ジオタグツイート分析に基づく 群衆の認知特性抽出および郷土品推薦システム

白数　紘之† 　王　元元†† 　河合由起子† 　アダムヤトフト†††

† 京都産業大学コンピュータ理工学部　〒 603–8555 京都市北区上賀茂本山
†† 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
††† 京都大学情報学研究科社会情報学専攻　〒 606–8501 京都市左京区吉田本町
E-mail: †{g1444666,kawai}@cc.kyoto-su.ac.jp, ††y.wang@yamaguchi-u.ac.jp, †††adam@dl.kuis.kyoto-u.ac.jp

あらまし　本研究では，ツイートの発信位置と言及言語の相違から群衆の認知特性を抽出することで，任意の場所において国民性に合わせてその地域の郷土品を提示可能な推薦システムの構築を目指す．本論文では，ヨーロッパにおけるジオタグツイートを対象とし，群衆を国民として各国における郷土品に対する認知特性を抽出する．抽出手法では，ジオタグツイートの発信位置，発信時刻，言及言語を分析し，任意の場所で任意の期間で発言された言及言語ごとに特徴語を抽出する．本論文では，ジオタグツイート分析に基づく群衆の認知特性抽出および郷土品推薦手法について述べ，抽出した各場所ごとに抽出した特徴語の相関性について検証する．

キーワード　認知特性抽出，ジオタグツイート分析，郷土品推薦

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，ソーシャルネットワークサービス（SNS）データ，センサデータといった大量のストリーミングデータ分析技術が，国内外で広く注目されている．ジオタグ SNS を対象として，特定の店舗等で Check-in するユーザの移動軌跡を分析し，その店舗等のトレードエリアを抽出する手法 [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内容位置との差異，発生時間とコンテンツ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 [5] [6]．

本研究では，ジオタグツイートデータから発信位置と言及言語の相違から群衆の認知特性を抽出することで，任意の場所において国民性に合わせてその地域の郷土品を提示可能な推薦システムの構築を目指す．認知特性とは，「ものや情報などを，それが何であるかを判断したり解釈したりする過程のこと」であり，本研究において群衆の認知特性は，任意の場所において，その場所の情報に対して同じ属性の人（国民）が喜ぶ嗜好性の高いものを解釈できる情報とし，特に，「情報」を「郷土品」とする．これにより，旅先で，その土地の人気の郷土品であり，かつ異なる場所（国）の国民の嗜好性に合った郷土品を推薦できる．また，本論文では，特に日本国外で多くの旅行者が 1 回の旅行で複数の国を訪問可能なヨーロッパを対象とする．

認知特性となる郷土品抽出は，まず，シソーラスより郷土品に関する語彙を含むツイートを選別し，ツイートの発信場所における言及内容の言語に対する特徴語を抽出する．具体的には，

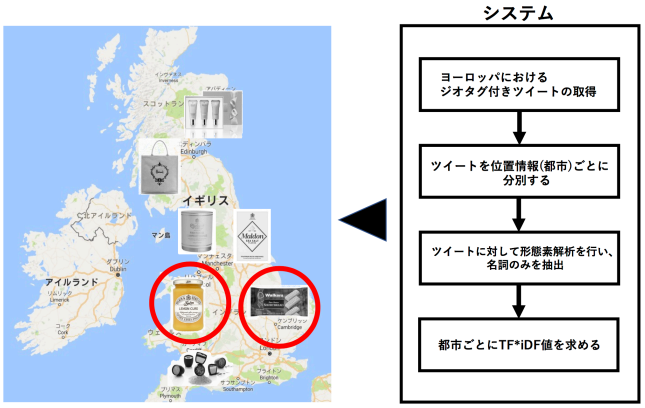

図 1　郷土品推薦システム（イギリスにおけるフランス人への推薦例）

まず，ツイートで言及されている言語を言及言語ごとに分類する（例えば，日本語で書かれた場合の言及言語は日本語）．次にツイートの発信位置の緯度経度より，発信場所を特定し，特定地域ごと言及言語ごとに分類する．そして，郷土品に関する語彙を抽出し，出現頻度を算出し，出現頻度の高い語彙をその地域の言語ごとの特徴語としてランキングする．最後に，任意の言及言語に対する各地域の特徴語の類似度を算出し，類似度を重みとし各特徴語の出現頻度に乗算しリランキングする．これにより，各地域において各国民の嗜好性の高い郷土品として推薦可能となる．さらに，例えば，日本人は紅茶の嗜好性が高いと判明した場合，ツイートの少ない地域においても紅茶に関する郷土品をシステムが検索，推薦でき，お土産として喜ばれることが期待できる．

図 1 は，郷土品推薦システムの一例であり，フランス語で言及されたツイートを提案手法により分析した結果，イギリスの地域の郷土品としてジャム（アフタヌーンティのマーマレード

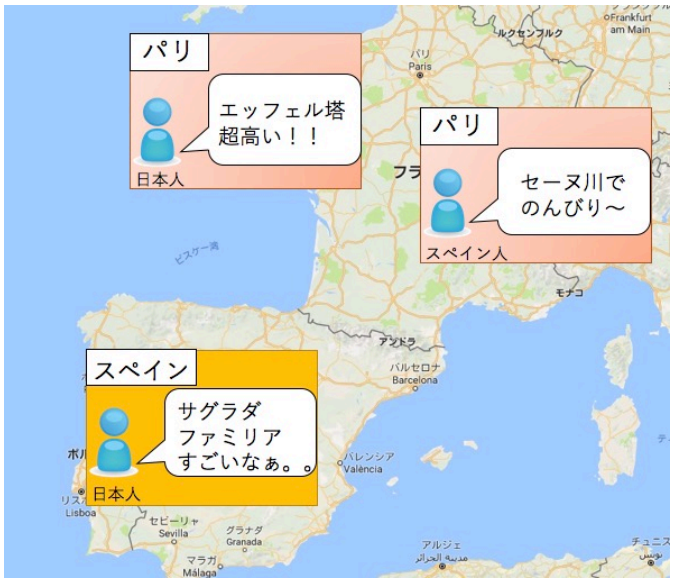

図 2　時空間における言語形態の違いによる認知特性

ジャム）やショートブレッドを推薦する．

本論文では，ジオタグツイート分析に基づく群衆の認知特性抽出および郷土品推薦システムを構築し，特徴語の相関性について検証する．

2.　郷土品推薦システム

2.1　システム概要

本研究は，ツイート発信位置，発信時刻，言及言語の違いによって生じる相違を分析し，ユーザの認知特性の一要因として特徴語を抽出し，お勧めの郷土品として提供することを指す．本研究における認知特性とは，言及言語における発信場所や時刻の相違から抽出される特徴語である（図 2）．具体的には，任意の言語で異なる場所や時間における発話内容の相違であり（例えば，日本人がパリやで発信する内容とスペインで発信する内容の相違），任意の場所にで異なる言語における発話内容の相違であり（例えば，パリにおける日本人の発信する内容とスペイン人の発信する内容の相違），これらの相違は異なる認識を示している（例えば，日本人にとってパリでは「エッフェル塔」であり，スペイン人にとっては「セーヌ川」がパリにおける代表的な観光名所という認識）．本稿では特に，日本人の郷土品に対する認知特性に焦点をあてる．

2.2　ツイートの発信国名の付与

提案する認知特性抽出手法では，言及言語の多様性が重要となる．そこで，本稿では多くの言語が使用されている欧州のジオタグ付ツイートデータを対象とする．

まず，欧州の指定地域から重複を除いたジオタグ付ツイートを The Streaming APIs (注1) を用いて取得する．次に，ツイートの発信位置に基づき，ツイートの緯度経度情報を Yahoo!ジオコーダ API を用いて変換し住所を取得する．住所に含まれる国名を各ツイートの発信国として付与する．以上より，ユーザ ID，緯度経度，発信時刻，言及内容，ユーザ設定言語（母国語），発信国，言及言語を管理する．

2.3　発信国における言及言語に基づく特徴語抽出

本研究では，これまで先行研究で行ってきた時空間である発信国と発信日時ごとの相違に基づく特徴語抽出に加え，言及言語ごとの特徴語を抽出する．

まず，郷土品に関する辞書を作成する．辞書は，国名とそれに対する郷土品に関する固有名詞の単語のペアであり，固有名詞以外となる「紅茶」、「コーヒー」「チョコレート」「土産」といったものは特定の国名とならず，all とする．本稿における辞書の作成は，Web より「土産」＋「国名」として検索された結果より，人手で選択した．

次に，任意のエリアの任意の時間における，ジオタグツイートを抽出し，郷土品辞書を用いて任意のエリアの国名の単語 i を含むツイートを取得し，下記の TF 式より特徴語 i の重要度を算出し，ランキングする．

$$\frac{d\text{ 期間中に }c\text{ 国で発信された }l\text{ 言語の単語 }i\text{ の出現回数}}{d\text{ 期間中に }c\text{ 国で発信された }l\text{ 言語における総単語数}} \quad (1)$$

これにより，任意の国における任意の期間での言及言語ごとの特徴語を抽出でき，例えば，フランスにおけるイタリア人（イタリア語）やドイツ人（ドイツ語）の特徴語（認知特性）を抽出できる．

2.4　言及言語に基づく典型的なお土産品抽出

次に，多くの場所で話題の特徴語 i を言及言語ごとの典型的な特徴語として，以下の DF を算出する．

$$\frac{\text{単語 } i \text{ の出現した期間数} \times \text{国数}}{D\text{ 期間} \times C\text{ 国総数} \times L\text{ 言語総数}} \quad (2)$$

前節より抽出した単語 i に上記の算出値を乗算し，その場所の郷土品のうち，言及言語の群衆にとって典型的なお土産品として推薦する．

さらに，郷土品やお土産品に関するツイートの少ない場所に関して，抽出された典型的なお土産品名を用いて，「場所名」＋「典型的なお土産品名」の検索結果のうち検索ヒット数の多い順にランキングし，推薦可能にする．これにより，例えばヨーロッパにおける日本人に喜ばれるお土産をどこでも推薦できる．

2.5　言及言語に基づくセンチメント付与と相違に基づく特徴語抽出

前節より，任意の場所における任意の国民にとって話題性の高い特徴語（人気の郷土品）を抽出した．しかしながら，この郷土品（特徴語）がその国民にとって喜ばれるものかまでは考慮できていない．そこで，各郷土品に対する国民ごとのセンチメント（Positive/Negative）値を付与する．

まず，ツイートに対するセンチメントの分類器を作成する．今回は，Stanford の 1.5 百万のツイートデータセット (注2) を用い，TFIDF 値を算出し，BoW（Bag of Words）を作成し，ニューラルネットワーク（75%の適合率）を用いて分類器を作成する．この分類器を用いて，各ジオタグツイートのセンチメントを判定する．

各ツイートのセンチメント値を用いて，前節で抽出した特徴

(注1)：https://dev．twitter．com/streaming/overview

(注2)：http://help.sentiment140.com/for-students/

表 1　ツイートストリーミングデータ

開始日時	経過日時	ツイート数	量 [KB]	日本語（%）
2016/04/01	2016/10/31	12,958,029	3,536,999	25,882（0.19%）

語を含む任意の場所の任意の言及言語のツイート集合のセンチメント値の相加平均を算出し，特徴語に対するセンチメントを算出する．

さらに，下記の場合における特徴語のセンチメント値の相違に基づき，特徴語を選別する．

- 同一の場所における異なる言語間の相違
- 異なる場所における同一言語間の相違

各場合の相関係数が高いほど，典型的な特徴語とし，また，低いほど特異な特徴語として選別する．

3. 実　　験

本研究では，言語形態を可視化することを目的としており，多様な言語が使用されている欧州を対象とする．本論文では，図 3(a) における 2016 年 4 月 1 日から 10 月 31 日の約 6ヶ月間，欧州の 32ヵ国に対するツイートを収集し，特に，図 3(b) から (h) の 7 都市において，表 1 に示す欧州における 6ヶ月間のツイートのうち，言及言語を日本語とした 25,882 ツイート（全体の約 0.19%）を検証対象とした．

3.1　欧州におけるツイートの特徴

本節では，取得した欧州におけるツイートの言及言語の特徴についてまとめる．

まず，ツイートで発言された言及言語は 53 種類であった．これら言及言語をツイート数の多い順にランキングすると，1 位が英語で半数を占め，次いで，スペイン語で約 13%，3 位のフランス語以下では数%であった．

また，下記の 2 種類をノイズとして除去した．

- 3ヶ月間で発信したツイートが 2 ツイート以下のユーザが発信したツイート
- bot ツイート：今回，3ヶ月間で発信したツイート数でランキングを作成し，上位 300 アカウントを bot とみなしノイズとして除去した．閾値は，100 アカウントごとに上位 10 アカウントを 2 名による目視により判定し，6 割の bot アカウントを含む 300 とした．なお 400 アカウントでは bot が 2 アカウント含まれていた．

ノイズを削除したツイートから取得できた言及言語は 49 種類となり，ノイズ除去前から減少した 4 言語は，マラーティー語，タミル語，パンジャーブ語，シンド語であった．これらをノイズ除去前同様，ツイート数の多い順に言及言語ごとにランキングした結果 1 位は英語で半数を占め．次いで，スペイン語で約 14%，3 位にフランス語で約 7 %となり，ノイズ除去前と同程度であったが，4 位以下では順位がドイツ語からイタリア語と順位が入れ替わった．オランダ語（6 位）とポルトガル語（7 位）はノイズ除去前同様の順位であった．以上のことから欧州における公用語は英語と言える．

3.2　欧州におけるユーザのツイートにおける多言語性

ユニークユーザ総数は 614,292 人であった．これらユーザの他言語性を言及言語から検証した．各週ごとに，Monolingual（単一言語話者）と bilingual（二言語話者）以上のユーザ数を取得した．その結果，最初の 17 週目を除いて単一言語話者が平均 83.4%を占めており，二言語話者以上は 16.6%となるが，そのうち二言語を使用しているユーザが平均 12.5%となった．前節の言及言語の公用語とそれ以外の差異を考慮すると，ツイートの内容や発信するときの状況によって公用語と母国語を使いわけている可能性を示唆しているといえる．

3.3　各都市における特徴語検証

本説では，ユーザ群の各場所ごとに抽出された特徴語を検証する．本稿では，ヨーロッパにおける公用語である英語の利用率の低い日本人を対象とし，言及言語を日本語とした．

表 2 に，抽出結果を示す．上段が各都市ごとで抽出した特徴語のうち出現頻度の高い上位 10 単語であり，下段は全体の都市における出現頻度を積算した共通性を考慮したリランキング結果である．また、下線の特徴語は，他の都市に出現していないその都市特有の単語である．

結果より，都市特有の単語は上段の共通性を考慮しない手法にのみ出現していることが分かる．また，11 位以下の上位 30 位では，パリでは 9 位の「フレンチ」以外に「ガレット」や「モノプリ」が抽出されており，ロンドンでは 5 位の「マーケット」と 10 位の「ジャック」以外に,「ケーキ」といった単語が抽出された．これらはその都市特有の特徴語と言え，現地ならではの郷土品を抽出できた．また，上下段両方でベルリンを除いて 1 位が「写真」であったことから，ベルリンの「ビール」は代表的な郷土品であることが明らかである．一方で，ベルリンやフィレンツェの「リップ」は郷土品として抽出されたが，共通性が低いため 11 位以下となり，お土産として推薦できなかった．

全都市の共通性を考慮した類似度結果となる下段は，全ての都市において 1 位は変化がなく，ロンドンとローマの 2 位と 3 位の結果が入れ替わるといったように，1 段階程度で順位変異が見られた．また，パリの「ソース」，ロンドンの「チョコレート」，ローマの「TEA」，ベルリンの「クリーム」「パスタ」といった特徴語がリランキングより新たに推薦された．

4. 関 連 研 究

大量のジオタグツイートの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 [1] は，レストラン等の特定の店舗で Check-in した際に発信されるジオタグツイートを分析し，ユーザの移動軌跡を抽出し，その店舗等のトレードエリアの発見を行った．また，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．さらに，自然災害や疾病の流行を検出する手法 [3] や，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 [4].

これまで著者らも，ユーザ行動分析として日本および米国の

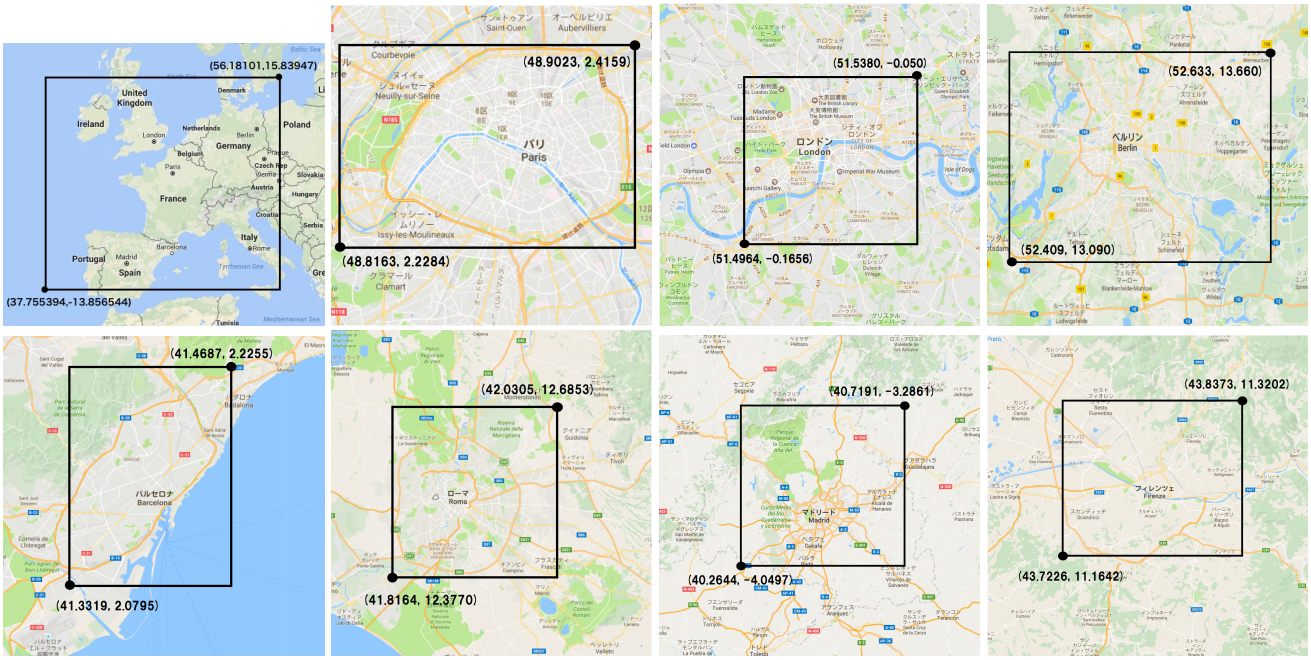

図 3　(a) ツイート取得対象領域（全体)、(b) パリ, (c) ロンドン, (d) ベルリン, (e) バルセロナ, (f) ローマ, (g) マドリッド, (h) フィレンツェ, における対象領域

表 2　各都市で日本語におけるツイートから抽出された郷土品ランキング結果（上段：共通性の考慮無し，下段：共通性の考慮有り）

都市	共通性	抽出された特徴語（上位 10 件）
(b) パリ	無	写真，チーズ，TEA，ティー，クリーム，コーヒー，ビール，ワイン，フレンチ，酒
	有	写真，チーズ，TEA，ティー，クリーム，ビール，コーヒー，ワイン，酒，**ソース**
(c) ロンドン	無	写真，酒，TEA，ティー，マーケット，ビール，コーヒー，紅茶，クリーム，ジャック
	有	写真，TEA，酒，ティー，ビール，コーヒー，クリーム，紅茶，**チョコ**，**チョコレート**
(d) ベルリン	無	ビール，写真，スープ，コーヒー，クリーム，パスタ，酒，菓子，リップ，ミルク
	有	ビール，写真，スープ，コーヒー，酒，菓子，**TEA**，**ティー**，**クリーム**，**パスタ**
(e) バルセロナ	無	写真，ビール，コーヒー，ワイン，生ハム，土産，オイル，TEA，スープ，チョコ
	有	写真，ビール，コーヒー，ワイン，生ハム，土産，TEA，菓子，酒，オイル
(f) ローマ	無	写真，パスタ，ティー，コイン，ビール，ワイン，クリーム，レモン，チーズ，コーヒー
	有	写真，ティー，パスタ,，ビール，ワイン，コイン，クリーム，チーズ，**土産**，**TEA**
(g) マドリッド	無	写真，生ハム，ティー，酒，ビール，土産，チーズ，ワイン，TEA，クリーム
	有	写真，生ハム，ティー，ビール，酒，土産，ワイン，チーズ，TEA，クリーム
(h) フィレンチェ	無	写真，ワイン，ティー，パスタ，コーヒー，ビール，チョコ，TEA，チョコレート，リップ
	有	写真，ワイン，ティー，ビール，パスタ，コーヒー，TEA，チョコ，チョコレート，**土産**

数ヶ月間のジオタグ付ツイートデータを分析し，データ発生位置とコンテンツ内容位置との差異，発生時間とコンテンツ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 [5] [6].

しかしながら，既存研究を含め，ジオタグの時間と場所，コンテンツの時間と場所に加え，言語形態を考慮した時空間における言語特性分析に関する研究は稀である.

また，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の収集の提案 [10] 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訪問目的を推定する手法の提案 [11] などの研究が行われている.

本研究では，ジオタグの場所と時間情報に加え，言語形態を含めた群衆における認知特性抽出および，認知特性に基づき，旅行先における郷土品推薦を行う点が，既存研究との特異点である.

5.　まとめと今後の課題

本論文では，ユーザ行動に対する認知特性の解明を目指し，ユーザ行動に対する認知特性として言語形態に着目し，任意の発信位置と時刻における言及言語に対する特徴語を言語（国）ごとの認知特性として抽出した．特に，本稿ではお土産推薦システム構築を目指し，任意の国における異なる国のユーザにとっての郷土品推薦，また，典型的な特徴語抽出によるユーザ（国）にとっての典型的なお土産品を考慮することで，郷土品

の中でもユーザに喜ばれるお土産品の推薦手法を提案，実験検証した．

今後，提案したセンチメント値に基づく特徴語抽出の検証，ならびに他言語との相関分析を行う予定である．

謝　　辞

本研究の一部は，JSPS 科研費 16H01722，15K00162 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] Y. Qu, J. Zhang: Trade Area Analysis using User Generated Mobile Location Data, WWW2013, pp. 1053-1064 (2013).

[2] J. Yuan, Y. Zheng, X. Xie: Discovering Regions of Different Functions in a City Using Human Mobility and POIs, KDD2012, pp. 186-194 (2012).

[3] T. Sakaki, M. Okazaki, Y. Matsuo: Earthquake shakes Twitter users: real-time event detection by social sensors, WWW2010, pp. 851-860 (2010).

[4] 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, SIGSPATIAL2014, pp. 163-172 (2014).

[5] S. Wakamiya, A. Jatowt, Y. Kawai, T. Akiyama: Analyzing Global and Pairwise Collective Spatial Attention for Geo-social Event Detection in Microblogs, WWW2016, pp. 263-266 (2016).

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

[7] D. Mocanu, A. Baronchelli, N. Perra, B. Gonçalves, Q. Zhang, A. Vespignani: The Twitter of Babel: Mapping World Languages through Microblogging Platforms, PLoS ONE 8(4): e61981 (2013).

[8] G. Neubig, K. Duh: ツイートの情報量について－情報理論に基づく多言語調査－，言語処理学会第 20 回年次大会発表論文集 (2014).

[9] 岡山愛，河合由起子，Muhammad Syafiq Mohd Pozi，Adam Jatowt: ツイート多言語分析に関する一検討，WebDBForum (2016).

[10] 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，WebDBForum (2016).

[11] 野沢悠哉，遠藤雅樹，江原遥，廣田雅春，横山昌平，石川博: マイクロブログを用いたユーザの訪問目的と動向の推定，WebDBForum (2016).