Twitterからの情報抽出

―感染症情報と被災文化財情報を例にして―

応用

荒牧 英治[1]　橋本 泰一[2]
[1]東京大学　[2]東京工業大学

Twitterとは

マイクロブログとは，ごく短い文章のみを投稿，閲覧できるコミュニケーション・サービスである．そのマイクロブログの先駆者的なサービスがTwitter☆1である．Twitterでは，ユーザは主に2つのことができる．1つは，「ツイート」と呼ばれる140文字以内の短文☆2を投稿することである．もう1つは，他人のツイートを購読すること（フォロー）である．ユーザはほかのユーザをフォローすることで他人のツイートを購読し，自分をフォローしている（自分のツイートを購読している）ユーザに対して自分のツイートを送ることができる．日常の何気ない一言を交換し合うことで緩やかなコミュニケーションを形成するのがTwitterの特徴である．

Twitterは自社のAPIを公開し，サードパーティによるアプリケーションの開発を促している．そうすることで，Twitterというサービスを中心としたアプリケーションやサービス圏を構築することに成功した．特に，スマートフォンなどのモバイル端末向けの専用アプリケーションの開発は目覚ましい．本来，Twitterは投稿可能な文字数が140文字以内に制限されているため，長文の入力や閲覧が困難であるモバイル端末とうまくマッチした．現在では，多くのユーザがモバイル端末を使ってTwitterを楽しんでいる．

ユーザはいつでもどこでも容易にツイートを投稿できるようになったおかげで，ある瞬間，人々が体験したことや思ったことがツイートとして流通するようになった．

ユーザのリアルタイムの生の声がつまったツイートから，その声を拾い集めることができれば，実世界で起きている現象を捉えることができる[1)]．このためTwitterは実世界を捉える重要な情報リソースとなる可能性を秘めている．Hubermanら[2)]による人間関係の解析，Boydら[3)]によるコミュニケーション活動の解析，Sakakiら[4)]による地震の探知などさまざまな活用法はその一例である．ここでは，Twitterからのインフルエンザや花粉症といった疾患の流行と東日本大震災における文化財の被災状況に関する情報抽出事例について紹介する．

☆1 http://www.twitter.com/
☆2 他のSNSでは「つぶやき」と呼ばれる．

Twitterから疾患の流行を見つける

ある疾患の症状を持つ人が多ければ多いほど，その疾患に関するツイートを投稿する人が増えるという仮説に基づき，3つの疾患の流行を把握するシステムを構築した（表-1）．Twitterを利用して疾患の流行を把握することには，次の2つの利点がある．

✓**【データ量の多さ】**インフルエンザを含んだツイートは平均1,000発言／日を超えている．
✓**【情報の即時性】**早い速度で直接ユーザから情報収集が可能．

これらの3つのシステムは同じ仕組みで疾患した患者の情報を抽出する．まず，疾患に関連したツイートを集め，ツイートの位置情報を推定する．次に，集めたツイートから疾患にかかったユーザを特定し患者数を推定する．

	カゼミル （エスエス製薬へ技術提供）	インフルくん[5]	花粉症なう （ニフティとの共同研究[1]）
			
対象	風邪とその6つの症状（喉の痛み，寒気，鼻水，咳，熱，頭痛）	インフルエンザ	花粉症
設置	http://kazemiru.jp/ 2010 年 11 月～現在	http://mednlp.jp/influ/ 2011 年 3 月～現在	http://mednlp.jp/kafun/ 2010 年 2 月～2010 年 6 月

表 -1　疾病把握システムたち

まず，ツイートを取集するために Twitter の API を利用する．2008 年 11 月から開始し，30 億件以上を収集した．次に「風邪」「インフルエンザ」などの各疾患と関連するキーワードを含むツイートを抽出する．そして，ツイートに付与された GPS 情報とユーザのプロフィール情報からユーザの位置情報を推定する．

次のようにキーワードを含んではいるが疾患にかかっている患者がいるとははっきり言えないツイートも多くある：

- 頭痛... インフルエンザかもしれない
- 今年はインフルエンザになってない！
- もしかしてインフルエンザじゃない？

このように疾患にかかった人物が特定できないツイートは全体の約 40% もあった．そこで，機械学習器を用いて，患者が特定できるツイートを判別する．これは，スパムメール・フィルタリングや評価表現分析といった文書分類タスクと類似している．ここでは，文書分類タスクでよく用いられる，キーワードの周辺文脈を素性とした Support Vector Machine（SVM）をベースとして分類器を構築した．分類器で患者を推定できたツイートから位置情報と患者数を推定し，疾患患者の分布図を生成する．

分類器の学習データを作成する上で，2 つのツイートの言語的特徴に注目した．そうすることで，疾患の患者がいることを伝えるツイートを判定することができる．

- **感染者情報はあるか？**
  ユーザもしくはユーザの近辺にいる人が疾患に感染しているという内容のツイートかどうか．
- **24 時間以内の情報か？**
  時制を表す表現に注目し，投稿から 24 時間以内の情報を含むツイートかどうか．
- **感染の事実を伝えているか？**
  仮定や疑問文など疾患の事実を損なう種類のモダリティを持つツイートでないかどうか．

本当に流行が分かるのか？（インフルエンザ編）

インフルくんを用いて，2008 年 11 月から 2009 年 7 月までのツイートで検証した．2009 年 4 月にパンデミック騒動が起きたため，インフルエンザについて過剰に報道されていた．次の 3 つの患者数予測を比較したのが図 -1 である☆3．

☆3　ただし，予測患者数は各手法とも平均値で正規化している．

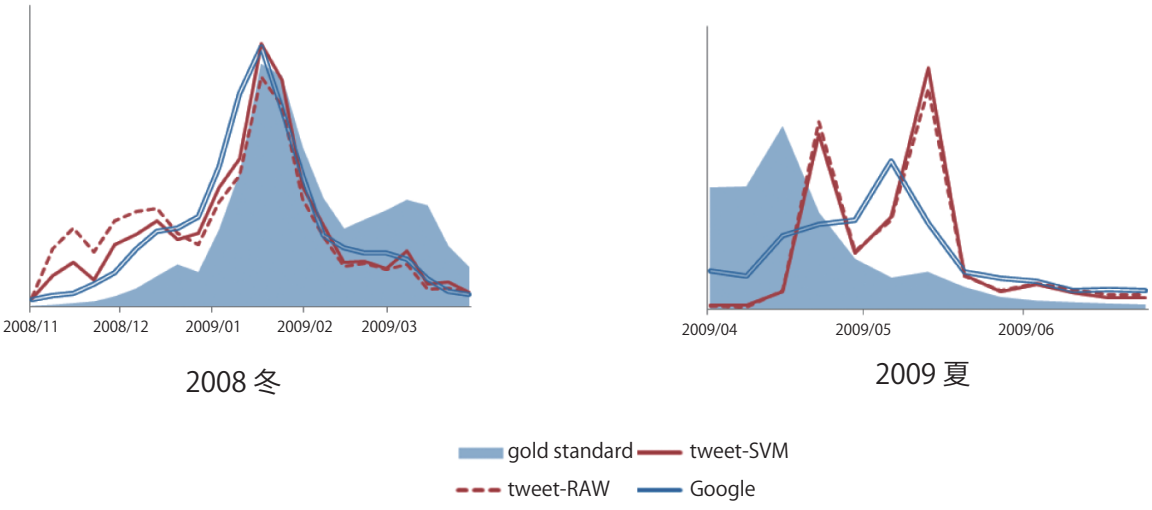

図 -1　インフルエンザの流行の可視化

- **tweet-SVM**：SVM を用いて患者数を推定
- **tweet-RAW**：キーワード「インフルエンザ」「インフル」を含むツイート数
- **Google**：Google flu trend（日本語版）☆4．Google Web 検索のクエリで患者数を推定 [6)]
- **gold standard**：国立感染症研究所から毎週報告される定点当たりの患者数

非過熱報道期（2008 年冬）では，どの手法も患者数の予測が正確であるのに対して，過熱報道期（2009 年夏）では患者数の推定が大幅に誤っている．つまり，過熱報道期でのバイアスを受けやすいことが分かる．特に，Web 検索を用いた手法（Google）よりも Twitter ベースの手法が劇的に精度を下げており，Twitter が報道の影響を受けやすいことが分かる．この理由の 1 つとして，Twitter はコミュニケーションツールとして使われることが多く，報道に反応したユーザのツイートがほかのユーザへ多大な影響を与えるからだ．

Twitter から被災文化財を見つける

2011 年 3 月 11 日 14 時 46 分に東北地方太平洋沖地震により，多くの人々が被害にあったと同時に，

☆4 http://www.google.org/flutrends/

茨城 6 鹿島神宮でもひどい崩壊が・・・
http://twitpic.com/48df9j

図 -2　東北地方太平洋沖地震発生 7 分後のつぶやき

寺院・神社・石仏・板碑など屋外にある文化財も多大な被害にあっている．たとえば，茨城県北茨城市の国登録有形文化財の五浦六角堂（岡倉天心設計）は津波により消失し，伊達家の菩提寺として知られる瑞巌寺の壁は剥落している．文化庁の調べによると，国登録文化財だけでも被害件数は 19 都道県で約 500 件以上にのぼるとされている．Twitter では，地震発生直後にリアルタイムで文化財の被災状況や復興への期待などが多数発信されていた（図 -2）．

Twitter の API，Google や Yahoo! といった検索エンジンを利用して，3 月 11 日から 4 月 10 日までの被災文化財名を含むツイート（5 万件以上）を収集した．そのうち，瑞巌寺（約 1,500 件）と鹿島神宮（約 3,700 件）に関するツイートから頻出する単語を抽出した（表 -2, 3）．

期間	3/11 ～ 12	3/13 ～ 14	3/15 ～ 4/10
1 位	瑞巌寺	瑞巌寺	瑞巌寺
2 位	松島	松島	松島
3 位	避難	国宝	被害
4 位	津波	津波	再開
5 位	情報	避難	拝観
6 位	無事	無事	国宝
7 位	被害	壁	宮城
8 位	場所	被害	復興
9 位	門前	ヶ所	津波
10 位	伊達	県	県

表 -2 「瑞巌寺」頻出上位 10 単語

期間	3/11 ～ 12	3/13 ～ 14	3/15 ～ 4/10
1 位	鹿島	鹿島	鹿島
2 位	神宮	神宮	神宮
3 位	鳥居	駅	駅
4 位	地震	バス	線
5 位	崩壊	東京	運転
6 位	運行	運行	遠方
7 位	石	間	間
8 位	駅	茨城	日
9 位	間	高速	地震
10 位	バス	鳥居	影響

表 -3 「鹿島神宮」頻出上位 10 単語

瑞巌寺は，宮城県宮城郡松島町にある寺であり，本堂や庫裏など建物群の一部が国宝に指定されている．津波が山門の前まで押し寄せたが，実質的被害はなかった．しかし，白壁にひびが入るなどの被害を受けた．文部科学省からは，3 月 15 日に " 国宝：瑞巌寺庫裏及び廊下（漆喰壁に一部崩落・亀裂を確認）" と発表された．

Twitter では 3 月 11 日から 3 月 12 日にかけて，「津波」「被害」「門前」という単語が頻出し，津波が瑞巌寺の門前まで迫っていたことが分かる．3 月 13 日から 3 月 14 日にかけて，「壁」「被害」という単語が頻出し，瑞巌寺の壁に被害があったことが伝えられていた．3 月 15 日から 4 月 10 日の上位には「再開」「拝観」「復興」という単語が頻出し，4 月 8 日に拝観が再開されるというニュースが広まるとともに，再開したら瑞巌寺にぜひ行きたいといった内容のツイートが目立つ．

鹿島神宮は茨城県鹿嶋市にある重要文化財である．本殿などは地震による被害はなかったが，鳥居が完全に崩壊し約 60 基の灯篭が倒れるといった被害を受けた．文部科学省からは被害の詳細の発表はなかった．

Twitter では，鹿島神宮の鳥居に被害があったことを把握することができ，地震発生 7 分後には被害状況を伝えるツイートが投稿されていた．3 月 11 日から 3 月 12 日の上位には「鳥居」「地震」「崩壊」という単語から，地震によって鳥居が崩壊し鹿島神宮に被害があったことが窺える．3 月 13 日以降の上位には「駅」「バス」「運転」という単語から，鹿島神宮から東京へのバスの運行再開が分かる．

リアルタイムの被災情報とは？

集めたツイートを調べてみると．新聞やテレビやインターネットのニュースサイトのニュースについての投稿，ほかのユーザのツイートの再投稿（リツイート）や返信が多く含まれていた．このようなツイートに含まれる情報は Web コンテンツから取得できるため重要ではない．むしろ，ユーザがその瞬間に経験したこと，感じたことについてをツイートから抽出したい．

そこで，集めた被災文化財に関するツイートを人手で分類して，リアルタイムの情報を含むかどうかを判定する分類器を構築した．素性としては，

- ツイートの文字数
- 文化財に関するキーワードの出現位置
- 単語

を用いた．分類器は SVM で動径基底関数カーネル（Radial Basis Function）を利用した．リアルタイム情報を含むツイートの分類精度は F-measure で約 84% 程度であった．

情報利得（Information Gain）を利用して，有効な素性を分析してみると，ツイートの文字数，文化財に関するキーワードの出現位置は有効的であった．また，単語では「なう」，URL，助詞や助動詞などがリアルタイム情報を含むかどうかの判定に影響を

与えることが分かった．

こういった特徴を踏まえて，ツイートを見なおしてみると，ユーザはある瞬間に経験したことや感じたことをツイートする場合には，文章を短くコンパクトにまとめる傾向にある．また，モバイル端末を用いて，写真などの付加情報をツイートに付与することも特徴的である．

まとめ

ここでは，Twitter からの情報抽出アプリケーションとして，インフルエンザや花粉症といった疾患の流行の可視化と災害時の文化財の被害情報抽出事例を紹介した．Twitter からある瞬間のユーザの体験や感情を抽出できることが分かるだろう．

しかし，現在では Twitter はコミュニケーションツールとして活用されているため，本当に欲しい情報を得る際には余分なツイートをいかに効率よくフィルタできるかが重要になる．そのためには，単なるキーワードによるフィルタだけではなく，日常的な文章における時制やモダリティの抽出や意味解析などの自然言語処理技術の発展が必須である．

参考文献

1）高橋哲朗，野田雄也：実世界のセンサーとしての Twitter の可能性，信学技報，Vol.110, No.400, NLC2010-38, pp.43-48 (2011).
2）Huberman, B. and Wu, D. R. F. : Social Networks that Matter : Twitter under the Microscope (2009).
3）Boyd, D., Golder, S. and Lotan. G. : Tweet, Tweet, Retweet : Conversational Aspects of Retweeting on Twitter, In Proceedings of HICSS43 (2010).
4）Sakaki, T., Okazaki, M. and Matsuo, Y. : Earthquake Shakes Twitter Users : Real-time Event Detection by Social Sensors, Proceedings of the 19th International Conference on World Wide Web（WWW）(2010).
5）Aramaki, E., Maskawa, S. and Morita, M. : Twitter Catches The Flu : Detecting Influenza Epidemics Using Twitter, Proceedings of Empirical Methods in Natural Language Processing（EMNLP2011）(2011).
6）Ginsberg, J., Mohebbi, M. H., Patel, R. S. and Brammer, L. : Detecting Influenza Epidemics Using Search Engine Query Data, Nature Vol.457, 19 (2009).

（2011 年 11 月 22 日受付）

荒牧 英治（正会員）▌aramaki@hcc.h.u-tokyo.ac.jp

2005 年東京大学大学院情報学研究科博士後期課程修了，博士（情報理工学）．自然言語処理（機械翻訳／翻字），医療情報（電子カルテ文章からの情報抽出）の研究に従事．言語処理学会，医療情報学会，ACL 各会員．

橋本 泰一（正会員）▌hashimoto.t.ab@m.titech.ac.jp

2002 年東京工業大学大学院情報理工学研究科計算工学専攻博士課程修了．現在，同大総合プロジェクト支援センター 特任准教授．自然言語処理，情報検索，テキストマイニングに関する研究に従事．言語処理学会，人工知能学会各会員．博士（工学）．