

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

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 〒135-8548 東京都江東区豊洲 3 丁目 7 番 5

^{††} 芝浦工業大学工学部 〒135-8548 東京都江東区豊洲 3 丁目 7 番 5

E-mail: [†]{ma13004, hirakawa, ohzeki}@shibaura-it.ac.jp

あらまし Twitter が持つ速報性に着目し、ツイートを収集することで特定のイベントの発生を検出する研究が近年増えている。これらは、Twitter を利用して迅速かつ正確にイベントを検出することを目的としている。本研究では、鉄道の遅延というイベントを対象を絞り、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。先行研究では、電車の遅れに言及しているツイートを 3 ルールに基づいて抽出し、直近 20 分間での抽出ツイート数が閾値を超えていれば、遅延と推定する手法を提案している。我々は、先行研究の手法に加え、ツイートの投稿に利用した端末の種類や投稿場所を推測し、それらを考慮した新たな抽出ルールを提案する。また、時間帯や路線によって閾値を変動させ、投稿時間に応じて抽出ツイートの重みづけを行う手法を提案する。そして、速報性と正確性の観点からシステムの有効性を評価する。

キーワード マイクロブログ、ソーシャルセンサ、抽出ルール、運行情報推定

1. はじめに

近年、Twitter^(注1)が注目を集めている。Twitter とは、マイクロブログサービスの 1 つであり、全世界の月間アクティブユーザー数は 2 億 4100 万人 (2013 年時点)^(注2)を超える。サービス利用者は、「ツイート (Tweet)」「つぶやき」と呼ばれる最大 140 字のテキストを投稿することで、他の利用者へ情報を発信することができる。投稿は即座に反映されるためリアルタイム性に優れること、投稿内容が短いため気軽に投稿できること、スマートフォンなどの携帯端末を用いて場所を問わず容易に投稿できることから、現在の身の回りの状況・体験が多く投稿されている。このことから、ツイートを常に収集し続けることで、実世界で起きている事象を迅速に観測できることが期待されており、関連する研究は 2007 年より増加傾向にある [1]。これらの研究は、Twitter を一種のセンサ (ソーシャルセンサ (Social sensor)) として捉え、特定のイベントの発生をより素早く、より正確に検出を行うことを目的としている。例えば、Twitter から地震の検出を目的とした研究 [2]、インフルエンザの流行予測を目的とした研究 [3]、犯罪情報の集約を目的とした研究 [4]、交通情報の収集を目的とした研究 [5] などが挙げられる。

本研究では、鉄道の遅延というイベントを対象を絞り、電車の遅れに言及しているツイートを常時抽出する。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「遅延」といった運行情報を発表し、鉄道利用者へ周知している。しかし、これらの運行情報は、ダイヤが乱れ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。刻一刻と状況が変化中、情

報が素早く伝われば、迂回ルートや他の交通機関の利用を検出するなど行動の選択肢も広がる。そのため、運行情報を早期に入手することの重要性は高いといえる。本研究では、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・配信するシステムを提案する。そして、既存手法および公式運行情報と比較し、迅速かつ正確に遅延の検出ができたかの評価実験を行い、システムの有効性を示す。

以降、第 2 章では Twitter を用いた鉄道の運行情報推定に焦点を当てた関連研究について述べ、第 3 章では実際にダイヤ乱れ時に投稿されたツイートの分析結果について報告する。第 4 章では提案するシステムの概要およびツイートの抽出手法及び運行情報の推定手法について述べ、第 5 章で評価実験の結果を提示し、有効性を示す。第 6 章では提案するシステムのインタフェースを示し、最後に、第 7 章で本研究のまとめと今後の課題について述べる。

2. 関連研究

ツイートをもとに鉄道運行情報の推定・予測を行っている関連研究を紹介する。

村山ら [6] は、第 84 回コミックマーケットを対象に、イベントのハッシュタグと会場最寄り駅名を含むツイートに対して、eki ラベル (駅に関する情報を含むツイート群) か event ラベル (イベントに関する情報を含むツイート群) のどちらかを付与し、鉄道の混雑状況の抽出を試みた。ラベルの付与には、Naive Bayes 分類器に、単語の出現回数と逆文書頻度 (IDF) の情報を追加した分類器を利用している。そして、ラベルを付与したツイートに対して各単語の番号を行、各ツイートの番号を列、TF-IDF を値とした行列を作成し、Ward 法による階層的クラスタリングを行った。その結果、駅関連の混雑情報を含む

(注1): <https://twitter.com/>

(注2): <https://investor.twitterinc.com/annuals-proxies.cfm>

ツイート群が抽出できたとしている。しかし、リアルタイムに情報抽出を行うことを研究対象としておらず速報性を考慮していないため、すぐに情報を知りたいという鉄道利用者の需要を満たしていない。

築瀬ら [7] は、JR 常磐線の運行状況を一般の Twitter 利用者のツイートから予測し、電車が遅延していると推測される場合には、メールで自動通知を行うシステムを提案した。この研究では、遅延が起これと、「常磐線」という語を含むツイート数が増加するという現象に注目した。直近 5 分間あたりに「常磐線」という語を含むツイートが 6 件以上投稿され、かつそれらに「遅延」「遅れ」「運転見合わせ」という語が含まれていると、運行状況に異常がある可能性が高いとしている。しかし、通知の対象としている路線が 1 路線で他路線の運行状況は考慮されていない、遅延には言及しているが「遅延」「遅れ」「運転見合わせ」の 3 語のいずれも含まないツイートは抽出漏れになる問題がある。

土屋ら [8] は、鉄道運行トラブル発生時の意思決定支援を目的に、Twitter を解析することで、東京メトロの路線の鉄道運行トラブルを検出し、遅延の継続時間と他路線への遅延の連鎖を予測した。この研究では、Support Vector Machine(SVM) を用いて路線名を含む運行トラブルに関連するツイートを抽出し、文献 [9] のバースト検知手法を用いてトラブルの検出を行っている。また、トラブル発生後 5 分間に投稿された運行トラブル関連ツイートを 1 つのドキュメントとみなし、このドキュメントの特徴量をもとに、作成した分類器を用いて遅延の継続時間と他路線への遅延の連鎖を予測している。しかし、ツイートの投稿間隔をもとにバーストの検知を行っているため、ツイート数が急激に伸びるトラブル発生直後はバーストとして検知できるが、ツイート数が少なくなるトラブル終盤はバーストとして検知できない。また、ツイートの絶対数が少なくなる早朝や深夜に発生した遅延を迅速に検出できない問題がある。

長野ら [10] は、交通情報、特に鉄道分野におけるソーシャルセンサとしての Twitter の有用性を確認するため、Twitter からの鉄道運行情報検出システムを Android アプリで開発を行った。この研究では、まず直近 N 分間に投稿された各路線名を含む最新のツイートを 3 分おきに最大 100 件収集する。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、当該路線が遅延していると推定している。ここで、運行情報語とは、平常時のツイートと比較して遅延時のツイートのみに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。

- ・ ルール 1：路線名の後に運行情報語が出現する。
- ・ ルール 2：引用でない発言部分に路線名が出現する。
- ・ ルール 3：返信ツイート (@ から始まる) ではない。

文献 [10] によると、ツイートの長さ・路線名・運行情報語・引用記号 (RT, QT)・ユーザ名のそれぞれ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 (適合

率と再現率の調和平均) を示したと報告している。運行情報の検出においては、 $N = 20$, $\theta = 3$ (1 日平均ツイート数が 1,000 件未満の路線) または 8 (1 日平均ツイート数が 1,000 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、bot やスパムによるツイートが誤抽出される [7] と同様に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。例えば、「東西線」という名称を持つ鉄道路線は全国に 5 路線存在するため、「東西線」という語を含むツイートを収集しても、どこかの東西線を指しているのか分からない。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 [8] と同様に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。そこで本研究では、文献 [10] を既存手法とし、この手法をもとにした新たな手法を提案することで、速報性と正確性の向上を目指す。

3. ツイートの分析

首都圏の 21 路線 (JR7 路線、地下鉄 7 路線、私鉄 7 路線) を対象に 2013/11/1 ~ 2014/1/6, 3/29 ~ 5/25, 10/4 ~ 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

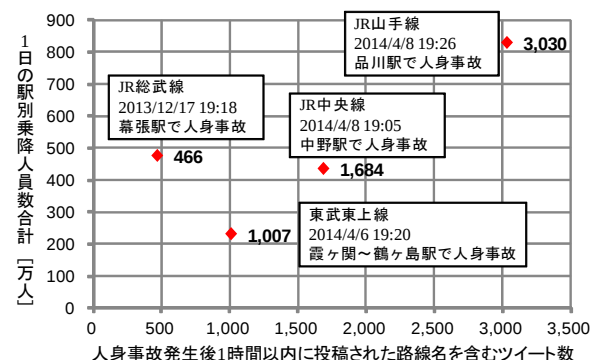


図 1 各路線の 1 日駅別乗降人数合計と発生後ツイート数 (19:00-19:29)

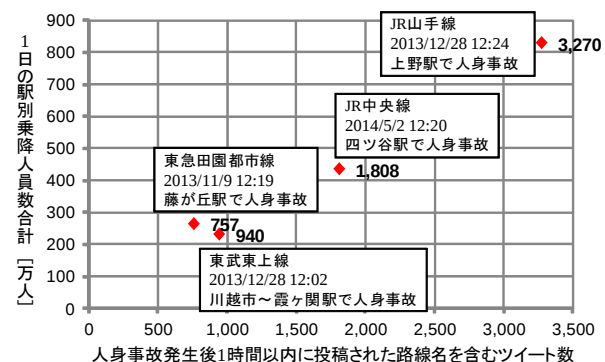


図 2 各路線の 1 日駅別乗降人数合計と発生後ツイート数 (12:00-12:29)

路線によって遅延発生後のツイート数に差異があるのかを確認するために、遅延発生路線の 1 日駅別乗降人数合計^(注3)と

(注 3): <http://www.mlit.go.jp/sogoseisaku/transport/>

遅延発生後 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。19 時 00 分～19 時 29 分に発生した人身事故 (標本数 $n = 4$) の例を図 1 に、12 時 00 分～12 時 29 分に発生した人身事故 ($n = 4$) の例を図 2 に示す。図 1 では、相関係数 $r = 0.78$ ，図 2 では $r = 0.99$ となり、ともに正の相関があることを確認した。従って、遅延発生時のツイート数は路線利用者数によって左右されるといえる。

3.2 時間帯利用者数とツイート数の相関

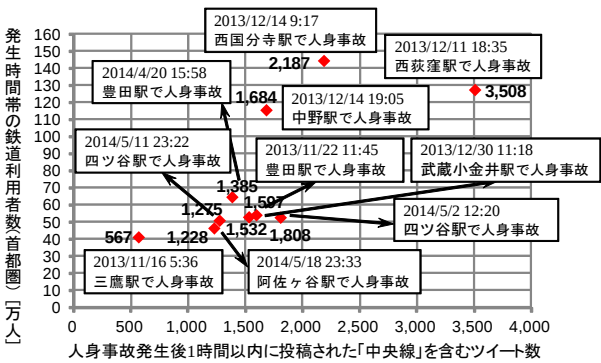


図 3 発生時間帯の鉄道利用者数と発生後ツイート数 (JR 中央線)

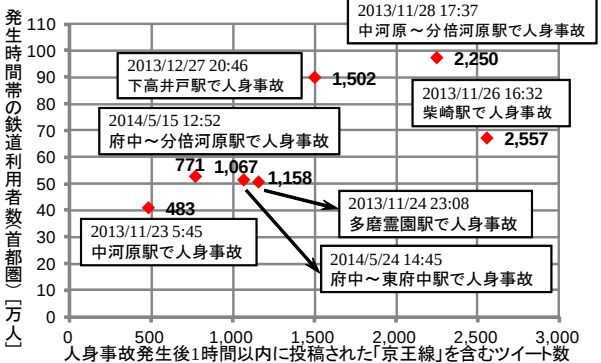


図 4 発生時間帯の鉄道利用者数と発生後ツイート数 (京王線)

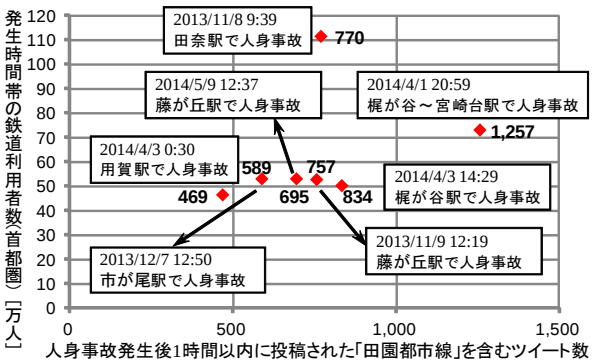


図 5 発生時間帯の鉄道利用者数と発生後ツイート数 (東急田園都市線)

遅延時間帯によって遅延発生後のツイート数に差異があるのかを確認するために、発生時間帯 (30 分刻み) における首都圏の鉄道利用者数の推計^(注3)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、JR 京浜東北線や小田急線など対象 21 路線中 11 路線で相関係数 $r = 0.6$ 以上を示し、正の相関がみられた。 $r = 0.73$ となった JR 中央線での人身事故 ($n = 10$) の例を図 3 に、 $r = 0.71$ となった京

王線での人身事故 ($n = 7$) の例を図 4 に示す。一方で、JR 山手線など 10 路線では $r = 0.6$ 未満となった。 $r = 0.34$ となった東急田園都市線での人身事故 ($n = 7$) の例を図 5 に示す。以上より、遅延発生時のツイート数が時間帯によって左右される路線が存在することが明らかになった。また、遅延発生時のツイート数が時間帯に影響されない路線が存在することも併せて確認した。

3.3 ツイートの投稿端末推測

2014/10/4～8 に発生した遅延 5 件について、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計 10,713 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 (以下、クライアント) の情報を Twitter API^(注4)から取得した。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 アカウントやスパムアカウントによる投稿のいずれかに手動で分類した。推測の際には、クライアントの情報を用いて bot アカウントの判別を行っている文献 [11] の手法を参考にした。例えば、投稿に利用したクライアントが「Twitter for iPhone」(iPhone 版 Twitter 公式クライアント) や「Twitter for Android」(Android 版 Twitter 公式クライアント) である場合は、携帯端末からツイートを投稿したと推測できる。また、「twitterbot.net」(自動投稿機能付きの bot 生成クライアント) である場合は、bot アカウントから投稿されたツイートであると推測できる。

分類結果を表 1 に示す。表 1 より、遅延発生時に投稿されるツイートは、携帯端末での投稿が多いことが明らかになった。

表 1 投稿に利用した端末の分類

投稿端末	ツイート数	割合	投稿に使われたクライアント (上位)	
			名称	ツイート数
(i) 携帯端末	7,174	66.9%	Twitter for iPhone	3,230
			Twitter for Android	1,789
			twicca	302
(ii) PC	837	7.8%	Twitter Web Client	423
			Janetter	88
			TweetDeck	78
(iii) bot・スパム	2,702	25.2%	twitterbot.net	392
			dlvr.it	184

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、ユーザがツイートを投稿した場所について推測を行った。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。位置情報が無いツイートに関しては、文献 [12] の手法を参考に、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」項目の記述内容をもとに行った。任意項目ではあるが、ユーザはここに普段の居住地や居場所を記載できる。文献 [12] によると、プロフィール欄を使った場所推定は、一定の精度を得られたと評価している。そのため、本研究の場所推測手法として採用した。

予備実験として、2014/10/15～17 の 2 日間に投稿された「山手線」を含むツイートに、どれほど位置情報が付与されたか

(注 4): <https://dev.twitter.com/>

を調査した．その結果，位置情報付きツイートは全 15,479 ツイート中 109 ツイート (0.7%) と非常に少ないことが判明した．従って，ほとんどのツイートは「場所」項目の記述を参照することになる．

次に「場所」の記述内容を解析し，以下の 6 パターンのいずれかに振り分けた．なお，ここでの「首都圏」とは，関東 1 都 6 県に山梨県と静岡県東部地方を加えた地域とした．

- 記述なし 記述が無い．
- 場所情報以外 記述はあるが場所についての情報が無い．
- 日本 「Japan」など日本を表す用語が記述されている．
- 首都圏 「関東」「東京都」「新宿」など関東地方を表す用語や首都圏の都道府県市町村名，駅名が記述されている．
- 首都圏以外 「関西」「大阪」「名古屋」など首都圏以外の日本の都道府県名，市町村名，駅名が記述されている．
- 推測不可能 「N 区」「駅前」「家」など推測できない場所や日本以外の地名が記述されている．

各ツイートに対し振り分けを行った結果を表 2 に示す．遅延発生時に投稿されたツイートの半数は「場所」項目の記述そのものが無い (a に該当)，または場所についての情報が記述されていなかった (b に該当) ため，場所推定ができなかった．一方で，「場所」項目に場所について記述があるツイート (c～f に該当) のうち，64.5% は首都圏の地名や駅名を含んでいる (d に該当)．以上より，首都圏の路線で遅延が発生すると，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．

表 2 各ツイートを投稿したユーザの「場所」項目の記述内容

記述内容	ツイート数	割合	記述エリア		
			エリア	ツイート数	割合
a. 記述なし	4,438	41.4%	-	-	-
b. 場所情報以外	867	8.1%	-	-	-
場所の記述あり	5,408	50.5%	c. 日本	261	4.8%
			d. 首都圏	3,491	64.5%
			e. 首都圏以外	409	7.6%
			f. 推測不可能	1,247	23.1%

4. 提案手法

4.1 システムの概要

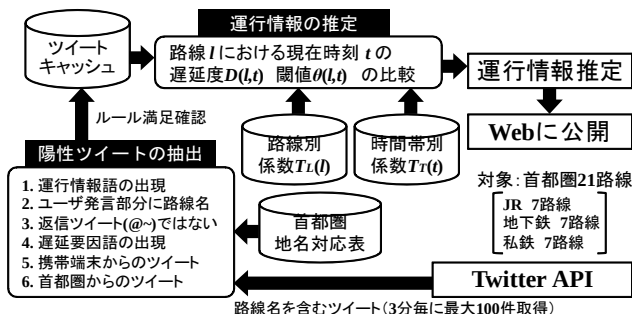


図 6 提案システムの概要図

提案するシステムの概要図を図 6 に示す．図中の遅延度 $D(I, t)$ ，閾値 $\theta(I, t)$ ，路線別係数 $T_L(I)$ ，時間帯別係数 $T_T(t)$ については，4.3 で解説する．ツイートの取得間隔や最大取得件

数は，既存手法に沿って設計にする．推定した遅延情報は Web 上に公開し，配信を行う．今回は，首都圏の 21 路線を配信の対象とし，サービス提供時間は 6:00 から 23:59 までとした．

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提案手法における陽性ツイートとする．

- ルール 4：路線名の後に遅延要因語が出現する．
- ルール 5：携帯端末から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．

4.2.1 ルール 4

遅延要因語とは，本研究で定義した 2013/11/1～2014/1/26，3/29～5/25，6/24～7/10 に対象路線で発生した公式発表があった遅延 682 件の遅延要因をまとめた語群である．「人身事故」「接触」「信号点検」「安全確認」「異音」「大雨」「濃霧」「停電」「沿線火災」など 58 語を遅延要因語とした．ルール 4 を定めたことで「埼京線，信号点検なう」や「京王線で沿線火災発生」といった遅延情報語は含まれていないが，運行に異常があることを示唆するツイートを陽性ツイートとして抽出できる．既存手法にルール 4 を適用することで，抽出漏れが削減されると期待できる．

4.2.2 ルール 5

3.3 にて出現した携帯端末用クライアントをあらかじめ登録し，照合を行う．登録した携帯端末用クライアントは「Twitter for iPhone」「Twitter for Android」「Twitter for Android Tablets」「twicca」「ついつづる for iPhone」など全 104 アプリケーションである．そして，PC から投稿されたと推測されるツイートや bot アカウントやスパムアカウントから投稿されたと推測されるツイートを除外する．既存手法にルール 5 を適用することで，誤抽出が削減されると期待できる．

4.2.3 ルール 6

位置情報付きツイートまたはプロフィール欄の「場所」項目の記述を利用して投稿場所の推測を行う．まず，文献 [12] の手法を参考に，首都圏の全都県名・市区町村名・駅名 7,278 語を収録した首都圏地名対応表を作った．記述内容は「とうきょう」「Yokohama」といった平仮名やアルファベットで書かれているケースも見受けられた．そのため，首都圏地名対応表は平仮名とアルファベットへの展開も行った．次に，首都圏地名対応表に含まれている地名が「場所」の記述内容にあるかを照合する．首都圏地名対応表に存在していたら，首都圏から投稿されたツイートであると推測した．そして，首都圏から投稿されたと推測されるツイート (3.4 における d)「場所」項目の記述がないツイート (3.4 における a)「日本」「Japan」といった日本を表す用語が記述されているツイート (3.4 における c) を陽性ツイートの抽出対象とする．それ以外のツイートは，抽出対象から除外する．2 つ目と 3 つ目を抽出対象としたのは，首都圏以外から投稿したツイートであると断定できないためである．特に，2 つ目はツイート全体の 4 割を占めるため，抽出対象が

ら除外してしまうと抽出漏れが多く起こることが予想される．以上より，ルール 6 を適用することで，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について言及していることが期待できる．

4.3 運行情報の推定

路線 l における現在時刻 t での閾値 $\theta(l, t)$ を式 (1)，遅延度 $D(l, t)$ を式 (2) 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し， $D(l, t)$ が $\theta(l, t)$ 未満であれば，路線 l は平常運行であると推定する．

$$\theta(l, t) = T_B T_L(l) T_T(t) \quad (1)$$

$$D(l, t) = \sum_{i=1}^N C_i P_{li} \quad (2)$$

T_B は基本とする閾値であり，5.2 で定める． $T_L(l)$ は路線 l によって変化する係数であり， $T_T(t)$ 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． C_i は t より i 分前 $\sim (i-1)$ 分前の 1 分間の重み値， P_{li} は t より i 分前 $\sim (i-1)$ 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．なお，本研究では，既存手法と同じく $N = 20$ と定めた．

4.3.1 路線と時間帯による閾値の変動

3.1 より，路線による遅延発生後のツイート数の違いを考慮し，係数 $T_L(l)$ を導入する． $T_L(l)$ は，路線 l の 1 日駅別乗降人員数合計^(注3)を，東京メトロ有楽町線の 1 日駅別乗降人員数合計で割った値とする．東京メトロ有楽町線は，対象 21 路線中で最も 1 日駅別乗降人員数合計が少ない．例えば， $T_L(l = \text{小田急線}) = 2.27$ ， $T_L(l = \text{丸ノ内線}) = 1.60$ ， $T_L(l = \text{有楽町線}) = 1.00$ となる． $T_L(l)$ が最大の路線は JR 山手線で， $T_L(l = \text{山手線}) = 4.85$ である．

また，3.2 より，時間帯による遅延発生後のツイート数の違いを考慮し，係数 $T_T(t)$ を導入する． $T_T(t)$ は，時刻 t の時間帯 (30 分刻み) の鉄道利用者数^(注3)を，16 時 30 分～16 時 59 分の鉄道利用者数で割った値とする．ただし，3.2 において，相関係数 $r = 0.6$ 未満となった 10 路線については，時間帯に関わらず常に $T_T(t) = 1.0$ とする．16 時 30 分～16 時 59 分の鉄道利用者数は，6 時 00 分～6 時 29 分から 23 時 30 分～23 時 59 分の各時間帯 (30 分刻み) それぞれの鉄道利用者数の中で中央値をとる．平均値でなく中央値を採用する理由は，朝のラッシュ時の鉄道利用者数は他時間帯に比べて極めて多く，その影響を減らすためである．例えば， $T_T(t = 19:00 \sim 19:29) = 1.71$ ， $T_T(t = 23:00 \sim 23:29) = 0.75$ ， $T_T(t = 16:30 \sim 16:59) = 1.00$ となる． $T_T(t)$ が最大の時間帯は 8 時 00 分～8 時 29 分で， $T_T(t = 8:00 \sim 8:29) = 3.38$ である．

このように，路線と時間帯の両方を考慮することで，実際のツイート数を反映した閾値を設定できる．例えば，JR 山手線における 8 時 00 分～8 時 29 分の閾値は， $\theta(l = \text{山手線}, t = 8:00 \sim 8:29) = 12.82 T_B$ と大きな値になる．一方で，東京メトロ有楽町線における 6 時 00 分～6 時 29 分の閾値は $\theta(l = \text{有楽町線}, t = 6:00 \sim 6:29) = 0.61 T_B$ と小さな値になる．

4.3.2 陽性ツイートの重みづけ

よりリアルタイムな運行情報をシステムに反映するために，

抽出した各陽性ツイートに対し，重み値 C_i をつける．直近に投稿された陽性ツイートほど情報が持つ価値を高め，時間が経過した陽性ツイートほど情報が持つ価値を低くする．これにより，遅延の検出開始までの時間が短縮されることが期待できる．

5. 評価実験

5.1 陽性ツイートの抽出精度

首都圏の路線で発生した遅延 (2014/10/22～27 に対象路線で発生した遅延 5 件) を対象に，発生 1 時間前から発生 1 時間後に投稿された計 4,582 ツイートを用いて，既存手法と提案手法の陽性ツイート抽出における精度を比較評価した．比較対象は，ルール 1～3 を適用した既存手法とそれにルール 4，ルール 5 をそれぞれ適用した手法，ルール 1～5 をすべて適用した手法の 4 手法である．既存手法では首都圏の路線で発生したと明確に分かる遅延において投稿されたツイートを対象として，陽性ツイートの抽出精度を評価している．従って，本研究においても対象を揃えるため，同一路線名による誤抽出を考慮したルール 6 を適用させない手法で比較を行った．評価指標には，以下の式の適合率，再現率，F 値を利用する．式中の TP は陽性ツイートを陽性ツイートとしてシステムが正しく抽出したツイート数，FP は陽性ツイートでないツイートを陽性ツイートとしてシステムが誤って抽出したツイート数，FN は陽性ツイートを陽性ツイートでないツイートとしてシステムが誤って抽出しなかったツイート数である．適合率は誤抽出の少なさを，再現率は抽出漏れの少なさを示す．F 値とは，適合率と再現率の調和平均である．

$$\text{適合率} = \frac{TP}{TP + FP} \quad \text{再現率} = \frac{TP}{TP + FN}$$

$$F \text{ 値} = \frac{2 \times \text{適合率} \times \text{再現率}}{\text{適合率} + \text{再現率}}$$

表 3 既存手法とルール 4,5 を適用した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 適用)	72.3%	67.6%	69.9%
既存手法 + ルール 4 適用	72.1%	72.2%	72.1%
既存手法 + ルール 5 適用	92.1%	61.5%	73.7%
ルール 1～5 適用	92.5%	64.9%	76.3%

評価結果を表 3 に示す．ルール 4 は再現率の向上，ルール 5 は適合率の向上にそれぞれ寄与した．ルール 1～5 を適用した手法では，既存手法と比較して再現率は低下したが適合率が大幅に向上しており，F 値も 6.4% 改善された．

ただし実際にシステムを運用する際は，既存手法では，同一路線名の他路線で遅延が発生した場合 (東西線，中央線など)，あるいは同一路線であっても首都圏以外の地域で遅延が発生し首都圏に影響がない場合 (東海道線，中央線など) でもその路線名を含むすべてのツイートに対してルール 1～3 を適用している．そのため，首都圏以外の路線に関する運行情報も陽性ツイートとして抽出し，上記のケースでは多くの誤抽出をしてしまう．そこで，提案するルール 6 を適用することで，誤抽出を

時刻	既存手法(ルール1~3)	ルール6適用
11:56	0	0
11:58	1	1
12:00	0	0
12:02	0	0
12:04	0	0
12:06	1	0
12:08	0	0
12:09	1	1
12:10	0	0
12:12	0	0
12:14	0	0
12:16	1	0
12:17	1	0
12:18	0	0
12:20	1	0
12:21	5	1
12:22	2	1
12:24	0	0
12:26	1	1

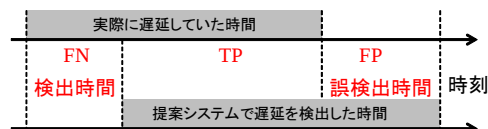
既存手法にルール 6 を適用した手法，ルール 1~6 をすべて適用した手法の 2 手法に対して，前述の陽性ツイート抽出精度の評価を行った結果を表 4 に示す．その結果，同一路線名による誤抽出を考慮したルール 1~6 を適用した手法と既存手法の F 値がほぼ同等であった．しかし，実際にシステム運用する際には，同一路線名による誤抽出が考慮されていない既存手法の抽出精度は，表 3 よりも下がるのが予想される．従って，既存手法に加え，提案手法の 3 ルールが有効だと考えられる．

	適合率	再現率	F 値
既存手法 + ルール 6 適用	71.4%	58.3%	64.2%
ルール 1~6 適用	91.6%	56.9%	70.2%

て遅延と検出した時間、FN は遅延の発生をシステムが正しく検出できなかった時間である。F 値は、1 分毎の推定結果をもとに計算を行うとし、実際の運行状況と推定結果がどの程度一致したかを示す。実際の運行状況は、公式運行情報の遅延発生時刻及び配信停止時刻を参考とした。

$$\text{適合率} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad \text{再現率} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$F \text{ 値} = \frac{2 \times \text{適合率} \times \text{再現率}}{\text{適合率} + \text{再現率}}$$



重み値 C_i は、直近の時刻ほど重みが大きくなり、時刻が経過すると重みが小さくなるようする。重み値 C_i の決定においては、今回は、以下の (A) ~ (E) の 5 パターンをあらかじめ用意し、その中でどれが有効かを比較評価する。

(E) $C_1 = 2.0, C_2 = 1.9, C_3 = 1.8, \dots, C_{19} = 0.2, C_{20} = 0.1$

C_i のパターン	(A)	(B)	(C)	(D)	(E)
最適な T_B の遅延 50 件中央値	2.0	2.5	2.0	2.0	1.5

はじめに、遅延 50 件各件に対して、(A) ~ (E) の各パターンにおいて最も F 値が高かったときの T_B を求めた。これが各パターンにおける最適な T_B といえる。その 50 件の中央値を表 5 に示す。 T_B は 0.5 から 50.0 の間で 0.5 ずつ変更した。次に、この 50 件に対し、 $T_B = 2.0$ のときの (A)、 $T_B = 2.5$ のときの (B)、 $T_B = 2.0$ のときの (C)、 $T_B = 2.0$ のときの (D)、 $T_B = 1.5$ のときの (E) をそれぞれ適用させたときの各件の F 値・検出時間・誤検出時間を求めた。表 6 にその中央値と平均値を示し、図 9 に各パターン及び既存手法における 50 件の検出時間の累積分布を示す。図中の既存手法は、陽性ツイート抽出にルール 1 ~ 3 を適用し、閾値を 8(JR 山手線、JR 中央線など 4 路線) または 3(それ以外の路線) とする手法である。中央値・平均値で重みづけなしの (A) をすべて上回ったのは $T_B = 2.5$ のときの (B)、 $T_B = 2.0$ のときの (C) となった。20 分以内に遅延を検出できた割合は、 $T_B = 2.0$ のときの (A) が 66%、 $T_B = 2.5$ のときの (B) が 72%、 $T_B = 2.0$ のときの (C) が 76%、 $T_B = 2.0$ のときの (D) が 70%、 $T_B = 1.5$ のときの (E) が 80%、既存手法が 56% となった。従って提案システムでは、検出時間及び誤検出時間を考慮して、パターン

(C)/ $T_B = 2.0$ を用いる．既存手法と比較を行うと，既存手法は各時刻での遅延を検出できた割合が $T_B = 2.0$ のときの (C) よりも低い割合で推移している．以上より，既存手法と比べて速報性の観点で改善されたことが確認できる．重みづけの有効性を示した例として，2014 年 10 月 30 日 14 時 02 分に JR 総武線錦糸町駅で発生した人身事故において，(A) 及び (C) の 14 時 02 分から 14 時 29 分の遅延度 $D(l, t)$ の推移を図 10 に示す．(A)，(C) とともに $T_B = 2.0$ であるから，閾値 $\theta(l, t)$ は $T_L(l = \text{総武線})T_T(t = 14:00 - 14:29) = 2.07$ より 4.14 である．図より，重みづけがない (A) は発生 8 分後に遅延を検出しているが，(C) は発生 5 分後に遅延を検出している．これは，重み値を付与したことによって検出時間が短縮されたことを意味する．

表 6 F 値，検出時間，誤検出時間の中央値・平均値の比較

	中央値		平均値	
	F 値		F 値	
(A), $T_B = 2.0$	F 値	62.7%	F 値	58.0%
	検出時間 [分]	11.0	検出時間 [分]	32.5
	誤検出時間 [分]	0.0	誤検出時間 [分]	13.5
(B), $T_B = 2.5$	F 値	63.2%	F 値	59.3%
	検出時間 [分]	9.5	検出時間 [分]	26.8
	誤検出時間 [分]	0.0	誤検出時間 [分]	12.9
(C), $T_B = 2.0$	F 値	63.2%	F 値	58.5%
	検出時間 [分]	8.5	検出時間 [分]	24.8
	誤検出時間 [分]	0.0	誤検出時間 [分]	12.0
(D), $T_B = 2.0$	F 値	64.1%	F 値	57.6%
	検出時間 [分]	9.5	検出時間 [分]	27.1
	誤検出時間 [分]	0.0	誤検出時間 [分]	10.2
(E), $T_B = 1.5$	F 値	63.8%	F 値	56.3%
	検出時間 [分]	8.0	検出時間 [分]	21.8
	誤検出時間 [分]	0.0	誤検出時間 [分]	12.6

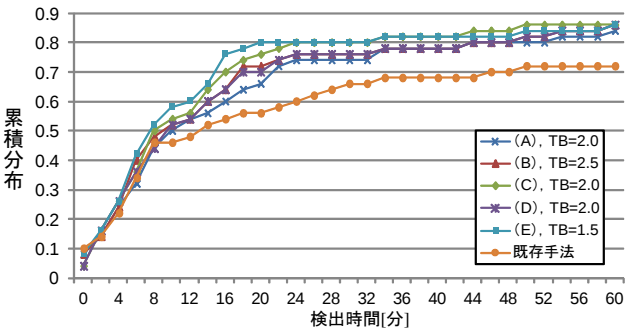


図 9 各パターンと既存手法における検出時間の累積分布 (60 分以内)

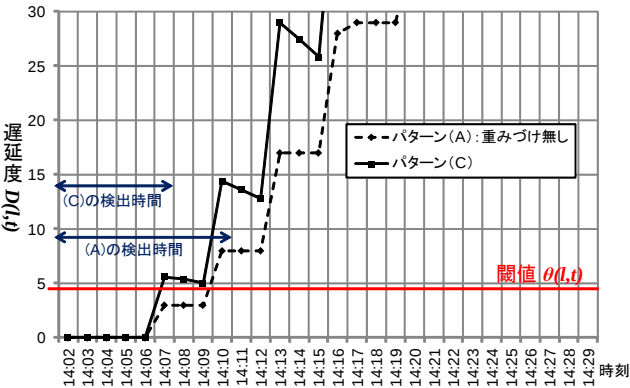


図 10 重みづけ無しの (A)，(C) の遅延度 $D(l, t)$ の推移

5.3 公式運行情報との比較

提案システムを 2014/11/30～12/15 に稼働させ，この期間に対象路線で起きた発生時刻が明確に分かる遅延 55 件の公式運行情報の第一報配信時刻と提案システムの遅延検出時刻を比較した．評価結果を表 7 に示す．表より，55 件中 45 件 (81.8%) で提案システムの方が早く遅延を検出できた．

表 7 公式運行情報の配信時刻と提案システムの遅延検出時刻の差

	該当件数	割合
提案システムの遅延検出時刻の方が早い	45 件	81.8%
公式運行情報の配信時刻の方が早い	8 件	14.5%
提案システムで遅延を検出できなかった	2 件	3.6%
合計	55 件	100.0%

提案システムの方が早く遅延を検出できた 45 件の中で，公式運行情報の第一報配信時刻と最も差があったのは，2014 年 12 月 9 日 17 時 04 分に東京メトロ日比谷線三ノ輪駅で発生した車両点検による遅延である．提案システムでは 17 時 16 分に遅延を検出したのに対し，公式運行情報の第一報配信時刻は 19 時 27 分であり，2 時間 11 分の差があった．また，2014 年 12 月 8 日 16 時 17 分に JR 中央線内で発生した人身事故の影響を受けた東京メトロ東西線の遅延は 2 時間 4 分，2014 年 12 月 4 日 7 時 39 分に東急東横線内での急病人看護の影響を受けた西武池袋線の遅延は 1 時間 29 分，2014 年 12 月 2 日 12 月 2 日 7 時 43 分に東京メトロ千代田線綾瀬～北千住駅間での車両点検による遅延は 44 分の差を確認した．一方で，2014 年 12 月 14 日 22 時 41 分に東京メトロ銀座線原町駅での乗客転落の影響による運転見合わせや 2014 年 12 月 8 日 10 時 45 分に JR 中央線豊田～日野駅間で発生した人身事故による運転見合わせは，公式運行情報の配信時刻と比較し 1 分早く検出できたが，ほとんど違いはみられなかった．この 45 件の平均差は，21.2 分だった．反対に，公式運行情報の配信時刻の方が早かった 8 件の中で，提案システムの遅延検出時刻と最も差があったのは，2014 年 12 月 4 日 19 時 03 分に小田急線渋谷～新松田駅間で発生した動物衝突による遅延である．公式運行情報は 19 時 25 分に第一報を配信したのに対し，提案システムでの遅延検出時刻は 19 時 38 分であり，13 分の差があった．しかし，この遅延を除く 7 件は，公式運行情報が第一報を配信してから提案システムで遅延を検出するまでの差がすべて 5 分以内に収まっている．この 8 件の平均差は 3.6 分だった．

以上の評価結果を踏まえ，提案システムの方が公式運行情報と比較して速報性に優れた鉄道運行情報の検出及び配信に成功しているといえ，一定の有効性を示した．

また，提案システムでは，公式運行情報を配信しないような中小規模の遅延に関しても検出することがわかった．例えば，2014 年 12 月 12 日 23 時 45 分付近に投稿された「田園都市線」を含むツイートを観察すると，「田園都市線，青葉台駅で接触事故のため運転見合わせ」「田園都市線止まって動かない」「田園都市線とまった，うごけ～」「田園都市線遅延して後ちょっとで終電逃すとこだった」といった陽性ツイートとなるツイートが多く出現する．このことより，この時刻付近に東急田園都市線

が遅延していることが推測できる。しかし、公式にはこの遅延に関する情報は配信されていない。以上より、公式運行情報が配信しないような中小規模の遅延に関しても、提案システムで検出できることを確認した。

6. インターフェース

提案するシステムのインターフェースを図 11 に示す。配信 Web ページより、対象 21 路線の最新の運行情報（「遅延・運転見合わせ」または「平常運行」）と直近 20 分間に投稿された抽出した陽性ツイートを確認できる。

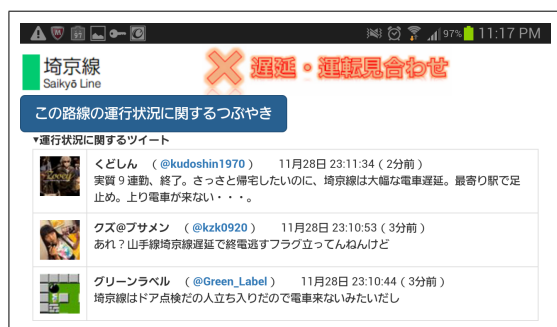


図 11 提案システムのインターフェース

7. まとめと今後の課題

本研究では、鉄道利用者により早い情報提供を行うことを目的とし、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。提案システムでは、先行研究の手法に加え、遅延要因語を新たに定義することやクライアントアプリケーションの情報を利用して投稿端末を推測することで電車の遅延に言及している陽性ツイートの誤抽出及び抽出漏れの削減を目指した。また、プロフィール欄を用いたツイートの投稿場所の推測を行うことで、同一路線名が全国に複数存在する場合の路線の特定を目指した。運行情報の推定においては、遅延時に投稿されるツイートの分析結果を踏まえて、路線や時間帯によって閾値の変動させ、陽性ツイート各件に重み値を付与することにより、速報性と正確性の向上を図った。

検討の結果、陽性ツイートの誤抽出・抽出漏れの削減に成功し、同一路線名の他路線で遅延が起きた場合に陽性ツイートの誤抽出を大幅に抑えることができた。また、最適な重み値と閾値を定めたことで、正確性の指標となる F 値の向上と遅延が発生してから検出までにかかる時間、ダイヤ回復後において誤検出の終了までにかかる時間の短縮に成功した。公式運行情報の配信時刻と提案システムの遅延検出時刻との比較では、8 割以上の遅延において提案システムの方が早く遅延を検出したという結果を得られた。以上より、既存手法と比較して速報性と正確性に優れた鉄道運行情報の配信ができたことと結論づけることができ、提案システムの有効性を明らかにした。

今後、検討すべき課題を 3 つ挙げる。1 つ目は、陽性ツイートの更なる抽出精度向上のため、運行情報語や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である。本研究では、遅延時に投稿されるツイートで高頻度で使われる語を中心に

リストを作成し、ツイート本文との照合を行うことで、陽性ツイートの抽出を行っている。しかし、ネットスラングや口語で鉄道の遅延を意味する語は対象外となっている。そのため、このような語を含むツイートは抽出漏れとなってしまう課題がある。2 つ目は、より詳細な運行情報の推定である。本研究では、推定する運行情報を「遅延・運転見合わせ」「平常運転」の 2 種類に限定している。文献 [8] では、「全線見合わせ」「一部運転見合わせ」「遅延」「平常運転」という細かい粒度で推定を行っており、詳細な運行情報を素早く知りたいという鉄道利用者の需要に応える必要がある。また、Twitter から遅延要因や遅延の発生区間の推定も行うことで、更なる実用的なシステムを目指したい。3 つ目は、運行情報推定における F 値の算出方法である。本研究では、公式運行情報に記載されている情報を実際の運行状況として評価に用いた。しかし実際は、5.3 のように中小規模の遅延に関しては公式運行情報が発表されないケースがある。そのため、5.2 において F 値を算出する過程で、小さい遅延を誤検出と受け取っていることがある。今後は、実際の運行状況は何を基準とすべきかを検討する必要がある。

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] Sakaki, T., Okazaki, M. and Matsuo, Y., “Tweet Analysis for Real-Time Event Detection and Earthquake Reporting System Development”, *IEEE Transactions on Knowledge and Data Engineering*, Vol.25, No.4, pp.919-931, 2013.
- [3] 荒牧英治, 増川佐知子, 森田瑞樹, “Twitter Catches the Flu: 事実性判定を用いたインフルエンザ流行予測”, 情報処理学会研究報告, Vol.2011-SLP-86, No.1, pp.1-8, 2011.
- [4] 古川忠延, 阿部修也, 安藤剛寿, 岩倉友哉, 志賀聡子, 高橋哲郎, 井形伸之, “Twitter からの犯罪情報抽出の可能性調査”, 情報処理学会研究報告, Vol.2011-DD-82, No.3, pp.1-6, 2011.
- [5] Sakaki, T., Matsuo, Y., Yanagihara, T., Naiwala P. Chandrasiri, Nawa, K., “Real-time Event Extraction for Driving Information from Social Sensors”, *Proceedings of the 2012 IEEE International Conference on Cyber Technology in Automation, Control and Intelligent Systems*, pp.221-226, 2012.
- [6] 村山敬祐, 佐伯圭介, 遠藤雅樹, 横山昌平, 石川博, “マイクロブログマイニングによるイベント時の駅混雑原因の特定と状況の把握”, 第 6 回データ工学と情報マネジメントに関するフォーラム (DEIM2014), B3-6, 2014.
- [7] 築瀬拓弥, 増田英孝, 山田剛一, 荒牧英治, 中川弘志, “Twitter を用いた電車遅延の自動通知”, 情報処理学会研究報告, Vol.2013-IFAT-110, No.1, pp.1-6, 2013.
- [8] 土屋圭, 豊田正史, 喜連川優, “マイクロブログを用いた鉄道の運行トラブルの発生検出および付帯情報の検出”, 第 6 回データ工学と情報マネジメントに関するフォーラム (DEIM2014), B3-2, 2014.
- [9] 蝦名亮平, 中村健二, 小柳滋, “リアルタイムバースト検知手法の提案”, 日本データベース学会論文誌, Vol.9, No.2, pp.1-6, 2010.
- [10] 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.
- [11] 湯田雅, 矢吹太郎, 佐久田博司, “Twitter 上の bot の判別による情報伝達の効率化”, 情報処理学会第 74 回全国大会, 6ZE-1, 2012.
- [12] 高橋哲朗, 野田雄也, “実世界のセンサーとしての Twitter の可能性”, 電子情報通信学会技術研究報告, Vol.110, No.400-NLC2010-38, pp.43-48, 2011.