6K-08

ジオタグツイートの多言語分析に基づく認知特性可視化システムの提案

岡山愛[†1]　河合由起子[†1]　*AdamJatowt*[†2]　秋山豊和[†1]

†1 京都産業大学　†2 京都大学

1　はじめに

近年，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によるユーザの振る舞い分析技術が，国内外で広く注目されいる．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 センサデータを含めて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た，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日本および米国の数ヶ月間のジオタグ付ツイートデータを分析し，データ発生位置と言及されている場所との差異，発生時間と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の分析および可視化に関する研究を行ってきた [3].

本研究では，ジオタグ付ツイートデータを時間と場所と言語に基づき分析し，ユーザ行動に対する認知特性の解明を目指す．特に，認知特性の一要素として発信場所と言語形態の相違に着目し，多言語である欧州のツイートを対象とし，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，まず，ツイートで言及した言語を言及言語として抽出する．次にツイートの発信位置の緯度経度より，発信国を特定し，発信国ごとの各言及言語における TF-IDF 値を算出し，特徴語として抽出する．本稿では，発信国における言及言語（各国）の特徴語の相関から任意の国に対する各国の認知特性について考察する．

2　認知特性可視化システム

本研究は，ツイート発信位置，母国語，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抽出し，可視化を目指す．本研究における認知特性とは，時空間における言語形態の違いから抽出される特徴語である（図 1）．具体的には，異なる場所で異なる言語間の発話内容の違いであり（例えば，京都弁のユーザが東京で発信する標準語と京都弁の内容の違い），抽出された言語に対する認識が異なる（例えば，京都出身ユーザにとって東京では「東京スカイツリー」が重要）．

図 1: 時空間における言語形態の違いによる認知特性

2.1　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 [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．

2.2　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，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，母国語と言及言語の相違に基づき特徴語を抽出する．

前節より取得した言及言語より，日時，発信国，言及言語の相違を考慮した単語 i の重要度は，下記の $TFiDF$ 式より算出する．

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

これにより，任意の国における任意の期間での言及言語の相違による特徴語を抽出でき，例えば，フランスにおける英語（公用語）やドイツ語の特徴語を提示できる．なお，欧州において約 50%と最も利用されている言及言語は英語であり，公用語（標準語）と言え

A Proposal of Crowd cognitive visualization system based on Multilingual Analysis of Geotagging Tweet
†1 Ai OKAYAMA　†1 Yukiko KAWAI　†2 Adam JATOWT　†1 Toyokazu AKIYAMA
†1 Kyoto Sangyo University
†2 Kyoto University

[1] https://dev.twitter.com/streaming/overview

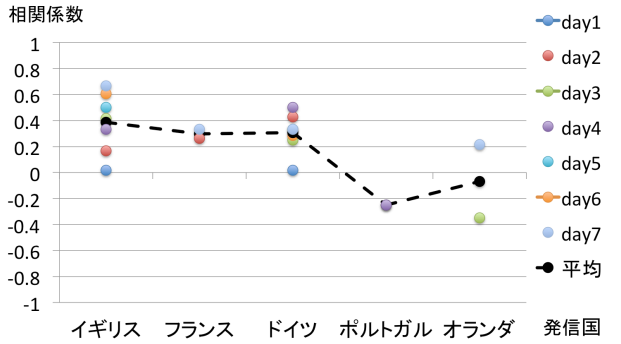

図 2: ドイツ語とフランス語における特徴語の順位相関

る [4]．また，母国語の相違による，日時，発信国，母国語の相違も式（1）同様に算出される．これにより，例えば，フランスにおけるイタリア人やドイツ人の特徴語（認知特性）を提示できる．

さらに，任意の国における任意の期間での言及言語が公用語の単語 i の重要度から，言及言語が母国語の単語 i の重要度の差分を算出し，正の場合（公用語＞母国語）と負の場合において重要度の重みを増減する．これにより，よりユーザの認知特性を反映した特徴語抽出が可能となる．

3 多言語分析に関する検証

発信国における言及言語および母国語の違いによる特徴語抽出結果を検証する．2016 年 4 月 30 日から 1 週間分の欧州の 36 万ツイートのうち，発信された主要 10ヵ国約 16 万ツイートを検証対象とした．抽出された言及言語，母国語種類はいずれも約 50 カ国分であった．

3.1 言及言語の違いによる特徴語検証

2.2 の式 1 を用いて母国語における特徴語抽出を行った．本稿では，そのうちの母国語がフランス語とドイツ語との相関を検証する．図 2 は，各発信国における（横軸）各日（7 日間）での，フランス語とドイツ語のスピアマンの順位相関結果である．順位相関は，$TFiDF$ 値をもとにランキングしており，1 だと単語間の類似性が高く，-1 だと単語間の類似性が低い．まず，フランスとポルトガルは日に関わらず相関結果が分散していないが，イギリス，ドイツ，オランダは日によって結果が分散していた．次に，イギリス，フランス，ドイツは相関係数の平均値が 0.3 から 0.4 で正の相関であったが，ポルトガルとオランダは平均値が 0 以下と負の相関であった．全体としては，0.4 を下回っており，相関は低いといえ，言及言語の違いによって抽出される単語の多様性が明らかとなった．

3.2 発信国の相違による特徴語検証

2.2 の式 1 を用いて言及言語における特徴語抽出を行った．本稿では，そのうちの言及言語が英語とドイツ語を検証する．表 1 に，各発信国の 7 日間における特徴語のうち，フランスとベルギーの上位 5 位（$TFiDF$）を示す．母国語同様に，英語とドイツ語の特徴語の重複は少なかった．また，ドイツ語において，フランスでは，Eiffel，Tour（エッフェル塔）や Betschdorf(ベッチドルフ)，ベルギーでは，Antwerpen（アントウェルペン），Limburg（リンブルフ），Schorre（De Schorre という公園），Hasselt（ハッセルト）など発信国ごとの観光地名が英語（公用語）と比較して多く特徴語として抽出された．

表 1: 英語とドイツ語の特徴語（上位 5 位）

発信国	言及言語	特徴語
フランス	英語	Paris，France，Land-Paris，Valbonne，#grenoble，Dauphin，Rouen，Disney，#IHeartFrance，**Eiffel**，Lac
フランス	ドイツ語	**Eiffel**，**Tour**，JED，Betschdorf，Dieppe
ベルギー	英語	**Antwerpen**，Brussels，Belgium，Brouwerij，Sfinks
ベルギー	ドイツ語	**Antwerpen**，Limburg，Schorre，Hasselt，Heusden-Zolder，#einfachso，Martens，#renntv

以上より，言及言語と母国語，公用語ごとの特徴語から，場所と時間と言語形態に基づき抽出した特徴語はユーザ認知特性の一つとして有用であると言える．

4 まとめ

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語の相違，母国語と言及言語との差異，発信位置と各言語の発祥場所（母国語）との差異，さらに発信位置と言及言語との差異を抽出し，場所や時間における各出身地ごとの言語形態を分析・可視化し，検証した．その結果，言及言語や母国語の違いにより抽出できる特徴語の多様性が示された．今後，抽出した特徴語を用いて認知特性に基づく推薦システムを構築し，検証する．

謝辞

本研究の一部は，JSPS 科研費 16H01722，15K00162 および総務省戦略的情報通信研究開発推進事業 SCOPE（150201013）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] Qu et al.: Trade Area Analysis using User Generated Mobile Location Data, WWW2013 (2013).

[2] Yuan et al.: Discovering Regions of Different Functions in a City Using Human Mobility and POIs, KDD2012 (2012).

[3] Émilien Antoine, Adam Jatowt, Shoko Wakamiya, Yukiko Kawai, and Toyokazu Akiyama.: Portraying Collective Spatial Attention in Twitter, KDD 2015, pp. 39-48, Sydney, Australia, August (2015).

[4] 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).