

(75) マイクロブログの解析による 道路事業に係わる情報抽出の研究

今井 龍一¹・中村 健二²・田中 成典³・藤本 雄紀⁴

¹正会員 東京都市大学准教授 工学部 (〒158-8557 東京都世田谷区玉堤1-28-1)

E-mail: imair@tcu.ac.jp

²正会員 大阪経済大学准教授 情報社会学部 (〒533-8533 大阪市東淀川区大隅2-2-8)

E-mail: k-nakamu@osaka-ue.ac.jp

³正会員 関西大学教授 総合情報学部 (〒569-1095 大阪府高槻市霊仙寺町2-1-1)

E-mail: tanaka@res.kutc.kansai-u.ac.jp

⁴非会員 関西大学大学院 総合情報学研究科 (〒569-1095 大阪府高槻市霊仙寺町2-1-1)

E-mail: k398881@kansai-u.ac.jp

公共事業の評価は、様々な分析や意見聴取などから定量的・定性的に行う必要がある。道路事業では、道路交通分析を実施して交通実態を定量的に把握されており、そこに国民の意見をきめ細やかに反映できると、道路事業評価の高度化が期待される。国民意見の収集にマイクロブログの活用が有効策として挙げられるが、投稿された意見が示す対象地点の特定が難しい。

本研究では、代表的なマイクロブログのTwitterのツイートと、既存の道路交通データとを組み合わせた道路事業に係わるツイート抽出のテキストマイニング手法を考案し、その有用性を検証した。

Key Words : microblog, road contract, information extraction, text mining

1. はじめに

公共事業は、各事業特性に応じた分析結果や国民の意見聴取結果などに基づいて、定量的・定性的に評価する必要がある。道路事業に着目すると、交通量や民間プローブデータを活用した旅行速度などを用いて道路交通分析を実施して交通円滑性や定時性向上の効果などを把握している。これらの調査・分析結果から得られた定量的な交通実態に対し、国民（道路利用者や周辺住民など）からの意見を反映できると、道路事業評価の高度化を図る支援策となることが期待される。

ここ数年を振り返ってみると、SNS（Social Networking Service）の浸透が目覚ましく、その蓄積されたデータは“ビッグデータ”と称され、様々な分野で活用が期待されている。SNSの一種であるマイクロブログのTwitterに着目すると、つぶやき（以下、ツイート）数は1日7,000万件とも言われている。ツイートは、大規模な災害対応のみならず、電車遅延などの日々の異常事象の情報共有でも活用されている。このツイートの中には、道路事業に対する国民の意見が潜在している可能性が極めて高い。

この動向を踏まえて、既存研究では、膨大なツイートに対してテキストマイニングを適用して道路事業評価に関するツイートの抽出が試行^{2,4}されている。その結果、道路事業に係わるツイートが存在し、日常生活に浸透している高速道路や国道の路線名のツイートは比較的抽出しやすいが、具体的な道路の区間や位置の特定は難しいことが明らかにされている。Twitterでは経緯度を付与したツイートを扱えるが、利用率は1%以下⁵となっている。また、ツイートに含まれる地域識別子を元にした位置推定手法も研究⁶されているが、地域識別子の含まれないツイートが多く、適用条件が限定される課題がある。

本研究は、Twitterのツイートと既存の道路交通データとを組み合わせた道路事業に係わるツイート抽出のテキストマイニング手法を提案する。具体的には、まず、膨大なツイートから高速道路に係わるツイートを収集する。次に、各ツイートの単語の出現頻度に基づき特徴量を算出する。最後に、クラスタリング手法を用いて交通現象を示すツイートと、道路利用に係わる意見のツイートとに分割する。

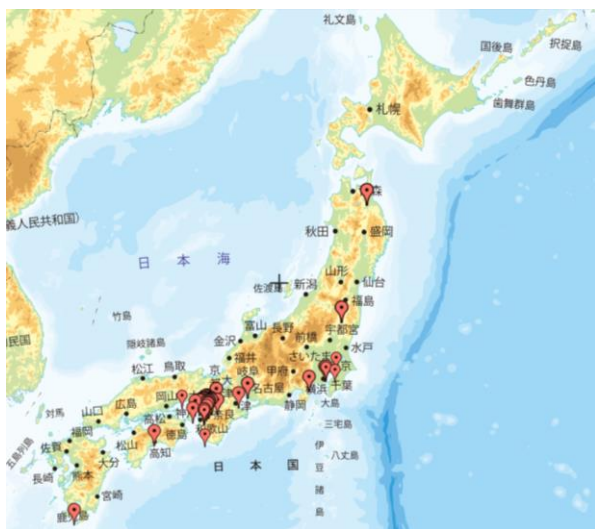


図-1 収集した 192 ツイートの地図への描画結果
(出典：国土地理院数値地図 25000 (2015 年 6 月))

2. 道路事業に関するツイートの収集

本章では、Twitter社のAPI⁷⁾を用いて、道路事業に関するツイートを収集する際に必要となる検索キーワード集（地名辞典）を作成する。次に、道路事業に関するツイートを収集し、そのツイートの位置情報を分析する。

(1) 検索キーワード集（地名辞典）の作成

道路事業に関するツイートを収集するには、当該事業に係わる検索キーワード集（地名辞典）が必要である。一般に道路交通に関する情報を含むツイートには、路線名、サービスエリア名やインターチェンジ名などの単語が含まれていることがある。また、これらの単語は略語や通称名などの類義語もよく利用されている。

本研究では、道路管理者が利用している単語や様々なWebサイト（SNS含む）で扱われている単語を収集対象とした。網羅的かつ自動的な単語収集の手法も確立すべき課題であるが、今回は試験的に近畿圏の高速道路および一般道を対象に手作業で単語を収集した。その結果、971の単語を検索キーワード集として収録した。

(2) ツイートの収集

本研究では、Twitter社のSearch API⁸⁾に、前項(1)で作成した検索キーワード集（地名辞典）を適用し、2014年8月24日から2015年6月28日の期間を対象にツイートを収集した。その結果、2,039,091件のツイートを収集できた。

次に、本研究では、この収集したツイートの位置情報の精度を確認する。今回は、既存研究で検索性の高いとされている高速道路のうち、管理区域が近畿圏内の阪神高速道路を対象とする。2014年8月24日から9月9日の間

のツイートの内、無造作に抽出した400ツイートをGoogle社のジオコーディングAPI⁹⁾を用いて地図に描画した（図-1参照）。その結果、次の2点の知見を得た。

一点目は、図に示すとおり、ツイートから適切な位置情報を取得できないことである。阪神高速道路に関するツイートであるにも関わらず、全国に拡散して描画されている。この原因の具体例を挙げると、「広島の実家に向かってののに阪神高速が渋滞だ」の目的地「広島」がジオコーディングの対象として処理されていた。

二点目は、道路事業に係わる多くのツイートには位置情報が含まれていないことである。今回対象とした400ツイートのうち、地図に描画できたのは192ツイート（48%）であった。したがって、多くのツイートは位置情報が含まれていないものと考えるのが妥当である。

以上のことから、収集した道路事業に関するツイートをそのまま利用すると、適切な位置を表現するのが難しい課題が明らかとなった。

3. 道路事業に関するツイートの抽出手法の考案

(1) 抽出手法の概要

第2章で得られた課題への解決策として、本研究では、ツイートに対して、道路管理者から一般提供されている道路交通データを組み合わせる手法（以下、本手法）を考案する。本手法を適用すると、位置情報の付与した交通現象を表すツイートと、道路利用に係わる意見のツイートとに分割できる。

(2) 抽出手法の手順

本手法の手順（図-2参照）は交通現象に係わるツイートの抽出機能、道路利用に係わる意見のツイート抽出機能と評価表現分析機能から構成される。

交通現象に係わるツイートの抽出機能は、道路交通データ（工事実績など）をキーワードとして、ツイートをクラスタ分割する。まず、道路管理者がWeb上で公開している事故、渋滞や工事規制のデータを収集し、それらが発生した時間帯と位置情報とを抽出する。次に、道路交通データから抽出した時間帯に近接し、かつ位置情報を含むツイート群を交通現象に関わる意見のツイートとして抽出する。この抽出したツイート群をMeCab⁹⁾を用いて形態素解析し、得られた各単語に対してtf-idf値を算出し、ツイート群の特徴ベクトルを生成する。

次に、道路交通データから抽出した時間帯に近接し、位置情報の含まないツイートを交通現象に関わるツイートクラスタに分割する。これらのツイート群も先述した手順で特徴ベクトルを生成する。そして、VSM (Vector

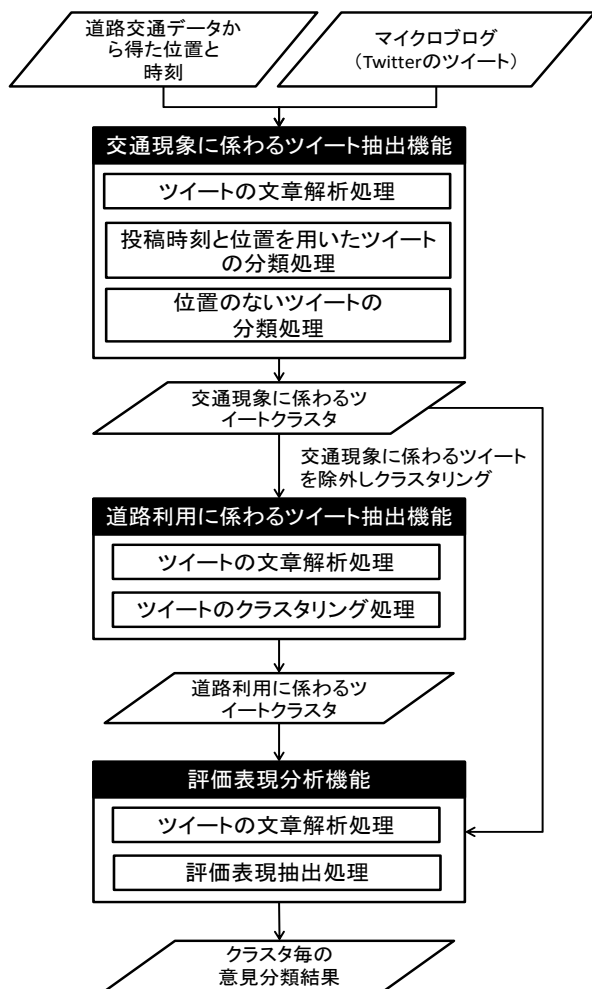


図-2 本手法の手順

Space Model)¹⁰⁾を用いて位置情報を含むツイート群の特徴ベクトルとの類似度を算出し、ある一定値以上の類似度のある場合は、そのツイート群に新規追加する。本研究では、この類似度の閾値をパラメータ α と設定する。

道路利用に係わる意見のツイート抽出機能は、前述の交通現象に係わるツイートの抽出機能にて抽出されなかったツイートを対象にk-medoids法¹⁰⁾を適用して、関係性のあるツイートをクラスタ分割し、道路利用に係わる意見として抽出する。なお、k-medoids法の適用の際に設定するクラスタの分割数をパラメータ β とする。

評価表現分析機能は、各機能によりクラスタ分割されたツイートに対し、既存研究の評価表現分析⁴⁾の手法を適用し、肯定的な意見と否定的な意見とに分割する。

4. 本手法の有用性の検証

(1) 検証内容

本研究では、第2章で収集したツイートに本手法を適用し、その有用性を検証する。そこでは、VSMとk-

medoids法を用いてクラスタ分割する手法（以下、既存手法）と、本手法を適用した結果とを比較する。

(2) 検証条件

(a) 検証対象の道路及び期間

本検証では、第2章(2)と同様、阪神高速道路のツイートを抽出対象とした。対象期間は、ツイートが通常よりも多くなり、肯定的・否定的な意見の遷移を確認できることが予想できる守口線フレッシュアップ工事の前後とした。具体的には、工事前の11月12日から終了1週間後の12月18日までの36日間とした。その内、工事期間は11月26日～12月4日の9日間であった。

(b) 検証データ

本検証では、第2章で収集した2,039,091件のツイートのうち、前項(a)の条件に該当する4,860件のツイートをを用いる。また、道路交通データは阪神高速道路の公式Webサイトで公開されている工事実績を用いる。

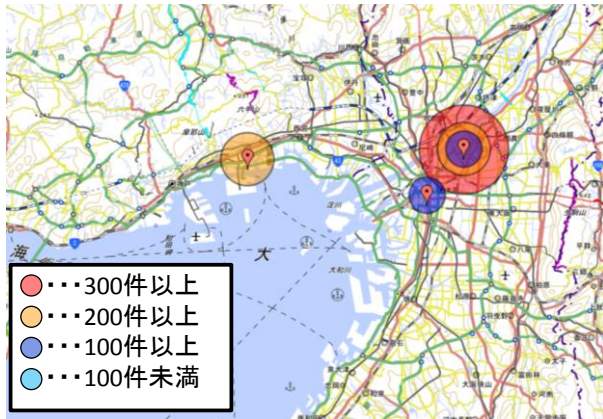
(c) 検証パラメータ

本検証では、交通現象に係わるツイートの抽出機能におけるパラメータ α を0.25とした。また、道路利用に係わる意見のツイート抽出機能におけるk-medoids法のパラメータ β を20とした。なお、 α 及び β は、検証作業の過程で調整し、一定の抽出結果が得られる上記の数値を採用した。

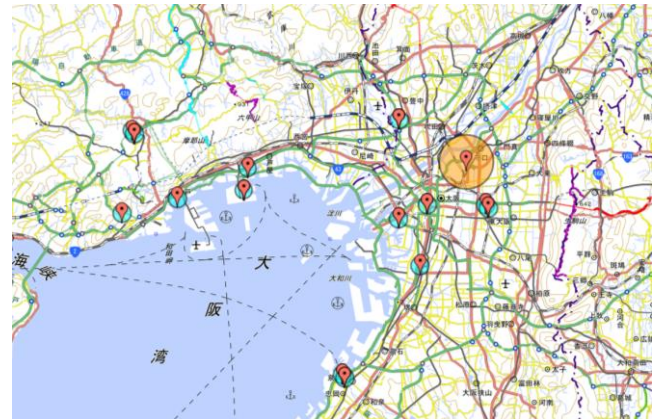
(3) 検証結果

図-3は、本手法の交通現象に係わるツイートの抽出機能および既存手法による抽出結果の内、クラスタの中心ツイートに類似する上位20件のツイートから単一の位置情報が特定できるクラスタを地図に描画して示している。図の詳細として、既存手法では1,200ツイートを5件のクラスタに分割して描画されているのに対し、本手法では533ツイートを66件のクラスタに分割して描画されている。また、本手法の抽出結果のクラスタの内容を確認すると、「この時間でも阪神高速神戸線は渋滞中」や「阪神高速守口線の工事で塀筋とかめっちゃ混んでる」などの交通現象に関するツイートが収録されていることが確認できた。

道路利用に係わる意見のツイート抽出機能の抽出結果を確認すると、「やっと明日解除か守口線。今日の阪神高速に塩カル撒いてテンションあがったーご苦労さまです。阪神高速。」や「阪神高速になるといきなり周りの車の運転が荒く感じる」などの道路利用者の主観的なツイートが見られた。この抽出結果を評価表現分析機能に適用したところ、工事中は「阪神高速守口線が使えないのは不便すぎる」などの否定的な意見が多数見られたが、工事後は「守口線の路面が綺麗で走りやすい」などの肯



(a)既存手法：5件



(b)本手法：66件

図-3 検証結果の地図描画 (出典：国土地理院数値地図 25000 (2015年6月))

定的な意見が見られた。

(4) 考察

本手法を用いると、ツイートに位置情報の付与が可能となり、ツイート分析に道路交通データ（今回は工事実績）を組み合わせる有効性を示唆する知見を得た。また、交通事象と道路利用に関する意見とに分割できるため、道路事業評価の多様な用途に対応できる。一方、ツイートの投稿時刻は、道路交通データの保持する時間（工事の時間帯）と合致しない場合も想定される。このため、工事や渋滞の時間帯などの道路交通データの特徴を踏まえたツイートの抽出手法の確立が課題として挙げられる。

守口線工事の前中後でツイートの内容が変化していることに加え、道路管理者の日常業務では把握しにくい道路利用者の意見も確認することができた。なお、今回の検証では、肯定的・否定的な意見の判定は手作業で行ったが、判定用の単語辞書を作成することで自動化できる。

5. おわりに

本研究では、Twitterのツイートと既存の道路交通データとを組み合わせた道路事業に係わるツイート抽出手法を提案した。本手法の有用性の検証結果から、位置情報を付与した交通事象に係わるツイートや道路利用に関する意見のツイートを抽出することがわかった。また、道路利用に関する意見のツイートを肯定的・否定的意見判定の自動化の可能性を確認した。

今後の課題として、ツイートの投稿時刻と投稿内容の指す時刻との違いを考慮した交通事象の分割手法の確立が挙げられる。また、一般論の課題になるが、ツイートの信憑性の評価や作為的なツイート拡散を想定したクリーニング処理の確立も挙げられる。

今後、これらの課題への解決策の確立に向けて取り組

む。さらに、抽出したツイートを既存の道路交通分析に組み合わせることによる有効性を研究する。

謝辞：本研究の遂行にあたり、阪神高速道路（株）の各氏から貴重なご意見を賜った。ここに記して感謝の意を表する。

参考文献

- 1) 橋本浩良，河野友彦，門間俊幸，上坂克巳：交通円滑化対策のためのプローブデータの分析方法に関する研究，平成22年度国土交通省国土技術研究会，2010。
- 2) 榊剛史，柳原正，那和一成，松尾豊：Twitterを用いた道路交通情報の抽出，電子情報通信学会論文誌(D)，電子情報通信学会，DOI：10.14923/transinfj.2014JDP7060，2015。
- 3) 矢野晋哉，伊藤秀昭，安田幸司：道路開通に関するツイッター情報の分析事例，土木計画学研究発表会・講演集，土木学会，Vol.48，2013。
- 4) 今井龍一，高橋哲朗，田嶋聡司，山影譲，重高浩一：道路事業に係わる行政相談資料及びTwitterのつぶやきに対するテキストマイニング技術の適用～道路事業評価の高度化支援に向けた一考察～，土木計画学研究発表会・講演集，土木学会，Vol.48，2013。
- 5) 橋本康弘，岡端起：都市におけるジオタグ付きツイートの統計，人工知能学会誌，人工知能学会，Vol.27，No.4，pp.424-431，2012。
- 6) 石田和成：地域特有の単語共起にもとづく位置推定と地域トピックの考察，情報処理学会研究報告，情報処理学会，Vol.117，No.2，pp.1-6，2015。
- 7) Twitter Inc.：Twitter Developers，<<https://dev.twitter.com/>>，(参照2015.7.30)。
- 8) Google Inc.：Google Maps Geocoding API，<<https://developers.google.com/maps/documentation/geocoding/intro?hl=ja>>，(参照2015.7.30)。
- 9) 工藤拓：McCab，<<http://mecab.sourceforge.jp/>>，(参照2015.7.30)。
- 10) Salton, G. and McGill, M.：Introduction to Modern Information Retrieval, McGraw-Hill, 1983。
- 11) Kaufman, L. and Rousseeuw, J.：Finding Groups in Data: An Introduction to Cluster Analysis, John Wiley & Sons, 1990。