

ソーシャルメディア上の混雑投稿と列車混雑率データとの 関係性に関する基礎検討

辰井 大祐[†] 國松 武俊[†] 武内 陽子[†] 横山 元紀[‡]
田中 峻一[†] 青柳 宗之[‡] 小西 勇介[‡] 坂入 整[‡]

[†]公益財団法人鉄道総合技術研究所 信号技術研究部

[‡]東日本旅客鉄道株式会社 JR 東日本研究開発センター フロンティアサービス研究所

1. はじめに

コロナ禍において、鉄道の混雑に関する関心が一段と高まっている。通常通りの列車運行状況であれば、旅客側である程度の混雑予測が可能であると考えられるが、運転見合わせが起こるような輸送障害発生時においては、旅客が必要とする情報を、適切なタイミングで提供することが輸送サービスの面で重要である。

そこで、混雑状況への高い関心の把握手法を検討するため、ソーシャルメディア（以下、SM）上の混雑に関する投稿（以下、定性データ）の活用可能性に着目した。本稿では、定性データが表現する混雑を混雑率に数値化し、鉄道事業者がもつ混雑率（以下、定量データ）と対応付ける分析を実施した。

2. 関連研究

SM 上の定性データを活用するためには、定性データが正しい情報であるという信頼性を確保する必要があり、信頼度の算出方法も提案されている^[1]。しかし、文献^[1]では、実際の定量データとの対応付けによる検証は実施されていない。また、SM 上の定性データから、混雑に関する表現の定量化する取り組み^[2]もあるが、対象が鉄道外であり列車内の混雑表現の数値化にはそのまま適用できないという課題がある。

3. 活用データ

定性データとして、ジョルダン株式会社開発、運営の運行状況投稿サイト「ジョルダンライブ!^[3]」の投稿、Twitter^[4]のつぶやきを用いた。ジョル

ダンライブ!（モバイル版）の投稿は（1）路線、（2）遅延状況（7段階）、（3）投稿時刻、（4）区間と方向、（5）混雑状況（5段階）、（6）コメントで構成されている。

また、定性データと対応付ける定量データとして、鉄道事業者がもつ実績ダイヤのデータと各列車の各駅間における混雑率データを用いた。

4. 定性データの列車・駅間の特定ルール

定性データと定量データを対応付けるためには、定性データが表現する列車、区間を特定する必要がある。

ジョルダンライブ!に関しては、遅延状況が「止まっている」の場合、投稿時刻以降、区間の始端駅を最初に発車した列車とする。それ以外の場合は、投稿時刻直前に区間の始端駅を発車した列車とする。

一方、Twitter に関しては、以下の手順で分析対象の投稿を抽出し、列車や駅間を特定する。

- ① 路線名をキーワードにして、輸送障害発生時の路線に関するつぶやきを抽出する。
- ② 混雑に無関係（宣伝等）のつぶやきを削除する。
- ③ 事前に用意した混雑表現リストにある表現が含まれているつぶやきを抽出する。
- ④ 駅名が含まれているつぶやきを抽出する。
- ⑤ 投稿の時刻、④で抽出した駅とともに、つぶやきの文の中に列車や方向の表現があれば、それを利用する。つぶやきの文から明確に判別できない場合には、実績ダイヤをもとに、投稿時刻直前に④で抽出した駅を発車した列車とする。

5. 定性データと定量データの対応付けと分析

5.1 定性データと定量データの対応付け

大都市圏内で発生した 12 件の輸送障害を対象に、ジョルダンライブ!と Twitter の定性データを抽出し、定量データと対応付けた。

ジョルダンライブ!について、5段階の混雑状況

Basic study on correlation between posts on social media and actual congestion data

Daisuke Tatsui[†], Taketoshi Kunimatsu[†], Yoko Takeuchi[†], Motoki Yokoyama[‡], Shunichi Tanaka[†], Muneyuki Aoyagi[‡], Yusuke Konishi[‡], Sei Sakairi[‡]

[†] Signalling and Operation Systems Technology Division, Railway Technical Research Institute

[‡] Frontier Service Development Laboratory, Research and Development Center of JR East group, East Japan Railway Company

の表現と、文献^[5]における混雑率の文章表現とを比較し、各混雑状況がどの区分の混雑率となるかを照らし合わせた。そして、各混雑状況の想定混雑率を、各区分の平均値の20%、70%、125%、175%、225%とし、4章の方法で抽出した165件の定性データと、定量データを対応付けた。その結果を、図1(左)に示す。定性データと定量データの相関係数は0.91であり、4章の対応付けルールによる列車や駅間の特定方法は、ある程度の信頼性を確保できたと考えられる。また、近似式の傾きが1より大きいことから、実際の混雑よりも混んでいるように感じる旅客が投稿している可能性が示唆された。

また、Twitterについては、4章③に含まれる混雑表現の混雑率を数値化するためのアンケートを実施し、回答者が考える混雑率の平均値を採用した。4章の方法で抽出した58件の定性データと定量データの対応付け結果を、図1(右)に示す。相関係数は0.93であり、近似式の傾きも1以上となり、ジョルダンライブ!と同様の傾向を確認することができた。

一方で、双方ともに、定性データと定量データとがずれている投稿も散見される。その理由は、以下のことが考えられる。

- 4章のルールで特定した列車や駅間が投稿対象の列車や駅間と異なる。
- 特定した列車と駅間は正しいものの、列車内で混雑率に偏りがあり、定量データの混雑率と投稿対象の混雑率が異なる。
- 特定した列車と駅間は正しいものの、混雑表現がどの混雑状況区分に該当するかの判断は個人差がある。

5.2 投稿数の分析

定性データと定量データを対応付けた結果を用いて、ジョルダンライブ!、Twitterについて混雑区分ごとの投稿数の分布を図2に示す。図2では、○の面積が投稿数を表す。

ジョルダンライブ!に関しては、混雑が高い状況だけではなく、混雑率が低い(100%未満)場合にも投稿数が多いことが確認できる。

一方、Twitterに関しては、空いている場合の投稿数が最も多かったものの、全体のデータ数が少ないため、特徴的な傾向を確認することができなかった。

6. まとめ

本稿では、SM上の混雑に関する定性データと、鉄道事業者がもつ定量データとの対応付けを行い、基礎的な分析を実施した。その結果、定量

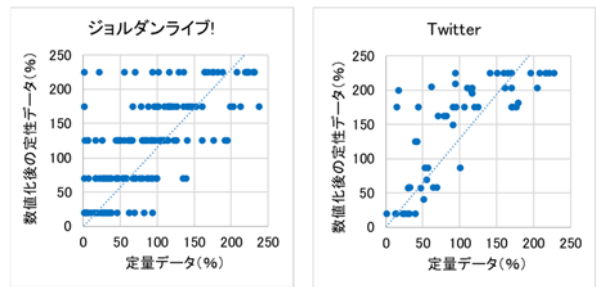


図1 定性データと定量データの対応付け結果

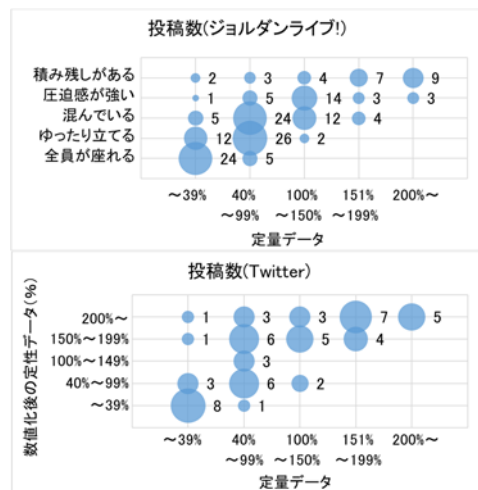


図2 投稿数の分布

データと定性データには高い相関があること、数値化後の定性データの方が定量データより混雑が高く出る傾向があることを確認した。また、混雑率が高い状況だけではなく、混雑率が低い状況にも投稿数が多いことを確認した。

今後は、定性データ分析の深度化やアンケート等を活用し、より正確な混雑状況への高い関心の把握手法について検討する。

参考文献

- [1] 長谷川陽平、足立進吾、藤原仁貴、米原三揮、”鉄道分野における SNS データ活用の利便性を向上する信頼度評価手法の提案 “、第20回情報科学技術フォーラム、2021年
- [2] 沖拓弥、”SNS への投稿情報に含まれる混雑度表現の抽出と定量化手法 東日本大震災発生前後の tweet を対象としたケーススタディ “、第30回人工知能学会全国大会、2016年
- [3] ”ジョルダンライブ!“、<https://www.jorudan.co.jp/unk/live.html>
- [4] ”Twitter“、<https://twitter.com/>
- [5] 国土交通省 鉄道局都市鉄道政策課、”鉄道利用者アンケート調査結果“、2021年1月