社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

予測型トラヒックエンジニアリングのための ソーシャルメディアデータからトラヒック変動に関する情報の抽出

河島　滉太†　　大下　裕一†　　村田　正幸†

† 大阪大学 大学院情報科学研究科
〒 565–0871 大阪府吹田市山田丘 1–5
E-mail: †{k-kawashima,y-ohsita,murata}@ist.osaka-u.ac.jp

あらまし　モバイル端末の普及に伴い、通信ネットワークに流れるトラヒック量は、量・変動ともに増大している。これに対し、変動するトラヒックを収容する技術として予測型トラヒックエンジニアリング (TE) に関する検討が進められている。予測型 TE では、予測されたトラヒック量に基づいて、ネットワーク資源を動的に割り当てる。そのため、トラヒックの予測精度が、予測型 TE の制御性能に大きな影響を与える。従来、過去のトラヒック変動の時系列からトラヒックを予測する手法が検討されてきた。しかしながら、そのような手法では、過去のトラヒック変動には予兆が含まれていないものを予測することができない。これに対し、そのような予兆は、現実世界を反映したソーシャルメディアデータに含まれている可能性がある。本稿では、ソーシャルメディアデータにそのような予兆が含まれているかについて確認する。確認するにあたり、ソーシャルメディアデータから情報を抽出し、抽出した情報に基づいて非日常なトラヒック変動が発生している時間帯を予知する手法を提案する。総トラヒック量に基づいて非日常的なトラヒック変動を予知する手法と比較し、評価を行うことで、現実世界に起因する非日常的なトラヒック変動の予兆がソーシャルメディアデータから抽出できることを示す。

キーワード　トラヒックエンジニアリング、トラヒック予測、非日常的トラヒック変動予知、ソーシャルメディアデータ、Twitter

Extracting Information on Traffic Changes from Social Media Data for Predictive Traffic Engineering

Kota KAWASHIMA†, Yuichi OHSITA†, and Masayuki MURATA†

† Graduate School of Information Science and Technology, Osaka University
Yamadaoka 1–5, Suita-shi, Osaka, 565–0871 Japan
E-mail: †{k-kawashima,y-ohsita,murata}@ist.osaka-u.ac.jp

Abstract　The amount of traffic through networks is increasing both in quantity and in fluctuation as the mobile terminals become popular. Predictive Traffic Engineering (TE) is one approach to accommodating the fluctuating traffic. Predictive TE allocates the resources in advance before the traffic changes by using the predicted traffic. For predictive TE, the accurate prediction of the future traffic is important. Thus, many methods to predict the future traffic from the traffic rates in the previous time slots have been proposed. However, these method cannot predict the traffic changes whose signs are not included in the previously observed traffic. On the other hand, the signs may be included in social media data reflecting the real world. In this paper, we investigate the signs of the traffic changes caused by the events in the real world included in the social media data. To investigate them, we propose a method to extract information related to the real-world events, and a method to forecast the unusual traffic changes based on the extracted information. We evaluate our forecasting method compared with a method to forecast based on the total traffic rate. Based on the results, we discuss the signs of the traffic changes caused by the events in the real world. The results indicate that the signs of the unusual traffic changes are included in social media data.

Key words　Traffic Engineering, Traffic Prediction, Forecasting Unusual Traffic Changes, Social Media Data, Twitter

1. はじめに

スマートフォンやタブレットといった高機能なモバイル端末の普及に伴い、ネットワークを流れるトラヒック量は、量・変動ともに大きくなっている。ネットワークを提供する通信事業者は、このような大きなトラヒック変動にも対応し、通信品質を維持することが求められる。従来、このような変動に対応した通信品質の維持は、ネットワーク内の各地点に十分な資源を用意することにより行われてきた。しかしながら、この方法では、実際のトラヒックを収容するのに必要な資源と比べて、著しく大きな資源を準備する必要があり、多大なコストがかかる。

上記の問題に対応するために、ネットワークを動的に制御する手法の検討が進められている [1], [2]。これらの手法では、ネットワークの各地点におけるトラヒック量の変化に応じて、ネットワーク上のリソースを調整することにより、資源が必要な箇所に十分な資源を割り当て、サービス品質の低下を防ぐ。

上記のようなネットワーク制御により、トラヒック変動による通信品質の劣化を防ぐためには、通信品質が劣化する前に、通信品質を劣化させるような変動の発生を予測し、制御を行う必要がある。我々の研究グループでは、トラヒック変動を予測して、前もって制御を行う予測型のネットワーク制御技術の研究を進めてきた [3]。トラヒック変動に関する予測値に基づいて、仮想化されたネットワーク機能の位置を徐々に変更することにより、急激な設定変更を行わなくてもトラヒック変動に追随可能であることを示している。

ネットワークに将来流れるトラヒックの予測方法に関しては、様々な検討が進められてきた [4]〜[6]。これらのトラヒック予測手法では、過去に観測されたトラヒックの時系列データをもとに、トラヒック変動モデルのパラメータを推定し、そのトラヒック変動モデルを用いて将来のトラヒック変動を予測する。しかしながら、これらの予測手法では、現実世界で発生したイベントに伴い、ネットワークを利用するユーザが局所的に集中するといった、過去のトラヒック変動には予兆が含まれないようなトラヒック増加を予測することはできない。このような現実世界で発生したイベントに起因する、非日常的なトラヒック変動は、高機能なモバイル端末の普及に伴い大きくなっている。

これに対し、非日常的なトラヒック変動の予兆は、現実世界を反映したソーシャルメディアデータに含まれている可能性がある。高機能モバイル端末の普及に伴い、ソーシャルメディアは著しく成長を続けており、人々の行動に関する多様なデータを取得することが可能である。現在、情報科学の分野では、ソーシャルメディアから得られるデータの活用に関して様々な検討が進められている [7], [8]。このことから、トラヒック予測の分野への応用も期待することが出来る。ソーシャルメディアから得られる情報をもとに、イベントを検出し、当該イベントに起因するトラヒック量を予測することによって、従来のトラヒック予測手法に比べて予測精度が向上し、予測型の動的ネットワーク制御の性能を向上させることが可能となる。

そこで、本稿では、ソーシャルメディアデータを解析することで、非日常的なトラヒック変動の発生を予知出来ることを確認する。確認するにあたり、現実世界で発生したイベントによって、非日常的なトラヒック変動が発生している時間帯を予知する手法を提案する。提案手法では、Twitter のツイート情報を解析し、予知に用いる。ユーザがツイートする内容は、現実世界で発生するイベントを反映したものである。提案手法では、ツイートを解析することにより、現実世界で発生しているイベントを検知し、当該イベントに関連するツイート数の伸びにより、当該イベントの規模を予測し、予測されたイベントの規模に基づき、非日常的なトラヒック変動の発生を予知する。

以降の本稿の構成は次の通りである。2. では、非日常的トラヒック変動の予兆となる情報をソーシャルメディアデータから抽出する方法と、抽出された情報を用いた非日常的トラヒック変動予知手法について述べる。3. では、提案手法を評価し、評価結果にもとづいて、非日常トラヒック変動の予兆について考察を行う。4. で、まとめと今後の課題について述べる。

2. ソーシャルメディアデータからの情報抽出

本章では、現実世界で発生したイベントに起因するトラヒック変動の予兆を含む情報をソーシャルメディアから抽出する手法について述べる。抽出された情報にそのような予兆が含まれている場合、非日常的トラヒック変動の予知精度を向上させることが可能となる。そこで、本稿では、抽出された情報を用いて非日常的トラヒック変動を予知する手法についても検討する。

2.1 ソーシャルメディアデータの概要

本稿では、ソーシャルメディアデータとして、Twitter Streaming API [9] から得られたツイートを用いる。Twitter では、各ユーザが投稿するものはツイートと呼ばれる短文であり、気軽に投稿できるため、当該ユーザの現在の状況に関連したツイートが多くある。そのため、現実世界で発生しているイベントを検知することが可能であると考えられる。また、Twitter では、ツイートを行う際に、当該ツイートを行った位置情報を付与することが可能であり、他のユーザと位置情報を共有することができる。本稿では、ソーシャルメディアデータから現実世界において各地域で発生するイベントに起因する非日常的なトラヒック変動を早期に予知することを目指しており、各地域のイベントを検知するためには、上述のツイートの付与された位置情報は有用である。ツイートに付与された位置情報をもとにすることにより、特定の地域のみで特定の単語の出現頻度が高くなると、当該単語に関するイベントが発生したと検知することが可能となる。以上の理由から、位置情報付きのツイートを用いて、非日常的なトラヒック変動を予知することとした。

2.2 イベント関連単語の抽出

ツイートに含まれるテキストには、現在もしくは将来発生する現実世界のイベントに関連した単語 (イベント関連単語) が含まれていると考えられる。そのため、テキストを解析すれば、非日常的なトラヒック変動の予兆を把握することが可能となる。イベント関連単語を抽出するにあたり、ツイートに含まれるテキストを形態素解析によって品詞分解し、その中から名詞を抽出する。そして、抽出された名詞の中から、イベント関連単語を抽出する。イベント関連単語は、情報探索やテキスト

マイニングの分野で利用される、TFIDF [10](Term Frequency Inverse Document Frequency) を用いて抽出する。

本節では、TFIDF の概要を記したのち、イベント関連単語抽出における TFIDF の適用方法について記す。

2.2.1 TFIDF の概要

TFIDF は、ある文書に出現した単語の重要度を表す指標である。ここで、文書 d に含まれる単語 t の TFIDF を $F(t,d)$ としたとき、$F(t,d)$ は以下のように定義される。

$$F(t,d) = tf(t,d) \cdot idf(t) \tag{1}$$

$tf(t,d)$ は、文書 d における単語 t の出現頻度であり、$idf(t)$ は、単語 t の逆文書頻度である。$tf(t,d)$ は、以下のように定義される。

$$tf(t,d) = \frac{n_{t,d}}{\sum_{s \in W_d} n_{s,d}} \tag{2}$$

$n_{t,d}$ は、文書 d における単語 t の出現回数を表しており、W_d は文書 d に含まれる単語の集合である。$\sum_{s \in W_d} n_{s,d}$ は、W_d の要素となっている全ての単語の、文書 d における出現回数の総和を表す。また、$idf(t)$ は、以下のように定義される。

$$idf(t) = \log \frac{N}{df(t)} \tag{3}$$

N は総文書数であり、$df(t)$ は単語 t が出現する文書数を表している。$idf(t)$ は単語 t が出現する文書数が増加するにつれて、値が減少する。すなわち、$idf(t)$ は、単語 t の重要度を表す。複数の文書にまたがって出現する単語は重要度が低い単語として IDF 値は小さくなり、特定の文書にのみ出現する単語は重要度が高い単語として IDF 値は大きくなる。すなわち、$F(t,d)$ が大きい単語は、文書 d に含まれる単語のうち、特徴的な単語とみなされることとなる。

2.2.2 ソーシャルメディアデータへの TFIDF の適用

イベント関連単語は、普段はツイートに含まれる頻度が少ない単語であるが、イベントが発生している場合はその出現頻度が上昇する。すなわち、イベント関連単語は、イベントが発生する時間帯の特徴的な単語であるといえる。

そこで、我々はイベント関連単語抽出に TFIDF を適用する。しかしながら、各ツイートに含まれるテキストの文字数は最大で 140 文字と制限されているため、1 つのツイートを文書をみなし、TFIDF を適用すると、特徴的な単語を抽出することが難しくなる。そこで、本研究では、予め定義した一定時間 (タイムスロット) の間に投稿された全ツイートに含まれる単語からなる文書を生成する。そして、生成された文書を用いて、当該文書内に含まれる単語の TFIDF を算出し、高い TFIDF を有する単語をイベント関連単語とみなす。

以下では、ツイートデータに対する TFIDF 適用手順を説明する。

日付 D のタイムスロット T で投稿されたツイートの集合を $G_{T,D}$ として表記したとき、$G_{T,D}$ は以下のように定義される。

$$G_{T,D} = \{v_t | S_{T,D} \leqq t < S_{T+1,D}\} \tag{4}$$

ただし、v_t は時刻 t で投稿されたツイートであり、$S_{T,D}$ は日付 D のタイムスロット T の開始時刻を表す。

また、日付 D のタイムスロット T で投稿されたツイートに含まれている名詞の集合を $N_{T,D}$ として表記したとき、$N_{T,D}$ は以下のように定義される。

$$N_{T,D} = \{w | w \in N(v), v \in G_{T,D}\} \tag{5}$$

ただし、$N(v)$ はツイート v に含まれている名詞の集合である。

ネットワークに流れるトラヒックに大きな影響を与える現実世界のイベントが発生した場合、イベント関連単語の出現頻度が急増する。そこで、本稿では、出現頻度が高い単語に焦点を当て、TFIDF を適用しイベント関連単語を抽出する。ここで、$N_{T,D}$ の部分集合 $N_{T,D,z}$ を以下に定義する。

$$N_{T,D,z} = \{w_1, w_2, \ldots, w_z \in N_{T,D}\} \tag{6}$$

ただし、w_i は $N_{T,D}$ に含まれた名詞であり、添字 i は $F_{T,D}(w_i) \geqq F_{T,D}(w_{i+1})$ になるように設定される。$F_{T,D}(w)$ は、日付 D のタイムスロット T で投稿されたツイートにおける w の出現回数である。すなわち、日付 D のタイムスロット T で投稿されたツイートについて、出現回数の上位 z の名詞からなる集合が $N_{T,D,z}$ で表される。

本稿では、日付 D のタイムスロット T に関して、$w \notin N_{T,D,z}$ を除外したうえで、当該タイムスロットでの全ツイートを連結した文書 $d_{T,D}$ を作成する。

そして、$N_{T,D,z}$ の要素である w について、当該単語の指標を TFIDF にもとづいて算出する。ここで、当該指標を $M_{w,T,D}$ としたとき、$M_{w,T,D}$ は以下のように定義される。

$$M_{w,T,D} = \frac{F_{T,D}(w)}{\sum_{s \in N_{T,D,z}} F_{T,D}(s)} \cdot \log \frac{|B_{T,D}|}{|B_{w,T,D}|} \tag{7}$$

ただし、$B_{T,D}$ は日付 D のタイムスロット T と比較するタイムスロットの集合であり、$B_{w,T,D}$ は $B_{T,D}$ の各要素について、w を含むツイートが存在するタイムスロットが要素となる部分集合である。また、$B_{T,D}$ は以下のように定義される。

$$B_{T,D} = \{(T,D)\} \bigcup_{i=1}^{x} B_{T,D-i,y} \tag{8}$$

ただし、$B_{T,D,y}$ は日付 D のタイムスロット $T-y$ から $T+y$ までのタイムスロットの集合である。

$M_{w,T,D}$ が高い単語が、日付 D のタイムスロット T で投稿されたツイートに含まれる特徴的な単語を表す。しかしながら、$M_{w,T,D}$ を算出する際の比較対象であるタイムスロットのうち、過半数で出現する単語については、本稿ではイベント関連単語とみなさないものと定める。すなわち、$N_{T,D,z}$ の要素である単語について、以下の式を満たし、かつ $M_{w,T,D}$ が高い上位 u の単語を、日付 D のタイムスロット T でのイベント関連単語と定める。

$$|B_{w,T,D}| \leqq \alpha \tag{9}$$

ただし、α はパラメータである。

2.3 抽出単語にもとづく非日常的トラヒック変動予知

非日常的トラヒック変動の予兆は、イベント関連単語及び当該イベント関連単語を含むツイート数であると推察している。そこで、本稿では、非日常的トラヒック変動の予兆について議論するにあたり、イベント関連単語にもとづき非日常的トラヒック変動を予知する手法 (CbFmethd; Contents based Forecasting method) を提案する。

現実世界のイベントに伴いネットワークの利用ユーザーが局所的に集中し、トラヒックが急増した場合、当該イベントに関連した単語を含むツイート数は増加する。そこで、本稿では、イベント関連単語を含むツイート数に着目し、非日常的トラヒック変動を予知する。

CbFmethod では、現在のタイムスロットを T としたとき、タイムスロット $T+1$ における非日常的トラヒック変動を以下の手順によって予知する。

まずはじめに、タイムスロット T のイベント関連単語を抽出する。ここで、タイムスロット T のイベント関連単語の集合を W_T と表記する。また、タイムスロット i において、W_T の要素であるいずれかの単語を含むツイート数を x_{i,W_T} と表記する。そして、過去の観測時系列 $x_{T-s+1,W_T}, x_{T-s+2,W_T} \ldots, x_{T,W_T} (s \geqq 1)$ から、タイムスロット $T+1$ における W_T を含むツイート数 x_{T+1,W_T} を予測する。本稿では、ツイート数を予測するにあたり、以下のような単純な予測手法を用いる。まず、x_{i,W_T} を $x_{i,W_T} = ai + b$ としてモデル化する。a, b は、過去の観測時系列 $x_{T-s+1,W_T}, x_{T-s+2,W_T} \ldots, x_{T,W_T}$ に対し最小二乗法を適用することによって算出される。そして、算出された a, b を用いて、x_{T+1,W_T} は $a(T+1)+b$ として得られる。

CbFmethod では、x_{T+1,W_T} にもとづき、タイムスロット $T+1$ での非日常的トラヒック変動を予知する。現実世界のイベントが発生に伴い、当該イベントを表す単語を含むツイート数が増加すると、x_{T+1,W_T} は増加する。しかしながら、上記で示した手法では、イベントが発生している間は x_{T+1,W_T} が減少するにも関わらず、非日常的トラヒック変動は発生し続けているという事態が生じる可能性がある。

そこで、CbFmethod では、以下の条件のいずれか一つを満たす場合、次タイムスロット $T+1$ において非日常的トラヒック変動が発生しているという予知を行う。

- 予測値 x_{T+1,W_T} が閾値 β 以上の値となるとき。

$$x_{T+1,W_T} \geqq \beta \tag{10}$$

- 非日常的トラヒック変動が現タイムスロット T 以前に既に予知されており、かつタイムスロット T で実際に非日常的トラヒック変動が発生しているとき。

一つ目の条件は、現実世界のイベントの開始に伴う非日常的トラヒック変動の予知に関連したものであり、x_{T+1,W_T} を予知に用いる。二つ目の条件は、非日常的トラヒック変動が過去に既に予知されている場合に適用される。トラヒック量が通常に戻る時、上記の条件によって非日常的トラヒック変動の終了が検知される。

3. 評　　価

本章では、CbFmethod の予知精度を評価する。そして、その評価にもとづき、現実世界のイベントに起因する非日常的トラヒック変動の予兆について考察を行う。

3.1 評 価 方 法

3.1.1 非日常的トラヒック変動の定義

本評価では、非日常的トラヒック変動が発生しているタイムスロットをトラヒックデータを用いて定義する。しかしながら、特定の地域から流入したトラヒックの時系列データは取得していない。これに対し、対象地域から投稿された位置情報付きツイートの総数を用いることによって、非日常的トラヒック変動を定義する。文献 [11] では、各エリアから得られた位置情報付きツイート数と、当該エリアからのトラヒック流入量との間には、ツイートの投稿後 2 時間以内であれば、高い相関が認められることが示されている。そこで、本評価では、対象地域からのトラヒック量を示す値として、当該地域から投稿された位置情報付きツイートの総数を用いる。

以下の条件を満たす場合、日付 D のタイムスロット T において、非日常的トラヒック変動が発生しているものとみなす。

$$x_{T,D} \geqq x^{\text{usual}}_{T,D} + \epsilon \tag{11}$$

ただし、$x_{T,D}$ は日付 D のタイムスロット T において、対象地域から投稿された位置情報付きツイートの総数である。$x^{\text{usual}}_{T,D}$ は日常的トラヒック変動時において対象地域から投稿された位置情報付きツイート数を表す値であり、ϵ はパラメータである。すなわち、非日常的トラヒック変動が発生しているタイムスロットを、日常的トラヒック変動が発生している場合に比べて ϵ 以上多い位置情報付きツイートが投稿されたタイムスロットと定義する。

$x^{\text{usual}}_{T,D}$ については、日付 D が休日 (土曜日・日曜日・祝日)、休前日 (金曜日・祝日の前日)・平日 (その他の曜日) に応じて区分する。すなわち、$x^{\text{usual}}_{T,D}$ は以下によって定義される。

$$x^{\text{usual}}_{T,D} = \begin{cases} \frac{1}{|D^{\text{h}}|} \sum_{d \in D^{\text{h}}} x_{T,d} & (D \text{ is a holiday}) \\ \frac{1}{|D^{\text{eve}}|} \sum_{d \in D^{\text{eve}}} x_{T,d} & (D \text{ is a day before holiday}) \\ \frac{1}{|D^{\text{o}}|} \sum_{d \in D^{\text{o}}} x_{T,d} & (otherwise) \end{cases}$$

ただし、D^{h} はデータセットに含まれる休日に該当する日付の集合である。同様に、D^{eve} 及び D^{o} はそれぞれ休前日、平日に該当する日付の集合である。

本評価では、タイムスロットの長さを 1 時間に設定する。

3.1.2 パラメータ設定

CbFmethod の評価を行うにあたり、制御パラメータである u, x, y, z 及び α を設定する必要がある。本評価では、u を 5、x を 10、 y を 2、 z を 50、α を 6 に設定した。

3.1.3 データセット

Streaming API では、定義されたフィルタに合致するツイートをリアルタイムで取得することができる。そのフィルタの一つである **location** では、位置情報が付与されたツイートのう

ち、指定した矩形領域内から投稿されたツイートを取得することが可能となる。矩形領域を指定する際には、2 つの経緯度を利用する。すなわち、2 つの経緯度が頂点となるような矩形領域内での位置情報付きツイートを取得することが出来る。本研究では、日本国内で投稿された位置情報付きツイートを収集するために、日本国内が収まる矩形領域を示す座標を **location** パラメータに設定し、ツイートの収集を行った。なお、位置情報付きツイートを以降ではツイートとして記述する。

ツイートに付与された位置情報からは、当該位置が含まれるような矩形の 4 つの頂点の経緯度座標と、その矩形がカバーする地域名を取得することが出来る。

本評価では、渋谷駅周辺で投稿されたツイートをデータセットとして用いる。当該ツイートをデータセットに含めるにあたり、(1) 当該ツイートに付与された地域名が渋谷区に合致する、(2) 当該ツイートの投稿位置を示す 4 点の座標が渋谷駅を中心とした 1-km^2 の矩形領域内に含まれる、のいずれか一方に当てはまるツイートを、データセットに含める。

我々の研究グループでは、2016 年 10 月 4 日 から 2016 年 12 月 23 日にかけてツイートの収集を行った。しかしながら、Streaming API とのコネクション切断等の通信エラーにより、下記の期間で投稿されたツイートについては取集できていない。

- 2016 年 11 月 7 日 14:00 - 2016 年 11 月 7 日 17:59
- 2016 年 11 月 9 日 04:00 - 2016 年 11 月 9 日 14:59
- 2016 年 11 月 22 日 06:00 - 2016 年 11 月 22 日 12:59
- 2016 年 12 月 1 20:00 - 2016 年 12 月 02 日 11:59

上記の期間で投稿されたツイートを用いて評価を行うが、イベント関連単語を抽出する過程において、過去 x 日分のツイートデータが必要となる。しかしながら、上記で示したように、ツイートの収集が行われていない期間が存在する。また、x として本評価では 10 に設定していることから、下記で示す期間における非日常的トラヒック変動の予知に関して、提案手法の評価を行う。

- 2016 年 10 月 14 日 00:00 - 2016 年 11 月 6 日 23:59
- 2016 年 12 月 13 日 00:00 - 2016 年 12 月 23 日 23:59

3.1.4 比 較 手 法

CbFmethod を評価するにあたり、総トラヒック量のみから非日常的トラヒック変動を予知する手法 (VbFmethod; Volume based Forecasting method) を比較手法として用意する。ソーシャルメディアから抽出された情報を用いて予知を行う手法と、総トラヒック量のみから予知を行う手法を比較することにより、ソーシャルメディアデータには、非日常的トラヒック変動の兆候が含まれていることを確認する。

VbFmethod は過去のトラヒックデータから将来のトラヒックを予測する。本評価では、CbFmethod においてツイート数の予測値を取得した方法と同様に、総トラヒック量の予測値を取得する。そして、以下の条件が成り立つとき、VbFmethod は非日常的トラヒック変動を予知する。

$$\hat{x}_{T,D} \geqq x^{\text{usual}}_{T,D} + \rho \tag{12}$$

ただし、$\hat{x}_{T,D}$ は日付 D のタイムスロット T における予測総ツイート数であり、ρ はパラメータである。

本評価では、CbFmethod および VbFmethod はともに、s を 2 に設定する。すなわち、現タイムスロットを T としたとき、タイムスロット $T-1$ と T で得られたツイート数に基づいて、タイムスロット $T+1$ の予測値を取得する。

3.1.5 評 価 指 標

本評価では偽陰性率 (False Negative Rate; FNR) および偽陽性率 (False Positive Rate; FPR) を評価指標として用いる。FNR は以下のように定義される。

$$\text{FNR} = \frac{m_n}{r_p} \tag{13}$$

ただし、 r_p は式 (11) を満たすタイムスロット数であり、m_n は式 (11) を満たすタイムスロットのうち、非日常的トラヒック変動が予知されなかったタイムスロット数である。FPR は以下のように定義される。

$$\text{FPR} = \frac{m_p}{r_n} \tag{14}$$

ただし、 r_n は式 (11) を満たさないタイムスロット数であり、m_p は式 (11) を満たさないタイムスロットのうち、非日常的トラヒック変動が誤って予知されたタイムスロット数である。

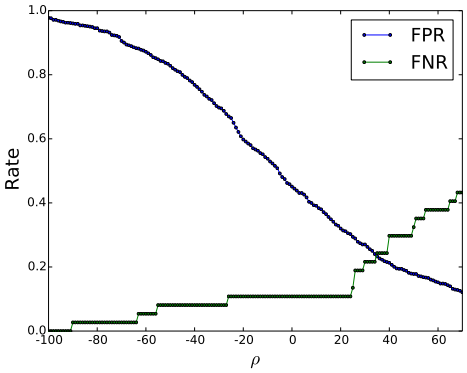
トラヒックエンジニアリングにおいて、非日常的トラヒック変動のすべてを予知できることが望ましい。そこで、本評価では、CbFmethod、VbFmethod それぞれについて FNR が 0 となるパラメータについて着目する。

3.2 評 価 結 果

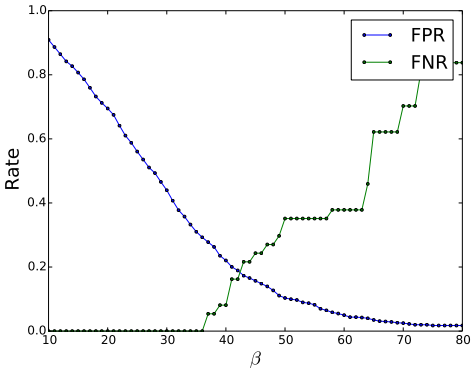
ϵ が 70 に設定されたときの FNR 及び FPR を図 1 に示す。図 1(a) は式 (12) 中の ρ を変化させたときの VbFmethod の結果を表しており、図 1(b) は β を変化させたときの CbFmethod の結果を表している。

図 1(a) に関して、$\rho = -91$ の箇所では、FNR が 0、FPR が 0.9614 という結果が得られている。すなわち、VbFmethod は非日常的トラヒック変動が発生していないタイムスロットのうち、96%のタイムスロットで、非日常的トラヒック変動という誤った予知が行われている。それに対し、図 1(a) の $\beta = 36$ に着目すると、FNR が 0、FPR が 0.2927 という結果が得られている。この結果からは、CbFmethod が、VbFmethod に比べて非日常的トラヒック変動の予知精度が高いことを明らかとなった。これは、CbFmethod は、ツイートに含まれる単語からイベント関連単語を抽出し、当該単語を含むツイート数を用いて予知を行ったためである。

続いて、偽陽性 (FP) が発生した原因について調査するために、β が 40 のときに、FP が発生したタイムスロットについて着目したところ、それらのタイムスロットのうち 60%のタイムスロットでは、スパムに含まれる単語がイベント関連単語として抽出されたことによって、誤った予知が行われていたことを確認した。スパムツイートが急増すると、それらのツイートに含まれる単語の 1 タイムスロットでの出現頻度が急増する。その結果、CbFmethod は、スパムツイートに含まれた単語をイベント関連単語と判断し、非日常的トラヒック変動が発生す



(a) VbFmethod



(b) CbFmethod

図 1　FNR and FPR ($\epsilon = 70$)

るという誤った予知が行われた。この点に関して、投稿されたツイートがスパムと判断する解析技術を用いることによって、スパムの影響による FP の発生は削減することが可能である。CbFmethod によって解析するツイートのうち、スパムツイートを除外することによって、CbFmethod の予知精度は更に向上すると考えられる。

3.3 考　　察

現実世界のイベントに起因する非日常トラヒック変動の予兆について考察する。我々は、現実世界のイベントに関連したツイート数は、非日常トラヒック変動の予兆の一つであるという仮説を立てた。そして、その仮説に基づき、非日常トラヒック変動を予知する CbFmethod を提案した。提案手法を評価したところ、総トラヒック量を用いては予知困難であった非日常的トラヒック変動を提案手法は予知可能であることを示した。すなわち、非日常的トラヒック変動の予兆は、ソーシャルメディアデータに含まれるということを確認することができた。ただし、CbFmethod で用いるイベント関連単語を抽出するにあたり、本稿では示した手法では、現実世界のイベントに関連したツイートとスパムツイートを区別することはできない。そのため、非日常的トラヒック変動の予兆を抽出する過程において、スパム等の影響を除外する必要がある。

4. ま　と　め

本稿では、ソーシャルメディアデータに含まれる、非日常的トラヒック変動の予兆について調査を行った。その調査にあたり、ツイートに含まれたテキスト情報から、非日常的トラヒック変動の予兆であると推察される単語を抽出する手法について提案している。加えて、抽出された単語にもとづき、非日常的トラヒック変動を予知する手法についても提案した。提案した予知手法を評価したところ、総トラヒック量にもとづく手法では予知できなかったトラヒック変動を予知できることが明らかとなった。この評価結果からは、ツイートから抽出された単語は、非日常的トラヒック変動の予兆であり、トラヒック予測へ適用した場合に有効であることを確認した。

今後の課題としては、ソーシャルメディアから抽出された非日常的トラヒック変動の予兆を、トラヒック予測分野へ応用することが挙げられる。

文　　献

[1] H. Abou-zeid, H.S. Hassanein, and S. Valentin, “Energy-efficient adaptive video transmission: Exploiting rate predictions in wireless networks,” IEEE Transactions on Vehicular Technology, vol.63, no.5, pp.2013–2026, June 2014.

[2] A. Beloglazov, J. Abawajy, and R. Buyya, “Energy-aware resource allocation heuristics for efficient management of data centers for cloud computing,” Future Generation Computer Systems, vol.28, no.5, pp.755–768, May 2012.

[3] K. Kawashima, T. Otoshi, Y. Ohsita, and M. Murata, “Dynamic placement of virtual network functions based on model predictive control,” Proceedings of IEEE/IFIP NOMS 2016 Workshop: International Workshop on Analytics for Network and Service Management (AnNet 2016), pp.1037–1042, April 2016.

[4] W. Liu, A. Hong, L. Ou, W. Ding, and G. Zhang, “Prediction and correction of traffic matrix in an IP backbone network,” Proceedings of The 33rd IEEE International Performance, Computing, and Communication Conference (IPCCC) 2014, pp.1–9, Dec. 2014.

[5] H.E. Hag and S.M. Sharif, “An adjusted ARIMA model for internet traffic,” Proceedings of AFRICON 2007, pp.1–6, Sept. 2007.

[6] B. Krithikaivasan, Y. Zeng, K. Deka, and D. Medhi, “Arch-based traffic forecasting and dynamic bandwidth provisioning for periodically measured nonstationary traffic,” IEEE/ACM TRANSACTIONS ON NETWORKING, vol.15, no.3, pp.683–696, June 2007.

[7] E. D’Andrea, P. Ducange, B. Lazzerini, and F. Marcelloni, “Real-time detection of traffic from Twitter stream analysis,” IEEE Transactions on Intelligent Transportation Systems, vol.16, no.4, pp.2269–2283, Aug. 2015.

[8] M. benKhalifa, R.P.D. Redondo, A.F. Vilas, and S.S. Rodríguez, “Identifying urban crowds using geo-located social media data: a Twitter experiment in New York City,” Journal of Intelligent Information Systems, pp.1–22, 2017.

[9] “Streaming API,” `https://dev.twitter.com/streaming/overview`.

[10] G. Salton and C. Buckley, “Term-weighting approaches in automatic text retrieval,” Information Processing and Management, vol.24, no.5, pp.513–523, 1988.

[11] B. Yang, W. Guo, B. Chen, G. Yang, and J. Zhang, “Estimating mobile traffic demand using Twitter,” IEEE Wireless Communications Letters, vol.5, no.4, pp.380–383, Aug. 2016.