D-021

鉄道の利用状況を用いた Twitter ユーザの行動の推定
Estimation of Twitter User's Behavior using Their Usage of Railway

河邉 拓也†　山田 剛一†　絹川 博之†
Takuya Kawabe　Koichi Yamada　Hiroshi Kinukawa

1. はじめに

近年，Twitter に代表されるマイクロブログが普及したことで，ユーザの状況を即座に投稿，共有することが可能となった．投稿内容は，ユーザの行動や属性，趣味嗜好を含む場合がある．人の行動や属性，趣味嗜好を知ることはマーケティングや鉄道の混雑予測を行う際に重要な情報となる．ユーザの行動を得るために鉄道の利用状況を用いることで，投稿時の位置とユーザの動線を得ることが可能である．

本研究では，マイクロブログのひとつである Twitter を用い，Twitter ユーザの鉄道の利用状況やその他の行動とツイートの特徴から路線または駅利用者の特徴と傾向を分析する．そのためには，まず Twitter ユーザごとに利用する路線・駅を抽出する必要がある．よって，本稿ではツイートの内容から Twitter ユーザごとの鉄道の利用状況の基本情報である利用路線・駅を抽出する手法を提案し，ユーザごとの行動の推定を行う．提案手法の流れを**図 1** に示す．

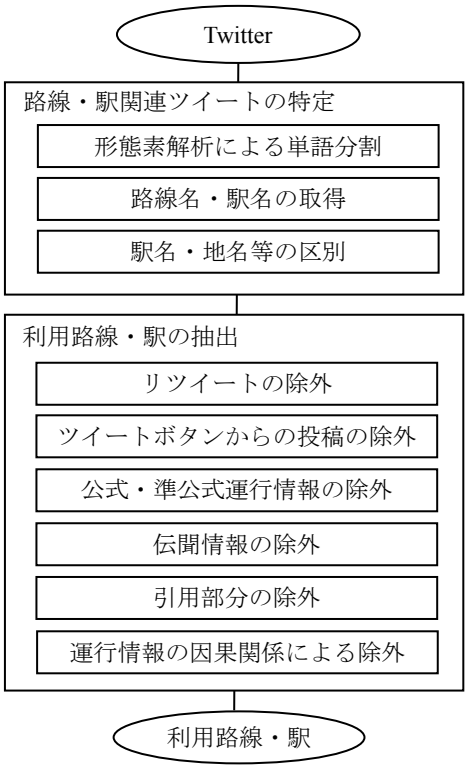

図 1　提案手法の流れ

†東京電機大学大学院　未来科学研究科
Graduate School of Science and Technology for Future Life, Tokyo Denki University

2. 路線・駅関連ツイートの特定

Twitter ユーザのツイートから路線・駅の記述があるツイートを特定する．そのために，形態素解析による単語分割，路線名・駅名の取得，駅の判別を行う．これにより，ツイート中に含まれる路線・駅を取得し，路線・駅関連ツイートの特定を行う．

2.1 形態素解析による単語分割

ツイートを形態素解析し，品詞が名詞であるものを取得する．形態素解析には， lucene-gosen [1] を用いる．解析辞書には， IPA 辞書に駅データ.jp [2] で提供されている全国の路線名，駅名を追加したものを用いる．

2.2 路線名・駅名の取得

2.1 で取得した名詞から駅情報データベース [2] を利用し，路線名もしくは駅名として使われている可能性のある語を取得する [3].

2.3 駅名・地名等の区別

路線名として使われている可能性のある語は必ず路線名であるのに対し，駅名として使われている可能性のある語は地名等のように駅名として使われていない場合がある．そのため，駅名として使われている可能性のある語が駅名であるかの判別を行う必要がある．

(1) ツイート中に路線名を含み，駅名として使われている可能性のある語がその路線に所属する駅の駅名であれば，その語は駅名である．
(2) 駅名として使われている可能性のある語の後ろの文字が「駅」であれば，その語は駅名である．
(3) 駅名として使われている可能性のある語の後ろが「北口」や「南口」等の出口を表す語であれば，その駅名として使われている可能性のある語は駅名である．
(4) ツイート中に「ホーム」や「電車」等のような鉄道に関連する語句を含む場合，そのツイートに含まれる駅名として使われている可能性のある語は駅名である．

上記の条件により，駅名として使われている可能性のある語が駅名として使われているかを判別する．

3. 利用路線・駅の抽出

2. で特定したツイートからユーザが利用している路線と駅の抽出を行う．その方法として，路線・駅の利用とはならないツイートに含まれる路線・駅，または利用していないと判断できる路線・駅を除いたものをユーザが利用している路線・駅として抽出を行う．

除外方法について，以下に述べる．

Copyright © 2014 by
The Institute of Electronics, Information and Communication Engineers and
Information Processing Society of Japan All rights reserved.

3.1 リツイートの除外

リツイートは自身が投稿したものではなく他のユーザが投稿したものであるので，これに含まれる路線・駅は利用していないとわかる．そのため，リツイートに含まれる路線・駅は利用路線・駅からは除外する．

3.2 ツイートボタンからの投稿の除外

ツイートボタンとは，ユーザが興味や関心をもった Web ページに関する情報を Twitter に投稿し，共有するシェアボタンであり，これにより投稿されたツイートは，ユーザの鉄道利用とは関係がないため，このツイートに含まれる路線・駅は利用路線・駅から除外する．ツイートには，どのクライアントから投稿されたかを示す source という情報が付与されている．source が「Tweet Button」であるツイートはツイートボタンからの投稿である．これを利用し，ツイートボタンからの投稿を検出する．

3.3 公式・準公式の鉄道運行情報ツイートの除外

鉄道会社のアカウントから投稿された鉄道運行情報ツイートや鉄道会社が発表した運行情報または鉄道運行情報の配信サービスを行っている会社からの情報をもとに生成されたツイートは，鉄道の利用を表していないため，これらのツイートに含まれる路線・駅は利用路線・駅から除外する．そのための公式または準公式の鉄道運行情報ツイートの検出方法として，公式または準公式の鉄道運行情報ツイートによくみられる語句や表現が，そのツイート中にどれくらいの頻度で現れるかを調べる．その頻度が閾値以上の場合，そのツイートは公式または準公式の鉄道運行情報ツイートとし，閾値未満の場合，そうでないツイートとする．これにより，公式または準公式の鉄道運行情報を検出する．

3.4 伝聞情報に含まれる路線・駅の除外

「～らしい」等の伝聞情報は投稿ユーザが実際に体験したことではないので，伝聞情報に含まれる路線・駅は鉄道の利用を表していない．よって，伝聞情報に含まれる路線・駅は利用路線・駅から除外する．そのためには，伝聞情報を検出する必要がある．路線・駅を含む伝聞情報には鉄道運行情報に関連するツイートが多く，このようなツイートにはいくつかのパターンがある．そのパターンに合致する部分を伝聞情報として検出する．

3.5 引用部分に含まれる路線・駅の除外

「RT」以降の文やダブルクォーテーションで括られた文等は他のユーザのツイートの引用部分になるので，そこに含まれる路線・駅は投稿ユーザの鉄道の利用を表していない．そのため，引用情報に含まれる路線・駅は利用路線・駅から除外する．

3.6 鉄道運行情報の因果関係による除外

鉄道運行情報を含むツイートは以下の例のように因果関係が存在する場合がある．

「京浜東北線で人身事故があって、中央線下り線が 7 分遅れ。やられたー」

この例の場合，「京浜東北線で人身事故」が原因であり，「中央線下り線が 7 分遅れ」が結果である．このような時，結果に含まれる路線・駅は利用路線・駅になり得るが，原因に含まれる路線・駅は利用路線・駅とはならない．そのため，原因に含まれる路線・駅は利用路線・駅から除外する．この時，因果関係が存在する鉄道運行情報は，ツイート中から「があって」等の文字列探し，存在した場合にその前後が鉄道運行情報であるかを判断することで検出を行う．

4. 考察

ツイートの中には，ひとつのツイートからでは駅名として使われている可能性のある語が駅名として使われているのか，地名等として使われているのかわからないが，前後のツイートを確認することでどちらとして使われているのかわかるものがある．本稿で提案した手法ではそのようなツイートに含まれる駅の取得に対応していない．よって，前後のツイートの情報を利用した新たな駅の判別方法が必要である．

本稿で提案した手法では，ニュース記事やまとめサイトを参照するツイートに含まれる路線・駅を利用路線・駅と抽出してしまう場合がある．ニュース記事やまとめサイトを参照するツイートは，ツイートボタンからの投稿の除外で対応できているものもあるが，source が「Tweet Button」でないために起こる誤抽出も存在する．ニュース記事やまとめサイトを参照するツイートについては，新たな検出方法が必要である．

5. おわりに

本稿ではツイートの内容から Twitter ユーザごとの鉄道の利用状況を抽出する手法を提案した．今後，本稿の提案手法の有効性について実験評価し，改良する必要がある．また，ユーザの利用駅から利用した可能性のある路線を推定する手法を提案していきたい．さらに，ユーザの特徴を取得する手法も提案し，本研究の目的である路線または駅の利用者ごとの特徴や傾向の分析をしていきたい．

謝辞

本研究に際して使用させていただいた lucene-gosen の開発者の方々，「駅データ.jp」開発者の方々に深く感謝いたします．

参考文献

[1] lucene-gosen : http://code.google.com/p/lucene-gosen/
[2] 駅データ.jp : http://www.ekidata.jp/
[3] 河邉拓也, 山田剛一, 絹川博之, “複数の鉄道運行情報を含む Tweet に対応した運行情報の抽出と提供”, 情報処理学会第 76 回全国大会講演論文集第 2 分冊, pp.103-104 (2014).

Copyright © 2014 by
The Institute of Electronics, Information and Communication Engineers and
Information Processing Society of Japan All rights reserved.