2M-01

速報性と正確性の向上を図ったTwitterからの鉄道運行情報検出システムの検討

新井 誠也† 平川 豊‡ 大関 和夫‡
†芝浦工業大学大学院理工学研究科 ‡芝浦工業大学工学部

1 はじめに

マイクロブログ「Twitter」が持つ速報性に着目し，ツイートを常に収集することで，実世界で起きている事象を迅速に観測する研究が増加している [1]．これらの研究は，地震や感染症の流行など，予めターゲットとするイベントを定め，そのイベントの発生をより正確に，素早く検出することに焦点を当てている．

一方，鉄道のダイヤが乱れると，鉄道事業者がWebサイト上で「運転見合わせ」や「遅れ」といった運行情報を発表する．しかし，これらの運行情報は，ダイヤの乱れ発生してからしばらく経たないと公開されない場合があり，鉄道利用者への情報伝達が遅れてしまう問題がある．情報が素早く伝われば，迂回ルートや別の交通機関の利用を検討するなど行動の選択肢も広がる．

本研究では，鉄道の遅延というイベントに対象を絞り，より早い情報提供を行うことを目的とし，鉄道利用者から投稿されたツイートをもとに各路線の運行情報を推定し，配信するシステムを提案する．そして，速報性と正確性の観点でシステムの有効性を評価する．

2 既存手法

長野ら [2] は，Twitterからの鉄道運行情報検出システムをAndroidアプリで開発した．この研究では，直近 N 分間に投稿された路線名を含むツイートを3分おきに取得する．その後，鉄道の遅延に言及しているツイートの中から以下の3ルールをすべて満足するツイートを陽性ツイートとして抽出し，陽性ツイート数が閾値 θ を超えれば遅延と推定している．

- **ルール1**：路線名の後に運行情報語が出現する．
- **ルール2**：ユーザ発言部分 (リツイートでない部分) に路線名が出現する．
- **ルール3**：返信 (@から始まる) ツイートではない．

運行情報語とは，遅延時に投稿されるツイートでのみ出現回数が多くなる運行情報に関連した14語 (「止まった」「ストップ」「遅延」「遅れる」「見合わせ」など) を指す．文献 [2] では，ツイートの長さ・路線名・運行情報語・リツイート記号 (RT,QT)・ユーザ名の出現位置を素性とするSVM(Support Vector Machine) による分類よりも，3ルールに基づいた手法の方が，陽性ツイート抽出において高いF値 (再現率と適合率の調和平均) を示したと報告している．運行情報検出においては，$N = 20$, $\theta = 3$(1日平均ツイート数が1,000件未満の路線) または8(1,000件以上の路線) のときに高いF値を示したとし，これらの条件でシステムを実装している．しかし，陽性ツイート抽出の際にbotやスパムによるツイートが誤検出される，遅延には言及しているが運行情報語を含まないツイートは抽出漏れになる，同一路線名が全国に複数存在する場合は遅延の発生路線が特定できない等の問題がある．また閾値を決定する際，路線によるツイート数の違いは考慮されているが，時間帯によるツイート数の違いなど他の要素は考慮されていない．

3 ツイートの分析

首都圏の21路線を対象に，2013/11/1〜2014/1/6, 3/29〜5/25，10/4〜10/8 に投稿された路線名を含む計3,523,294ツイートを収集し，以下の分析に用いた．

3.1 路線利用者数とツイート数の相関

路線によって遅延発生後のツイート数に差異があるかを確認した．そして，発生路線の1日駅別乗降人員数合計と発生後1時間以内に投稿された路線名を含むツイート数の相関を調べた．12:00〜12:30 に発生した人身事故 (標本数 $n = 4$) の例を**図1**に示す．相関係数 $r = 0.99$ となり，強い正の相関があった．有意水準0.05，自由度 $v = n - 2$ でt検定を行ったところ，検定統計量 $= r\sqrt{v}/(1 - r^2) = 9.92$ がt分布のパーセント点 $= 4.30$ を超えたため，相関係数の有意性が認められた．

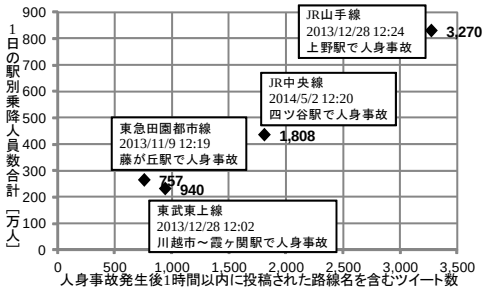

図1 各路線の1日駅別乗降人員数合計と発生後ツイート数

3.2 時間帯利用者数とツイート数の相関

遅延発生時間帯によって遅延発生後のツイート数に差異があるか検証した．そして，発生時間帯 (30分刻み) における推計鉄道利用者数と発生後1時間以内に投稿された路線名を含むツイート数の相関を調べた．その結果，東急東横線やJR中央線など，21路線中11路線で $r = 0.6$ 以上を示し，正の相関がみられた．$r = 0.71$ となった京王線での人身事故 ($n = 7$) の例を**図2**に示す．

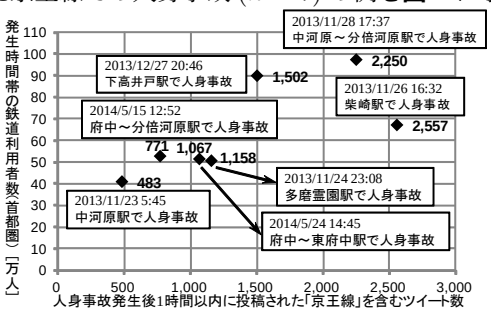

図2 発生時間帯の鉄道利用者数と発生後ツイート数

3.3 ツイートの投稿端末推測

遅延発生時刻から平常運行に戻った時刻までに投稿された路線名を含むツイートを対象に，ユーザが投稿に利用したクライアントアプリケーションを調べた．その情報から各ツイートの投稿端末を推測し，(i) 携帯端末 (スマートフォン，タブレットなど) からの投稿，(ii)PCからの投稿，(iii)botによる投稿のいずれかに分類した．その結果，それぞれの割合は，(i)66.9%，(ii)7.8%，(iii)25.2% となった．以上より，遅延発生時のツイートは，携帯端末での投稿が多いことが明らかになった．

3.4 ツイートの投稿場所推測

3.3で利用したツイートを対象に，ユーザが投稿した場所についての推測を行った．場所推定は，ツイートに付与されている位置情報をもとに行った．位置情報が無

Train-status detecting system from Twitter improving reporting delay and accuracy

†Seiya Arai ‡Yutaka Hirakawa ‡Kazuo Ohzeki
†Graduate School of Engineering and Science, Shibaura Institute of Technology
‡College of Engineering, Shibaura Institute of Technology

いツイートに関しては，ツイートを投稿したユーザのプロフィール欄にある「場所」の記述をもとに場所推定を行った．その結果，「場所」の記述があるツイートのうち 55.6% は首都圏の都県名・市区町村名・駅名を含んでいた．以上より，首都圏の路線で遅延が起こると，首都圏から多くのツイートが投稿されることを確認した．

4 提案手法

4.1 システム概要

本研究で提案するシステムの概要図を示す (**図 3**)．推定した運行情報は，Web 上に公開し，配信を行う．

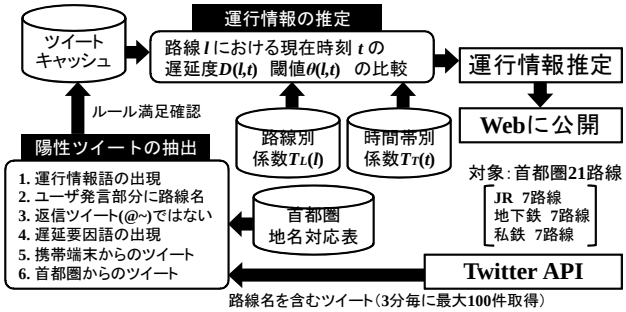

図 3　提案システムの概要図

4.2 陽性ツイートの抽出

ツイートの分析結果を踏まえ，既存手法の 3 ルールに加えて新たにルール 4〜6 を提案する．これによる誤抽出及び抽出漏れの削減を図る．ルール 1 またはルール 4 を満たし，かつルール 2,3,5,6 をすべて満足するツイートを陽性ツイートとする．遅延要因語とは，2013/11/1〜2014/1/6，3/29〜5/25，6/24〜7/10 に対象路線で起きた遅延 682 件の遅延要因をまとめた 58 語 (「人身事故」「信号点検」「異音」「大雨」「停電」など) を指す．

- **ルール 4**：路線名の後に遅延要因語が出現する．
- **ルール 5**：携帯端末で投稿したツイートである．
- **ルール 6**：首都圏から投稿したツイートである．

ルール 4 を定めたことで，「埼京線，信号点検なう」など，運行情報語は無いが運行に異常があることを示唆するツイートを抽出できる．ルール 5 は，3.3 で出現した携帯端末用のクライアントアプリケーション一覧をあらかじめ登録し，照合を行う．そして，PC 端末と bot から投稿されたと推測できるツイートを除外する．ルール 6 は，位置情報付きツイートまたは投稿ユーザのプロフィール欄「場所」の記述を利用して，投稿場所の推定を行う．首都圏の全都県・市区町村名と全駅名 7,278 語を収録した首都圏地名対応表を作り，表に含まれる地名が「場所」の記述にあるかを照合する．含まれていたら，首都圏から投稿したツイートと推測する．そして，首都圏以外から投稿したと推測されるツイートを抽出対象から除外する．なお，「アジア」「日本」「Japan」といった記述の場合や，記述が無い場合は首都圏以外から投稿したと断定できないので，抽出対象に入れた．以上より，同一路線名が全国に複数存在しても，陽性ツイートは首都圏の路線の運行情報を述べていることが期待できる．

4.3 運行情報の推定

路線 l において，ある時刻 t での閾値 $\theta(l,t)$ と遅延度 $D(l,t)$ を以下のように定義する．$D(l,t)$ が $\theta(l,t)$ 以上であれば，路線 l は遅延していると推定する．

$$\theta(l,t) = T_B T_L(l) T_T(t) \qquad D(l,t) = \sum_{i=1}^{N} C_i P_{li}$$

T_B は基本となる閾値である．$T_L(l)$ は路線 l，$T_T(t)$ は時刻 t の属する時間帯 (30 分刻み) によって変化する係数である．$T_L(l)$ は，3.1 より，鉄道利用者が多い路線 (JR 山手線など) ほど値が大きい．$T_T(t)$ は，3.2 より，鉄道利用者が多い時間帯 (ラッシュ時) ほど値が大きい．3.2 において，相関係数 $r = 0.6$ 未満 となった 10 路線では，常に $T_T(t) = 1$ とする．C_i は t より i 分前〜$(i-1)$ 分前の 1 分間の重み値，P_{li} は t より i 分前〜$(i-1)$ 分前の 1 分間に投稿された路線 l の陽性ツイート数である．重みをつけることで，よりリアルタイムな運行情報を反映できると期待できる．今回は，$N = 20$ とした．

5 評価実験

5.1 陽性ツイートの抽出精度

対象路線で遅延発生後に投稿された計 4,582 ツイートを用いて，陽性ツイートの抽出精度を評価した (**表 1**)．比較対象は，既存手法とそれにルール 4,5,6 をそれぞれ適用した手法，ルール 1〜6 を適用した提案手法の 5 つである．ルール 4 は再現率，ルール 5 は適合率の向上に寄与した．ルール 6 は再現率・適合率ともに低下したが，同一路線名の他路線で遅延が起きた場合に大幅に誤抽出を抑えており，正確性の向上に一定の効果があった．

表 1　既存手法と提案手法の陽性ツイート抽出精度

	適合率	再現率	F 値
既存手法 (ルール 1〜3)	72.3%	67.6%	69.9%
既存手法 + ルール 4	72.1%	72.2%	72.1%
既存手法 + ルール 5	92.1%	61.5%	73.7%
既存手法 + ルール 6	71.4%	58.3%	64.2%
提案手法 (ルール 1〜6)	91.6%	56.9%	70.2%

5.2 既存手法と公式運行情報との比較

2014/10/28〜11/7 に起きた発生時刻が明確に判明した遅延 50 件について，発生時刻 10 時間前〜ダイヤ回復 10 時間後に投稿されたツイートを対象に，最適な重み値 C_i と閾値 T_B を求めた．陽性ツイートの抽出は，ルール 1〜6 を適用する．1 分毎の推定結果が実際の運行状況とどの程度一致するか (F 値)，遅延検出開始までの時間 [分](以下，DST) を評価指標とし，各件で最適となる C_i, T_B を求めた．その結果，遅延 50 件の中央値・平均値で既存手法 (重みづけ無し) をすべて上回り，かつ最良だったのは，$C_1, ..., C_5 = 2.0$, $C_6, ..., C_{10} = 1.5$, $C_{11}, ..., C_{15} = 1.0$, $C_{16}, ..., C_{20} = 0.5$, $T_B = 2.0$ であった (**表 2**)．システムの実装には，この値を用いた．

表 2　既存手法と提案手法の遅延 50 件の F 値，DST の比較

	F 値 (中央/平均)	DST(中央/平均)
既存手法 (C_i 無し)	62.7%/58.0%	11.0/32.5
C_i 有り，$T_B = 2.0$	63.2%/58.5%	8.5/24.8

次に，システム稼働中 (2014/11/30〜12/15) に，対象路線で起きた発生時刻が明確に分かる遅延において，公式運行情報と提案システムの第一報配信時刻を比較した．結果，55 件中 45 件 (81.8%) で提案システムの方が早く遅延の発生を検出し，配信できた．例えば，東京メトロ千代田線の綾瀬〜北千住駅間の車両点検による遅延 (12/2 7:43 発生) は，提案システムでは 7:51 に検出したが，公式運行情報の第一報配信時刻は 8:35 だった．

6 まとめと今後の課題

本研究では，Twitter から鉄道運行情報を迅速かつ正確に検出するシステムを提案した．投稿端末や投稿場所を利用した新たな陽性ツイート抽出ルールの設定，各陽性ツイートの重みづけ，時間帯や路線による閾値の変動により，速報性と正確性の向上に成功した．

今後は，陽性ツイート抽出において，運行情報語・遅延要因語の網羅性を上げて様々な表現に対応すること，より詳細な運行情報の推定が課題に挙げられる．

参考文献

[1] 榊剛史, 松尾豊, “ソーシャルセンサとしての Twitter”, 人工知能学会誌, Vol.27, No.1, pp.67-74, 2012.
[2] 長野伸一, 上野晃嗣, 長健太, “ソーシャルセンサからの鉄道運行情報検出システムの開発”, 電子情報通信学会論文誌, Vol.J96-D, No.10, pp.2262-2273, 2013.