6P-05

ジオタグ付ツイートの多言語相関性に基づく POI 推薦システムの提案

井上憲汰朗[†1]　神原亮弥[†1]　王元元[†2]　*Panote Siriaraya*[†1]　河合由起子[†1,3]　下條真司[†3]
†1 京都産業大学　†2 山口大学　†3 大阪大学

1　はじめに

近年，ユーザの行動分析および可視化に関する研究において，ソーシャルネットワーク上のジオタグ付き SNS データ分析に関する研究開発が盛んに行われている [1][2]．これまで我々も，ユーザ行動分析としてデータ発生位置とコンテンツで言及されている位置との差異，発生時間とコンテンツ言及時間との差異分析 [3]，さらにユーザプロファイルの言語および言及言語を考慮したユーザ特性（国民性）抽出に関する研究を行ってきた [4][5]．これにより，時空間およびユーザの言語に基づき多くの店舗（レストラン）の中から，旅先の異国でも各国民にあった店舗の推薦が可能になった．しかしながら，レストラン等の店舗と比較して少数となる観光名所（POI）ではその偏りは大きく，言語ごとの訪問数（ツイート数）の少ない POI に対して有効でないという問題があった．そこで本研究では，母国語と異なる異国の場所ごとで各言語に POI に対する人気度を算出し，各言語の POI に対する嗜好性として抽出し，それら嗜好性と他言語との相関性を算出することで，訪問数の少ない POI 集合に対しても嗜好性にあった POI の推薦を目指す．これにより，例えば，ニューヨークでの「Chelsea Market」などのイタリア語の少ない POI に対して，イタリア語と相関性の高いフランス語により新たに POI として推薦可能となる．さらに，抽出した言語の相関性から，例えばフランス語の少ない別の都市となるシアトルにおいて，「PiKE Place Market」の未知の POI 推薦が可能となる．

本論文では，米国の都市おける複数の欧州言語を対象に各言語間の相関性を算出し，相関性に基づき POI をランキングする．また，提案手法より抽出した POI のうちツイート数が少ない POI および全言語での評判は高くないマイナーな POI に対する評価実験を行い，有効性を検証する．

2　位置と言語分析に基づく POI 推薦

本章では，任意の場所における言語特性の抽出ならびに言語の相関性に基づく POI 推薦手法について述べる．

POI 推薦システムの処理の概要は，まず取得したツイートから POI 名を抽出し，次に，発信位置（国）ごとに同一の言語（国）のツイートを分類し，POI ごとに出現頻度（TF）を算出し，各言語国間の相関係数を類似度として算出し，最後にユーザ指定の地域内のツイートの POI の出現頻度をツイートから算出し，値の高い POI をマップ上に提示する．

表 1: 各言語のジオタグ付ツイート数と New York と Boston 上位 30 件の POI に対するツイート総数.

Language	#Tweets	#I'm @ (%)	NY	Boston
Italian	390,957	18,253 (4.7%)	827	185
French	906,572	36,938(4.1%)	1,095	57
Spanish	4,019,581	274,974(6.8%)	11,652	1,053
Portuguese	642,671	26,423 (4.1%)	3,065	331
Total	3,456,608	282,167 (8.2%)	16,639	1,626

2.1　発信場所と言語に基づく POI 抽出

まず，ジオタグツイートの発信位置，発信時刻，母国語および言及言語を抽出し，任意の期間と地域と言語に基づきツイートを分類する．ここで母国語とは，ユーザがツイート利用登録時に設定する言語とし，言及言語はツイートの内容に用いられている言語とする．この母国語と言及言語より，任意の言語 l は {母国語 $_l$} ∨（言及言語 $_l \subseteq$ 母国語$_l$）として分類される．

次に，分類された言語ごとの POI 辞書を作成する．POI 辞書は，言語，緯度経度，地物名のタプルであり，ツイートの定式文となる “I'm at” とマッチングしたツイートの定式文以降に記載される単語を地物名（POI）として抽出する．

各言語の POI 辞書に基づき，全言語 L に対して任意の言語 l_x の都市 p でのみ発信された各 POI j に対する嗜好性となる評価値を，出現頻度 $TF_{\{x,j\}}$=（l_x における POI j 出現回数）／（l_x における POI 総出現回数）から算出する．算出した言語 l_x の POIj に対する評価値 $TF_{\{x,j\}}$ と他言語 l_y の評価値 $TF_{\{y,j\}}$ より，x 国と他国 y 間の類似度 $sim(x,y)$ を相関係数より算出する．

最後に，任意の地域 p の POI を含むツイートを取得し，ツイート数が閾値以上の場合は下記 TFiDF よりランキングした POIj を抽出する．

$$\frac{l_y \text{ 言語の POI } j \text{ の出現回数}}{l_y \text{ 言語における POI 総数}} \cdot \log \frac{\text{言語総数 } L}{\text{POI } j \text{ の出現言語数}}$$

2.2　ツイート数の少ない都市における各言語との類似性に基づいた POI 抽出

地域 p におけるツイート数が閾値未満の場合は，言語 l_x にとっては訪問頻度の少ない地域となる．本手法は，各都市における他言語との POI の類似性を考慮することで，他言語の都市 p における POI j に対する評

A POI Recommender system based on Multilingual Analysis of Geotagged Tweets
†1 Kentaro INOUE　†1 Ryoya KANBARA　†2 Yuanyuan WANG
†1 Panote SIRIARAYA　†1,3 Yukiko KAWAI　†3 Shinji SHIMOJO
†1 Kyoto Sangyo University
†2 Yamaguchi University
†3 Osaka University

表 2: 言語 l_x の New York の POI に基づいた類似度

l_x	FR	ES	IT	PT	Avg.
FR (French)	1	**0.77**	0.74	0.71	0.74
ES (Spanish)	0.77	1	0.86	**0.92**	0.85
IT (Italian)	0.74	0.86	1	**0.87**	0.82
PT (Portuguese)	0.71	**0.92**	0.87	1	0.83
Average	0.74	0.85	0.82	0.83	0.81

表 3: 言語 l_x の Boston の POI に基づいた類似度

l_x	FR	ES	IT	PT	Avg.
FR (French)	1	**0.46**	-0.03	0.16	0.20
ES (Spanish)	**0.46**	1	0.01	0.44	0.30
IT (Italian)	-0.03	**0.01**	1	-0.09	-0.04
PT (Portuguese)	0.16	**0.44**	-0.09	1	0.17
Average	0.20	0.30	-0.04	0.17	0.16

価値 $TF_{\{x,j\}}$ を言語間の類似度 $sim(x,y)$ を用いて下記の式（1）より言語 l_x の POIj に対する評価値を抽出する．

$$\sum^{D}\Big(sim(x,y)\ \cdot\ TF_{\{x,j\}}\Big)\Big/\sum^{D}TF_{\{x,j\}} \qquad (1)$$

D は言語数であり，場所 p における言語 l_x の POIj に対する評価値が算出される．

3 実験

本稿において，2016 年 11 月 12 日から 2019 年 1 月 7 日の約 2 年分の欧州領域のツイートを対象に，4 言語を対象とした POI 推薦システムを構築し，米国の New York と Boston における POI 抽出結果について検証する．表 1 に New York と Boston における分類した 6 言語のツイート数，“I’m at” 数，POI 総数を示す．

3.1 各言語における POI の多様性検証

提案手法より New York において算出した各言語の言語間の類似性を表 2 に示す．表の太字は l_x に対して他言語で最も類似度が高い結果を示している．表より，最も高い類似度はスペイン語とポルトガル語の 0.92 となり，最も低いのはフランス語とポルトガル語の 0.71 であった．また全体ではスペイン語の平均が 0.85 と他言語との相関性が高く，New York では全ての言語間で高い相関性がある結果となった．

次に，表 3 に Boston における相関結果を示す．最も高い類似度はフランス語とスペイン語の 0.46 であったが，高い相関性ではなかった．また，フランス語およびスペイン語の平均は低い相関となり，それ以外の言語は相関がない結果となった．

3.2 各言語の POI 抽出のユーザ評価による検証

前節の結果に基づき，Boston で低いが相関のあるフランス語とスペイン語のうち，ボストンでの POI に対するツイート数が最小のフランス語を対象とし，フランス在住のフランス語のネイティブスピーカーのうち New York と Boston の各々に訪問経験のある 16 人と 11 人に，POI の評価を行ってもらい，有効性を検証した．評価方法は，対象都市において抽出した POI 上位 30 件（Boston は 27 件）のうち下位 10 件に対してランキングによる評価を行ってもらった．ベースラインは全言語対象ではあるが一般的な推薦サービスである Foursquare の rating を用いたランキングとし，ユーザ評価とのスピアマン順位相関より検証した．また，類似度（Sim）は，表 2 の New York のみを用いた場合（NY）と New York と Boston の両方を用いた場合（Both）とした．

表 4: 推薦された POI に対するフランス人の評価結果

City	Sim	Rating average	Proposed average	gain(%)
New York	NY	-0.22	0.22	+44.2%
New York	Both	0.79	0.83	+3.6%
Boston	Both	0.00	0.55	**+55.0%**
Average	—	0.54	0.14	+39.5%

表 4 に，スピアマン順位相関による評価結果を示す．全結果において，rating による推薦と比べて提案手法による推薦が良好であった．また，フランス語のツイートの少ない Boston では rating の評価結果より 55% と最も高い向上が見られた．以上より，提案する言語相関に基づく POI 推薦手法の有効性が確認できた．

4 おわりに

本論文では，各国民の嗜好性の解明を目指し，場所と言語情報に着目し，訪問先の都市においてツイートの少ない POI でも言語相関性から言語ごとに嗜好性の高い POI 抽出手法を提案した．実験よりフランス語のツイートの少ないボストンでも提案手法の推薦によりベースライン（rating）より 55% もの向上が見られた．今後，対象都市と言語国数を拡大した評価を行う．

謝辞

本研究の一部は，JSPS 科研費 16H01722，17H01822，17K12686 および Society 5.0 実現化研究拠点支援事業の助成を受けたものである．ここに記して謝意を表す．

参考文献

[1] T. Hu, et. al.: Mining Shopping Patterns for Divergent Urban Regions by Incorporating Mobility Data, Proc. of CIKM2016, pp. 569-578 (2016).

[2] Chen, S. et. al.: Social Context Awareness from Taxi Traces: Mining How Human Mobility Patterns Are Shaped by Bags of POI, Adjunct Proc. of UbiComp/ISWC’15 Adjunct, pp. 97-100 (2015).

[3] É. Antoine, A. Jatowt, S. Wakamiya, Y. Kawai, T. Akiyama: Portraying Collective Spatial Attention in Twitter, Proc. of KDD2015, pp. 39-48 (2015).

[4] P. Siriaraya, Y. Nakaoka, Y. Wang, Y. Kawai: A Food Venue Recommender System Based on Multilingual Geo-Tagged Tweet Analysis. Proc. of IEEE/ACM2018 ASONAM, pp.686-689（2018）.

[5] 中岡 佑輔，パノット シリアーラヤ，王 元元，河合 由起子，秋山 豊和，ジオタグツイートの多言語性と評判に基づく Venue 推薦，WebDB2018，Vol. 2018-IFAT-132, No. 5, pp.1-6（2018）.