ジオタグツイートの文体分析に基づく話題抽出と可視化

深谷　大樹† 　有馬　直也†† 　河合由起子† 　湯本　高行†††

† 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
†† 兵庫県立大学工学部　〒 671–2201 兵庫県姫路市書写 2167
††† 兵庫県立大学大学院工学研究科　〒 671–2201 兵庫県姫路市書写 2167
E-mail: †{g1445133,kawai}@cc.kyoto-su.ac.jp, ††eo13t002@steng.u-hyogo.ac.jp,
†††yumoto@eng.u-hyogo.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、任意の場所における地域性に基づいた話題および信頼性の高い話題発見を目指す。本論文では、日本国内のジオタグツイートを対象とし、群衆を 8 地域分ごとの地方単位とし、地域ごとの話題に対する認知特性を抽出する。抽出手法は、まず、新聞の文体分析より作成した辞書に基づき標準語とそれ以外に分別し、標準語以外に対して方言判定し、ツイートの言及言語として地域名を付与する。次に、任意の場所で発言されたツイートを付与された言及言語の地域ごとに特徴語を抽出し，その場所に対する地域ごとの話題として提示する。さらに，同じ場所において標準語に対する特徴語を抽出し，標準語と方言ごとの特徴語の相関性を考慮し、地域ごとの嗜好性の高い話題を発見する。これによりツイートの少ない地域においても地域性に基づいた話題発見が可能となる。本論文では、ジオタグツイートの文体分析に基づく群衆の認知特性となる話題を抽出し，実験検証する．

キーワード　文体分析，ジオタグツイート分析，認知特性抽出，可視化

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．

本研究は，ジオタグ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、本研究では、ユーザ行動に対する認知特性として，言及言語に着目する．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，日本国内の日本語ジオタグツイートを対象に標準語と方言に分類し，ツイートの発信場所における言及内容の言語ごとに特徴語を抽出する．具体的には，任意の発信位置と時刻における言及言語の差異（例えば北海道における九州弁と関西弁）を抽出し，場所や時間における各出身地ごとの認知特性として検出する．これにより，ユーザが特定の場所でツイートを発信する際に，いずれの言語を選択して何の話題について言及しているかが明らかとなり，ユーザ行動との関連性や各言語ごとの特性抽出につながる．さらに，標準語との相関性（例えば北海道における九州弁と標準語）を抽出することで，より九州地方の特異な話題を発見できる．

提案手法では、まず、新聞データの文末表現を分析し、辞書を作成する。作成した辞書に基づき標準語とそれ以外に分別し、さらに、国内を北海道地方から九州地方までの 8 地域単位とした方言辞書を作成する。方言辞書を用いて、標準語以外に対して各 8 地域に基づきツイートに対して地域名を付与する。そして、任意の場所に任意の期間に発言された地域ごとの言及言語集合から、特徴語を TF・iDF 値の高い順に抽出し、場所に対する各地域の嗜好性の高い話題である群衆の認知特性として推薦する。また，標準語に対して，抽出された特徴語のみに対する TF 値を抽出し，シグモイド関数の重みとして算出し，地方ごとの特徴語とする．さらに、他の場所でも同様に特徴語を抽出し、それら特徴語の類似性を算出することで、地域性による嗜好性の高い話題等を群衆の認知特性として抽出する。これによりツイートの少ない場所においても地域性に基づいた話題発見も可能となる。

本論文では、ジオタグツイートの文体分析に基づく群衆の認知特性抽出を構築し、検証する。

2. 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].

図 1　時空間における言語形態の違いによる認知特性

図 2　ここにフローチャート

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。

3.　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）.

提案システムでは、まず、各ツイートの方言を判定する。方言判定手法は、新聞データベースより文体表現を分析し、標準語と非標準語を判定する。これにより標準語、非標準語辞書を作成し、ツイートを判定する。さらに非標準語と判定されたツイートに対して、8 地域単位とした方言辞書を用いて判別し、言及言語として各ツイートに付与する。次に、ジオタグの発信位置、発信時刻、言及言語から時空間と言語形態に基づいた話題を抽出する。具体的には、任意の場所に任意の期間に発言された地域ごとの言及言語集合から、特徴語を TF より抽出し、推薦提示する。さらに、他の発信場所でも同様に特徴語を抽出し、それら特徴語の類似性を抽出することで、地域性による嗜好性の高い話題を群衆の認知特性として抽出する。

4.　文体分析と話題抽出

本説では、ジオタグツイートの時空間および言語形態の相違に基づく話題抽出および可視化システム構築手法について述べる。図 2 に処理の流れを示す。

4.1　文 体 分 析

本研究では，時空間である発信場所と発信日時ごとの相違に基づく特徴語抽出に加え，言及言語の相違に基づき特徴語を抽出するため，時空間と言及言語の多様性が重要となる.

まず，ツイートの文体が標準語か否かを分類するための、標準語に対する文末表現辞書を構築する。また、ツイートを形態素解析し、抽出した文末表現がこの辞書に含まれていた場合は標準語のツイートに、含まれていなければそれ以外のツイートに分類する。本手法では文末表現を、文末が「動詞」、「形容詞」、「判定詞」、「助動詞」、「助詞」(接続助詞、終助詞)、「接尾辞」(形容詞性述語接尾辞、動詞性接尾辞) のみで構成される表現と定義する。

標準語に対する文末表現辞書の構築は、新聞記事に対して形態素解析を行い、文末表現を抽出する。その際、文末の形態素が 1 文字のとき、(a) その直前の形態素が名詞なら 1 文字のまま抽出し、(b) そうでなければ直前の形態素と合わせて抽出する。(c)2 文字以上の場合もそのまま抽出する。たとえば、「これ/は/いい/記事/だ」では、判定詞「だ」の直前が名詞「記事」であるため「だ」のみを文末表現として抽出する。「これ/は/おいしい/です/ね」では、終助詞「ね」の直前が動詞「です」であるためそれらを合わせて「ですね」を文末表現として抽出する。「雨/が/降る/らしい」では、助動詞「らしい」は 3 文字なのでそのまま抽出する。また、新聞記事内にも「～るわ」や「～わい」のように標準語でない表現も存在する。これを除去するため、抽出した文末表現のうち出現頻度の高い表現のみを使用する。

2014 年 1 月 1 日～20 日の毎日新聞の記事 4445 件に対して上記の手法により文末表現辞書を構築した。なお、形態素解析には Juman を用い、予備実験で最も精度の良かった出現頻度 21 以上の表現を辞書に使用した。その結果、辞書内には 169 種類の文末表現が得られた。抽出した文末表現の例を表 1 に示す.

表 1　抽出した文末表現例

文末表現	出現頻度	文末表現	出現頻度
した	3837	だ	947
いる	2985	ある	944
ない	1430	いう	935
する	1098	れた	817
いた	995	だった	775

この辞書を用いてツイートの文体の分類を行う。具体的には、ツイートに対しても形態素解析を行い、同様の文末表現を抽出し、文末表現辞書内に存在するかどうかで分類を行う。ただし、文末が「名詞」で終わっているツイートについては体言止めとして標準語のツイートに分類する。また、ツイートには複数の文が存在している場合もある。そのような場合には、文ごとに文末表現を抽出し、すべての文が標準語であると判断された場合のみ、そのツイートを標準語のツイートに分類する。

さらに、非標準語のツイートに対して、方言辞書を用いて方言を判定する。方言辞書は、辞書サイト Weblio の方言辞書 (注1) を用いて、8 地域（北海道、東北、関東、中部、近畿、中国、四

(注1)：http://www.weblio.jp/cat/dialect

表 2　方言辞書に登録された方言例

地域名	方言
北海道	なまら, あずましい, わやくちゃ, おだつ, しばれる
東北	っぺ , がおる, しったげ, くっちゃべる, おしょすい
関東	ござんす, さむしー, だすけ, ばーか, だでね
中部	ずら, えらかった, まんず, おぞい, ぼっけえ
近畿	ほな, あんごー, しはる, けったいな, おおきに
中国	ぶち, おぞい, ぼっけえ, やねこい, たいぎい
四国	けんど, うちんく, おもっしょい, ひやこい, かいね
九州	ばい, 何ば, きつか, したとね, よだきー

国、九州）分を作成した。表 2 に方言辞書例を示す。

4.2　ツイートの場所と時刻と言語の付与

まず，日本国内のジオタグストリーミングツイートを The Streaming APIs (注2) を用いて、指定地域から重複を除いて取得する．指定地域は，1 度以上異なる南西および北東を指定することで，その 2 点に囲まれた矩形領域のストリーミングツイートを取得できる．次に取得したジオタグツイートの言及内容に対して、形態素解析し、名詞と形容詞を取得する。また、前説の辞書を用いて、言及内容に対して標準語、方言を判定し、言及言語として付与する。

以上より，ユーザ ID，緯度経度，発信時刻，言及内容，言及言語を管理する．

4.3　発信場所における言及言語に基づく特徴語抽出

本研究では，時空間である発信場所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 $TF \cdot iDF$ 式より算出する．

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

これにより，任意の場所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，北海道における九州弁や関西弁ごとの特徴語を抽出できる．

4.4　標準語における特徴語の出現頻度を考慮した特徴語抽出

各地域ごとに標準語のツイートに対して式（1）を用いて単語 i における重み x として算出し，下記の式（2）を用いて，任意の地域に対する特徴語として抽出する．

$$TF_i \cdot iDF \cdot \left(1 - \frac{1}{1+e^{-x}}\right) \quad (2)$$

5.　実　　装

本研究では，時空間および言及言語の認知特性として特徴語を抽出し、可視化する。今回，実装で用いた日本語ツイートは、

(注2)：https://dev.twitter.com/streaming/overview

図 3　実験で収集したジオタグツイートの取得領域

表 3　ツイートストリーミングデータ

場所	開始日時	経過日時	ツイート数 [個]	量 [GB]
日本	15-07-13	16-10-01	360,000,000	9

表 4　国内における方言を用いたツイート数と割合

地域名	総数	割合
標準語	1,653,122	6.0
北海道	277,572	1.0
東北	136,050	0.35
関東	104,612	0.5
中部	12,424	0.05
近畿	47,917	0.2
中国	351,646	1.2
四国	100,694	0.4
九州	75,074	0.3
計	2,759,111	10.0

図 3 に示す領域の本研究では，時空間および言及言語の認知特性として特徴語を抽出し，可視化する．今回，実装で用いた日本語ツイートは，図 3 に示す領域の 2015 年 07 月 13 日から 2016 年 10 月 01 日の計約 15ヶ月間の約 360 万ツイートを対象とする（表 3）．

5.1　言及言語の分類結果

本節では，日本国内におけるツイートの言及言語の多様性について検証する．表 4 に取得した標準語と各言及言語の種類を示す．表より，ツイートには多様な方言による記述がなされていることが明らかとなった．今後，特定の場所における方言の利用と特徴語抽出を行う予定である．

6. まとめと今後の課題

本研究では、任意の場所における地域性に基づいた話題発見を目指し、ツイートの発信位置と言及言語から特徴語を抽出し、群衆の認知特性として抽出した。提案手法では、新聞の文体分析より作成した辞書に基づき標準語とそれ以外に分別し、標準語以外に対して方言判定し、ツイートごとに地域名を付与した。また、任意の場所で発言された言及言語の地域ごとに特徴語を TF・iDF より抽出した。実験では、ジオタグツイートの文体分析に基づく群衆の認知特性抽出を構築し、抽出した特徴語の相関性の低さから、特徴語の多様性が明らかとなった。

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] 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] Sakaki et al.: Earthquake shakes Twitter users: real-time event detection by social sensors, WWW2010 (2010).
[4] Magdy, A., Alarabi, L., Al-Harthi, S., Musleh, M., Ghanem, T. M., Ghani, S., and Mokbel, M. F.: Taghreed:A System for Querying, Analyzing, and Visualizing Geotagged Microblogs, SIGSPATIAL 2014, pp. 163-172 (2014).
[5] Shoko Wakamiya, Adam Jatowt, Yukiko Kawai and Toyokazu Akiyama.: Analyzing Global and Pairwise Collective Spatial Attention for Geo-social Event Detection in Microblogs, WWW 2016, ACM Press, Montreal, Canada, demo paper pp. 263-266 (2016).
[6] 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).
[7] Delia Mocanu, Andrea Baronchelli, Nicola Perra, Bruno Gonçalves, Qian Zhang, AlessandroVespignani.: The Twitter of Babel: Mapping World Languages through Microblogging Platforms, PLoS ONE 8(4): e61981. doi: 10.1371/journal.pone.0061981.
[8] Graham Neubig, Kevin 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).