7K-06

ジオタグ付ツイートの多言語相関性に基づく Venue 推薦システムの検討

白数紘之[†1]　先原進之介[†1]　中岡佑輔[†1]　*Panote Siriaraya*[†1]　河合由起子[†1]　*Adam Jatowt*[†2]
†1 京都産業大学　　†2 京都大学

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，ジオタグ付きのソーシャルネットワークサービス（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]．これまで著者らも，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，さらに位置と時間の関係性を考慮した時空間差異分析および可視化に関する研究を行ってきた [3]．これにより，ユーザの関心を時空間の観点から俯瞰することが可能となったが，ユーザ特性（年齢や性別，人種）までは考慮しておらず，群衆の嗜好性に基づいた情報推薦までには至っていなかった．また，ジオタグツイートがツイートに占める割合は数パーセントと低く，都市部以外では適応が困難という根本的問題が残る．

そこで，本研究では，ジオタグツイートから時空間情報となる場所と時間以外に，発信ユーザが登録する母国語および内容に記述されている言及言語の言語情報を考慮することで，発信位置（国）と言語（国）との同一性から群衆（国民）の嗜好性を抽出し，各国民間の類似性を抽出することでツイートの少ない地域も含めたいずれの場所でも嗜好性の高い情報の推薦を目指す．例えば，スペイン人のツイートが少ない「ローザンヌ」において，類似度の高いイタリア人の嗜好と類似度は低いがツイート（情報）の多いドイツ人の嗜好も考慮した Venue 推薦が可能となる．

本論文では，対象領域を多言語性の高いヨーロッパ 19 カ国における飲食店（Venue）推薦システムを構築し（図 1[1]），提案手法より抽出した飲食店に対するフランス人による評価実験を行い，有効性を検証する．

2　位置と言語分析に基づく Venue 推薦

本章では，任意の場所における言語（国民）の嗜好性抽出ならびに Venue 推薦，可視化手法について述べる．

Venue 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 Venue 名を抽出し，Venue 名と発信位置から Venue の属性情報となるジャンル名を取得する．ジャンル名は「BAR」や「CAFE」など 100 種類程度の統一形式となるため，数十万以上の固有の Venue 名を用いた言語国の類似度抽出（次のステップ）で生じるコールドスタート問題を回避できる．次に，発信位置（国）

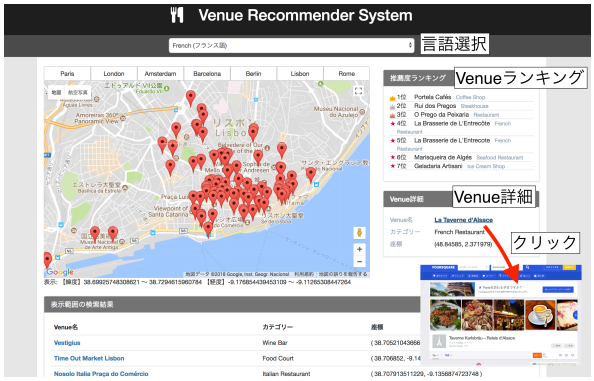

図 1: レストラン推薦システムのインタフェース

ごとに同一の言語（国）のツイートを分類し，それらのジャンル名の出現頻度を算出し，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 Venue の出現頻度をツイートから算出し，値の高い Venue をマップ上に提示する．

2.1　発信場所と言語に基づく Venue 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 $_l \subseteq \overline{\text{母国語}_l}$）として分類される．

次に，分類された言語ごとの Venue 辞書を作成する．Venue 辞書は，言語，緯度経度，地物名，属性情報のタプルであり，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（Venue）として抽出する．属性情報は，抽出した Venue 名を用いて Swarm API から取得したカテゴリとジャンルとし，ジャンルはカテゴリの下位層になる．

各言語の Venue 辞書に基づき，全言語 L に対して言語 l_x の言語国の都市 p でのみ発信された各ジャンル j に対する嗜好性となる評価値を出現頻度 $TF_{\{x,j\}}$=（l_x におけるジャンル j 出現回数）／（l_x におけるジャンル総出現回数）から算出する．

算出した言語 l_x のジャンル j 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 Venue を含むツイートを取得し，ツイート数が閾値以上の場合（ツイート数が多い場合）は下記よりランキングした Venue を抽出する．

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

Venue Recommender system for Regions with Dense and Sparse Geo-tagged Tweets based on Multilingual Analysis
†1 Hiroyuki Shirakazu　†1 Shinnosuke Sakihara　†1 Yusuke Nakaoka
†1 Panote Siriaraya　†1 Yukiko Kawai　†2 Adam JATOWT
†1 Kyoto Sangyo University
†2 Kyoto University

[1] http://yklab.kyoto-su.ac.jp/~sirakazu/VenueRecomender/

表 1: ジオタグ付ツイートと各都市におけるツイートされたユニーク Venue 数 (下線は言語国に対する首都).

Language	#Tweets	#Total venues(%)	London	Rome	Paris	Barcelona	Berlin	Lisbon	Amsterdam
All	25,993,771	-	-	-	-	-	-	-	-
Italian	2,251,204	36,940(1.6%)	2,914	6,203	369	1,706	81	39	153
French	2,430,737	29,851(1.2%)	1,568	363	16,445	797	5	157	209
Spanish	4,801,999	34,813(0.7%)	3,624	3,419	868	20,614	117	240	464
German	2,041,920	55,414(2.7%)	1,454	367	211	820	873	44	276
Portuguese	881,874	22,359 (2.5%)	634	115	479	373	131	2,127	313
Dutch	1,671,522	158,517 (10.5%)	197	67	368	261	68	101	3,165
Total	14,079,256	337,894 (2.3%)	10,391	10,534	18,750	24,571	**1,275**	2,708	4,580

表 2: 言語 l_x のジャンルに基づいた類似度 ($sim(x, y)$).

l_x	FR	ES	DE	IT	PT	NL	Average
FR (French)	1	0.50	0.53	0.47	0.36	**0.62**	0.50
ES (Spanish)	0.50	1	0.59	0.55	0.47	**0.71**	0.56
DE (German)	0.54	**0.70**	1	0.63	0.69	0.67	0.65
IT (Italian)	0.70	0.55	**0.70**	1	0.57	0.63	0.63
PT (Portuguese)	0.37	0.48	0.50	0.39	1	**0.54**	0.46
NL (Dutch)	0.62	**0.72**	0.70	0.63	0.54	1	0.64
Average	0.50	0.59	0.60	0.53	0.63	0.63	0.58

表 3: 推薦された飲食店に対するフランス人の評価結果

City	TF average(SD)	Similarity average(SD)	gain(%)
Berlin	2.75 (0.62)	3.44 (0.46)	**+25.19%**
Lisbon	3.96 (0.50)	3.82 (0.27)	-3.67%
Amsterdam	3.29 (0.40)	2.98 (0.89)	-10.5%
Rome	3.51 (0.60)	3.61 (0.55)	**+2.78%**
Barcelona	3.07 (0.47)	3.6 (0.68)	**+14.81%**
Average	3.32	3.49	**+4.99%**

2.2 ツイート数の少ない地域における各言語との類似性に基づいたジャンル抽出

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，他言語とのジャンルの類似性を考慮することで，他言語の l_y におけるジャンル j に対する評価値 $TF_{\{y,j\}}$ 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 のジャンル j に対する評価値を抽出する．

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

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

3 実験

本稿において，2016 年 4 月 1 日から 2017 年 4 月 30 日の約 13ヶ月間の欧州領域のツイートを対象に，6 言語を対象とした飲食店推薦システムを構築し，6 言語の首都とそれ以外の 1 都市の 7 都市における飲食店抽出結果について検証する．表 1 に 7 都市における Venue のうち「Food」カテゴリの各言語ごとの総数を示す．

3.1 各言語における Venue の多様性検証

提案手法より算出した各言語の言語間のジャンルに対する類似度を表 2 に示す．表の太字は l_x に対して他言語で最も類似度の高い結果を示す．

表より，最も高い類似性はオランダ語に対するスペイン語の 0.72 であった．また全体ではドイツ語が平均が 0.65 と他言語との類似性が高かった．

3.2 各言語のジャンル抽出の検証

本実験では，対象都市に訪問したことのある人またはその都市に在住のフランス人とパリ在住のフランス人計約 50 人に対して，フランスとロンドンを除く 5 都市で表 2 の類似度を用いた提案手法より推薦した飲食店に対して 5 段階評価を行ってもらい，有効性を検証した．比較手法はツイート割合（TF）とした．

表 3 より，フランス語以外の言語の類似性を考慮した推薦手法により，全体では 5%程度の向上がみられた．また，フランス語のツイート数の最も少ないベルリンでは 25%の向上が見られた．リスボンとアムステルダムに対する減少は，評価者が各々5 人と 8 人と少なかったことが影響したと考えられる．今後より多くの評価者による検証により，提案手法を適用する閾値となるツイート数の割合を検証する．

4 おわりに

本論文では，群衆（国民）の嗜好性の解明を目指し，場所と言語情報に着目し，各言語における Venue 抽出手法を提案し，フランス人 50 名による他国の各都市に対する飲食店評価実験を行い，フランス語のツイートの少ないベルリンで提案手法の推薦によりベースライン（TF 値）より 25%の向上が見られた．今後，ツイートの少ない地域判定の閾値検証ならびに言語国数を拡大した評価を行う．

謝辞

本研究の一部は，総務省 SCOPE（受付番号 171507010），JSPS 科研費 16H01722，15K00162，17K12686 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).