2N1-OS-10d-2

Urban Sensing: ソーシャルメディアからの都市情報構造化に向けた試み

Urban Sensing: An Approach to Organizing Urban Information from Social Media

長野 伸一
Shinichi Nagano

(株) 東芝 研究開発センター
Corporate R&D Center, Toshiba Corporation

1. はじめに

世界各地でスマートコミュニティの取り組みが進められている．安全で快適な街づくりに向けて，環境に設置したセンサーや，利用者のプローブを活用し，電気，水道，道路，鉄道といったインフラを効率よく運用する技術開発や実証実験が行われている．こうしたインフラは市民生活に大変密着しており，一部のインフラで発生した災害・事故が他へ波及し，都市全体でインフラ機能が低下するリスクを有している．インフラに関する災害・事故の関連情報を広く活用し，リスクの回避や軽減が求められている [3].

こうした情報源の一つとして，ソーシャルメディアの活用が注目されている．ソーシャルメディアには，市民やインフラ利用者が見聞したり経験した出来事に関連する情報がほぼリアルタイムに投稿されている．インフラ利用者に対するインタラクティブな接点の 1 つとして，センサー情報だけでは網羅が難しい災害・事故の状況把握や因果関係の分析，インフラ利用に関する不満・要望の収集などへの活用が期待されている [2, 4].

本稿では，都市のインフラ情報の構造化に向けて，ソーシャルメディアの一つである Twitter からインフラ利用に関する情報を収集し，ナイーブな手法によりインフラ利用者の関心事を表す語彙を抽出するフィージビリティについて述べる．

2. ツイート数の推移

Twitter 上の鉄道利用に関する発言を題材として，ツイート数の変化を観察するために，Twitter API を利用して「山手線」を含むツイート (2012 年 1 月 1 日から 12 月 31 日，50,638 件) を収集した．投稿日ごとにツイート数を集計したものを図 1 に示す．1 年間を通して，ほぼ毎日 150 から 200 件のツイートがある他，ツイート数が 300 件を超える日が 17 日あることが分かった．また，1,000 件を超えた日が 3 日あり，1 月 25 日は，新宿-新大久保間での火事発生，および携帯電話の大規模な通信障害が重なったことによるもの，6 月 20 日は台風 4 号の影響による線路内への資材の撹乱，10 月 4 日は人身事故の影響によるもので，いずれも十万人規模の鉄道利用者に影響があったとされる．このように，平時と震災時とでツイート数は大きく異なり，自然災害や人的災害の発生による影響や混乱の大きさによって，ツイート数は大きく変動する．

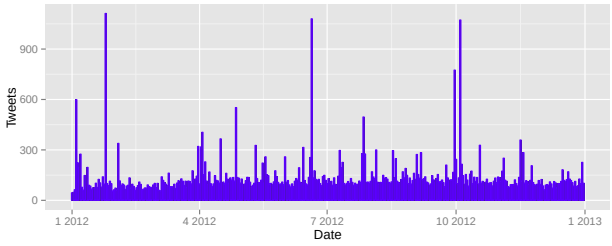

図 1: 「山手線」に関するツイート件数の時系列変化

表 1: 各指標による上位 10 件

tf	df	df'	tf-idf	tf'-idf'
電車	電車	電車	電車	人身
運転	運転	時間	運転	事故
時間	時間	今日	時間	新駅
今日	今日	ホーム	今日	見合わせ
事故	事故	一周	事故	運転
人身	人身	外回り	ホーム	影響
ホーム	ホーム	徒歩	人身	再開
徒歩	外回り	中央	徒歩	驚愕
遅延	遅延	内回り	情報	線路
外回り	内回り	車内	遅延	結果

3. 語彙の出現頻度

本節では，鉄道に関するツイートに現れる一定頻度の語彙を，関心事を表す語彙とみなしてツイートから抽出することを試みる．災害など鉄道沿線で発生した事象の抽出には，人名や地名，施設名などの固有名詞が重要な手がかりとなるが，本稿では特定の事象に限定せず，広く抽出することを目的とするため，一般名詞に限定して調査を行う．前節のツイートを対象に，MeCab[*1] を利用して形態素解析を行う．ツイートおよび形態素解析結果には以下の処理を施しておく．

- リツイートを除外
- ツイート内から，HTML タグ，ハッシュタグ，スクリーン名を除外
- 形態素解析結果から，1 文字の形態素，記号のみから構成される形態素，ひらがなのみから構成される形態素，固有名詞を除外

連絡先: 神奈川県川崎市幸区小向東芝町 1

*1 http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html

まず，ナイーブな手法として，語彙 w に対する単純な出現回数 $tf(w)$，出現ツイート数 $df(w)$，および $tf*idf(w)$ を，それぞれ上位 200 件を求める．$tf*idf(w)$ の定義を式 (1) に示す．W, D は，ぞれぞれ語彙の種類数，総ツイート数を表す．また，高頻度語の影響を抑制するために，出現頻度は W に対する $tf(w)$ の比率で表している．

$$tf*idf(w) \quad = \quad tf(w) \times log\frac{D}{df(w)} \tag{1}$$

上位 10 件を表 1 に示す．ツイートは最大で 140 文字の短い文書であるため，大部分の語彙は 1 つのツイートに 1 度しか出現しない．実際，$tf(w)$ と df に並ぶ語彙は概ね同じものが多く見られる．両指標を利用して算出される $tf*idf(w)$ も同様である．

次に，平時と災害時とで異なるツイート数を考慮するために，上記 $tf*idf(w)$ を一部変更したした指標 $tf'*idf'(w)$ を導入する．すなわち，語彙 w の出現ツイート数 $df(w)$ に代わり，出現日数 $df'(w)$ を導入する．$tf'*idf'(w)$ の定義を式 (2) に示す．$df'(w)$ の逆数を利用するため，災害時など極希に発生する事象に対して，高頻度で出現する語彙に高いスコアを付与できる．

$$tf'*idf'(w) \quad = \quad df(w) \times log\frac{D'}{df'(w)} \tag{2}$$

$df'(w)$ および $tf'*idf'(w)$ の抽出結果を表 1 に示す．$df'(w)$ に並ぶ語彙は $tf(w)$，$df(w)$ と概ね同じものが多く見られるが，$tf'*idf'(w)$ には，災害時に見られる語彙が抽出されている．

表 2: $df'(w)$ の上位 200 語彙のクラスタリング結果

番号	語彙
1	人身，事故，発生，緊急，見合わせ，他 20 語
2	階段，途中，帰宅，無事，ホーム，他 19 語
3	アナウンス，出口，方面，乗り換え，他 19 語
4	不動産，物件，賃貸，マンション，他 16 語
5	広告，ラッピング，車両，編成，他 16 語
6	イベント，募集，参加，希望，他 13 語
7	男性，女性，話題，学校，大学，他 10 語
8	最終，案内，終電，行き，ルート，他 6 語
9	トレイン，サービス，ネット，開始，他 4 語
10	改札，便利，付近，ホテル，近く，他 3 語
11	運行，運休，環状，状況，ツイート，他 2 語
12	絶対，発見，無理

表 3: $tf'*idf'(w)$ の上位 200 語彙のクラスタリング結果

番号	語彙
1	マンション，物件，賃貸，不動産，他 18 件
2	遅延，火災，転落，影響，台風，他 17 件
3	運転，見合わせ，人身，事故，振替，他 13 件
4	緊急，停止，ボタン，混雑，他 11 件
5	車内，放送，広告，無線，他 11 件
6	自殺，死亡，女性，ニュース，他 4 件
7	運行，運休，ツイート，環状，状況
8	原因，障害，故障，通信
9	途中，帰宅，ラッシュ，通勤

4. 抽出語彙のクラスタリング

前節において df' および $tf'*idf'$ それぞれで抽出した語彙 200 件に対し，ツイート中の共起にもとづいてクラスタリングを行う．共起性の指標として相互情報量を利用して，各語彙ペアに対する隣接行列を作成し文献 [1] の手法を利用して，共起語グラフの Modularity を計算する．なお，相互情報量の下限値を $log2$，クラスタ要素数の下限値を 3 に設定した．

クラスタリング結果を表 2，3 に示す．表 2 から，毎日のように投稿されているツイートでは，通勤や日常生活に関する話題が最も関心が高いことが分かる．また，沿線の不動産やホテルに関する情報，ダイヤへの乱れや終電に関する情報が見られる．表 3 についても似た傾向が見られるが，表 2 と異なり，事象の種類ごとにクラスタが構成されている．自然災害および人的災害の発生に関する情報や，その原因・影響に関する情報を整理することで，減災施策へと結びつけ，鉄道利用者へ提供するサービスの改善へつなげられると期待される．

5. おわりに

都市インフラ情報の構造化に向けて，実世界センサとしての Twitter から，インフラ利用に関する情報抽出に関するフィージビリティについて述べた．ナイーブな手法によりインフラ利用者の関心事を表す語彙を，ある程度抽出できることを確認した．今後は，様々なインフラを対象とした分析を行い，事象オントロジの構築を進める．

なお，本論文に掲載の商品，機能等の名称は，それぞれ各社が商標として使用している場合がある．

参考文献

[1] V. D. Blondel, et al.: Fast unfolding of communities in large networks, Journal of Statistical Mechanics: Theory and Experiment (2008).

[2] T. Sakaki, et al.: Earthquake Shakes Twitter Users: Real time Event Detection by Social Sensors, Proc. of WWW2010, pp.851-860 (2010).

[3] K. Sasaki, et al.: Feasibility Study on Detection of Transportation Information Exploiting Twitter as a Sensor, Proc. of WCMCW2012 (2012).

[4] A. Sheth: Citizen Sensing, Social Signals, and Enriching Human Experience, IEEE Internet Computing vol.13, no.4, pp.87-92 (2009).