound가 소개되었는데, r에 대한 멱 법칙과 F-measure의 곱으로 표현되는 목적함수가 최대가 되는 r이 해가 된다.

본 레터에서는, POI 중심을 업데이트하는 것을 허용함으로써 기존 연구[3]에서의 추정 성능을 더욱 개선하는 방법을 제안한다. 구체적으로, 목적 함수가 최대가 되는 반지름 r과 업데이트된 POI 중심 c를 찾는 동시 최적화 문제 및 이를 해결하는 효율적인 알고리즘인 반복적 GeoSocialBound (I-GeoSocialBound)를 소개한다.

※ 본 연구는 2017년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업(2015R1A2A1A15054248)이며, 과학기술정보통신부 및 정보통신기술진흥센터의 SW중심대학지원사업(no. 2017-0-00091)의 연구결과로 수행되었고, 2017학년도 단국대학교 대학연구비 지원으로 연구되었음.

◆ First Author : Department of Computer Science and Engineering, Dankook University, wyshin@dankook.ac.kr, 정회원
° Corresponding Author : Department of Computer Science and Engineering, Dankook University, choisi@dankook.ac.kr, 정회원

논문번호 : KICS2017-09-252, Received September 18, 2017; Revised September 23, 2017; Accepted September 25, 2017

Ⅱ. 트위터 데이터 수집

본 장에서는 POI 및 해당 위치에 관련한 트위터를 수집하는 방법을 설명한다. 먼저, POI 집합과 해당 POI 중심을 얻기 위해 많은 양의 공간 정보를 가진 오픈소스 데이터베이스인 Geonames를 사용한다. 빌딩, 농장, 길, 언덕 등과 같이 데이터베이스 안에 다양한 공간 특성 클래스가 존재하지만, 간단히 그 중에서 운동장 (S.STDM)과 기념물 (S.MNMT)에 해당하는 Dodger Stadium과 Hollywood Sign을 고려한다 (두 가지 POI는 기존 연구 결과[5] 대비 사회적 POI 경계선 추정 성능 차이를 나타내는 경우와 그렇지 않은 경우를 보여주기 위해 대표로 선택한 것이며 추정 결과는 IV장에서 보인다).

다음으로, 트위터 데이터를 수집하기 위해 Twitter Streaming API에서 수집한 데이터셋을 사용한다. 데이터셋은 미국에서 2015년 7월 29일부터 2015년 8월 29일까지 대략 한 달간 트위터 사용자들로부터 기록된 방대한 양의 공간 태그된 트윗을 포함한다. 각 트윗은 수 많은 귀속된 필드를 갖지만, 본 연구에서의 데이터 분석을 위해 아래의 세 가지 필수 필드만을 채택한다.

- *text*: 상태 업데이트의 실제 UTF-8 텍스트
- *lat*: 트윗 위치의 위도
- *lon*: 트윗 위치의 경도

사용자들은 POI에 대한 그들의 관심도를 표현하기 위해 트윗 안에 POI 이름을 태그하거나 이름을 삽입하는 경향이 있기 때문에, POI 이름을 포함하는 트윗과 같이 모든 관계된 레코드를 쿼리할 수 있다. 따라서, 쿼리 처리를 통해 text 필드가 POI 이름과 관계된 공간 태그만 필터링을 수행한다. 이를 통해 데이터셋은 POI 이름을 가진 공간 태그된 트윗 (즉, 관련 레코드)과 POI 이름이 없는 공간 태그된 트윗 (즉, 무관 레코드) 이렇게 두 개의 부분 집합으로 나뉜다.

Ⅲ. 개선된 추정 방법

본 장에서는 데이터 분석을 위해 사용되는 중요한 용어를 정의한 후 새로운 동시 최적화 문제를 소개한다. 그리고, 제안한 I-GeoSocialBound를 설명한다.

3.1 정의

사회적 POI 경계선을 규명하기 위해 POI 중심 및 POI 중심으로부터 도달할 수 있는 최대 가능한 거리를 찾는 것에 초점을 맞춘다.

정의 1: POI 중심 c와 특정 반지름 $r>0$을 갖는 원 (c,r)이 주어질 때, 원 (c,r) 안에 POI 이름을 포함한 모든 공간 태그된 트윗 (관련 레코드)의 집합을 $D(c,r)$라 하자. 이 때, 사회적 POI 경계선은 $D(c,r)$ 안의 주어진 점들을 포함하는 convex 다각형으로 정의된다.

정의 2: POI 이름이 주어질 때, 원 (c,r) 안에서의 추정 질은 아래와 같이 정의된다.

$$\mathrm{Precision}(c,r)=\frac{|D(c,r)|}{|D_{all}(c,r)|} \tag{1}$$

여기에서, $\mathrm{Precision}(c,r)$은 모든 predictive positives에 대한 true positives의 비율이고, $D_{all}(c,r)$은 원 (c,r) 안에서의 모든 공간 태그된 트윗의 집합을 나타낸다. 이 때, predictive positives는 true positives와 false positives의 합으로 표현되는데, true positives는 원 (c,r) 안의 관련 레코드의 개수인 $|D(c,r)|$이고 false positive는 동일 안 안의 무관 레코드의 개수인 $|D_{all}(c,r)\backslash D(c,r)|$이다. 이 경우 predictive positives는 $|D_{all}(c,r)|$에 해당한다.

정의 3: 원 (c,r)이 주어질 때, F-measure $F(c,r)$는 다음과 같다.

$$F(c,r)=\frac{2\mathrm{Precision}(c,r)Recall(c,r)}{\mathrm{Precision}(c,r)+Recall(c,r)}$$

이 때, $F(c,r)$는 $\mathrm{Precision}(c,r)$과 $Recall(c,r)$의 조화 평균을 의미하고, $Recall(c,r)$은 모든 actual positives에 대한 true positives의 비율이다.

3.2 새로운 최적화 문제

기존 연구[5]에서는 F-measure $F(c,r)$이 POI 중심으로부터의 반지름 r에 관해 특정 임계점까지 거의 감소하지 않음을 관찰하였다. 이러한 이유로, 가장 높은 F-measure를 갖는 사회적 POI 경계선을 찾는 것이 좋음에도 불구하고, 본 연구의 목적이 사회적 POI 경계선 추정 정확도를 거의 훼손시키지 않으면서 최대한 넓은 공간 영역을 포함하는 경계선을 찾는 것이라면 약간의 감소된 $F(c,r)$ 값과 함께 사회적 POI 경계선을 더욱 확장하는 것 또한 좋을 것이다. 본 연구에서는 다음과 같이 목적 함수 $r^{\alpha}F(c,r)$가 최대화되는 원의 반지름 r과 업데이트된 POI 중심 c를 찾는 새로운 동시 최적화 문제를 한다.

$$(c^*,r^*) = \arg\max_{c,r} r^{\alpha} F(c,r) \tag{2a}$$

$$\text{subject to } 0 < d(c,c^{(0)}) + r \le \bar{r} \tag{2b}$$

$$\Pr ecision(c,r) \ge \eta \tag{2c}$$

여기에서, $\Pr ecision(c,r)$과 $\eta \in (0,1)$은 (1)에서 원 (c,r) 안에서의 추정 질 및 목표 질 임계값을 나타낸다. 또한, $\alpha \ge 0$는 공간 범위의 다른 레벨 사이에 균형을 맞추는 반지름 지수, $d(c,c^{(0)})$는 좌표 c와 $c^{(0)}$ 사이의 공간 거리, $c^{(0)}$는 Geonames에 의해 제공된 초기 POI 중심, $\bar{r}$는 $\Pr ecision(c,r) \simeq 0$을 만족하는 임의의 큰 반지름을 나타낸다. (2b)는 변수 (c,r)에 대한 공간 검색 영역을 초기 POI 중심인 $c^{(0)}$로부터 최대 반지름 $\bar{r}$로 제한하기 위함이고, (2c)는 원 (c,r) 안에 포함된 모든 공간 태그 수 대비 POI 이름을 포함한 관련 레코드의 최소 수 비율을 보장하기 위함이다. $\alpha = 0$일 때에는 특수한 경우로 제약 조건 (2b)와 (2c) 하에서 $F(c,r)$를 최대화하는 문제로 해석될 수 있다.

3.3 I-GeoSocialBound

본 절에서는 (2)에서의 제한된 최적화 문제를 푸는 I-GeoSocialBound 알고리즘을 소개한다. 제안한 알고리즘은 반복적으로 수행하도록 설계되는데, 전체적인 과정은 표 1에서 보여 진다. 각 반복 단계에서, 함수 GeoSocialBound$(D(c^{(k)},\bar{r}), D_{all}(c^{(k)},\bar{r}))$가 호출된다. 이 함수는 기존 연구[5]에서 소개된 사회적 POI 경계선 추정 알고리즘으로써 주어진 POI 중심 $c^{(k)}$에 대해 최적의 반지름 $r^{(k)}$를 찾는데 사용 되며, 출력으로 $r^{(k)}$, $D(c^{(k)},r^{(k)})$, $F(c^{(k)},r^{(k)})$를 반환한다. 표 1의 4번째 줄에서, p_l은 $l^{th} \in D(c^{(k-1)},r^{(k-1)})$가 되는 그러한 l번째 공간 태그의 좌표이다. 알고리즘은 6번째 줄에서의 종료 기준 하에서 동작하며, $\epsilon > 0$ 은 허용 오차를 나타낸다.

표 1. 제안 알고리즘 (I-GeoSocialBound)
Table 1. Proposed algorithm (I-GeoSocialBound)

Input: $D_{all}(c^{(0)},r)$, $D(c^{(0)},r)$, $\triangle r$, η, α, ϵ, $c^{(0)}$, $\bar{r}$
Output: $D(c^*,r^*)$
Initialization: $k \leftarrow 0$; $r^* \leftarrow 0$; $c^* \leftarrow c^{(0)}$; $L^{(0)} \leftarrow 0$; $D(c^*,r^*) \leftarrow \varnothing$; $(r^{(0)}, D(c^{(0)},r^{(0)}), F(c^{(0)},r^{(0)})) \leftarrow$ GeoSocialBound$(D(c^{(0)},\bar{r}), D_{all}(c^{(0)},\bar{r}))$
1: do
2: $k \leftarrow k+1$
3: $L^{(k)} \leftarrow |D(c^{(k-1)},r^{(k-1)})|$
4: $c^{(k)} \leftarrow \dfrac{\sum_{l=1}^{L^{(k)}} p_l}{L^{(k)}}$
5: $(r^{(k)}, D(c^{(k)},r^{(k)}), F(c^{(k)},r^{(k)})) \leftarrow$ GeoSocialBound$(D(c^{(k)},\bar{r}), D_{all}(c^{(k)},\bar{r}))$
6: while $|r^{(k)\alpha} F(c^{(k)},r^{(k)}) - r^{(k-1)\alpha} F(c^{(k-1)},r^{(k-1)})| \ge \epsilon$
7: $c^* \leftarrow c^{(k)}$
8: $r^* \leftarrow r^{(k)}$
9: return $D(c^*,r^*)$

Ⅳ. 추정 결과

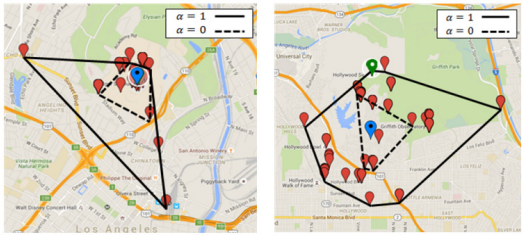
본 장에서는 Ⅲ장에서 소개한 I-GeoSocialBound 알고리즘의 추정 결과를 보인다. 실험을 위해 간단히 $\eta = 0.5$, $\epsilon = 10^{-4}$를 가정하였는데, 이 값들은 추정 질과 수렴 속도를 조절하기 위해 다른 값으로 변경될 수 있다. 또한, 모든 POI에 대해 $\triangle r = 10$m를 가정한다. 또한, $\bar{r} = \gamma r_{cover}$ 를 가정하였는데, 이 때 γ는 양수이고 r_{cover}는 각 POI의 공간 영역을 덮은 POI 중심으로부터의 근사화된 반지름인데, Google Maps Geocoding API로부터 얻어진다. γ는 POI 유형에 따라 다르게 설정될 수 있는데, 본 실험에서는 모든 POI에 대해 $\gamma = 10$을 가정한다.

먼저, I-GeoSocialBound 알고리즘의 계산 복잡도에 대해 설명한다. 표 1에서 보인 제안 알고리즘의 반복 횟수는 두 가지 POI에 대해 모두 2회로 주어짐을 확인하였다. 따라서 기존 GeoSocialBound 알고리즘은 $D_{all}(c,r)$ 크기에 선형적으로 증가한다[5]는 사실로부터, I-GeoSocialBound 알고리즘의 계산 복잡도도 역시 $D_{all}(c,r)$ 크기에 비례하며 기껏해야 GeoSocialBound 알고리즘의 두 배 정도 복잡도를 가짐을 알 수 있다.

표 2. 추정 성능
Table 2. The estimation performance

POI name	α	$r^{*\alpha} F(c^*,r^*)$	Improvement ratio
Dodger Stadium (S.STDM)	0	0.912	0.66%
	0.5	32.977	0.77%
	1	1268.55	0.08%
Hollywood Sign (S.MNMT)	0	0.833	1.46%
	0.5	37.735	8.66%
	1	1,553.4	0.48%

$\alpha \in \{0, 0.5, 1\}$일 때 두 가지 POI에 대한 추정 성능은 표 2에서 요약되는데, 표의 마지막 열은 기존 연구[5]에서 보인 기존 GeoSocialBound 알고리즘 대비 개선율을 나타낸다. 표 2에서는 $\alpha \in \{0, 0.5, 1\}$에 대해 제안 알고리즘의 최적화된 목적 함수 (또는 성능 지표)인 $r^{*\alpha}F(c^*, r^*)$ 값 및 기존 GeoSocialBound 알고리즘 대비 $r^{*\alpha}F(c^*, r^*)$ 값의 개선율을 보인다. 표로부터 α가 증가함에 따라 알고리즘은 추정 공간 영역을 확장하는 것에 초점을 맞추어 결과를 제공하는 경향이 있기 때문에 $r^{*\alpha}F(c^*, r^*)$는 증가함을 알 수 있다. 제안한 기술을 더 잘 이해하기 위해, 그림 1에서 추정된 사회적 POI 경계선을 도시하였는데, 붉은색 핀은 POI 이름을 가진 공간 태그된 관련 레코드, 녹색 핀은 초기 POI 중심, 푸른색 핀은 업데이트된 POI 중심을 나타낸다. 표 2와 그림 1로부터 다음과 같은 흥미 있는 사실을 관찰할 수 있다. 업데이트된 POI 중심이 초기 중심으로부터 멀리 떨어져 있을 때 (예: Hollywood Sign), POI 중심의 업데이트를 통해 기존 GeoSocialBound 대비 I-GeoSocialBound의 추정 성능은 8.66%까지 증가하게 된다. 반면, 두 개의 POI 중심이 매우 가까울 경우 (예: Dodger Stadium), 성능 개선은 미미하게 된다.



(a) Dodger Stadium (b) Hollywood Sign

그림 1. 추정된 사회적 POI 경계선
Fig. 1. Estimated social POI boundaries

V. 결 론

본 논문에서는 POI 중심을 업데이트하는 것을 허용함으로써 기존 GeoSocialBound 알고리즘의 성능을 개선한 I-GeoSocialBound 알고리즘을 소개하였다. 또한 트위터 데이터셋을 통해 제안 알고리즘의 사회적 POI 경계선 추정 성능의 우위를 검증하였다.

References

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

[2] W.-Y. Shin and D. D. Vu, "Density-based estimation of POI boundaries using geo-tagged tweets," *J. KICS*, vol. 42, no. 2, pp. 453-459, Feb. 2017.

[3] 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.

[4] 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. 4, pp. 226-231, 1996.

[5] D. D. Vu, H. To, W.-Y. Shin, and C. Shahabi, "GeoSocialBound: An efficient framework for estimating social POI boundaries using spatio-temporal information," in *Proc. ACM SIGMOD Workshop GeoRich'16*, pp. 1-6, San Francisco, CA USA, Jun. 2016.