The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2018

乗換案内ログとSNSの融合による未来に発生する混雑原因の特定

Why does the station get crowded irregularly in the near future?

山下 達雄[*1] 坪内 孝太[*1] 丸山 三喜也[*1] 山浦 優樹[*1] 岡田 宏一朗[*1]
Tatsuo Yamashita Kota Tsubouchi Mikiya Maruyama Yuuki Yamaura Koichiro Okada

[*1] ヤフー株式会社
Yahoo Japan Corporation

In previous work of future congestion prediction using transit search logs, whether target station will get crowded on target time or not is predictable, but we cannot specify the cause of irregular congestion. In order to specify the irregular congestion cause, we analyzed posts containing information on future events from SNS. By combining congestion prediction information from the transit search log and event information from SNS, it was possible to extract the cause of congestion accurately.

1. 序論

世の中においての人の動きを予測することは重要である。特に、いつどこで混雑するのか、何が原因で混雑するのかを知ることができれば、様々な応用が考えられる。その応用の一つとして、乗換案内サービスがある。乗換案内サービスは、移動元と目的地の 2 つ駅と移動日時を入力として路線を検索し最適な経路を提示する。その際、該当経路で異常な混雑が予測される場合、注意喚起を行うことでユーザの利便性が向上する。

このような観点から我々は、路線検索ログから未来の異常混雑予測を行う手法を提案した[坪内 17]。これにより、「いつどこで(どの駅で)」予期せぬ混雑が発生しそうかの予測できるようになった。しかし、この方法では「何が原因で」混雑しそうなのかは分からない。実用上の観点から、ユーザに原因も提示することで安心感を与えたい。そこで、本研究では、この未来に発生すると予測される異常混雑の原因を特定する手法の提案を行う。

混雑原因を特定するには、イベント開催予定情報やイベントについて述べられている SNS 投稿などの言語リソースが必要となる。我々は、SNS である Twitter に着目し、未来のイベントについての内容が含まれる投稿(ツイート)を収集・整理して、注目度の高いイベントを抽出した。そして、これを既存の路線検索ログからの異常混雑予測と融合することにより、未来の日時に混雑が予想される駅に対してその混雑原因を提示することが可能となった。

本稿では、まず、融合要素の一つである、路線検索ログから未来の異常混雑予測を行う手法を説明する。次に、もう一つの融合要素である、Twitter から混雑場所と混雑原因を取り出す手法を説明する。そして、この二つの要素を結びつけることによる、混雑原因イベントを抽出する方法について述べ、その有効性を確認するための実験と評価について述べる。

2. 路線検索ログによる混雑駅の予測

本研究は、乗換案内の検索ログから異常混雑予測を行う坪内らの手法[坪内 17]をベースとしている。

坪内らの研究では、まず過去の路線検索データを集め、時間ごとの検索量の通常パタンを導出した。たとえば「A 駅から B 駅まで、火曜日の 10 時 00 分から 10 時 10 分の間はいつもであれば 80 回検索される」というパタンである。このパタンの導出にはバイリニアポアソン回帰モデルを用いた。

連絡先:山下達雄, ヤフー株式会社,
tayamash@yahoo-corp.jp

次に、近い未来の検索量を自己回帰モデルで推測した。つまり、「このままの様子で検索量が増えた場合に、最終的に何件になるか」という推定である。ユーザは移動の前に予めスケジュールを検索することが多い。その検索量を集め、数日後の検索量を予測した。

最後に導出された「いつものパタン」から推定される検索量と、「このまま増えたらどうなるか」という観点から推定された検索量を比較し、両者に乖離がある場合は異常混雑が発生する可能性が高いとした。

このように検索ログから未来に発生する異常混雑を予知することができる。しかし、この手法では混雑する駅と日時は予測できるが、混雑原因は分からないという問題がある。

3. Twitter による混雑スポットの予測

本節では、SNS (Twitter)からのイベント情報抽出について述べる。これは前節の手法では不可能である「未来の異常混雑の原因」を取得するための前段処理である。

未来のイベントについて言及しているツイートには、イベントの開催日(例:1 月 15 日)や開催場所(例:東京ドーム)が含まれているものがある。そのイベントに言及しているツイートすべてにこれらの情報が含まれているわけではないが、ツイート数の多い人気のイベントではそれらを含むものが一定数存在する。

異常混雑予測のためには人気のイベントを補足できればよいので、我々は開催日や開催場所に着目して情報抽出を行う。

まず、第 1 段階の処理として、未来の日付を含むツイートを収集する。例えば、今日が 7 月 24 日だとして、未来の日付である「7 月 25 日」について言及しているツイートを収集するには、本文中に「7 月 25 日」または「7/25」という文字列が含まれているもののみを選別する。これにより、日付をラベルとしたツイート集合ができる。

次に、第 2 段階の処理として、各日付を含むツイート集合からランドマーク(イベントの開催場所となりうるもの)を含むツイートを選別する。ツイートからのランドマークの抽出には、エンティティリンキングツールを用いる。具体的には、ランドマーク辞書のエントリ文字列にマッチさせ、曖昧性解消処理を経て、最終的に抽出結果を得る。ランドマーク辞書は、ランドマーク名と最寄り駅(複数)情報を持つ。本研究は、乗換案内の利便性向上を目的としているため、どの駅からも遠いランドマークは使用しない。

これら 2 段階の処理により、特定日付と特定ランドマークを同時に含むツイートの集合ができる。ツイート集合中のツイート数

が多ければ、その日時にそのランドマークで開催されると思われるイベントの人気が高い、つまり、当日そのランドマークが混雑すると考えられる。

しかし、当然、ノイズが多く含まれる。特定日付と特定ランドマークが含まれたツイートが多いという事実だけでは、イベントへの言及であるとは限らない。また、イベントへの言及であったとしても、必ずしも参加者が多いことを意味するわけではない。

4. 路線検索ログによる予測と Twitter による予測の融合

前節で説明した、ツイッターからの未来イベント抽出は、そのままで利用するにはノイズが多いという問題がある。しかし、本来の用途である、路線検索ログからの異常混雑予測に対する原因抽出という観点であれば、つまり、路線検索ログによる予測と融合すれば有効に利用できる。

前述の通り、ランドマーク辞書の各エントリには最寄り駅情報が含まれている。これを用いて、ツイッターによる未来日付イベント予測と路線検索ログによる未来日付駅異常混雑予測とを結びつける。未来日付イベント予測はツイート数を閾値として、ある一定の値を超えるもののみ利用する。

これらの処理によって、未来の日時に異常混雑が予想される駅に対してその混雑原因を予測することが可能となる。

4.1 評価

まず、Twitter からの未来日付イベント予測結果について、混雑の有無の判定を人手で行い、精度を出した(twitter only)。次に、駅名と日付を介して検索ログによる未来日付駅異常混雑予測と結びつくものに絞って、精度を出した(proposed)。また、未来日付イベント予測のツイート数を閾値として、精度の変化を見た(Threshold)。

データの期間は 2016 年 12 年 29 日～2017 年 1 月 18 日。路線検索ログによる混雑予測 209 件のうち、ツイッターによるイベント予測(ツイート数 5 以上)に紐づくものは 129 件であった。

評価は 3 名の判定者により行った。判定基準は、未来日付イベント予測のツイート集合から混雑要因となるイベント情報が得られるか、である。ただし、イベントが実際に混雑したかについては、実地での観測データ(正解データ)がないため、判定者各自がインターネット上の情報で確認を行った。最終的な判定は、3 名の判定の多数決とした。

結果を図 1 に示す。未来日付イベント予測の際のツイート数が多いほど、混雑イベント抽出能力が高い傾向にあるが、検索ログによる異常混雑予測と結びつけることにより精度が上がるこ

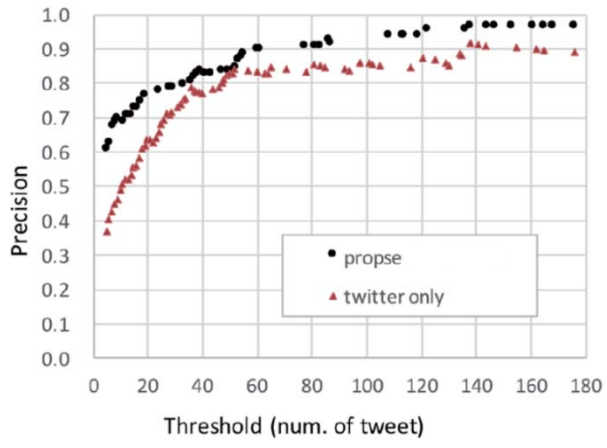

図 1

図 2

https://blog-transit.yahoo.co.jp/congestion/

とが分かる。つまり、融合により、高い Precision で混雑要因が抽出されていると言える。

4.2 サービスへの適用

融合により十分な精度が得られることが分かったが、実際の応用(サービス適用)に際しては、混雑要因をユーザに提示する必要がある。つまり、ツイート集合からイベント名やアーティスト名などの情報を抽出する必要がある。

我々はツイート集合に多く含まれるハッシュタグに着目した。Twitter による未来日付イベント予測のツイート集合(検索ログによる予測と紐づいたもの)17 件から取り出したハッシュタグの頻度上位 3 件を、判定者 1 名がイベント内容を表しているか否かを評価した結果、精度は 0.94 (16/17)であった。人手による NG ワードフィルタ等を併用すれば実用上十分な精度である。

以上の結果を踏まえ、乗換検索アプリに新機能として実装し、2018 年 2 月にリリースした(図 2)。

5. 関連研究

Twitter のテキスト(ツイート)から直接イベント情報を抽出する研究として[山田ら 16]がある。ランドマークで検索して集めたツイートを用い、パターンマッチングと機械学習で、イベント日付とイベント名を抽出している。これは近い未来に発生するイベントの予知ではあるが、混雑状況は正確に推測できない。

6. 結論

本研究では、路線検索ログからの異常混雑予測だけでは分からなかった混雑原因を、SNS (Twitter) データを解析して得たイベント情報と組み合わせることにより、抽出可能とした。その成果を踏まえて実際のサービスとしてリリースした。今後は、主にイベント情報の抽出精度向上を行っていく。

参考文献

[坪内ら 17] 坪内 孝太, 下坂 正倫, 小西 達也, 丸山 三喜也, 山下 達雄. 乗換案内データを用いた未来の混雑予測の研究, 2017 年度 人工知能学会全国大会講演集, 4I1-4in1, 2017.

[山田ら 16] 山田悠, 落合桂一, 菊地悠. Twitter を用いた地域イベント発見技術, NTT ドコモ R&D テクニカル・ジャーナル, Vol.23, No.4, 2016.