多言語類似性に基づくジオタグ付ツイート低密度地域における Venue 推薦手法の検証

中岡　佑輔[a)]　Panote Siriaraya[a)]　王　元元[b)]　河合由起子[a)]　Adam Jatowt[c)]

a) 京都産業大学コンピュータ理工学部　〒 603–8555 京都府京都市北区上賀茂本山
b) 山口大学大学院創成科学研究科　〒 755–8611 山口県宇部市常盤台 2-16-1
c) 京都大学大学院情報学研究科　〒 606–8501 京都市左京区吉田本町
E-mail: *a*){g1444936, kawai}@cc.kyoto-su.ac.jp, spanote@gmail.com, *b*)y.wang@yamaguchi-u.ac.jp, *c*)adam@dl.kuis.kyoto-u.ac.jp

あらまし　本研究では，ジオタグ付ツイートの発信位置と言語の相違を分析し，群衆（国民）の嗜好性を抽出することで，ツイートの少ない低密度地域でも国民性に合わせた Venue を推薦可能なシステムの実現を目指す．ジオタグツイートが相対的に少ない低密度地域における Venue 推薦をアイテム推薦におけるコールドスタート問題と捉え，提案手法では国民（ユーザ）のツイート数（レビュー数）が少ない Venue（アイテム）の評価値を，その Venue のジャンル（例えばインド料理店）に対する他の国民の嗜好性との類似度から算出する．各国民の各ジャンルに対する嗜好性はツイートの発信位置（国）とツイートの言語情報から算出する．この抽出されたジャンルに基づき Venue を選出し推薦提示する．なお，各言語ごとにツイート数の多い地域では従来手法の出現頻度より Venue の評価値を算出し，国民ごとに Venue を抽出推薦する．本稿では，特に多様な国民性が共存するヨーロッパを対象とし，ジオタグ付ツイートの時空間情報と言語情報分析に基づく群衆の嗜好性抽出および Venue 推薦手法について述べ，欧州の複数の国民となる被験者約 60 名を対象とした評価実験を行い，提案手法より抽出した Venue 推薦精度の有効性を検証する．

キーワード　ジオタグ付ツイート分析，ツイート多言語分析，Venue 推薦，MTurk 検証

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，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（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．

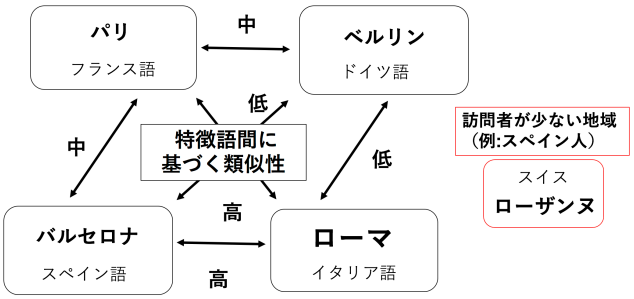

図 1　発信位置と言語情報に基づく情報推薦概念図：各国民の嗜好性抽出およびジオタグツイートの少ない地域における情報推薦

本論文では，対象領域を多言語性の高いヨーロッパ 19 カ国とし，指定言語に応じた Venue 推薦システムを構築し，検証する．具体的には，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）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（TF）を評価値として算出する．ジャンルの出現頻度より，各言語国間の相関係数を類似度として算出し，ユーザ指定の Venue 検索地域内でユーザ指定の言語のツイート数が閾値以上の場合は，ツイートから Venue の出現頻度を算出し，値の高い Venue をマップ上に提示する．また，指定地域のツイート数が閾値未満の場合，ジャンルより抽出した各言語との類似度と他言語のジャンルの評価値との積の総和から指定言語のジャンルの評価値を算出し，その上位ジャンルの Venue をマップ上に提示する．

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 Venue 推薦手法を提案し，欧州の 13ヶ月分のジオタグツイートを用いて 6 言語に対応した Venue 推薦システムを構築する．また，評価実験では，抽出した Venue およびジャンルの分類結果および各言語の相関性の検証，さらに欧州のネイティブスピーカーのユーザ 60 人による推薦された Venue に対するユーザ評価を実施し，言語に基づく Venue 推薦の有用性を検証する．

2. 関 連 研 究

大量のジオタグツイート（以下，ツイート）に対する時空間分析に関する研究が，国内外で広く取り組まれている．

Qu ら [4] は，レストランや店舗などの特定の店舗で Check-in した際に発信されるツイートを分析し，ユーザの移動軌跡を抽出し，そのレストランや店舗などのトレードエリアの発見を行った．

また，一定領域の分析結果を地図の LOD に同期し可視化することで効果的な時空間解析が実証されている [5]．さらに，地域に特色のある語と位置情報に新たな地域ユーザを手がかりとして付け加えた口コミ収集の提案 [6] や，観光客に関する情報を抽出する研究の 1 つとして Twitter に投稿されたツイートの位置情報と本文を用いることで，ユーザの観光地での訪問動向より訪問目的を推定する手法の提案 [7] などの研究が行われている．

一方で，地域に特色のある語と位置情報より新たな地域ユーザを手がかりとして付け加えた口コミの収集の提案 [6] や，観光客に関する情報を抽出する研究の 1 つとして Twitter に投稿されたツイートの位置情報と本文を用いることで，ユーザの観光地での訪問動向や訪問目的を推定する手法の提案 [7] などの研究も行われている．

これまで著者らも，ユーザ行動分析として日米両国の数ヶ月間のツイートを分析し，データ発生位置とコンテンツ内容位置との差異，発生時間と内容時間との差異の分析，さらに位置と時間の関係性を考慮した時空間差異の分析および可視化に関する研究を行ってきた [8]．また，ツイートの時間と場所と言語に基づき分析し，ユーザ行動に対する場所と言語の相違の可視化に関する研究を行ってきた [9]．

以上，既存研究を含めジオタグの時間および位置情報分析に関する研究は広く行われているが，これらに加えて言語情報から群衆（国民）の特性を抽出し，さらに群衆間の類似性および位置特性に基づき任意の場所のいずれにおいても Venue（地物）推薦を可能にする研究開発は稀である．

Algorithm 1 ジオタグツイートによる言語に基づいた地域の Venue 抽出

Require: Locations (cites) P, Geo-tagged tweets T, Languages L, Venues I, Genres J
Ensure:
for All geo-tagged tweets T in P **do**
  for All venues I in T **do**
    Compute an evaluation score of a genre j in the language l_x of a country x using TF: $TF_{\{x,j\}} \leftarrow \#j, \#genres$
    Compute a similarity between two countries (x and y): $sim(x,y) \leftarrow TF_{\{x,j\}}, TF_{\{y,j\}}$
    for All cities p in P **do**
      if $\#venues \geqslant \alpha$ **then**
        Compute an evaluation score of a venue i in the language l_y of the country y in a city p: $S_{\{i,p\}} \leftarrow |T \in p : l_y \in T : i \in I_t|, |T \in p : l_y \in T|, |L|, |l \in L_i|$
        Recommend venue i
      else
        Compute an evaluation score of a language l_x in a genre j in a city p: $S_{\{j,p\}} \leftarrow sim(x,y), TF_{y,j}$
        Recommend venue i
      end if
    end for
  end for
end for

3. 位置と言語分析に基づく Venue 推薦手法

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（ステップ）を以下に示す．

（1） 各言語国の Venue のジャンルに対する評価値抽出
（2） 言語国間のジャンルの評価値に基づく類似度抽出
（3） 任意地域の各言語国の Venue に対する評価値算出
（4） 任意地域の各言語国のジャンルに対する評価値算出
（5） Venue 数が閾値以上の場合は（3）の Venue 抽出
（6） Venue 数が閾値未満の場合は（2）および（4）を用いたジャンル抽出に基づく Venue 抽出
（7） マップ上に任意地域の言語毎の Venue を推薦提示

3.1 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 $_l\} \vee$（言及言語 $_l \subseteqq \overline{\text{母国語}_l}$）として分類される．たとえば，フランス人の嗜好性抽出では，任意の言語 $l_{\text{フランス}}$ は，母国語がフランス語の全てのツイートおよび母国語がフランス語以外で言及言語がフランス語のツイートが分類される．

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，

y	$sim(x=$フランス$, y)$	ジャンル$_1$の評価値		ジャンル$_2$の評価値	
スペイン	0.5	0.9	0.45	0.2	0.10
イタリア	0.85	0.3	0.26	0.9	0.77
合計	1.35		0.52		0.64

図 2　各言語との類似性に基づくジャンルに対する評価値算出例

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(注1) 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．たとえば，カテゴリは「公共施設」や「フード」などで,「フード」の下位層のジャンルには「中華」や「喫茶店」などが含まれる．

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．例えば，l_x=フランス語の母国フランスの都市 p=パリ周辺で発信されたツイートのジャンル j=カフェの出現頻度から，フランス人（この場合はパリ人）のカフェに対する嗜好性となる評価値が算出される（ステップ 1）．

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 $sim(x,y)$ を下記の相関係数より算出する（ステップ 2）．

$$\frac{\sum^{J}(TF_{\{x,j\}}-\overline{TF_{\{x,j\}}})(TF_{\{y,j\}}-\overline{TF_{\{y,j\}}})}{\sqrt{\sum(TF_{\{x,j\}}-\overline{TF_{\{x,j\}}})^2\sum(TF_{\{y,j\}}-\overline{TF_{\{y,j\}}})^2}} \quad (1)$$

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）は式（2）よりランキングした Venue を抽出する（ステップ 3，5）．

$$S_{\{i,p\}} = \frac{|T \in p : l_y \in T : i \in I_t|}{|T \in p : l_y \in T|} \cdot \log \frac{|L|}{|l \in L_i|} \quad (2)$$

上記の式を自然言語で書き直すと以下のようになる:

$$S_{\{i,p\}} = \frac{A}{B} \cdot \log \frac{C}{D}$$

A　言語 l_y を使用して地域 p から発信された i（Venue）に関するツイート数

B　地域 p から言語 l_y を使用して送信された Venue ツイートの総数

C　言語 L の総数

D　i（Venue）に対する言語総数

3.2　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であり，これは未知のアイテム推薦と捉えられる．そこで，他言語とのジャンルの類似性（ステップ 2）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 $TF_{\{y,j\}}$ を用いて下記の式（3）より言語 l_x のジャンル j に対する評価値を抽出する（ステップ 4）．

$$\sum^{D}\left(sim(x,y)\cdot TF_{\{y,j\}}\right) \Big/ \sum^{D} TF_{\{y,j\}} \quad (3)$$

D は言語数であり，式（3）は場所 p における言語 l_x のジャンル j に対する推薦度を算出しており，各言語 l_y との類似度 $sim(x,y)$ に言語 l_y のジャンル j に対する評価値を乗算した値の総和を全言語の類似度の総和で割った値である．

例を図 2 に示す．任意の地域でフランス人の訪問数が少なくツイート数が閾値以下の場合，フランス人のジャンル $_1$ に対する評価値は，まず，スペイン人との類似性 (0.5) と評価値 (0.9) から 0.45 を算出し，同様にイタリア人の 0.26 を算出し，総和 0.71 を類似度の総和で割った値 0.52 が算出される．なお，ジャンルに対する他言語（例えばイタリア人）の評価値がない場合は，類似度の総和ではその言語の類似度は 0 となる．

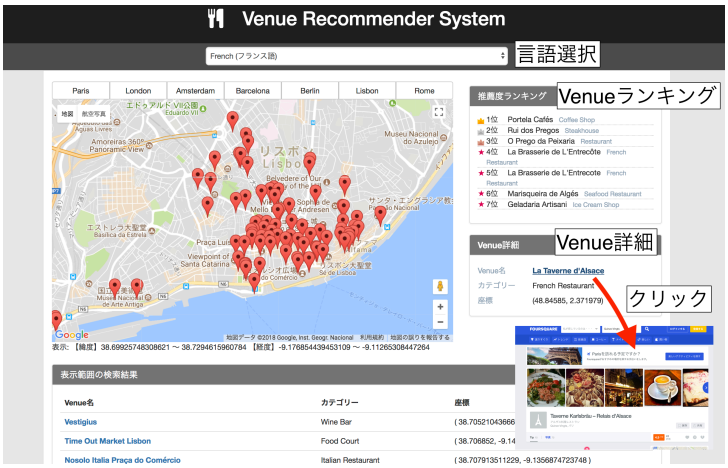

図 3　Venue 推薦システムのインタフェース．

3.3　Venue 抽出・提示

地域 p におけるツイート数が閾値未満のツイート数の少ない地域では，前節より抽出された全ジャンルのうち推薦度の高いジャンル j を用いて場所 p の周囲 r 内における同一ジャンルの全言語の Venue を Venue 辞書より選出し，出現頻度の高い順にランキング付けて Venue を抽出する（ステップ 6）．ただし，Venue 辞書の p における Venue 数が少ない場合は，ジャンル j と位置情報 p と r を用いた Swarm API の逆引きによる Venue 名検索，またはジャンル名 j と位置情報 p と r を用いた Web 検索より Venue 情報を取得する．

ツイート数が閾値以上の場合は，ステップ 4，5 で抽出した Venue 情報を取得する．最後に，Venue 辞書から抽出した緯度経度に基づき地域 p における言語 l_x に対するお勧めの Venue として，地図上にピンをプロットする（ステップ 7）．ユーザはピンにマウスオーバーすることで Venue 名とジャンル名を確認できる．またピンをクリックすると Venue に関するページへ遷移できる．

(注1)：https://developer.foursquare.com/

表 1　収集したジオタグツイートの概要と各都市で分析した独特な Venue 数 (下線の引かれている値はその都市における母国語の Venue 数を示す).

Language	#Tweets	“I’m at” tweets(%)	#Total venues(%)	London	Rome	Paris	Barcelona	Berlin	Lisbon	Amsterdam
All	25,993,771	1,231,980(4.7%)	342,992(1.3%)	-	-	-	-	-	-	-
Italian	2,251,204	98,488(3.6%)	36,940(1.6%)	2,914	6,203	369	1,706	81	39	153
French	2,430,737	36,163(1.4%)	29,851(1.2%)	1,568	363	16,445	797	5	157	209
Spanish	4,801,999	40,367(0.8%)	34,813(0.7%)	3,624	3,419	868	20,614	117	240	464
German	2,041,920	216,242(8.6%)	55,414(2.7%)	1,454	367	211	820	873	44	276
Portuguese	881,874	24,585 (2.8%)	22,359 (2.5%)	634	115	479	373	131	2,127	313
Dutch	1,671,522	257,383(15.4%)	269,413 (16.1%)	197	67	368	261	68	101	3,165
Total	14,079,256	673,228(4.8%)	448,790 (3.2%)	10,391	10,534	18,750	24,571	**1,275**	2,708	4,580

表 2　それぞれの都市の各言語の嗜好性が最も高い上位 10 ジャンル (下線の引かれている値はフランス語と重複しているジャンルを表す).

Language	Top 10 Food Genres
French (in Paris)	French Restaurant, Cafe, Fast Food Restaurant, Bakery, Bistro, Bar, Restaurant, Italian Restaurant, Coffee Shop, Japanese Restaurant
Italian (in Rome)	Italian Restaurant, Cafe, Pizza Place, Fast Food, Restaurant, Ice Cream Shop, Bar, Pub, Cocktail Bar, Bakery
Spanish (in Barcelona)	Cafe, Tapas Restaurant, Spanish Restaurant, Mediterranean Restaurant, Bar, underlineRestaurant, Bakery, Fast Food Restaurant, Sandwich Place, Italian Restaurant
German (in Berlin)	Cafe, German Restaurant, Seafood Restaurant, Salon, Fast Food Restaurant, Irish Pub, Italian Restaurant, Restaurant, Vegetarian Restaurant, Bar
Portugese (in Lisbon)	Portuguese Restaurant, Cafe, Restaurant, Burger, Joint, Bakery, Bar, Seafood Restaurant, Coffee Shop, Ice Cream Shop, Brazilian Restaurant
Dutch (in Amsterdam)	Cafe, Bar, Restaurant, Coffee Shop, Fast Food Restaurant, Bakery, Snack Place, American Restaurant, French Restaurant

表 3　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
French	(FR)	1	0.50	0.53	0.47	0.36	**0.62**	0.50
Spanish	(ES)	0.50	1	0.59	0.55	0.47	**0.71**	0.56
German	(DE)	0.54	**0.70**	1	0.63	0.69	0.67	0.65
Italian	(IT)	0.70	0.55	**0.70**	1	0.57	0.63	0.63
Portuguese	(PT)	0.37	0.48	0.50	0.39	1	**0.54**	0.46
Dutch	(NL)	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

4. データ収集と分類と実装

4.1 Twitter からのデータ収集と分類

我々の提案する推薦手法の実現可能性を調査するために，Twitter のジオタグ付ツイートを収集した．本研究では特に，母国語の多様性の高いヨーロッパのツイートに注目した．具体的には，次の 7 都市: (1) ロンドン (2) ローマ (3) パリ (4) バルセロナ (5) ベルリン (6) リスボン (7) アムステルダム、とこれらの都市における 6 言語:(1) イタリア語 (2) フランス語 (3) スペイン語 (4) ドイツ語 (5) ポルトガル語 (6) オランダ語のツイートを収集した．次に，世界で最も使用される共通言語、すなわち母国語が異なるユーザ同士がコミュニケーションを取るのに使用する言語である英語を，今回は分析対象の言語から除外した．

データは 2016 年 4 月 1 日から 2017 年 4 月 30 日までの 13ヶ月間で収集した。これらの 13ヶ月間のジオタグ付ツイートデータに対して，前章の提案手法に基づいて Venue 推薦システムを構成した（図 3）．各対象都市の中心から半径 20km 以内のジオタグ付ツイートの分類結果を表 1 に示す．全体で約 2,600 万のツイートから延べ約 34 万件の Venue が抽出された．

各対象 7 都市ごとで各 6 言語で発信されたユニークな Venue 総数は，7 万 3 千件であった（表の右半分）．表よりツイートに使用される言語と母国語が同じ場合 (例えばローマのイタリア語、パリのフランス語など) は高密度であったが，ツイートに使用される言語と母国語が異なる場合 (パリのイタリア語，ローマのポルトガル語など) は低密度であることが分かる．また，ロンドンなどの都市では異なる言語のツイートに対してかなりの密度があることも確認された．特筆すべきは，ベルリンにおけるフランス語では Venue が 5 件しかなく，本研究の他言語ユーザにとってツイート数の少ない低密度地域であり，Venue 推薦精度検証の対象言語および地域として有用なことである．

これらの Venue から全体で合計 155 の異なるジャンルを検出した．提案手法によって抽出された各言語に対する嗜好性の高い上位 10 ジャンルの例を表 2 に示す．フランス語ではフレンチやカフェ，スペイン語ではカフェやタパスレストラン，オランダ語ではコーヒーショップ等が上位に抽出され，概ね国民性にあったジャンルが抽出されていることが確認できる．

また，抽出されたジャンルに基づいた他言語ユーザとの類似度を表 3 に示す．6 言語間の類似度は，スペイン語とオランダ語が最も高く，ポルトガル語とフランス語が最も低い結果となった．全体平均は 0.58 となり，ドイツ語とオランダ語が平均して高い結果となった．

4.2 言語に基づく Venue 推薦システム

出現頻度および類似度を用いた提案手法よりツイートの位置および言語特性を用いた Venue 推薦システム (注2) を実装した（図 3）．実装したプロトタイプは，ユーザが言語を選択すると対象地域上に提案手法より抽出された指定言語ごとのレストラ

(注2)：http://yklab.cse.kyoto-su.ac.jp/~sirakazu/VenueRecommender/

表 4　異なる 3 都市の 2 言語で推薦された上位 10 件の Venue(下線部分は 2 言語で重複しているものを表す).

City	Language	Top 10 Food Venues
London	Spanish	Caffe Nero, Starbucks, All Bar One, McDonald's in London, Pret A Manger, Wasabi, EAT., Coffee Republic, Bill's Restaurant, Costa Coffee
	Italian	Caffe Nero, Starbucks, Pret A Manger, Caffe Nero Express, Franco Manca, Caffe Nero Central St Giles, Bill's Restaurant, EAT., The Breakfast Club, Adelaide Ice Cream And Hot Dogs
Rome	Spanish	McDonald's in Roma, Queen's Chips Amsterdam, Pompi, Old Wild West, Sant'Eustachio Il Caffe, Irish Pub, Royal Art Caffe Restaurant, Castroni, C'Era Una Volta Il Caffe, McDonald's galleria
	Italian	McDonald's in Roma, Old Wild West, La Piazzetta, La Bottega del Caffe, La Romana, Pompi, Osteria La Sol Fa, la Mela d'Oro, La Casetta, Farine La Pizza
Barcelona	Spanish	McDonald's in Barcelona, Pans & Company, Viena, La Tagliatella, La Paradeta Passeig, de Grocia, Bracaff, Granier, Farggi, La Muscleria, Foster's Hollywood
	Italian	La Tagliatella, McDonald's in Barcelona, Granier, Il Cafe di Roma, Il Cafe di Francesco, Vivari, 100 Montaditos, La Paradeta Passeig, de Grocia, Il Cafe di Roma, Fresc Co
	French	100 Montaditos, McDonald's in Barcelona, La Paradeta Passeig, de Grocia, Marco Aldany, Central Cafe, Hard Rock Cafe Barcelona, Gran Cafe, Hidden Cafe Barcelona, El Tastet de la Mar, El Merendero de la Mari
	German	Pans & Company, Granier, El Fornet d'en Rossend, 100 Montaditos, El Fornet, Costa Coffee, 365.cafe, Dehesa Santa Marta, El Fornet d'en Rossend, Restaurante Barceloneta
	Portugese	Pans & Company, Subway, Jamaica Coffee Shop, SandwiChez, Lizarran, Farggi, Taller de Tapas, Camp Nou Dinner Terrasse, Tapas24 Camp Nou, Audrey Brunch & Coffee
	Dutch	Granier, El Fornet d'en Rossend, Tapas24 Camp Nou, Camp Nou Dinner Terrasse, Restaurante Park Guell, Faborit, Granier Pans Artesans, El Arbol Brunch, Otto Sylt, Piscolabis

ンを推薦提示する．提示方法は，1）地図上にプロット，2）評価値よりランキングし上位 7 件を提示（図中の右上），3）地図の中心距離から近い順にランキング提示（図下）の 3 種類とした．Venue を選択クリックすると，詳細情報が別ウィンドウにて提示される．

5. 推薦手法の評価検証

前節の witter から得られたデータに基づいて，提案した推薦手法の複数の実験検証を行った．

5.1 他言語のユーザ間における Venue の嗜好性に対する多様性の検証

ユーザへの Venue 推薦における言語の有用性を検証するために，まず，各言語間において Venue に対する嗜好性の違い（多様性）があるか否かを相関性より検証した．各地域における他言語間の相関検証は，各言語に対して推薦された上位 20 件の Venue に対して，それら Venue の評価値より算出しランキングした結果をスピアマン順位相関係数より算出した．高い相関性 (つまり Venue の嗜好の多様性が低い) は，言語に関わらず類似した Venue の種類を好む可能性がある．

相関結果は，全てのペアの平均相関は 0.32 であり，相関は低いと考えられる．つまり Venue に対する嗜好性は比較的多様性が高い結果となった．このうち，最も相関の高いペアはローマにおけるスペイン語とイタリア語の 0.46 であり，最も相関の低いペアはバルセロナにおけるスペイン語とイタリア語の 0.18 であった．表 4 に，高い多様性 (つまり低い相関性は低い) 言語のペアに対して抽出された Venue 例を示す．

以上より，ツイートの言語を考慮したレストラン抽出結果より言語ごとに嗜好の多様性があることが確認できた．これはツイートの多様な言語特性を用いた Venue 推薦に有用である可能性があることを示している．

5.2 ジャンルに対する推薦手法の検証

次に，データが低密度な地域の推薦手法について検証した．本検証では，対象言語の実際のツイートを考慮せずに他言語との類似度に基づいて推薦されたジャンル（3.2 節で議論した提案手法）と，ツイートから特定された情報に基づいて推薦されたジャンルとを比較する（3.1 節で議論した TF 値による推薦手法）．今回はフランス語のツイートに注目した．その主な理由は，対象都市でフランス語を使用するユーザは，都市ごとの Venue 数が多様なため (例えば，ロンドンは Venue 数が四桁だが一方でベルリンは一桁)，後述のユーザによる評価検証にも有用であることが挙げられる．実験では，両方の推薦手法における嗜好性の評価値の関係をスピアマン順位相関係数によって検証した．

表 5 に結果を示す．アムステルダムとバルセロナで推薦されたジャンルの嗜好性の評価値は，リスボンとローマが強い正の相関がある一方で，2 つの推薦手法は弱い正の相関を示した．しかし，ベルリンでは強い負の相関関係であった．これはベルリンのフランス語のツイート数が少なかったためと考えられる．

5.3 Venue 推薦のユーザ評価

提案した推薦手法の精度を検証するために，訪問または住んだことにあるユーザに提案手法により推薦されたレストランに対する好みを評価する被験者実験を行った．本実験では，クラウドソーシングマーケットプレイスの Mechanical Turk(Mturk)[注3]

（注3）：https://www.mturk.com/mturk/welcome

表 5　異なる都市のフランス語の話し手における嗜好性の評価値に基づいた TF と類似度に基づく推薦手法の相関.

City	Genre correlation
Berlin	strong negative (-0.83)
Lisbon	positive (0.69)
Rome	strong positive (0.98)
Amsterdam	weak positive (0.31)
Barcelona	weak positive (0.37)

表 6　アンケートへの回答数.

Speaker	City	Number of responses
French	Berlin	7
	Lisbon	5
	Amsterdam	8
	Rome	11
	Barcelona	9
	London	3
	Paris	7
Italian	Rome	11
Dutch	Amsterdam	2
German	Berlin	2
Total		**65**

にて実施した．このプラットフォームは多くの研究領域の調査研究において，良質なデータを収集するのに効果的な方法であると示されている [10], [11].

TF に基づく推薦手法 (3.1 節の式 (2)) と類似性に基づく推薦手法 (3.2 節の式 (3)) より抽出された Venue に対して好みがあるかどうかを短いアンケートを記入してもらうよう回答者に求め，回答者には各都市で推薦された Venue に行く可能性があるかどうかを 7 段階の尺度 (0=可能性が低い 7=可能性が高い) 評価を求めた．提示する Venue 情報は Venue 名と取得した Venue に関する Foursquare へのリンク情報を提示した．アンケートに回答した人には$0.15 の報酬 (これは所要回答時間量に対する MTurk の報酬の方針に相当する) を与えた．

最初のユーザ評価実験のためのデータセットは，対象都市に住んだことのあるまたは訪問したことのある人々の中で，その都市のネイティブスピーカーではない人を対象とした．我々は，(1)Mechanical turk の予想サンプルサイズと (2) 前述した異なる都市における多様なツイート総数となるフランス語に焦点を当てた．表 7 に各都市における Mechanical Turk のアンケートへの回答数を示す.

もう一つのユーザ評価実験のためのデータセットは，対象都市とは同一国に住んでいるネイティブスピーカーに焦点をあて，パリのフランス人，ローマのイタリア人，ベルリンのドイツ人，アムステルダムのオランダ人からアンケートを回収した．7 名のパリのフランス語が母国語のユーザ，11 名のローマのイタリア語，2 名のベルリンのドイツ語，2 名のアムステルダムのオランダ語が母国語のユーザが回答した．今回ベルリンのドイツ語とアムステルダムのオランダ語のユーザ数は少なかったため，ユーザ評価対象外とした.

5.3.1　TF と類似度に基づくの Venue 推薦手法の比較

低密度地域における類似度に基づいた推薦手法の精度の検証は，最初のデータセットにおけるユーザ評価より行った.

表 7　フランス語の話し手による Venue 推薦の平均評価値

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%**

表 7 にフランス語の話し手による（在住していたあるいは訪問していたことがある）各都市に対する，TF に基づき推薦された上位 10 件の Venue と類似度に基づき推薦された上位 10 件の Venue に対する評価値の平均値を示す．なお，今回ロンドンに対する回答数は少なかったため ($N = 3$)，評価の対象外とした.

全体では，全対象都市における TF に基づく推薦手法の平均値の合計は 3.32 であり，類似度に基づく推薦手法の平均値の合計は 3.49 となり，類似度に基づいた推薦手法を使用した場合の性能が 5%向上した．ただし，フランス語のツイート数の比較的少ないリスボンとアムステルダムでは，類似度に基づいた推薦手法は TF と比較して平均-7.1%減少していた．つまり，ツイート数は多くはないが TF による分析のための十分な Venue 数があったと考えられる．結果で特筆すべきことは，類似度に基づいた推薦手法は，ツイート数が極端に少ないベルリンで 25.2%向上したことである.

次に、ツイートが高密度な領域で類似度に基づく Venue 推薦アプローチを検証するために，ユーザと母国語が同じ場合 (例えばパリのフランス人、ローマのイタリア人) のユーザ評価のデータセットを検証した．パリのフランス語が母国語のユーザの TF に基づいた Venue 推薦の平均値は 2.94（標準偏差=0.48）となり、類似度に基づいた推薦の平均値である 2.74（標準偏差=0.54）よりも若干高い結果となった．また，ローマのイタリア語が母国語のユーザの TF の推薦結果への評価値の平均値は 3.51（標準偏差=0.77）となり，類似度に基づく推薦の 3.61 より若干低い結果となった．高密度な両領域でのネイティブスピーカーによる評価値の平均は，TF に基づいた評価値は 3.23 となり，類似度に基づく評価値の 3.18 よりも 1.55%と若干ではあるが優位であった.

5.3.2　nDCG 測定を用いた推薦システムの評価

提案手法により抽出され順位付けされた Venue に対する精度検証を行った．検証では，フランス語を母国語とするユーザの評価に基づいて Venue を順位付けし，正規化減価累積利得 (nDCG) を算出した．提案する Venue 推薦手法は，ツイートが多い場合は $TFIDF$ に基づいた式 (2) より算出し，ツイートが少ない場合は類似度に基づいた式 (3) より算出しているため，両手法の評価を行った

各ヨーロッパのフランス語が母国語のユーザ評価結果を表 8 に示す．全体的な結果として，ローマが最も高い 0.973 となり，アムステルダムが最も低い 0.822 となり，全都市における順位

表 8　各都市におけるフランス語の話し手の Venue の順位付け精度に対する nDCG 測定値の結果．太字の数字は提案した推薦アルゴリズムを使用した値を示す．

City	#venues by tweets	Similarity	TFIDF-based
Berlin	5	**0.941**	0.898
Lisbon	157	**0.969**	0.913
Amsterdam	209	**0.822**	0.967
Rome	363	**0.973**	0.890
Barcelona	797	**0.965**	0.944
London	1,568	0.841	**0.968**
Paris	16,445	0.918	**0.961**

付け精度の平均は 0.94 となり，提案手法により高い順位付け精度が確認できた．また，結果よりツイート数が 1,500 件以上の都市では $TFIDF$ が優位であり，800 件未満では類似度を用いる方が優位であることが示された．今後，得られた結果を高密度と低密度の閾値として検討し，評価検証する．

6. 議　　論

本研究では，次の 3 つの研究課題に取り組んだ．

最初の課題は，ジオタグ付ツイートの言語特性をどのように抽出し，Web システムとしてどのように利用できるかである．本稿では，ユーザが設定している言語，さらに内容に記載されている言語に対して，ツイートの発信位置との関係性，ツイートの内容（Venue）との関連性から，言語ごとの発信場所の Venue となる特徴語を評価値と共に抽出した．また，母国語のツイートの少ない場所でも Venue を推薦可能な Web システムを構築した．

二つ目の課題は，ツイート数の少ない低密度の領域で，どのように Venue を発見し推薦するかである．本稿では，他言語との類似性を用いたジャンルの抽出手法を提案した．

三つ目の課題は，提案する推薦精度の検証である．複数の都市の複数の言語を用いて複数の実験を行った．ヨーロッパにおけるユーザ評価は，ジオタグツイートの言語特性を用いた Venue 推薦は，全体を通して有用であることが示された．特に，ツイートの低密度な領域では，推薦手法はフランス語が母国語のユーザにとってスピアマン順位相関係数の評価では約 25% の精度向上がみられ，さらに nDCG 測定による順位付け精度は 0.94 となり，良好であった．本稿では，ツイートの Venue 名となる固有名詞のみを対象としており，内容の詳細な分析には至っていない．しかしながら，処理コストとのトレードオフを考慮すると十分な精度であったと言える．今後，形容詞を分析対象とすることで，さらなる精度向上を目指す．

7. お わ り に

本研究では、ジオタグ付ツイートの言語特性を利用した新しい Venue 推薦システムを提案した．ジオタグ付ツイートの位置と言語，ツイート発信者の母国語に関する情報を抽出し，言語と場所に基づいたユーザに対する Venue 推薦を実現した．Venue に関するツイート数が十分にあった場合は $TFIDF$ に基づき推薦し，ツイート数の少ない低密度な領域ではジャンルの嗜好性の観点から異なる言語との類似度を算出し評価値を算出した．また，提案手法に基づき推薦システムを実装し，ツイートの少ない地域でも言語にあった Venue 推薦を実現し，それら Venue の詳細情報をユーザに推薦提示した．

提案手法の有用性を検証するために，約 2,600 万のジオタグツイートを収集し約 34 万件の Venue を抽出した．実験では 7 都市と 6 言語に焦点を当てた．結果，推薦された上位 20 件の Venue 間に弱い正の相関があり，言及言語を考慮した場合，ユーザの Venue の嗜好性に多様性があることが示された．第 2 の分析では，類似度に基づくジャンルの推薦手法がツイートの低密度な都市で使用されるときの相関係数を検証し，結果，フランス語のツイートが極端に少ないベルリンでは負の相関，リスボンやローマなどの都市で正の相関を示した．第 3 のユーザ評価による検証では，nDCG において高い順位付け精度を示し，全体を通して良好な推薦精度となったことを示せた．

今後，Venue 推薦精度のさらなる向上を目指し，低密度領域の判別閾値の適切な決定法を検討する．また，ツイート内容の形容詞を分析対象に加える予定である．さらに，Food（レストラン）以外の Venue への応用やヨーロッパ以外の地域 (アジアやアメリカなど) に拡大し検証する．

謝　　辞

本研究の一部は，総務省 SCOPE（受付番号 171507010），JSPS 科研費 16H01722，17K12686 の助成を受けたものである．ここに記して謝意を表す．

文　　献

[1] T. Hu, R. Song, Y. Wang, X. Xie, J. Luo: Mining Shopping Patterns for Divergent Urban Regions by Incorporating Mobility Data, Proc. of the 25th ACM International on Conference on Information and Knowledge Management (CIKM2016), pp. 569-578 (2016).

[2] J. Chen, S. Yang, W. Wang, M. Wang: Social Context Awareness from Taxi Traces: Mining How Human Mobility Patterns Are Shaped by Bags of POI, Adjunct Proc. of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2015 ACM International Symposium on Wearable Computers (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 the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD2015), pp. 39-48 (2015).

[4] Y. Qu, J. Zhang: Trade Area Analysis using User Generated Mobile Location Data, Proc. of WWW2013, pp. 1053-1064 (2013).

[5] A. Magdy, L. Alarabi, S. Al-Harthi, M. Musleh, T. M. Ghanem, S. Ghani, M. F. Mokbel: Taghreed: A System for Querying, Analyzing, and Visualizing Geotagged Microblogs, Proc. of SIGSPATIAL2014, pp. 163-172 (2014).

[6] 長島里奈，関洋平，猪圭: 地域ユーザに着目した口コミツイート収集手法の提案，DEIM Forum 2016，B4-3（2016）.

[7] Y. Nozawa, M. Endo, Y. Ehara, M. Hirota, S. Yokoyama, H. Ishikawa: Inferring Tourist Behavior and Purposes of a Twitter User, Proc. of AI4Tourism2016 (2016).

[8] S. Wakamiya, A. Jatowt, Y. Kawai, T. Akiyama: Analyzing

Global and Pairwise Collective Spatial Attention for Geo-social Event Detection in Microblogs, Proc. of WWW2016, pp. 263-266 (2016).

[9] M. S. Mohd Pozi, Y. Kawai, A. Jatowt, T. Akiyama: Sketching Linguistic Borders: Mobility Analysis on Multilingual Microbloggers, Proc. of WWW2017, pp. 825-826 (2017).

[10] Bentley, Frank R. and Daskalova, Nediyana and White, Brooke: Comparing the Reliability of Amazon Mechanical Turk and Survey Monkey to Traditional Market Research Surveys, Proc. of Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems, (CHI EA '17), pp.8 (2017).

[11] Buhrmester, Michael and Kwang, Tracy and Gosling, Samuel D: Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data?, Proc. of Perspectives on psychological science, pp. 3–5 (2011).