1I2-NFC-01-4in2

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
Extraction and quantification of the expressions of congestion status in information posted to SNS – A case study for tweets just after the Great East Japan Earthquake –

沖　拓弥[*1]
Takuya Oki

[*1] 東京工業大学
Tokyo Institute of Technology

It is the important issues how to prevent accidents caused by congestion, safely guide evacuees and reasonably support people returning home on foot in the aftermath of a large earthquake. The author has focused on information posted to SNS, and attempted to construct the method to grasp the location and degree of congestion based on text data quickly and in detail. In this paper, using text data posted to Twitter just after the Great East Japan Earthquake (11 March 2011 at 02:46 PM JST), the expressions of congestion status were extracted manually. Based on the extracted expressions, the characteristics were analyzed, and the method to quantify the degree of congestion was discussed.

1. はじめに

SNS への投稿情報から，どこが，どれほど混雑しているかを迅速かつ詳細に把握できれば，混雑による事故の予防や，大地震時における安全な避難誘導，徒歩帰宅者の合理的な支援等に役立つと考えられる．

混雑度を把握可能なデータやサービスには，NTT ドコモの「モバイル空間統計」[注1)]や，ゼンリンデータコムが公開する「混雑度マップ」[注2)]，Yahoo! JAPAN の「混雑レーダー」[注3)]等があり，それぞれ携帯端末を利用しているユーザの位置情報に基づき混雑度を推定している．しかし，データの時空間単位が粗いため（250m×250m メッシュ，1 時間ごと），施設・道路単位での混雑度推計や，リアルタイムでの混雑状況把握は困難である．

また，非常時に生じる混雑のメカニズムを分析し，混雑の抑制策について事前に検討しておくことは，将来の首都直下地震や南海トラフ地震に備える上でも重要である．しかし，首都圏の広範囲で，徒歩帰宅者や交通機関の運行再開を待つ人々による大規模な混雑が生じたことで知られている東日本大震災（2011 年 3 月 11 日 14 時 46 分発生）時についても，いつ，どこで，どの程度の混雑が生じたかを定量的に示している既往研究やデータは少ない．

上記の背景から，筆者は，東日本大震災時における首都圏の混雑状況に関する tweet（以下，混雑ツイート）に着目し，混雑状況の時間推移の把握に対する有効性や，tweet 中に共起する固有名詞・地名を用いることで，ジオタグ情報に頼ることなく，混雑ツイートの空間分布を把握できる可能性を示した[沖 2016]．混雑ツイートの多さは，ある時刻・地点における混雑状況に対する人々の関心の高さを示していると考えられる．しかし，実際の混雑の程度を把握するには，tweet 数だけではなく，tweet 本文の内容そのものをより詳細に分析する必要がある．

そこで本稿では，東日本大震災の発生前後に投稿された混雑ツイートを対象に，混雑度表現を抽出した上で，アンケートや画像付き tweet の情報を利用した混雑度表現の定量化手法について検討する．

連絡先：沖　拓弥，東京工業大学大学院理工学研究科，
〒152-8552　東京都目黒区大岡山 2-12-1-W8-10，
03-5734-3162，oki.t.ab@m.titech.ac.jp

2. 混雑ツイートの概要

2.1 混雑ツイートの定義

2011 年 3 月 11 日 14 時から 12 日 14 時までの 24 時間に投稿された日本語 tweet のうち，10%をランダムに抽出した．抽出したツイート（10%ツイート）の中に，混雑状況を表現する際に使用されやすいと予想される「混」という文字を含む tweet は，0.95%含まれていた．ただし，これらの tweet の中には，「混ざる」，「混じる」，「混ぜる」，「混線」，「混合」，「混入」等，群集や車両による混雑とは無関係な語が含まれている．そこで，KH Coder の「抽出語検索」機能を用いて，「混」という文字の用途を調べたところ，混雑に関連する可能性の高い語は，概ね，「混む」（変化語も含む），「混乱」，「混雑」の 3 語に限定可能であることが判明した（表 1）．本稿では，これら 3 語を本文中に含む tweet を「混雑ツイート」と定義し，分析を行う．

表 1 「混」という文字を含む tweet の細分類

抽出語	抽出語 2	品詞 / 活用	活用	頻度	頻度 2
混む		動詞		31.6%	
	混ん		連用タ接続		20.2%
	混み		連用形		10.2%
	混む		基本形		1.0%
	混ま		未然形		0.2%
	混も		未然ウ接続		0.0%
混乱		サ変名詞		30.4%	
混雑		サ変名詞		20.3%	
人混み		名詞		0.7%	
混		未知語		0.4%	
混沌		名詞		0.3%	

※抽出には KH Coder の「抽出語検索」機能を使用した．
※「頻度」および「頻度 2」は，10%ツイートに対する割合を表し，上記の抽出語が 1tweet 中に複数回あるいは複数種類用いられている場合には，重複してカウントしている．

2.2 混雑ツイートの時間推移

混雑ツイート数の時間推移を図 1 に示してある．全体の tweet 数は，本震の発生（11 日 14 時 46 分）直後に，5 倍以上に急増するものの，その後は次第に減少し，翌 12 日の tweet 数はさほど多くない．これに対し，混雑ツイートの数には，本震発生当日の夕方 18 時以降と，翌日の 8 時以降に特に多くなるという特徴が見られる．全体の tweet 数に占める混雑ツイートの割合は，11 日よりも，むしろ 12 日の方が大きい．これらの時間帯は，多数の人々が外出先から帰宅を試みたとされる時間帯と対応しており，混雑ツイートの時間推移が，混雑状況の時間推移を良く表現できている可能性を示唆している．

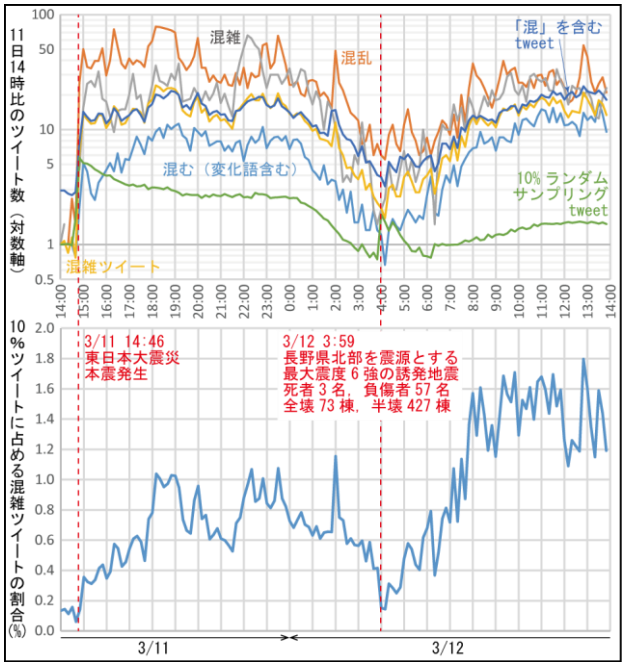

図 1　混雑ツイートの時間推移（10 分単位）

3. 混雑度表現の抽出

3.1 混雑度表現の抽出における問題

混雑状況に関する言語表現を抽出するために，混雑ツイート内で「混む」（変化語も含む），「混乱」，「混雑」の 3 語と共起する語に着目する．本稿では，共起語の抽出に KH Coder を利用した．抽出した共起語を利用するにあたっては，以下に挙げる問題を解決する必要がある．

(1) ポジティブ・ネガティブ，および，程度の判別

混雑状況の推定精度に直接関わる重要な問題である．

(2) 係り受け関係の確認

抽出した共起語は，「混む」（変化語も含む），「混乱」，「混雑」の 3 語のいずれかに直接係っているとは限らない．そのため，係り受け解析（構文解析）が必要となる．

このとき，一般的には，動詞である「混む」（変化語も含む）は副詞，名詞である「混乱」および「混雑」は形容詞によって修飾されると考えられる．しかし，CaboCha を用いた係り受け解析を試みたところ，上記以外の修飾事例が多数見られた．例えば，接頭詞名詞接続（「大」混雑など），名詞サ変接続（「激」混みなど），動詞非自立（「半端ない」，「ガン」，「まくる」など）が挙げられる．

3.2 混雑度表現の教師データ作成

単に係り受け解析を行うだけでは，前節で挙げた問題を解決することは難しい．そこで本稿では，2.1 節で定義した混雑ツイートから 1,000tweet をランダムに抽出した上で，目視にて混雑度表現の抽出を行い，その特徴を分析することとした．

混雑度表現が全く含まれないと判断した tweet は，1,000tweet 中 625tweet であり，混雑度表現と共起しない混雑ツイートが過半数を占めることが判明した．十分なサンプル数を確保可能かどうかは，今後の検証課題である．

次に，抽出した混雑度表現を，混雑「している」ことを伝えたいと考えられる表現（表 2）と，逆に，混雑「していない」ことを伝えたいと考えられる表現（表 3）に分類した．抽出結果に基づけば，混雑「している」ことを伝える表現の方が，混雑「していない」ことを伝えない表現に比べ，種類・登場回数ともに多いことから，目前の混雑が tweet のトリガーとなりやすいことがわかる．なお，表 2 の集計を行うにあたっては，tweet 間の表記ゆれを目視で修正している．以下に，その一例を示す．

表 2　混雑「している」ことを伝えたいと考えられる表現

表現	登場回数（構成比）	表現	登場回数（構成比）	表現	登場回数（構成比）
大	68（22.6%）	劇	3（1.0%）	恐ろしく	1（0.3%）
大変	33（22.0%）	ちょっと	3（1.0%）	あんなに	1（0.3%）
かなり	31（10.3%）	マジ	3（1.0%）	ビッシリ	1（0.3%）
すごい	29（9.6%）	あまりの	2（0.7%）	非常に	1（0.3%）
激	21（7.0%）	鬼	2（0.7%）	さらに	1（0.3%）
めちゃ	20（6.6%）	異常	2（0.7%）	何気に	1（0.3%）
すぎる	14（4.7%）	ごった返す	2（0.7%）	まくる	1（0.3%）
半端ない	10（3.3%）	少し	2（0.7%）	ほどほどの	1（0.3%）
ひどい	10（3.3%）	あふれる	2（0.7%）	やや	1（0.3%）
ものすごい	6（2.0%）	結構	2（0.7%）	ありえない	1（0.3%）
超	6（2.0%）	尋常ではない	2（0.7%）	若干	1（0.3%）
やばい	5（1.7%）	こんな	1（0.3%）	えらい	1（0.3%）
激しい	4（1.3%）	とんでもない	1（0.3%）		
相当	4（1.3%）	ハイパー	1（0.3%）	合計	301（100.0%）

表 3　混雑「していない」ことを伝えたいと考えられる表現

表現	登場回数（構成比）	表現	登場回数（構成比）	表現	登場回数（構成比）
そんなに…ない	4（13.8%）	ない	3（10.3%）	驚くほどではない	1（3.4%）
それほど…ない	4（13.8%）	全然…ない	2（6.9%）	ゼロ	1（3.4%）
大きな…ない	4（13.8%）	そこまで…ない	1（3.4%）		
あまり…ない	4（13.8%）	大して…ない	1（3.4%）		
特に…ない	3（10.3%）	さほど…ない	1（3.4%）	合計	29（100.0%）

(1) 漢字表記のゆらぎ

「凄い」「過ぎる」「溢れる」 → 「すごい」「すぎる」「あふれる」

(2) 平仮名・片仮名表記のゆらぎ

「ヤバい」「まじ」 → 「やばい」「マジ」

(3) 口語表現のゆらぎ

「めっちゃ／むっちゃ／めちゃくちゃ／無茶」 → 「めちゃ」
「ハンパない／ハンパねぇ」 → 「半端ない」
「尋常じゃない」 → 「尋常ではない」

また，上で挙げたような，「混む」（変化語も含む），「混乱」，「混雑」を直接修飾する語の他に，比喩や比較を用いて混雑度を伝えていると考えられる表現が散見された．例えば，交通に関わる表現（「乗れない」「降車できない」「終電」「ラッシュ」「満員電車」「通勤時」等），特定の日時に例える表現（「金曜の夜 11 時過ぎ」「土曜朝でない」等），平常時との比較（「いつも通り」「未だかつてない」「いつもの夕方」等），行動に関する表現（「な

かなか動けない」「入場制限」「将棋倒し」「大行列」等）のほか，「カートがほぼ無い」「映画のよう」「200m超」等の表現があった．

本節で得た知見を教師データとすれば，混雑度表現を自動的に，かつ，精度良く抽出できる可能性が高い．今後，混雑ツイート全体や，他の日時に適用し，精度検証を行う予定である．

4. 混雑度表現の定量化手法の検討

程度を表す言語表現をどのように定量化するかは，様々な領域において重要な課題であり，これまでも多くの研究がなされてきた．例えば，柴田らは，程度副詞（「ほんのちょっと」「少し」「けっこう」等）を用いた速度調整指示に応じて，被験者がどの程度，機械の速度を調整するかを解析することで，知能機械のための，程度副詞に対応した速度調整量決定モデルを構築している[柴田 2011]．熊本は，楽曲の印象表現を強める/弱めることに用いられる程度語119語の強弱関係を被験者実験により決定した上で，各程度語を点数化し，印象語に与える影響を，印象尺度を用いて評価している[熊本 2004]．また，騒音の程度を表現する「非常に」や「だいぶ」等の語と，実際の騒音の種類との関係については，矢野ら[矢野 2002]や宮川・青野[宮川 2002]などが分析を試みている．

実際の混雑度と言語表現を関連付ける方法としては，テキスト本文とあわせて投稿される画像情報を利用することも考えられる．混雑ツイートとともに投稿された画像は，混雑状況を撮影したものである可能性が高い．2011年8月以前はTwitter公式の画像投稿サービスが存在しなかったため，外部の主要な画像投稿サービスの一つであった「TwitPic」に投稿された画像のURL（「http://twitpic.com/」から始まるURL）を含むtweetの割合を調べたところ，混雑ツイート全体の約0.5%に過ぎないことが判明した．混雑度表現を含むものに限定すると，その割合はさらに減少することになる．

こうした背景から，筆者は，混雑度表現の序列化，および，混雑状況下において選択される混雑度表現の把握を目的として，Webアンケートを実施する予定である．混雑状況は，あらかじめ収集した混雑画像（混雑状況を撮影した写真）を被験者に示すことで，仮想的に設定する．このとき，可能な限り多様な属性（性別・年齢等）の被験者に協力を依頼することで，混雑状況別，属性別，個人ごとに異なる混雑度表現の選択傾向を定量化できると考えている（表4）．

表4 混雑度表現と実際の混雑度の対応表のイメージ

混雑度表現	実際の混雑度[人/m^2]
非常に	2.0 ～
すごく	1.5 ～ 2.0
とても	1.3 ～ 2.5
だいぶ	1.0 ～ 1.2
多少	0.5 ～ 1.0
・・・	・・・

5. まとめ

東日本大震災の発生前後に投稿された混雑ツイートから，1,000tweetをランダムに抽出し，目視で混雑度表現を抽出した上で，その特徴や自動抽出における課題を整理した．第3章の分析結果に基づけば，混雑ツイートの中には，ある程度，混雑度表現を含んだtweetが存在し，その表現方法は多様であるといえる．投稿されたtweet情報の信憑性については，別途検討する必要があるものの，tweetに含まれる混雑度表現を定量化することができれば，各地点の混雑度を推定するための携帯端末利用者を十分な数確保することが難しい状況下や，建物内部や地下空間などの電波が不安定な場所においても，混雑度を推定可能となることが期待される．そのためには，それぞれの混雑度表現が，どの程度の混雑度を表すことに使用されているかを，あらかじめ把握しておく必要がある．そこで，第4章では，抽出した混雑度表現を定量化するための手法について，多角的に検討を行った．今後，実際にWebアンケートを実施し，その結果を報告する予定である．

謝辞

本研究は，JSPS科研費15K18176の助成を受けて行った研究成果の一部である．なお，本稿で使用しているtweetデータは，株式会社NTTデータから提供を受けたものである．

注

1) https://www.nttdocomo.co.jp/corporate/disclosure/mobile_spatial_statistics/ （2016/03/28参照）
2) http://lab.its-mo.com/densitymap/ （2016/03/28参照）
3) http://map.yahoo.co.jp/maps?layer=crowd&v=3&lat=35.681277&lon=139.766266&z=15 （2016/03/28参照）

参考文献

[沖 2016] 沖拓弥: 東日本大震災時における首都圏の混雑状況に関するTweetの特徴，情報処理学会第78回全国大会講演論文集，No.2，pp.5-6，情報処理学会，2016．

[柴田 2011] 柴田論・山本智規・神代充: 程度副詞を用いた知能機械の速度調整に関する基礎的研究，人間工学，Vol.47，No.4，pp.155-159，2011．

[熊本 2004] 熊本忠彦: 程度語の序列化と自然言語感性検索への応用，信学技報，NLC2004-23，電子情報通信学会，pp.1-6，2004．

[矢野 2002] 矢野隆ほか11名: 騒音の社会反応の測定方法に関する国際共同研究―日本語のうるささの尺度の構成―，日本音響学会誌，Vol.58，No.2，pp.101-110，2002．

[宮川 2002] 宮川雅充・青野正二: 環境音に対する印象の尺度構成に関する再検討，日本音響学会誌，Vol.58，No.3，pp.151-164，2002．