

Twitter からの列車遅延情報収集手法の検討

新井 誠也[†] 平川 豊[†] 大関 和夫[†]

芝浦工業大学 工学部 情報工学科[†]

1. はじめに

通信サービスの発達により、駅に行かなくても Web 経由で簡単に各鉄道路線の列車運行情報を入手することが可能となった。しかし、従来サービスの多くが各鉄道会社から公式に提供される運行情報を引用して運営されており、それらは即時性に欠けるといった問題がある。刻一刻と運行状況が変化する中、即時性に優れた運行情報の提供が求められる。

一方、近年 Twitter が注目を集めている。Twitter 上では、ツイート (Tweet) と呼ばれる最大 140 字の短文のメッセージによって、即時性に優れた情報や新鮮な話題が常に発信され続けている。

本研究では、より早い情報提供を行うために、主に鉄道利用者から投稿された生の声であるツイートに焦点を当て、それらを情報源として、Web 上で列車運行情報を提示するシステムを検討する。

2. 関連研究

Twitter が持つ特長の 1 つとして、即時性に優れていることが挙げられる。近年では、その特長を生かした研究が行われている。

Sakaki らは地震の発生に言及したツイートを抽出し、位置推定を行い、それをもとに地震発生速報を流すシステムを開発した [1]。小林らはスポーツ中継の実況ツイートに着目し、盛り上がり区間を重要シーンとして検出することで、スポーツ映像の自動要約を行った [2]。

[1], [2] とともに Twitter が持つ即時性を生かした研究である。本研究では、その特長を列車の運行情報の収集に対して適用を試みる。

3. 列車遅延発生時におけるツイート件数の推移

図 1 は 2012 年 7 月 12 日に J R 京葉線で発生した風速計故障事故に関連して、キーワード「京葉線」を含んだツイート件数の推移である。ただし、リツイート (ReTweet (以下 RT): 他ユーザのツイートをそのまま引用したツイートのこと) は除いた。図 1 より、以下のことがわかる。

- ・事故発生及び運転見合わせ (21 時 12 分) 後から急激な伸びを示している。
- ・J R 東日本の公式運行情報の配信時刻 (21 時 35 分) よりも前に、J R 京葉線に関する多くのツイート (21 時 12 分から 21 時 34 分の間に計 90 ツイート) が投稿されている。

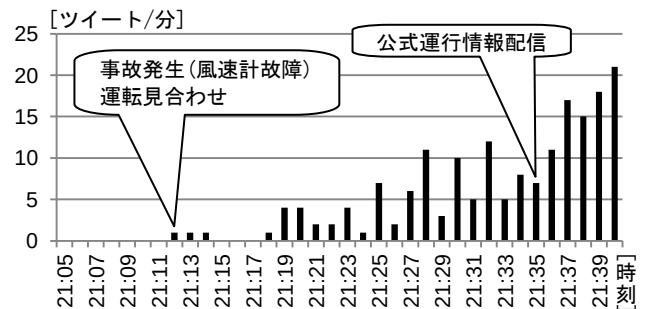


図 1 事故発生直後のツイート件数の推移

4. 提案手法

4. 1. システム概要

図 2 に提案するシステムの概要を示す。システムは、Web アプリケーションとして実装する。利用者からのアクセスにより、「〇〇線が止まった」「××線が遅れているな」といった列車遅延に関するツイートを収集する。その後、それらのツイートを解析し、各鉄道路線の運行状況を判断する。最後に、判断結果を利用者に提示する。

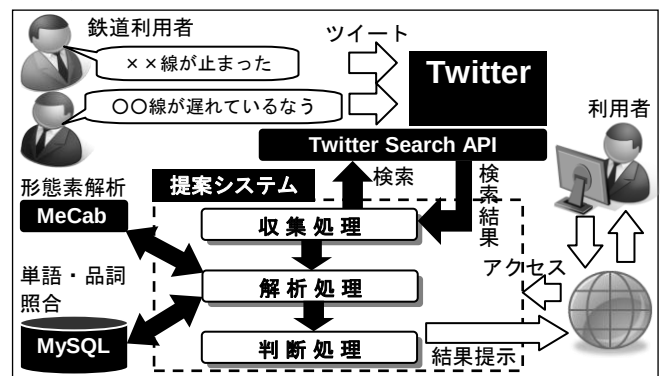


図 2 提案システムの概要

4. 2. 収集処理

Twitter Search API [3] を用いて、各鉄道路線に関するツイートを収集する (ただし、RT を除く)。検索キーワードとして各路線名を用いる。例えば、J R 山手線の場合は、「山手線」で検索する。

4. 3. 解析処理

検索でヒットした各ツイートを形態素解析にかける。形態素解析エンジンは MeCab [4] を使用する。その後、各ツイートを表 1 に示す 4 つのカテゴリのいずれかに分類する。

表 1 分類するカテゴリー一覧

運転見合わせツイート	「〇〇線止まっている」という運転見合わせに関するツイート
列車遅延ツイート	「〇〇線遅れている」という列車遅延に関するツイート
平常運転ツイート	「〇〇線は平常運転に戻りました」という遅延が解消して平常運転に戻った旨のツイート
その他のツイート	上記 3 カテゴリーのいずれにも当てはまらない列車の運行情報とは無関係のツイート

A method to gather delay information of the train in Twitter
[†]Seiya Arai, Yutaka Hirakawa and Kazuo Ohzeki,
 Department of Information Science & Engineering, Shibaura
 Institute of Technology

なお、分類手法は、ルールベースによる手法を採用した。運行状況を示すツイートにおいて出現頻度が高い単語(「止まっ」「見合わせ」「遅延」「遅れ」など)計 275 語とそれらの品詞をあらかじめデータベース(MySQL)に登録し、照合することで、分類作業を行う。表 2 に分類の性能評価結果を示す。性能評価実験は、2012 年 10 月 23 日～25 日に発生した 5 件の列車遅延を対象として遅延発生後に投稿された計 5,574 ツイートに対し、分類を試みた。

表 2 作成した分類器の性能評価結果

	精度(適合率)	再現率	F 値
運転見合わせツイート	0.95	0.79	0.86
列車遅延ツイート	0.98	0.85	0.91
平常運転ツイート	0.68	0.95	0.79
その他のツイート	0.80	0.96	0.87

4. 4. 判断処理

30 秒おきに各路線の運行状況の判断を行う。判断に際しては、運転見合わせツイート件数(T_s)と列車遅延ツイート件数(T_D)の合計が設定した閾値 θ を超えていれば、「運転見合わせ(ただし $T_D > T_s$ のとき、列車遅延)」(以下○)と判断し、超えていなければ「平常運転」(以下×)と判断するものとする。

また、路線や時間帯によって列車遅延発生時のツイート件数にばらつきがあることが確認できたため、 θ は式 1 をもって決定する。ここで、 $\alpha \in N$ とする。

$$\theta = T \times L \times \alpha \quad (\text{式 1})$$

T : 時間帯係数(時間帯ごとに定義), $0 < T \leq 1$
 L : 路線係数(路線ごとに定義), $0 < L \leq 1$ α : 基本閾値

T, L の設定は[5]を参考にした。例えば、 T は朝夕ラッシュ時には高い値、閑散時には低い値とした。

なお、基本閾値 α は以下の手順で決定した。

<定義>

列車遅延期間…公式配信に基づく列車遅延があったとされる期間

正受入: 列車遅延期間に○と判断した

正拒否: 列車遅延期間外に×と判断した

誤受入: 列車遅延期間外に○と判断した

誤拒否: 列車遅延期間に×と判断した

誤受入率: 誤受入数/(正拒否数+誤受入数)

誤拒否率: 誤拒否数/(正受入数+誤拒否数)

<実験>

- ・2012年11月18日8:00～19日20:00に投稿されたJR山手線に関するツイート(キーワード「山手線」を含むツイート、RTを除く)計7,417ツイートを収集・分類した。期間内に発生した公式配信があった列車遅延は2件である。
- ・上記を実験材料とし、|誤受入率-誤拒否率|が最小となる値を α として採用した。

判断材料として、まず直前30秒間に投稿されたツイートをを用いる手法(手法1)の検討を行った。しかし、この手法では毎回の判断をそれぞれ独立な事象として扱っているため、ノイズの誤検出が多いという問題があった。そこで次に、直前10分間に投稿されたツイートをを用いる手法(手法2)の検討を行った。これは前に起こった事象を一定程度考慮しながら、毎回の判断を行うものである。

その結果、手法2の適合率に大きな改善がみられ、正受入の割合を示した正受入率($=1-\text{誤拒否率}$)も手

法1より手法2の方が高いことがわかった(表3)。従って提案システムでは、手法2を採用した。

α は図3より、24とした($\alpha=24$ のとき、誤拒否率 $=0.0764$, 誤受入率 $=0.0793$)。

表3 手法1, 手法2の適合率・再現率・ α における正受入率

	適合率	再現率	F 値	正受入率
手法1	0.15	1.00	0.26	0.85
手法2	0.63	1.00	0.77	0.92

※公式配信があった列車遅延各件における検出の有無を評価した。評価に使用したのは、公式配信があった列車遅延時に投稿されたツイート群(A)と平常運転時に投稿されたツイート群(B)の各50時間分のデータである。Bは最初の○判断(誤受入)から30分間を同伴とみなすことで、誤検出件数を求めた。

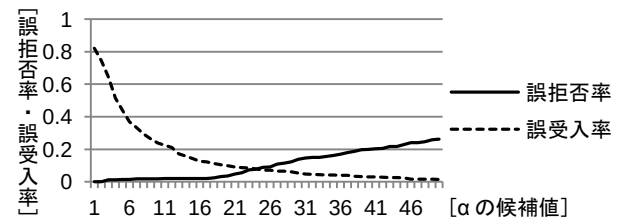


図3 手法2における誤拒否率・誤受入率

5. 評価実験

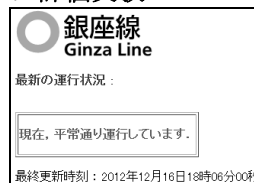


図4 インターフェース

以上を踏まえて提案システムを実際に稼働させた(図4)。稼働期間内に発生した公式配信があった列車遅延 41 件を対象として公式運行情報と提案システムとの配信時刻の差を検証した。対象とした路線は、首都圏を運行するJR、地下鉄、私鉄の全 21 路線である。評価結果を表4に示す。

表4 配信時刻の差の評価結果

項目	該当件数(%)	最大差	平均差
提案システムの方が早い	31 件(75.6%)	76 分	13.22 分
公式配信の方が早い	6 件(14.6%)	10 分	4.83 分
同時刻の配信	1 件(2.4%)	0 分	0 分
提案システムで検出できず	3 件(7.3%)	—	—
合計	41 件(100.0%)	—	—

※稼働期間: 2012 年 12 月 10 日 13:00～24 日 13:00

6. おわりに

本研究では、鉄道利用者のツイートを情報源として列車運行状況の判断を行うシステムを提案した。

その結果、全体の 75.6%で公式運行情報配信時刻よりも前に、提案システムにおいて列車遅延を検出・配信することができ、一定の有効性を示した。

今後は、機械学習による分類器の導入及び性能比較、ノイズの誤検出を更に低減させるための手法の検討が課題となる。

参考文献

- [1] Sakaki, Okazaki, Matsuo, "Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors", Proc. WWW'10, pp.851-860, 2010.
- [2] 小林, 野田, 高橋, 井出, 村瀬, 出口, "Twitter の実況書き込みを利用したスポーツ映像の要約", 信学技報, Vol.110, No.457, MVE 2010-162, pp.165-169, 2011.
- [3] GET Search <https://dev.twitter.com/docs/api/1/get/search>
- [4] MeCab <http://mecab.sourceforge.net/>
- [5] 国土交通省, "第 11 回大都市交通センサス", 2010.