ARG WI2 No.10, 2017
WI2-2017-9

ジオタグツイートの多言語分析に基づくレストラン推薦手法の提案

先原　進之介[†,a]　白数　紘之[†,b]　王　元元[††,c]　河合　由起子[†,d]
アダム　ヤトフト[†††,e]

† 京都産業大学　†† 山口大学　††† 京都大学

a) *g1444602@cc.kyoto-su.ac.jp*　b) *g1444666@cc.kyoto-su.ac.jp*　c) *y.wang@yamaguchi-u.ac.jp*
d) *kawai@cc.kyoto-su.ac.jp*　e) *adam@dl.kuis.kyoto-u.ac.jp*

概要　本研究では，ジオタグツイートの発信位置と言語の相違を分析し，群衆（国民）の嗜好性を解明することを目的に，任意の地域における国民性に合わせたレストラン推薦手法を提案する．本論文では，母国語の多様性が高いヨーロッパを対象とし，まず，ジオタグツイートの発信位置，発信時刻，言及言語を抽出し，任意の場所で任意の期間で発言された任意の言及言語に基づきツイートを分類する．次に，場所に関するツイートとなる”I'm at”以降の地物（Venue）名を抽出し，それら地物の属性情報（例えば，図書館やインド料理店）を取得し，レストラン（飲食店）に関する属性情報を含む Venue 名を抽出し，各言語の TF・iDF 値を算出し，言語ごとに上位の Venue 名をその国民に人気のレストランとして推薦する．さらに，全ての場所における任意の言語のツイートの属性情報に基づいた TF・iDF 値を算出することで，国民に人気のレストランの種類（インド料理店や中華店）を抽出でき，これによりジオタグツイートの少ない地域においても各国民へのレストランが推薦可能になる．本論文では，ジオタグツイートの時空間情報と言語分析に基づく群衆の嗜好性抽出およびレストラン推薦手法について述べ，抽出した各場所のレストランの相関性について検証する．

キーワード　ジオタグツイート分析，ツイート多言語分析，レストラン推薦

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．都市に存在する店舗や施設などで Check-in するユーザの移動軌跡を分析し，その都市の特徴を抽出する手法 [1] や，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施設のカテゴリ情報を用いて地域の機能性を発見する手法 [2] が実証されている．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，ユーザ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も残る．

そこで本研究では，ジオタグツイートデータから時空間情報となる場所と時間以外に，発信ユーザの母国語および内容に記述されている言及言語の相違を抽出することで，群衆の嗜好性の解明を目指す（図 1）．本論文では，任意の場所における各言語（国民）の嗜好性の高い特徴語抽出ならびに場所に依存しない嗜好性の高い特徴語抽出手法を提案する．特に，対象領域は多言語性の高いヨーロッパ 19 カ国とし，特徴語はレストラン（飲食店）およびレストランの種類（ジャンル）とする．これにより，各国民にとって人気のレストランを推薦でき，また，ジオタグツイートデータが少ない地域においても国民性に基づいたレストラン推薦が可能となる．

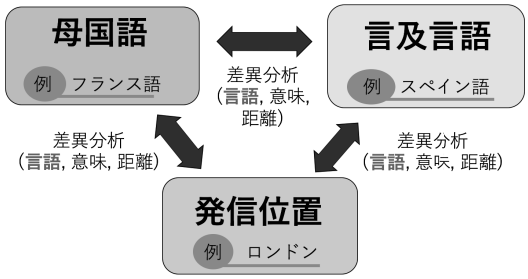

図 1 母国語，言及言語，発信位置の差異分析

本論文では，ジオタグツイートの時空間ならびに言語分析に基づく群衆の嗜好性抽出および推薦手法を提案し，実データを用いた実験より抽出した特徴語となるレストランおよびその種類の相関性について検証する．

2　国民の嗜好性抽出手法

本章では，任意の場所における言語（国民）毎に嗜好性の高い特徴語（レストラン）抽出ならびに場所に依存しない嗜好性の高い特徴語のジャンル抽出法を述べる．

Copyright is held by the author(s).
The article has been published without reviewing.

2.1 各地域の嗜好性の高い 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 $_l \vee$（言及言語 $_l \subseteq \overline{\text{母国語}_l})\}$ として分類される．

次に，分類された言及毎の Venue 辞書を作成する．Venue 辞書は，都市名，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として抽出する．属性情報は，Swarm (Foursquare) API[1]から地物名を用いて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．例えば，カテゴリは図書館や飲食店となり，飲食店の下位層のジャンルにはインド料理店や喫茶店等が含まれる．今回は，カテゴリを「food」とし，Venue 辞書を作成した．

作成した各言語の Venue 辞書に基づき，任意の地域に含まれるレストラン名を含むツイートを取得し，下記の式より閾値以上のレストランを推薦する．

$$\frac{d \text{ 期間中に } c \text{ 国で発信された } l \text{ 言語の単語 } i \text{ の出現回数}}{d \text{ 期間中に } c \text{ 国で発信された } l \text{ 言語における総単語数}} \cdot \log \frac{D \text{ 期間} \times C \text{ 国総数} \times L \text{ 言語総数}}{\text{単語 } i \text{ の出現した期間数} \times \text{国数}} \quad (1)$$

2.2 各言語の嗜好性の高いジャンル抽出

各言語の Venue 辞書に登録されている都市名を含まないツイートの少ない地域 p におけるレストラン推薦手法は，全ての場所における言語 l のツイートの属性情報に基づきジャンルの人気度を下記より算出し，上位のジャンルと p の地域名からレストラン検索し，検索結果上位を推薦する．なお，属性情報のうち，カテゴリの下位のジャンル（インド料理店や中華店）を用いる．

$$\frac{l \text{ 言語のジャンル } j \text{ 出現回数}}{l \text{ 言語におけるジャンル総数}} \cdot \log \frac{L \text{ 言語総数}}{\text{ジャンル } j \text{ の出現国数}}$$

3 実験

本稿では，表 1 に示す 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，ロンドン（中心市街地半径 20km）におけるスペイン語とフランス語に対するレストランおよび全領域における両言語に対するジャンルの相関性について検証した．なお，欧州全体における”I’m at”を含む数は 5%以下で，カテゴリ「food」の Venue 名の総数は 1%程度であった（表 1）．このうちロンドンに関しては，スペイン語が 3,624 店舗，フランス語が 1,568 店舗であった．

[1]https://developer.foursquare.com/

[2]全てのカテゴリの Venue 名のうち重複を省いた Venue 名

表 1 ツイートストリーミングデータの分類結果

言語	tweet 総数	“I’m at” 含む数（%）	Venue 総数[2]（%）
全言語	25,993,771	1,225,072（4.7%）	299,577（1.1%）
スペイン	4,801,999	94,384（1.9%）	34,812（0.7%）
フランス	2,430,737	102,330（4.2%）	29,850（1.2%）

3.1 各言語のレストラン抽出の検証

提案手法より抽出したスペイン語とフランス語の各上位 20 店舗の Venue 名の順位相関を同順位を含むスピアマン順位相関係数より算出した結果は 0.28 となり，正の相関ではあるが 0.7 以下と低い相関であった．また，式（1）の第一項目（TF 値）のみを用いて算出した Vanue 名の順位相関も 0.40 と低い相関となった．これより，母国以外の土地（ロンドン）におけるスペイン人とフランス人のレストランの嗜好性は異なり，本手法の位置および言語分析に基づいた，国民毎の多様性あるレストラン推薦の可能性が示唆された．

3.2 各言語のジャンル抽出の検証

提案手法より抽出した上位 20 ジャンルの順位相関をスピアマン順位相関係数より算出した結果は 0.6 となり，また，上位 10 ジャンルでは 0.81 となった．両結果とも 0.7 付近と比較的高い相関となった理由として，今回対象が 1 都市であったことが原因と考えられる．今後，複数都市におけるジャンルの相関を検証する予定である．

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，言語情報の相違に着目し，各言語における人気の Venue 名（レストラン）抽出手法を提案し，実験より各言語の相関が低いことを確認した．また，ツイート数の少ない場所でも言語ごとに嗜好性の高いレストランの推薦手法を提案した．今後，対象地域および言語種類を拡大した検証ならびにレストラン推薦システムの定性的評価を行う予定である．

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, R. Song, Y. Wang, X. Xie, and J. Luo: Mining Shopping Patterns for Divergent Urban Regions by Incorporating Mobility Data, Proc. of CIKM2016, pp. 569-578 (2016).

[2] J. Chen, S. Yang, W. Wang, and M. Wang: 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).