2N-03

ジオタグツイートの言語相関性分析による観光スポット推薦手法の検討

豊島亮真[†1]　阪本翔太[†1]　*Panote Siriaraya*[†2]　王元元[†3]　河合由起子[†1,4]
†1 京都産業大学　†2 京都工業繊維大学　†3 山口大学　†4 大阪大学

1 はじめに

携帯端末から発信されるデータには位置情報となるジオタグが付与されることが一般的となっており，ジオタグ付き SNS データを分析し，ユーザ行動に基づいたスポット推薦に関する研究が活発に行われている．Chen ら [2] のタクシーに設置した GPS から取得した人々の移動パターンと地域に存在する POI のカテゴリ情報を用いることで，地域の機能性を発見する手法の有用性の実証や，小原ら [1] の SNS データを分析し，都市の人気スポットを名前とともに抽出する手法が提案されている．これまで我々も大量のジオタグツイートから時空間の差異を分析することで特徴を抽出し，可視化することを行ってきており，さらに発信されるデータの言語による分析を行ってきた [3] が，発信位置と言語との相関性の相違検証は十分にされていなかった．

本研究では，ジオタグツイートから時空間情報ならびにツイートの言語情報に基づき，各言語・各地域のスポットに対する嗜好性（TF 値）を抽出し，さらにそれら TF 値から各言語間の類似度を算出し，類似度から地域間の相関性を検証する．また，相関性より地域での各言語のスポットに対する評価値を算出するスポット推薦手法を検討する．本稿では，対象地域を京都，東京等の 5 地域，対象言語を 12 言語とし，日本語を対象に各地域および全地域で抽出した類似度と相関性を検証する．さらに各地域ごとの相関性に基づくスポット抽出手法の検証をする．日本語ユーザによる各スポットに対する評価値を定性的評価値とし，ツイート数によるスポット評価値，google の rating におけるスポット評価値ならびに提案手法である相関性に基づくスポット抽出手法から，nDCG により本手法を検証する．

2 位置と言語分析に基づくスポット推薦

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スポット推薦手法について述べる．提案手法の概要は，まず緯度経度に基づき指定した矩形内の取得したツイートを言語ごとに分類し，次に各スポットの半径 nm 内ごとに分類しスポットごとの出現頻度（TF 値）を算出する．算出した TF 値から言語 x と他言語 y 間のコサイン類似度および相関性を求め，最後に TF 値，類似度，相関性から言語 x の各スポットに対する評価値を算出する．

2.1 言語間の類似度算出

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 $_l\} \vee$（言及言語 $_l \subseteq \overline{\text{母国語}_l}$）として分類される．

次に，各言語ごとに分類されたツイートをスポットの中心座標に基づき，半径 nm 内のツイートを抽出する．抽出したツイートより，任意の地域 p で発信された全言語 L に対して任意の言語 l_x の各スポット s に対する出現頻度 $TF_{\{x,s\}}$ を，（l_x におけるスポット s のツイート数）／（スポット s における全言語 L のツイート総数）として算出する．最後に，算出した言語 l_x のスポット s に対する $TF_{\{x,s\}}$ と他言語 l_y の $TF_{\{y,s\}}$ より，x 国と他国 y 間のコサイン類似度 $sim(x,y)$ を算出する．

2.2 地域よび言語間の相関性に基づくスポット抽出

各地域ごとに算出した他言語との類似度に基づき，地域 p におけるスポット s に対する言語 l_x の評価値を記の式（1）より算出する．

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

$TF_{\{x,s\}}$ は言語間の類似度 $sim(x,y)$，D は言語数である．ここで，任意の言語における他都市との相関性を算出することで，相関性が閾値以上の他都市の類似度 $sim_j(x,y)$ を用いて，平均値を $sim_{AVG}(x,y)$ を算出し，式（1）の $sim(x,y)$ として用いる．

以上より，例えば，京都におけるドイツ語を発信するユーザに対して各スポットの評価値算出では，京都と九州，京都と東北のドイツ語の類似度の相関係数を算出し，京都と東北との相関係数が閾値以上だった場合は，京都の類似度だけでなく京都と東北の類似度を用いてスポットに対する評価値が算出される．

3 実験

地域と言語に基づく相関分析によるスポット推薦手法を検証する．実験では，日本語を対象に各地域および全地域で抽出した類似度を算出し，日本語と他言語との相関を考察する．また，東京と他地域の類似度を用いたスポット推薦をユーザ評価値を Baseline として，google の rating におけるスポット評価値と比較検証する．

A Method of spot recommendation for tourism based on language correlation analysis by geo-tagging tweets
†1 Ryoma TOYOSHIMA　†1 Shota SAKAMOTO　†2 Panote SIRIARAYA　†3 Yuanyuan WANG　†1,4 Yukiko KAWAI
†1 Kyoto Sangyo University
†2 Kyoto Institute Of Technology
†3 Yamaguchi University
†4 Osaka University

表 1: 5 都市の 10 スポットに対するツイート数

言語	東京	名古屋	京都	大阪	福岡	合計
de	45	27	436	**361**	117	**986**
es	283	217	3,510	2,576	367	6,953
fr	53	102	1,102	882	149	2,288
id	**6**	93	**346**	789	**102**	1,336
it	24	**23**	1,058	559	136	1,800
ko	435	231	1,155	2,488	2,833	7,142
pt	13	322	589	653	104	1,681
th	399	251	1,821	4,588	355	7,414
zh	440	424	1,847	2,765	390	5,866
en	199	69	61,849	533	180	62,830
ja[2]	5,879	48,319	115,555	92,714	414,526	676,993
Total	7,776	50,078	189,268	108,908	419,259	775,289

実験では，2016 年 6 月 22 日から 2019 年 12 月 26 日の約 3 年半分のツイートのうち，提案手法より分類した 12 言語を対象に，東京，名古屋，京都，大阪，福岡[1]の 5 都市における主要な 10 スポット推薦のランキング結果を用いた．なお，各都市の主要な 10 スポットは博物館や寺社仏閣など 7 カテゴリに分類し選定した. 表 1 に 5 都市における 12 言語のツイート数を示す．最小数となる言語は下線太字，最大数となる言語は下線で示しており，英語を除いて最大総数はタイ語（th），最小総数はドイツ語（de）であった．なお，英語に関しては，他言語ユーザの多くが英語を用いてツイートしているため，今回は検証から除いた．

3.1 言語間の相関性検証

提案手法より抽出した類似度のうち，日本語に対する各都市の各言語に対する類似度を表 2 に示す．最小値となる言語は下線太字，最大値となる言語は下線で示しており，5 都市の平均で最も類似していた言語は韓国語（ko）の 0.88 となり，最も低かったのは表 2 のツイート総数と同じくドイツ語（de）の 0.64 であった．都市ごとに類似度とツイート数の最大値の言語を比較すると，東京ではツイート数が最も多いのが中国語だったの対し，類似度では韓国語が 0.95 と最大となり，逆に中国語は類似度は最小となった．また，他の全ての都市でも類似度とツイート数の最大となる言語は異なる結果となった．

3.2 地域間の相関性に基づくスポット推薦の検証

前節の類似度を用いて，東京における日本語に対するスポットの評価値を算出し検証した．実験では，京都の大学生 12 人が 12 スポットのうち未訪問のスポットに対して 5 段階のリッカート尺度で評価した平均を正解データとした．Baseline は google と foursquare の rating の平均値を用い，ユーザ評価との nDCG およびスピアマン順位相関より比較検証した結果，nDCG は 0.85 となり，スピアマン相関係数は 0.10 となった．

[1]今回はスポット数を同一にするため北九州も含む
[2]日本語は I'm at のみのツイート数とした

表 2: 日本語と他言語のスポットに基づいた類似度

l_y	東京	名古屋	京都	大阪	福岡	Avg.
de	0.61	0.89	0.56	**0.20**	0.92	**0.64**
es	0.68	0.78	0.40	0.84	0.92	0.72
fr	0.90	0.92	0.37	0.84	0.88	0.78
id	**0.48**	0.92	0.61	0.78	**0.76**	0.71
it	0.60	**0.57**	0.44	0.81	0.98	0.68
ko	0.95	0.93	0.75	0.87	0.89	0.88
pt	0.50	0.95	**0.32**	0.86	0.81	0.69
th	0.64	0.74	0.36	0.84	0.93	0.70
zh	**0.48**	0.88	0.65	0.89	0.87	0.75
Avg.	0.65	0.84	0.49	0.77	**0.88**	-

表 3: 東京における推薦スポットに対する nDCG

City	Speaman	gain(%)	nDCG	gain(%)
東京のみ	0.297	+6.29%	0.906	+16.7%
&名古屋	**0.311**	+7.69%	**0.908**	+17.0%
&京都	**0.311**	+7.69%	**0.908**	+17.0%
&大阪	0.262	+2.80%	0.906	+16.8%
&福岡	0.262	+2.80%	0.906	+16.8%
Average	0.289	+5.45%	0.907	+16.8%

表 3 に，評価結果を示す．表より，提案手法のうち，東京と名古屋，東京と京都の言語ごとの類似度を用いた結果が nDCG およびスピアマン相関係数の両方において最も良好な結果となった．また，Baseline より平均で 5.45%の向上が見られた．以上より，提案する言語相関に基づくスポット推薦手法の有効性が確認できた．

4 おわりに

本論文では，ツイートの発信位置と言語に基づくスポット推薦手法を提案し，都市ごとの言語相関によるスポット推薦精度を検証した．実験よりツイート数と類似度の相関はあるが，都市ごとに異なる類似度となり，発信位置と言語の両方の相関を考慮した推薦手法が nDCG では最大で 17.0%向上し，提案手法の有効性を確認できた．

謝辞

本研究の一部は，JSPS 科研費 16H01722，17K12686，19K12240 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] 小原基季, 森田和宏, 泓田正雄, 青江順一,Twitter 本文を用いた観光情報抽出及び分析システムの構築，第 29 回全国大会, 人工知能学会全国大会論文集 29 巻 pp. 1-3 (2015).

[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] M. S. Mohd Pozi, et.al: Sketching Linguistic Borders: Mobility Analysis on Multilingual Microbloggers, Proc. of WWW2017