路線トポロジを考慮した ジオタグツイート解析に基づくトピック伝搬分析

丸山　直樹† 安井　豪基† 王　元元†† 河合由起子† 秋山　豊和†

† 京都産業大学　〒 603-8555 京都府京都市北区上賀茂本山
†† 山口大学　〒 755-8611 山口県宇部市常盤台 2-16-1
E-mail: †{i1458076,akiyama}@cse.kyoto-su.ac.jp, ††gxyasui@gmail.com, †††y.wang@yamaguchi-u.ac.jp, ††††kawai@cc.kyoto-su.ac.jp

あらまし　マイクロブログから実世界のイベントを検出する研究が活発に行われており，イベントとして公共機関や道路における混雑状況や遅延検知に関する研究も注目されている．本研究は，電車の駅における実空間と仮想空間のトポロジを考慮した各駅間のトピックの伝搬の状況を機械学習解析し，事故等による遅延の予測を目指す．具体的には，電車の路線を実空間のトポロジとして利用し，仮想空間のトポロジとして，各駅にいるユーザから発信されたツイートをニューラルネットワークを用いて分析する．これにより，ある駅で遅延などのトピックが発生した際の他の駅への影響を観測することができ，混雑や遅延を避けたアクセスの推薦につながると考えられる．本論文では，路線トポロジを考慮したジオタグツイート解析に基づいた事故遅延予測手法を提案し，実際の遅延情報と比較検証する．

キーワード　ツイート分析，トポロジ形成，トピック伝搬

1. はじめに

近年，Twitter や Foursquare などのマイクロブログから実世界における特定の場所でトピックを検出する研究が活発に行われている．例えば，位置情報が付与されたソーシャルメディアへの投稿から催し物などのトピックを分析する研究が行われている [1]．また，通勤・通学などの交通手段として電車が多く用いられているが，Twitter へのツイート投稿数の推移から，列車の遅延などを検知する研究も行われている [2]．しかし，これらの研究では，トピックが発生している駅の分析のみが行われ，トピック（事故）が発生した際の他の駅への影響については十分に考慮されておらず，そのため遅延の影響に関する予測にまでは至っていない．

我々はこれまで，高さ情報のないジオタグツイートに対して複数の機械学習に基づき各フロアごとに分類することで，時間とフロアごとに関連あるトピックの抽出を行ってきた [3]．　例えば、東京スカイツリー周辺で発信されているジオタグツイートには異なるフロアの発言が含まれているが，緯度経度のみで高さ情報が含まれていない．先行研究より，各フロアごとにツイートを取得でき，フロアごとの話題を提供できた．また，テーマパーク等の複数のエリアの分類に関する評価も行い，時空間におけるトピック分類の有効性を示せた．しかしながら，各フロア間の関係性まで考慮しておらず，各フロア間のトピックや時間帯が及ぼす影響は発見できなかった．

そこで，本研究では，実空間における各エリア間の関係性を考慮することで，時空間でのトピック伝搬によるイベント発見手法を提案する．本論文では，特にフィジカル空間におけるエリア間の関係性が理解容易な路線上の各駅をエリアとし，イベントを事故等における遅延状況とし，トピックの伝搬を解析することで事故における各駅の遅延の影響発見手法を提案し，検証する．提案手法により，ツイートの少ない駅においても遅延状況を発見でき，さらに間接的なツイート表現でも遅延状況を発見できる．これは，電車遅延のように周辺への影響が大きいイベントの発生時は，直接的な表現でなくても，イベントに関連するツイートが発生している可能性があることから，影響力

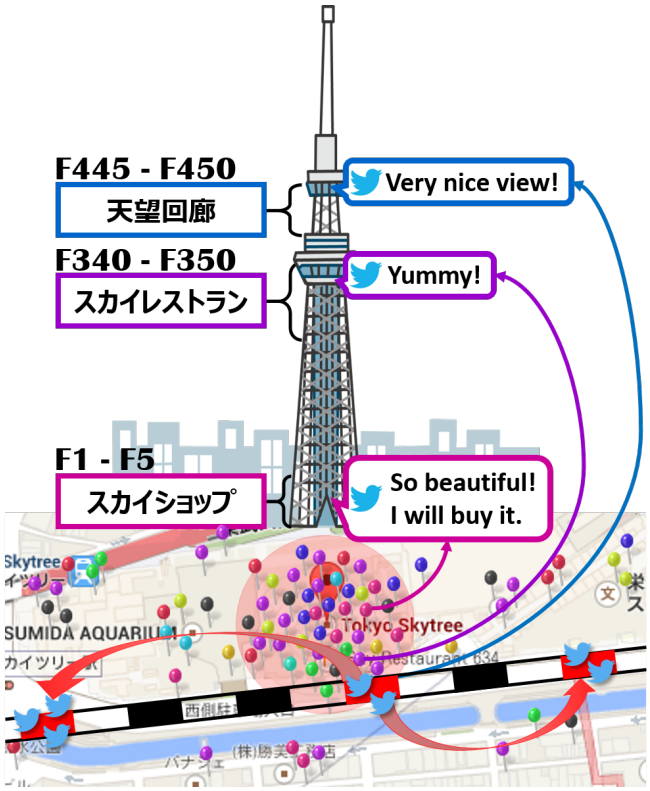

図 1　フロアおよび駅のトピック抽出とトピック伝搬

表 1　システム出力例（入力：上野駅，山手線）

ランク	駅	路線	予想影響度
1	池袋駅	山手線	大（92%）
2	東京駅	山手線	大（90%）
3	銀座駅	山手線	大（90%）
4	青山一丁目駅	東急田園都市線	中（81%）
5	中目黒駅	東急東横線	小（67%）
6	銀座駅	東京メトロ銀座線	小（63%）

が大きなイベントに対してイベント発生時間帯（今回は遅延時間帯）という属性のみでラベル付けすることにより，どの程度該当イベントをツイートから検知できるかを明らかにすることを目指している．

本論文では，具体的に以下の点について述べる．

- 各駅で発信されたジオタグツイートの解析
- 実空間の路線トポロジに基づく遅延影響（トピック）の駅の発見
- 遅延情報と提案手法との比較検証

提案手法では，まず駅近辺のツイートデータを収集し，辞書を生成する．辞書はツイート集合を形態素解析し，単語ごとに単語 ID を付与したものであり，学習に用いられるデータはその ID をもとに BOW（Bag of Words）を用いることで，単語の出現回数のベクトルとして表される．生成されたベクトルをニューラルネットワークを用いて，トピックの影響（事故の遅延）の駅を学習し，発見する．本論文では，関東 488 の駅を対象に提案手法による事故遅延の発見に関する有効性検証を行う．

本論文の構成は以下のとおりである．次章で提案システムの概要について述べ，3 章でツイートデータの分析手法を説明する．4 章でツイートに基づいて形成される仮想空間トポロジについて述べる．5 章で評価実験の結果を示し，6 章に関連研究を示す．最後に 7 章で結論と今後の課題について述べる．

2.　システム概要

本研究では，電車での事故や遅延などが発生した場合，各駅で発信されたジオタグツイートを解析し，実世界トポロジと仮想空間トポロジの双方を考慮したトピック伝搬を分析することを目的としている．まず，電車の各駅周辺で発信されたジオタグツイートを収集し，それらに含まれる単語集合を形態素解析によって分かち書きにする．そして，ニューラルネットワークに学習させ，過去の運行情報から事故や遅延などが発生した時刻と駅を抽出し，その時刻の周辺の駅でのツイートを分析することにより，トピックによる影響の伝搬状況を観測する．具体的には，発信されたツイートを駅ごとに集計し，ネットワークは各駅をラベルとして *BOW* (Bag Of Words) を用いてベクトル化した単語を学習する．ベクトル化には以下の手法を用いてベクトルを最適化している．

- *TF-IDF* による重み付け
- *LSI* (Latent Semantic Indexing) による次元削減

システムの概要を図 1，出力例を表 1 に示す．上野駅（山手線）で遅延が発生したと仮定する．システムでは、随時全ての

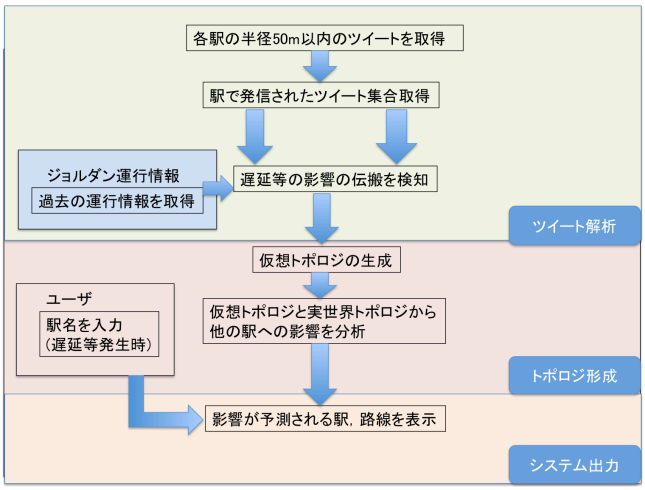

図 2　システム構成と処理の流れ

駅ごとのツイート集合を入力としおり，学習器より各駅ごとに遅延発生の推定を随時行い，閾値以上であれば遅延と判定する．この判定結果より，ユーザは遅延の影響が少ない駅や路線を発見でき，事故や遅延などのトピックによる混雑状況の予測につなげることが可能である．システム構成と処理の流れを図 2 に示し，次節以降で順に説明を述べていく．

3.　ツイート解析

本研究では，各駅で発信されたジオタグツイート集合をベクトル化することにより，駅同士の仮想的な結びつきを算出することで，仮想空間トポロジを形成する．はじめに，ニューラルネットワークに学習させるツイートに関する辞書を作成する．今回作成する辞書には，東京周辺の全ての駅で発信された 2 年分のツイートを用いた．そして，ツイート集合を形態素解析し，出現する全ての単語の中から，全ての文書に出てくるような頻出語や，一つの文書にしか出現しないような単語を除外する．ここで，作成された単語辞書にもとづいて，学習データを単語の出現頻度でベクトル化し，*TF-IDF* を用いて各単語のベクトルに重みを付ける．そして，*LSI* による次元削減を行うことにより，駅同士のトポロジを算出する上で重要なベクトルのみを抽出することが可能となる．次節以降ではツイートの取得範囲について具体的に説明する．

3.1　位置情報に基づくツイート取得

まず，指定地域から重複を除いた緯度経度情報を含むストリーミングツイートを Twitter Developers の The Streaming APIs(注1) を用いて取得する．指定地域は，1 度以上異なる南西および北東を指定することで，その 2 点に囲まれた矩形領域のストリーミングツイートを取得できる．このとき，各駅の半径 50m の範囲で取得している．

3.2　*TF-IDF* による重み付け

3.1 節で生成された辞書にもとづき．学習データは *BOW* を用いてベクトル化され，下記の式 *TF-IDF* により重み付けを行う．

(注1)：https://dev.twitter.com/streaming/overview

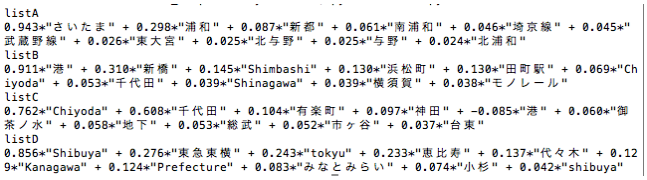

図 3　次元削減例

$$TF = \frac{\text{単語 } i \text{ の出現回数}}{\text{すべての単語の出現回数}}$$

$$DF = \frac{\text{単語 } i \text{ が出現した文書数}}{\text{総文書数}}$$

これにより，コサイン類似度を用いて駅同士の類似度を算出する上で重要な単語を抽出することができる．この時点でベクトルの次元数は単語の総数となっており，次元数が膨大になっているため，次元削減を行う．次項にて *LSI* による次元削減について述べる．

3.3　*LSI* による次元削減

次元数の削減は，ニューラルネットワークにおける過学習の緩和，学習コストの削減を目的としている．*LSI* による次元削減は，単語に含まれる潜在的な意味によりインデキシングを行うことにより，類義語や，同義語を一つのベクトルに圧縮することが可能となる．本論文では，*LSI* を用いて，*TF-IDF* により重み付けが付与されたベクトルの次元数を，入力の次元数として 300 次元になるように削減した．一つのベクトルとして圧縮された単語の集合の例を図 3 に示す.「Chiyoda」と「千代田」のように同じ意味をもつ単語や,「浦和」と「南浦和」や「北浦和」のような語句が，一つのベクトルとして圧縮されていることがわかる．

4.　仮想空間トポロジの形成

3 章で得られた各駅の特徴ベクトル集合から，仮想空間トポロジの形成を行う．事故や遅延などの発生した路線上の各駅への影響を抽出し，それ以外の駅との関連性も同時に算出することにより，事故や遅延など発生時の各駅への影響を検知することが可能であると考えられる．具体的には，過去の運行情報から，事故の発生した路線の各駅とそうでない路線の各駅で発信されたツイート集合を特徴ベクトル集合に変換し，それぞれに事故影響有り，事故影響無し，の 2 値のラベルを付与し，ニューラルネットワークにて学習させる．それにより得られた結果から，事故トピックの伝搬状況を分析し，仮想トポロジを生成する．本論文では，妥当性の検証のため，過去の遅延の生じた日と，遅延の生じていない日，各数日分のツイート集合を用いて実験を行った．

5.　実　　験

本章では，遅延情報と提案手法との比較検証として，事前にジョルダン (注1) より取得した運行情報に基づき，実際に事故があった時間帯の各駅における遅延判定結果について検証する．次章にて実験について述べる．本実験はイベントの影響度と機械学習に基づく判定の関係について調査することを目的としている．

(注1)：http://www.jorudan.co.jp/unk/

表 2　学習用データ 1

	日付	事故発生時間帯	事故発生路線
1)	2016 年 10 月 8 日	9 時〜11 時	山手線
2)	2016 年 10 月 30 日	13 時〜16 時	山手線
3)	2016 年 11 月 6 日	6 時〜8 時	山手線

表 3　学習用データ 2

	日付	事故発生時間帯	事故発生路線
1)	2016 年 8 月 7 日	9 時〜11 時	湘南新宿線
2)	2016 年 8 月 9 日	7 時〜9 時	湘南新宿線
3)	2016 年 10 月 17 日	18 時〜20 時	南北線
4)	2016 年 11 月 22 日	6 時〜8 時	南北線

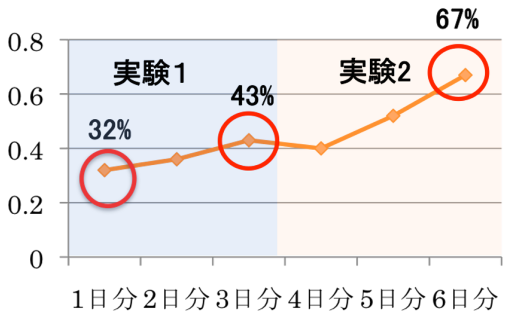

図 4　実験 1 と実験 2 の正解率比較（山手線の場合）

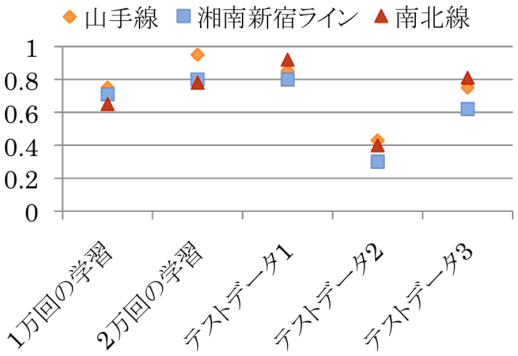

図 5　複数の路線の正解率比較

5.1　実 験 環 境

本実験では，取得したツイートのうち，2014 年 1 月 1 日〜2016 年 1 月 31 日の 2 年間にて東京駅周辺の各駅の半径 50m 以内で発信されたツイートから辞書を作成した．学習用データは，事故遅延のトピックが発生した日付けの発生時刻の時間帯の 1 時間ごとのツイート集合を用いた．

実験では、路線間の遅延発生状況抽出の検証を目的とし、以下の 4 つの実験を行った．

- 任意の路線に対する学習器の精度検証
- ツイートデータサイズに対する精度向上の検証
- 複数路線での検証
- 運行情報が提示される直前のツイートデータの検証

はじめに，任意の路線に対する学習器の精度検証として山手線で発生した事故に対する影響の検証を行った．学習データは，

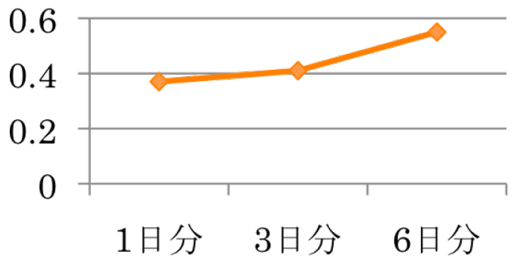

図 6　運行情報提示 1 時間前の正解率

4 つの実験全てにおいて，関東の東京周辺 488 駅近辺で発信されたツイート集合を用いた．そしてこの実験では，表 2 において山手線上の駅で発信されたツイートに，影響ありのラベルを付与し 3 つの検証を行った．

（1）　学習データに含まれる事故発生時のツイート
（2）　学習データに含まれない事故発生時のツイート
（3）　学習データに含まれない事故非発生時のツイート

の精度を検証した．

結果は，(1) の正解率は 85%，(3) の正解率は 75% と高い精度となった．しかしながら，(2) の学習データに含まれない事故発生時のツイートは正解率 43% と低い結果となった．そこで，学習データを増加させ，それに伴う正解率の精度向上に関する検証を行った．山手線にて，学習データに用いる遅延情報のデータサイズを増加させた際の，正解率の結果を図 4 に示す．3 日分では 43% であった．正解率は 5 日分，6 日分とデータサイズが増加すると向上し，6 日分のデータを使うと 67% となった．以上のことより，学習データを増やす事で学習データに含まれないツイートからも高い正解率が見込まれることが確認できた．

次に，他の路線（湘南新宿線，南北線）でも同様に検証を行った．表 3 のデータを影響ありとして，学習させた．それぞれの結果を図 5 に示す．ツイートが比較的少ない路線でも，山手線と同等程度の正解率が得られることを確認した．

最後に，運行情報提示前のデータとして，実験 2 と同様の学習データを用いて，山手線における事故遅延情報が提示された各 1 時間前のデータをテストデータとし，精度検証を行った．結果を図 6 に示す．1 日分の学習データでは 37%，3 日分の学習データでは 41%，6 日分では 55% が正解率となった．

6. 関 連 研 究

Twitter などのマイクロブログから実世界における特定の場所でのトピックを検出する研究が活発に行われている．例えば，位置情報が付与されたソーシャルメディアへの投稿から催し物などのトピックを分析する研究が行われている [1] [10] [11]．また，Twitter へのツイート投稿数の推移から，列車の遅延等を検知する研究も行われている [2] [4]，しかし，これらの研究では，事故や遅延などが発生している場所（駅など）での分析のみが行われ，事故発生による周辺の場所への影響については考慮されていない．他にも，駅構内の無線センサを用いて，駅内の混雑度や，人の流動を誘導するような研究が行われている [5]．この研究が駅内の人の流れに着目しているのに対し，我々は駅同士の遅延などのトピックの流れに着目している点が異なる．

また，ソーシャルメディアへ投稿された，ジオタグが付与された写真を解析することで，実世界での人流を解析するような研究がある [6]．この研究では，観光地等で人々がどれくらいの時間滞在するのかを予測しているが，それによる混雑度の推移や，交通機関への影響は考慮していない．ツイートから注目される話題を抽出する研究 [7] では，現在起こっている事故や遅延などに関するトピックを抽出することを目的としているが，我々はそのトピックの伝搬や，それによる影響に着目している点で異なる．

遅延の連鎖に着目した研究も存在する [8]．この研究では，同じ路線での遅延の影響のみを考慮しているため，我々の研究とは分析の範囲が異なり，ソーシャルメディアへ着目していない点も異なる．他にも，バスの到着時刻を予測するような研究もある [9]．この研究では，乗り降りする人の数や過去の運行の実績を基にしているが，電車とは他の路線から受ける影響が大きく違うため，我々の研究とは異なる．

7. ま　と　め

本論文では，実空間のトポロジと仮想空間のトポロジ双方に基づいた実空間でのトピックの伝搬状況を分析することを目指し，ジオタグツイートの解析をおこない，仮想空間のトポロジを形成するための礎を築いた．また，実空間のトピックの伝搬をツイート分析から読み取ることの可能性を示した．今後は，実世界のトポロジとトピックの内容による伝搬の速度について考察し，時間単位，分単位でトピックの伝搬を分析することが必要である．また，過去の運行情報用いて，路線を隔てたトピックの伝搬についてもっと考慮が必要である．そして，トピック伝搬による各駅の混雑や遅延の影響を推測し，影響の少ないルート推薦へと繋げていくことが課題となる．

謝　　辞

本研究の一部は，JSPS 科研費 15K00162，16H01722 の助成を受けたものである．ここに記して謝意を表す．

文　　献

[1] 伊藤 貴明，遠藤 雅樹，加藤 大受，江原 遥，廣田 雅春，横山 昌平，石川 博，Twitter を用いた駅イベント検出．第 8 回データ工学と情報マネジメントに関するフォーラム（DEIM Forum 2016），B2-3，2016.
[2] 簗瀬 拓弥，増田 英孝，山田 剛一，荒牧 英治，中川 裕志，Twitter を用いた電車遅延の自動通知．IPSJ SIG Technical Report，Vol. 2013-IFAT-110，No. 1，2013.
[3] 安井 豪基, 坪井 結香, 岡山 愛, 河合由起子, 王 元元 , 秋山 豊和 , 複合施設におけるツイート分析に基づくタグクラウド生成および可視化，第 8 回データ工学と情報マネジメントに関するフォーラム（DEIM Forum 2016），F1-6，2016.
[4] 新井 誠也，平川 豊，大関 和夫，Twitter からの列車遅延情報収集手法の検討．情報処理学会第 75 回全国大会，5N-2，2013.
[5] 荻原 崇，白 迎玖，垣 良宏，清木 康，無線センサを用いた駅構内における大規模人流誘導を対象とする流動意味解析システム，第 8 回データ工学と情報マネジメントに関するフォーラム（DEIM Forum 2016），F6-2，2016.

[6] 遠藤 雅樹，山中 光定，廣田 雅春，大野 成義，石川 博，位置情報付きツイートを利用した地域ごとのトレンド分析手法の検討．観光情報学会第 12 回全国大会，2015.
[7] 木原 大志，白木 原渉，越村 三幸，藤田 博，長谷川 隆三，Twitter の時系列解析による注目話題の抽出．情報処理学会第 74 回全国大会講演論文集，pp. 625-627，2012.
[8] 岩倉 成志，高橋 郁人，森地 茂，都市鉄道の遅延連鎖予測のためのエージェントシミュレーション，学術研究論文，Vol.15，No.4，2013.
[9] 前川 裕一，中島 秀之，白石 陽，乗降者数データと運行実績データを用いたバス到着時刻予測，情報処理学会第 76 回全国大会，2V-3，2014.
[10] 渡辺 大貴，相場 亮，Twitter を用いた開催中のソーシャルイベントの状況把握に関する研究，情報処理学会第 77 回全国大会，2M-05，2015.
[11] 山本 修平，佐藤 哲司，Twitter からの実生活情報の抽出法の提案，第 4 回データ工学と情報マネジメントに関するフォーラム（DEIM Forum 2012），F3-4，2012.