# ソーシャルセンサからの情報抽出技術

## Technology for Real-Time Information Extraction from Social Sensors

長野　伸一
■NAGANO Shinichi

**スマートフォンの普及に伴い，利用者が移動中や外出中にソーシャルメディアに投稿する利用形態が広まってきている。**
**東芝は，ソーシャルメディアへの利用者の投稿をソーシャルセンサとみなし，そのテキスト内容を解析して社会インフラに関する情報を抽出する技術の開発を進めている。**
**今回当社は，鉄道の運転見合わせや遅延などの運行情報を対象として，Twitter(†)への投稿から情報を抽出する技術を開発し，首都圏の鉄道路線を対象とした評価実験でその有用性を確認した。今後も，社会インフラ分野でのソーシャルメディア活用の可能性を追求していく。**

With the wide dissemination of smartphones in recent years, the posting of messages to social media in various situations while out of the office or away from home has become popular.

Toshiba has been researching a novel technology to extract information related to social infrastructure systems from the text contents of social media by handling them as social sensors. As part of this research, we have developed a method of extracting train status information from Twitter(†) data to show suspensions or delays in train services. Through verification tests applying this method to commuter train lines in the Tokyo metropolitan area, we have confirmed its effectiveness in terms of high accuracy and real-time characteristics and are now working toward its application to the social infrastructure field.

## 1　まえがき

様々なICT（情報通信技術）ソリューションをシステムとして有機的に結び付けることで，エネルギーや資源の効率的利用を中心とする新しい街づくり“スマートコミュニティ”への取組みが各地で進められている[1]。スマートコミュニティを構成する社会基盤の一つである交通分野では，環境やコストに配慮する一方で，円滑で快適な移動手段の提供が求められている[2]。その実現に向けて，GPS（Global Positioning System）やカメラなどの物理センサに加えて，ソーシャルセンサの活用も検討されている。

ソーシャルセンサは，ソーシャルメディア上で利用者が投稿した発言をセンサ情報の一種とみなし，発言内容を解析して社会インフラに関する情報を抽出するものである[3], [4]。代表的なソーシャルメディアの一つにTwitter(†)がある。Twitter(†)は，利用者がツイートと呼ばれる最大140文字のテキストを投稿できるサービスで，2011年10月時点で，日本国内での利用者数が約1,400万人以上に上る[5]。スマートフォンの普及とあいまって，利用者が外出中に投稿する利用形態が増えており，身の回りで見聞きしたり経験したりしたでき事も数多く投稿されている。利用者の目線で観察された情報が，物理センサの情報を補完するものとして期待されている。

東芝は，鉄道分野におけるソーシャルセンサの活用に着目し，Twitter(†)の利用者による鉄道関連のツイートの中から，鉄道の遅延や運転見合わせなど運行情報を抽出する技術を開発した。ここでは，その技術の概要と有用性の評価，ソーシャルセンサの可能性について述べる。

## 2　ソーシャルメディアに見られる実社会の動き

### 2.1　狙い

鉄道事業者は，事故などが発生して鉄道ダイヤの乱れが生じると，各事業者のWebサイト上で，遅延や運転見合わせ，及び運転再開に関する情報を順次発表しているが，利用者へ情報が伝わるのが遅れたり，適切に伝わらないことが起こっている。当社が確認したところ，事故から1時間以上経過してから第1報が発表された場合があった。また，鉄道事業者によっては，30分以上の遅延の発生やその見込みがある場合だけ発表をしている。一方，鉄道事業者側の責任で30分程度の遅延時間が発生した場合，それを許容できる利用者はわずか11.6％しかいないとの調査結果もある[6]。

利用者の視点で考えると，事故の影響の大きさに関わらず，鉄道の運行情報に対するアクセシビリティを高めることが望ましい。利用者が運行情報をいち早く知ることができれば，う回経路を利用する，別の移動手段を利用する，あるいは予定を変更するなど，状況に応じた行動を選択することができる。

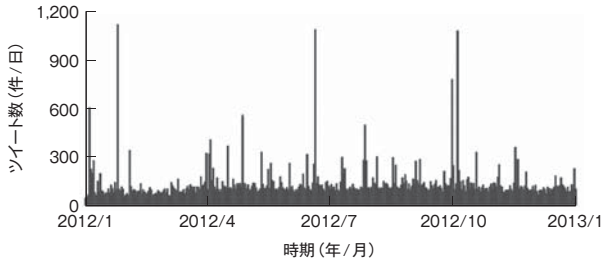


**図1.“山手線”を含むツイート数の時系列変化** ― 自然災害などによりダイヤの乱れが生じると，ツイート数は大きく変動する。

Time-series variations in number of tweets related to Yamanote Line

### 2.2　ソーシャルメディアへの投稿数の推移

Twitter(†)上の鉄道利用に関するツイートを題材としてその変化を観察するために，Twitter社がWeb上で公開している検索API（Application Programming Interface）を利用して，“山手線”を含むツイートを，2012年1月1日から12月31日にわたり50,638件収集した。投稿日ごとに集計したツイート数の推移を**図1**に示す。

1年間を通して，ほぼ毎日150～200件のツイートがあったほか，ツイート数が300件を超えた日が17日あることがわかった。また，1,000件を超えた日が3日あり，そのうち1月25日は，新宿駅−新大久保駅間での火災発生と携帯電話の大規模な通信障害が重なったことによるもの，6月20日は台風4号の影響による線路内への資材の散乱によるもの，そして10月4日は人身事故の影響によるもので，いずれも十万人規模の鉄道利用者に影響があったとされている。このように，平常時と災害時とでツイート数は大きく異なり，自然災害や人的災害の発生による影響や混乱の度合いによって，ツイート数は大きく変動する。

### 2.3　ソーシャルメディア投稿数と鉄道輸送人員の相関性

JR山手線のように，利用者数の多い鉄道路線ほど，ソーシャルメディアへの投稿が多いことは容易に予想される。そこで，鉄道路線ごとの平均輸送人員とツイート数の相関関係を調べた。平均輸送人員は，関東交通広告協議会が公表している2010年度の統計データ(7)を利用し，平均ツイート数は，2011年7月28日から10月11日の期間を対象にして，路線名をキーワードとして収集したツイートから算出した。首都圏の68路線に対する，1日当たりの路線ごとの平均輸送人員と平均ツイート数の散布図を**図2**(a)に示す。平均輸送人員と平均ツイート数の相関係数は0.83で，正の相関があることがわかった。

一方，災害や事故が発生したときの例として，爆弾低気圧と呼ばれる急激に発達した低気圧が日本各地を襲った，2012年4月3日に着目した。この日は，首都圏での強風のピーク時間が帰宅ラッシュの時間帯と重なり，大規模な鉄道の運行制限や運休が発生した。その散布図を図2(b)に示す。平均輸送人員とツイート数の相関係数は0.13で，相関性は見られなかった。路線ごとの平常時の平均ツイート数と比較すると，災害発生日のツイート数は影響の大きかった相鉄本線で約700倍も多く，平常時の平均ツイート数が非常に少ない東武鬼怒川線でも約10倍もあった。このことから，鉄道路線の利用者数に関わらず，平常時に対するツイート数の増加率を調べることにより，鉄道路線で発生したダイヤの乱れを観測できる。

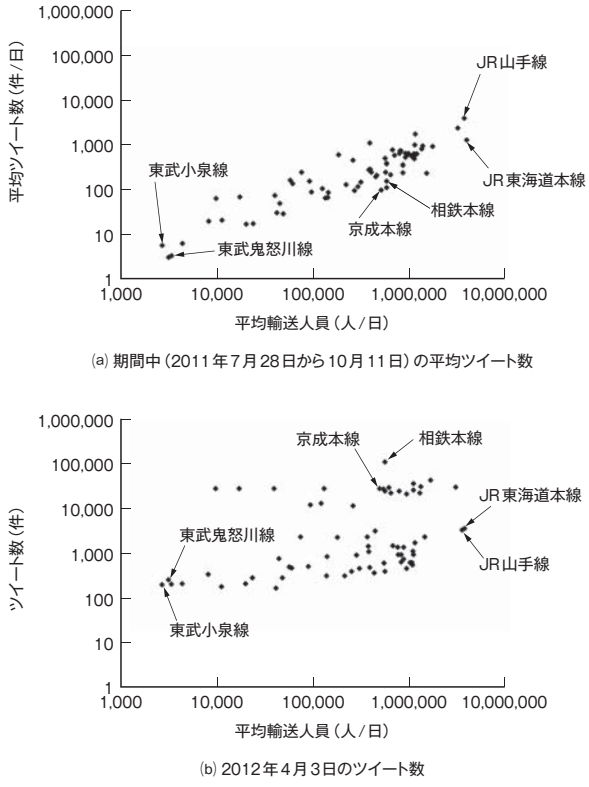


**図2．路線ごとの平均輸送人員とツイート数の散布図** ― 調査期間中の1日当たりの平均輸送人員と平均ツイート数の間には，正の相関があることがわかった。一方，自然災害が発生した日には，相関性が見られなかった。

Scatter diagrams to compare average number of tweets during testing period and number of tweets on one day when natural disaster occurred

このように，2.2節と2.3節の調査結果から，鉄道分野においては，実社会で起こったでき事に応じてツイート内容が変化しており，Twitter(†)は，実社会の状況をある程度反映するソーシャルセンサとして活用できる可能性があると考えられる。

## 3　ソーシャルメディアからの事象抽出

### 3.1　鉄道運行情報抽出の課題

Twitter(†)から鉄道運行情報を抽出するにあたり，大きく二つの課題が存在する。

一つ目の課題は，情報の不確実性である。例えば，“山手線”と“止まった”を含むツイートを検索すると，次のような例が収集される。

(1)　山手線が止まったので帰れない

(2) 埼京線が止まったので，山手線に乗り換えた
(3) 最近，よく止まるな RT @taro 山手線が止まった
(4) @hanako 山手線が止まったので，カフェで時間潰す

(1)からは，そのときに山手線が止まっているようすが読み取れるが，(2)，(3)，(4)からはそのようなことはわからない。(2)では，止まっているのは埼京線であり，山手線は通常運行している可能性が高い。(3)と(4)では，山手線が止まっていたのは過去のことであり，そのときは通常運行に戻っているかもしれない。ツイートから鉄道の運行情報を抽出する場合，両者を区別することが重要であり，路線のリアルタイムな運行情報に言及しているツイートを“運行情報ツイート”と呼ぶ。(1)～(4)のうち，(1)だけが運行情報ツイートである。

二つ目の課題は，路線によってツイート数が異なる点にある。首都圏の主要路線では，数分の遅延が発生しただけで，1分当たり数十件から数百件ものツイートが投稿される。一方，平常時から鉄道利用者が少ない路線では，数十分の遅延が発生しても，その事を話題にしてTwitter(†)に投稿する人は非常に少ない。ツイートから鉄道運行情報を抽出する場合，路線ごとのツイート数の違いを考慮することが必要となる。

### 3.2 抽出技術の概要

Twitter社の検索APIを利用して収集した，テキスト内容に対象鉄道路線の名称が含まれるツイートを，事象抽出処理の入力とする。抽出技術は，運行情報ツイートの抽出及び運行情報の推定から構成される。

まず，運行情報ツイートの抽出では，テキスト内容に含まれるキーワードとその出現位置に着目し，運行情報に言及している運行情報ツイートを抽出する。キーワードには，“遅延”や“見合わせ”など，運行情報発生時のツイートだけに多く見られる単語を選定している。また，キーワードの出現位置に関しては，路線名とキーワードが同じ文節に出現するものを抽出している。(2)では，止まったのは埼京線であり，山手線ではない。また，(3)や(4)のように，キーワードが引用部分や返信メッセージに出現するツイートは除外する。これらのツイートは，ほかの利用者とのコミュニケーションの中でキーワードが出現しているもので，発信した利用者本人が必ずしもその運行状況に遭遇したとは限らないからである。

次に，運行情報の推定では，各鉄道路線に対して，運行情報ツイートの件数を所定時間$N$分を単位として集計する。平常時のツイート数をもとに各鉄道路線に設定したしきい値$T$を参照して，運行情報ツイートの件数が$T$を超えていれば，運行情報ありと判定する。

## 4 有用性の評価

### 4.1 評価指標

運行情報ツイートを抽出するときの評価指標には，情報抽出の研究分野で一般に利用される，再現率，適合率，及びF値を利用する。再現率は抽出漏れの少なさを，また，適合率は抽出誤りの少なさを表す。また，F値は再現率と適合率の調和平均で定義され，抽出漏れの少なさと抽出誤りの少なさをバランスした指標として，広く利用されている。

運行情報推定の精度は，運行情報の抽出時間の短さで評価する。抽出時間とは，運行情報が実際に発生した時刻と，この技術で最初に抽出されたツイートの投稿時刻の差分である。

**表1. 評価結果**
Results of verification tests

| 路　線 | 平均ツイート数（件/日） | 抽出時間（分） | 再現率（%） | 適合率（%） | F値（%） |
|---|---|---|---|---|---|
| JR山手線 | 3,761 | 1.0 | 82.4 | 96.3 | 88.8 |
| JR中央線 | 2,262 | 3.3 | 92.9 | 94.5 | 93.7 |
| JR総武線 | 1,212 | 5.0 | 93.6 | 89.9 | 91.7 |
| JR横須賀線 | 567 | 1.0 | 98.8 | 61.8 | 76.1 |
| JR高崎線 | 283 | 4.0 | 97.5 | 49.8 | 65.9 |
| JR川越線 | 85 | 6.0 | 88.9 | 88.0 | 88.4 |
| 京王井の頭線 | 473 | 2.0 | 97.4 | 71.4 | 82.4 |
| 東京メトロ副都心線 | 438 | 6.0 | 60.0 | 33.3 | 42.9 |
| 東武東上線 | 618 | 4.0 | 97.4 | 79.4 | 87.5 |
| 東急池上線 | 128 | 3.0 | 88.1 | 68.9 | 77.3 |
| 全体平均 | − | 3.5 | 92.1 | 79.5 | 85.3 |

### 4.2 評価結果

首都圏10路線を対象に収集したツイートを使って実施した評価結果を**表1**に示す。

各路線に対して，1日当たりの平均ツイート数をもとに最適な$T$を設定した。再現率，適合率，及びF値のいずれも高く，特に再現率の平均は90％を超えており，抽出漏れの少ない技術であることがわかる。また，抽出時間は平均で3分台程度となり，この技術の有用性を確認できた。

### 4.3 Android(†)アプリケーションの開発

この技術を組み込んだAndroid(†)アプリケーション“路線実況”を開発し，（株）駅探と共同で，2012年3月29日から5月29日に一般公開の実証実験を実施した。アプリケーションは，ツイートの中から，鉄道の運行情報に関連するものだけをリアルタイムで抽出し，時系列（タイムライン）で表示する。首都圏の158の鉄道路線を対象として，利用者が路線一覧から手動で選ぶほか，Android(†)端末に搭載の加速度センサとGPSを利用して自動推定された，乗車中又は近隣の路線の中から選択できる。これにより利用者は，自身の状況に応じて，少ない操作回数で最新の運行情報へアクセスできる。

アプリケーションの画面例を**図3**に示す。路線一覧では，運行情報が抽出された路線名欄の右側に“！”マークが表示される。路線名を一つ選択すると，その路線に関する最新の運行情報ツイートを閲覧できる。

実証実験の期間中に計1万件以上のダウンロードがあった。


**図3. Android㈱アプリケーション“路線実況”の画面例** ― 路線一覧画面では，運行情報が抽出された路線名欄の右側に“！”マークが表示される（左）。路線名を一つ選択すると，その路線に関する最新の運行情報ツイートを閲覧できる（右）。

Example of screenshot of "Rosen-jikkyō" (real-time train status) application

また，アプリケーションのメニュー内に設置した利用者アンケートからは，「現地の詳しい状況を早く入手できる」や「悪天候時に役だつ」など，利便性を評価するコメントが寄せられた。また，「乗換えで次に利用する路線の情報も知りたい」，「運行状況の変化がわかると，待つべきか代替手段を利用するべきかの判断ができる」といった機能改善に対する要望も見られた。このように，鉄道運行情報の抽出機能に対して一定の有用性を確認することができた。

## 5　ソーシャルセンサの可能性

技術面においては，このようにテキスト内容の解析に基づく手法では，解析のための知識（単語辞書や抽出ルール）の整備が不可欠であり，その整備状況が解析精度に直結する。ソーシャルメディアにおいては日々新しい単語や省略語が登場するため，知識の定期的な更新が必要となる。流行や地域などに左右されず，同じ表現のキーワードが利用される分野においては，比較的低いコストで情報抽出が可能である。そのような例としては，ここで述べた鉄道運行情報や，地震の発生地域，インフルエンザの流行地域の特定などが挙げられる。抽出した情報をフィードバックし，利用者の利便性を高めることが，ソーシャルメディアに投稿される情報の質と量の向上につながっている。

鉄道以外の交通インフラへの応用では，道路の渋滞や事故などの情報抽出が考えられる。現在のソーシャルメディアは，テキストデータを投稿するものが中心であり，自動車の運転中には携帯電話を操作できないため，道路情報に関するツイート投稿は極めて少ない。今後，運転手の音声や視線などの情報を活用した新しいサービスの開発や，自動車のプローブ情報や道路に設置したカメラ映像などの物理センサと，ソーシャルセンサを融合させた状況把握技術の開発が望まれる。

これまでソーシャルメディアは，パソコンからの投稿が一般的で，一般消費者向けに販売された商品や開催されたイベントに対する効果を測定する，クチコミ分析の情報源として位置づけられてきた。近年，スマートフォンの普及に伴い，移動中や外出中に身の回りで見聞きしたり経験したりしたでき事について，その場で投稿する利用形態が広まっている。

社会インフラ向けの様々な製品を開発し製造している当社にとって，ソーシャルメディアはインフラ利用者とのチャネルとなるものである。インフラ利用者が抱える課題や求める価値をソーシャルメディアから獲得することで，“モノ”から“コト”へシフトした新しい社会インフラサービスを実現していく。

## 6　あとがき

鉄道分野でのソーシャルセンサの活用について述べた。今後，社会インフラを利用する市民の声をソーシャルメディアから収集し分析することで，社会インフラの運用やサービスを改善し，効率性と利便性の向上に取り組んでいく。

## 文　献


(1) Li, X. et al. Smart Community: An Internet of Things Application. IEEE Communications Magazine. **49**, 11, 2011, p.68－75.
(2) 野村総合研究所. ビッグデータ革命. アスキー・メディアワークス, 2012, 208p.
(3) Sheth, A. Citizen Sensing, Social Signals, and Enriching Human Experience. IEEE Internet Computing Magazine. **13**, 4, 2009, p.87－92.
(4) 榊　剛史 他. ソーシャルセンサとしてのTwitter. 人工知能学会誌. **27**, 1, 2012, p.67－74.
(5) ネットレイティングス. 2011年10月の日本の主要SNSサイトの動向. <http://www.netratings.co.jp/news_release/2011/11/sns-report-Oct-2011.html>, (参照2014-05-08).
(6) NTTデータ経営研究所. 社会インフラにおける停止許容時間についての調査. <http://www.keieiken.co.jp/aboutus/newsrelease/090907/index.html>, (参照2014-05-08).
(7) 関東交通広告協議会. 平成22年度1日平均乗降人員・通過人員. <http://www.train-media.net/report/1110/1110.html>, (参照2014-05-08).





**長野　伸一　NAGANO Shinichi, Ph.D.**
研究開発センター システム技術ラボラトリー主任研究員，博士（工学）。Linked Data，セマンティックWebの研究・開発に従事。電子情報通信学会，情報処理学会，人工知能学会会員。
System Engineering Lab.