6K-06

鉄道の混雑検出センサとして機能する Twitter ユーザの推定

亘理 湧† 豊田 哲也† 大原 剛三†

†青山学院大学 理工学部

1. はじめに

近年，都市部では，鉄道網の発達と利用者の増加に伴い混雑が頻発し，それは利用者の行動選択に影響を及ぼしている．そのため，混雑情報の早期取得は利用者にとって有益であるといえ， 既存サービスの 1 つである NAVITIME の「こみれぽ」[1]では，利用者が乗車中の電車の混み具合を投稿し，その情報を共有できる．しかし，混雑情報は投稿数が少なく，能動的な情報収集が効果的である．そこで本研究では，SNS の 1 つである Twitter を対象に，鉄道の混雑検出センサとして機能する Twitter ユーザの推定を行う．

2. 提案手法

2.1 提案手法の概要

提案手法の流れを図 1 に示す．今回は遅延による混雑を対象とし，遅延情報サイトから収集した遅延関連ツイートから混雑関連ツイートを抽出する．そして，その混雑関連ツイートを用いて，混雑状況を表現する単語・単語の組合せを抽出し，混雑表現辞書を作成する．実際の利用時には，その辞書に基づいて新たに取得した遅延関連ツイートから混雑検出に有用なツイートを抽出し，それらの投稿者を混雑検出センサとする．

2.2 遅延関連ツイートの分類

本研究では，混雑関連ツイートの収集元として，「電車遅延なう」[1]を用いる．ただし，ここから収集したツイートは遅延関連のツイートであるため，これを混雑関連のツイートとその他のツイートとに分類する．具体的には，遅延関連ツイートから MeCab[2]を用いて名詞，動詞，形容詞を抽出し，それらの頻度を要素とするベクトルで各ツイートを表現し，分類器を作成する．本研究では，遅延関連ツイートから無作為に抽出した正例（混雑関連ツイート）と負例（その他のツイート）各 250 件を訓練データとし，分類器として SVM を用いた．実際に本研究で用いた分類器の 10 分割交差検定による F 値は 0.829 となった．

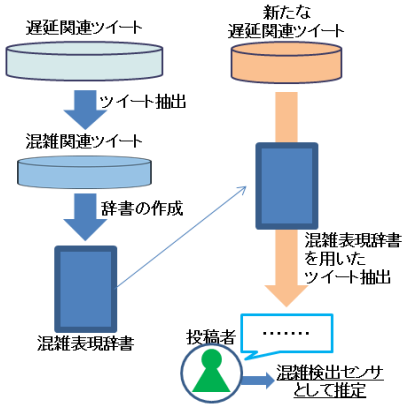

図 1. 提案手法の概要

2.3 混雑表現辞書

本研究で構築する混雑表現辞書は，混雑ワード，駅名ワード，比較ワードの 3 種類の単語から構成される．以下，それぞれの定義について述べる．

2.3.1 混雑ワード

本研究では，「混雑」，「混む」などの電車内もしくは駅ホーム上の混雑を表す単語を混雑ワードとする．また，「人」といった乗客を表現する単語と，「多い」といった乗客数を表現する単語は，出現する順序に関係なく，その組合せも鉄道の混雑を表すことができるため，そのような単語の組合せを 1 つの混雑ワードとして利用する．

2.3.2 駅名ワード

今回は首都圏 20 路線（JR16 路線，私鉄 4 路線）の駅の中でも，第 11 回大都市交通センサス[3]によって首都圏と定義された駅の駅名とその略称を駅名ワードとする．

2.3.3 比較ワード

「比較的」，「予想外」などの通常時と現在の混み具合を比較する単語を比較ワードとする．また，「普段」，「いつも」といった時間を表現する単語が先に出現し，「通り」，「以上」といった比較を表現する単語が後に出現すれば，

Estimation of Twitter Users Serving as Railway Crowdedness Sensors

Yu WATARI†, Tetsuya TOYOTA† and Kouzou OHARA†

†College of Science and Engineering, Aoyama Gakuin University

[1] http://feed.fkoji.com/train/

[2] http://www.mlit.go.jp/sogoseisaku/transport/sosei_transport_tk_000034.html

[3] http://www.mlit.go.jp/sogoseisaku/transport/sosei_transport_tk_000034.html

その組合せも通常時と現在の混み具合を表すことができるため，そのような単語の組合せを 1 つの比較ワードとして利用する．

2.4 混雑表現辞書の作成

混雑ワード，駅名ワード，比較ワードから構成される混雑表現辞書を作成するために，本研究では，混雑関連ツイートに含まれる全単語を Word2Vec[2] を用いてベクトル表現する．Word2Vec は単語の語順や意味を考慮して，文書中の単語すべてをベクトル化するため，Twitter 特有の表記ゆれにも対応できる．ここでは，各ワードに対する初期ワードを与え，それらとコサイン類似度が高い単語を抽出することで，混雑表現辞書を作成する．

2.5 混雑検出において有用なツイートの定義

抽出した混雑関連ツイートのうち，混雑検出において有用であると考えられるツイートを次のように定義する．

(1) 「混雑ワード」と「駅名ワード」を含んだツイート（以下，駅名ツイート）
(2) 「混雑ワード」と「比較ワード」を含んだツイート（以下，比較ツイート）

駅名ツイートは，特定の駅の混雑状況を知る手がかりとなる．一方，比較ツイートは，日常的にその路線を利用しているユーザによる投稿であることが期待でき，混雑度合いを知る手がかりとなり得る．したがって，これらのツイートを投稿するユーザは混雑検出センサとして機能することが期待できる．

2.6 駅名ツイートと比較ツイートの抽出

ツイート中に駅名ワードと混雑ワードがこの順番で含まれていれば駅名ツイート，比較ワードと混雑ワードがこの順番で含まれていれば比較ツイートとして抽出する．ただし，複数の単語から構成される混雑・比較ワード中の各単語間，および駅名ツイートにおける駅名ワードと混雑ワード間，比較ツイートにおける比較ワードと混雑ワード間の位置的な距離が離れすぎている場合は，それぞれの単語が別の意味で使われている可能性が高いため，これらの距離が 2 以内のツイートのみを抽出対象とする．

3. 評価実験と考察

実際のツイートを用いて提案手法の有効性を検証した．評価データには，「電車遅延なう」において 2016 年 12 月 1 日から 31 日までに収集した遅延関連ツイート（70,013 件）を用いた．混雑表現辞書の作成には，「電車遅延なう」において 2016 年 8 月 16 日から 11 月 30 日までに収集した遅延関連ツイートを 2.2 節の分類器で抽出した混雑関連ツイート（48,086 件）を用いた．混雑表現辞書に含まれる混雑ワードは 200 個，駅名ワードは 984 個，比較ワードは 601 個となった．評価データから抽出した駅名ツイート（91 件）と比較ツイート（92 件）が本来の意味で抽出できているかを評価する．評価指標には適合率を用いて評価した．なお，70,013 件のツイートすべてに正解ラベルを付与するのは困難であるため，今回は再現率による評価はしていない．

表 1．提案手法の抽出精度

	駅名ツイート	比較ツイート
正解データ数	68	85
適合率	0.747	0.924

実験の結果を表 1 に示す．表 1 より駅名ツイートは 7 割以上，比較ツイートは 9 割以上の精度で抽出できていることがわかる．駅名ツイートについて誤抽出が起こった原因としては，「駅」の文字を省いた駅名が地名として抽出されたためである．駅名ツイートの抽出精度を上げるために，誤抽出されたツイートから地名と断定できる特徴を調べ，その場合において例外処理を設けるなどの方法が考えられる．

4. おわりに

本稿では，鉄道の混雑検出センサとして機能し得るユーザの特徴的なツイートとして駅名ツイートと比較ツイートを定義し，それらの抽出手法を提案した．実験結果より，駅名ツイートに関しては 0.747，比較ツイートに関しては 0.924 の適合率で混雑検出に有用なツイートが抽出できることを確認した．今後の課題としては，駅名ツイートや比較ツイートを投稿したユーザの過去の投稿にそれらのツイートがどれだけ存在するかの検証，および鉄道路線を表現する単語の辞書構築とそれによる一般ツイートからの混雑検出センサユーザの同定が挙げられる．

謝辞

本研究に際して，「電車遅延なう」の開発者の方々に深く感謝いたします．

参考文献

[1] NAVITIME: こみれぽ, http://products.navitime.co.jp/service/komikomi/
[2] Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean: Efficient Estimation of Word Representations in Vector Space, arXiv cs.CL, Vol.2, No.4630, pp.1-12 (2013).