2G1-3

ソーシャルメディアからの人物目撃情報抽出システムの試作

Research on Real-time Event Detection from Social Media Prototype of Sighting Information Detection System

榊 剛史[*1] 松尾 豊[*1]
Takeshi Sakaki Yutaka Matsuo

[*1]東京大学
The University of Tokyo

In recent years, information on the Internet provided by indivisual users is more increasing and varied with the developments in social media. It includes not only text information but location information, time stamp information. Some of them refer to real-time events occurred in the real world, which means that we have hidden potential to extract a variety of events in the real world. In this research, we propose a method to extract people sighting information from social media using SVM and pattern matching, make prototype of sighting information detection system, and validate the possibility to extract real -time evens in the real world from social media.

1. はじめに

近年, ソーシャルメディアの普及に伴い, ネット上に流通するユーザーによって発信される情報の量は増え続けており, その種類も多様化している. 例えば, 最新のニュースやそれについての意見, 書籍や商品に関する感想, レストランや観光地に対するクチコミ, ユーザー同士のチャット・ディスカッション, ユーザー行動に対する言及, などがあげられる. これらのようなソーシャルメディア上の情報は, 既存のマスメディアで伝えられる情報と比べ, 信頼性は低く, 玉石混交ではある反面, その量・多様性・リアルタイム性に富んでいる. そのため, 既存のマスメディアからは得られない, 新たな種類の情報が得られる可能性があると考えられる.

実際,2011 年 3 月 11 日に発生した東日本大震災では, 地震発生情報や安否情報, 被災地の被害や被災地で不足している物資についてなど, 実に様々な情報が Twitter 上で発信され, 多くの人々の助けとなった. しかし, その反面, 原子力発電所事故の影響や被災者の情報についてデマが流布し, ソーシャルメディアの負の側面についてもクローズアップされたと言える.

ソーシャルメディアを対象とした研究やサービスが数多く行われている. Twitter を対象としたネットワーク分析 [Huberman 08], Twitter 上のデータを用いた株価予測やヒット映画予測 [Bolle 10, Asur 10], 選挙予測, 地震情報の抽出などの研究が行われている [Sakaki 10, Tumasjan 10].

本研究では, ソーシャルメディアのリアルタイム性を生かし, ソーシャルメディアから実世界において, リアルタイムに起きているイベント[*1]を抽出することを目指す. 本研究の最終的な目的は, 実世界において起きている様々なイベントを抽出する汎用的な手法の提案であるが, 最初の一段階として, イベントを絞って抽出することを目指す. 本論文では, リアルタイムなイベントとして「人物目撃情報」を対象とする. 人物目撃情報とは, ある特定の人物が「いつ」「どこで」目撃されたかについての情報である. これをリアルタイムに知ることで, 実世界での人の動きを把握することが出来る.

連絡先: 榊剛史, 東京大学工学系研究科, 東京都文京区弥生 2-11-16 工学部 9 号館, 03-5841-1161, sakaki@biz-model.t.u-tokyo.ac.jp

*1 ここでのイベントとは, 時間的近接性, 空間的近接性を持ち, かつ主体もしくは客体となる人間が共通している事物の集合と定義する.

表 1: 検索キーワード

遭遇	発見	見かけ	いた

本論文では,2. において, 人物目撃情報について言及した tweet の抽出方法についての述べ, 3. において, 収集した人物目撃情報 tweet から地名を抽出する方法について述べる. 4. において, 人物目撃情報 tweet の抽出精度, 地名抽出精度について実験を行う. 5. では実際に人物目撃情報検出システムを構築し実際に運用した実績について検証する. 6. について手法の問題点, 今後解決すべき問題点について述べる.

2. Twitter からの人物目撃情報の抽出

本章では,Twitter から人物目撃情報を抽出する方法について述べる. ある出来事に関する tweet を抽出する方法について榊らは,Twitter からのキーワード検索と Support Vector Machine を用いて地震に言及している tweet の抽出について 86%の精度を達成している [Sakaki 10]. また特定の分野の情報抽出には人手もしくは自動生成によるパターンマッチングの手法が用いられる. 近年では特に, 生物情報学, 医療情報学の分野でこの手法が見られる. 本研究では,1. キーワード検索により目撃情報の候補を収集し, 2. 収集候補から著名人の人名・通称を含む tweet を絞り込み, 3.SVM による手法とパターンマッチングによる手法の 2 つの手法により, 収集した tweet から目撃情報を抽出する.

1. キーワード検索で用いたキーワードは表 1 の通りである.
2., 3. については以下で説明する.

2.1 人物名を含む tweet の抽出

表 1 のキーワードを含む tweet から人物名を含む tweet のみを抽出する.

人物名リストは Wikipedia から収集した. 具体的には,Wikipedia タイトルより表 2 のような Wikipedia カテゴリに分類されており, かつ没年情報が存在しない（＝存命中である）人物名を収集した. 本研究では, 合計 195128 個の人物名を収集した.

これらの人物名リストを各 tweet 内からマッチングする際には,trie 木を作成し, 高速な探索を行った.

表 2: 使用した Wikipedia カテゴリ

女優	モデル	グラビアアイドル	アイドル	アナウンサー	レースクイーン
歌手	俳優	お笑い	ネット・ブロガー	サッカー選手	野球選手
フィギュアスケート	オリンピアン	プロレス	作曲家	作詞家	映画監督
ギタリスト	脚本家	声優	社長	アナウンサー	AV 女優
小説家	漫画家	タレント	アーティスト		

表 3: 抽出対象とする tweet

P/N 判定	tweet 内容
×	教習所でまさかのはるなとの遭遇ー@(・●・)@きゃぴ
×	留守番終わり〜. 出かけてた人は定食屋から出てきた西川きよしを見かけたらしい.
○	叶美香さん発見?見たいけど見れない〜後ろにいるのに〜直接見れない〜（;￣O￣）
○	渋谷なう. 近くの席に芸人のザブングル加藤発見w

2.2 SVM による人物目撃情報抽出

Twitter において, 表 1 のキーワードを用いて検索すると, 表 3 のような tweet が得られる. これらの tweet のうち表 3 にあるように, リアルタイムに人物を目撃しているユーザーによる投稿のみを抽出することを目指す.

榊らは下記の 3 つの特徴量を SVM に適用し機械学習することで,「地震」「揺れる」というキーワードを含む tweet からリアルタイムに地震に言及している tweet を抽出する手法を実現している.

- キーワード tweet 中に出現するキーワードを特徴量とする
- 語数 tweet 中の語数
- 文脈情報 tweet 中の検索キーワードの位置

本研究でも同様に 3 つの特徴量を用いる.

2.3 パターンマッチングによる人物目撃情報抽出

SVM による情報抽出と同様に, パターンマッチングより表 3 にあるようなリアルタイムな人物目撃情報の抽出を目指す.

このようなパターンマッチングには, 人手でパターンを生成する方法と自動でパターンを生成する方法があるが, 今回は人手でパターンを生成した. 用いたパターンは表 4 の通りである.

3. 位置情報の抽出について

２章のような手法により抽出した目撃情報を含む tweet から, その tweet の投稿者の位置情報を取得することを目指す. ここでの位置情報とは, 緯度・経度の組み合わせを指す. 目撃情報という性質上, 位置情報がなければユーザーにとっての情報の価値が半減すると考えられる. 位置情報の取得方法については, 大きく分けて 2 種類, 詳細に分けて 3 種類の方法が考えられる.

- tweet に付加された GPS 情報の利用
- tweet に含まれる地名情報を抽出
  - 形態素解析により「地名」と判断される地名を利用
  - パターンマッチにより「地名」と推測される語を利用

本章ではこれらの手法について検討を行う.

3.1 GPS 情報を利用した位置情報抽出

Twitter では各 tweet に GPS 情報を付加することが出来る. この GPS 情報を利用すれば, 目撃情報から実際にどこで目撃されたかを正確かつ容易に知ることが出来る. そこで, 予備調査として収集した tweet のうち GPS 情報が付加されている tweet の割合を調べた. 対象とした tweet は表 1 のキーワードを含む tweet を 2010 年 9 月から 2011 年 5 月まで約 9ヶ月間に渡って収集したものである. (ただし, 収集期間に投稿された, 表 1 のキーワードを含む tweet の全てを収集しているわけではない) 結果は, 表 5 である.

表 5: GPS 情報が付加された tweet 件数 (検索キーワード:表 1)

全 tweet 数	tweet 数 (GPS 有り)	GPS 付加割合
1,957,725	2688	0.001

表 5 より, 表 1 のキーワードを含む tweet のうち, GPS 情報が付加されている tweet の割合は 0.1%程度であり, 実用とするには非常に困難であることがわかる.

そのため, 本研究では,GPS 情報が付加されている tweet についてはその情報を使うものの, 他の手法により位置情報を抽出する必要がある.

3.2 tweet 中の地名を利用した位置情報抽出

今回収集した目撃情報には, 下記の様に GPS 情報は付加されていないものでも tweet 中の地名から位置情報が判定できるものがある.

羽田空港で, 寺門ジモンさん?に遭遇

このように tweet 中から地名情報を抽出することができれば, そこから位置情報を得ることが出来る. 本研究では地名-位置情報 (緯度, 経度) 変換には Google MAPS API を用いた[*2]. Google MAPS API を用いることで, 地名から容易に緯度・経度を取得することができる.

本研究では,1. 形態素解析による品詞情報, 2. パターンマッチングの 2 つの手法を用いて tweet 内からの地名情報の抽出を行う.

3.2.1 形態素解析の品詞情報による地名抽出

形態素解析器によっては, 詳細な品詞情報として「地名」を検出することができる. 本手法は, 大字・小字などの住所等に

*2 http://code.google.com/intl/ja/apis/maps/

表 4: 人物目撃情報抽出に用いたパターン

< 対象動詞 >	(発見 | 遭遇 | 見かけ | 見掛け | みかけ | 見た)
< 人物名 >	人物名リスト
< 人物接尾辞 >	(氏 | さん | ちゃん | くん | 君)
< 格助詞 >	(を | と | に)(, |)
パターン	< 人物名 >< 人物接尾辞 >< 格助詞 >< 対象動詞 >

表 7: 目撃情報抽出実験 SVM

手法	精度	再現率	F-value
キーワード	0.756	0.817	0.785
語数	0.627	0.696	0.659
文脈情報	0.684	0.722	0.702
全て	0.762	0.791	0.776

表 8: 目撃情報抽出実験 総合比較

手法	精度	再現率	F-value
SVM 全て	0.762	0.791	0.776
パターン	0.880	0.352	0.503
パターン/SVM	0.803	0.827	0.813

使われる地名については高い精度で検出することができるが,「六本木ヒルズ」などランドマークの名称や通称などの検出に対応することが困難である. 本研究では, 形態素解析器として MeCab を利用した[*3].

3.2.2 パターンマッチングによる地名抽出

日本語の文書においては,「格助詞+動詞」の組み合わせで, 格助詞に係る名詞の意味が限定されることが知られている. 例えば,「< 場所 | 理由 > で + 会う」や「< 場所 | 道具 > で見る」などである [Kawahara 02]. 今回は表 1 のように動詞が限定されているため, それに係る格助詞を絞り込んだパターンを用いることで, 地名情報の抽出を試みる. 実際に用いたパターンは表 6 の通りである.

4. 実験

本論文の提案手法について実験により検証を行う. まず, 人物目撃情報の抽出手法について検証を行った後, 位置情報の抽出手法について検証を行う.

4.1 人物目撃情報抽出の評価実験

2 章での提案手法について評価実験を行う. キーワードを特徴量とした SVM, パターンマッチング, 及びそれらを組み合わせた手法である. 対象とした tweet は人手で正解のタグ付けを行った 450tweets(正解:225tweets, 不正解:225tweets) とし, キーワードとしては MeCab で得られる形態素の原形全てを利用した.

結果は表 7, 8 である.

まず,SVM を利用した場合, キーワードによる特徴量のみが有効であり, tweet 語数および文脈情報は目撃情報の分類にはほとんど寄与していないことがわかる. 実際に分類に大きく影響しているキーワードは「さん」「偶然」「氏」「人」「今日」「にて」「駅」などである.

*3 http://mecab.sourceforge.net/

表 9: 地名情報抽出実験

手法	精度	再現率	F-value
品詞情報	0.423	0.600	0.496
パターン	0.853	0.700	0.770

図 1: デモシステム

次にパターンマッチングを用いた場合, さらに SVM とパターンマッチングを用いた場合を比較する. すると, パターンマッチングを用いた手法は精度は非常に高いものの, 再現率が小さいために F 値では SVM による手法の F 値を大きく下回っている. これは頻出パターンについては有効であるものの, 例外的なパターンを検出できないというパターンマッチングの性質と良く一致している. さらに SVM とパターンマッチングを組み合わせると, 一番高い値が得られている. これより, パターンが有効な場合はパターンマッチング, パターンマッチングでは検出できない tweet については SVM を組み合わせることで, 両方の長所を生かした高精度な検出が可能となっている.

4.2 地名情報抽出の評価実験

3 章での提案手法について評価実験を行う. 本実験では地名が含まれているを 50 tweets, 含まれていない tweets を 50 tweets, 合計 100 tweets を用意し, 品詞情報による手法とパターンマッチングによる手法の精度を比較した. 結果は表 9 の通り.

表 9 より, 精度, 再現率共にパターンマッチングを用いる手法の方が良い数値を示している. これよりパターンマッチングを用いる方が良いと考えられる. ただし, 詳細に中身を見てみると, パターンマッチングで抽出可能な地名と品詞情報を用いて抽出可能な地名は共通しておらず, 品詞情報を用いた場合しか抽出できない場合もある. そのため, 品詞情報を有効に活用するような手法も今後考えていく必要がある.

5. デモシステムの作成

提案手法を利用し, 実際にデモシステムを構築した. デモシステムは, 図 1 の通りである.

今回利用した手法は提案手法を, 以下のように組み合わせて

表 6: 地名抽出に用いたパターン

＜格助詞＞	(で | にて)
＜付加名詞＋格助詞＞	(付近で | 付近にて | 周辺で | 周辺にて | のあたりで | の辺りで | 辺りで | あたりで | を歩いてたら
パターン	(＜地名＞なう) ＜地名＞(＜格助詞＞ | ＜付加名詞＋格助詞＞)＜対象動詞＞

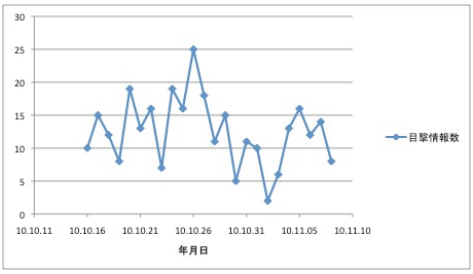

図 2: 目撃情報数の推移 (日単位)

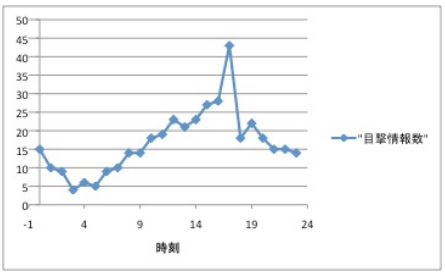

図 3: 目撃情報数の推移 (時間単位)

デモシステムを構築した

- 目撃情報の抽出には,SVM ＋パターンマッチングの抽出を利用
- 地名情報の抽出には, 品詞情報とパターンマッチングの両方を利用 (パターンマッチングを優先)
- 地名情報が抽出できたもののみを表示

実際に得られる目撃情報の数精度について 2010 年 9〜10 月にかけて実施した. まず日毎の目撃情報の数の推移は図 2 の通りである. 毎日一定の目撃情報が得られていることが分かる. 平均としては 12.5 件/day である.

次に時間帯ごとの目撃情報数の推移は図 3 である. これをみると, 夕方から夜にかけて多く目撃されていることが分かる.

6. 考察

まず, 目撃情報の検出精度であるが 8 割程度であり, 実用的な精度が得られているといえる. 今回の結果より, 口語的な文体であるクチコミ情報から, 特定の情報を抽出するには, パターンマッチングと機械学習の組み合わせが有効な手段の一つであることが言える. 今後の課題としては, 今回はよくあるパターンを人手で作成したが, このようなパターンを自動的に作成していくことを考えていきたい.

また位置情報抽出については, 形態素解析とパターンマッチングによりある程度の精度が得られたもののまだ十分な精度が得られてない. 今後の課題としては, 地名情報辞書の充実が考えられる. また, 今回地名検出には成功したものの, 地名-位置情報変換が行えない「自宅近く」や「近所のローソン」といった地名について, 各ユーザー毎の行動範囲をマイニングすることで地名-位置情報変換を行うことを考えていきたい.

7. 結論

本研究においては, ソーシャルメディアからの知識発見の一つのアプリケーションとして人物目撃情報を抽出する手法を提案するとともにデモシステムの構築を行った. これらを通して, クチコミのような口語的な文体からの知識発見に対するパターンマッチングと機械学習の組み合わせによる手法の有効性を示す共に, ソーシャルメディアからのリアルタイムなイベントを抽出する可能性を示せたと考えている. 本研究では, 取り組みやすいイベントとして人物目撃情報を取り扱ったが, 今後はより社会的に意義の深いイベントやユーザーに取っての価値が高いイベントで, かつソーシャルメディアからしか検出できないようなイベントの検出に取り組んで行きたい. 取り組んできたい.

参考文献

[Asur 10] Asur, S. and Huberman, B.: Predicting the Future with Social Media, *CoRR*, Vol. abs/1003.5699, (2010)

[Bolle 10] Bolle, J., Mao, H., and Zeng, X.-J.: Twitter mood predicts the stock market, *CoRR*, Vol. abs/1010.3003, (2010)

[Huberman 08] Huberman, B., Romero, D., and Wu, F.: Social networks that matter: Twitter under the microscope, *ArXiv e-prints* (2008)

[Kawahara 02] Kawahara, D. and Kurohashi, S.: Case Frame Construction by Coupling the Predicate and its Closest Case Component, *Journal of natural language processing*, Vol. 9, No. 1, pp. 3–19 (2002)

[Sakaki 10] Sakaki, T., Okazaki, M., and Matsuo, Y.: Earthquake shakes Twitter users: real-time event detection by social sensors, in *Proceedings of the 19th international conference on World wide web*, WWW '10, pp. 851–860, New York, NY, USA (2010), ACM

[Tumasjan 10] Tumasjan, A., Sprenger, T. O., Sandner, P. G., and Welpe, I. M.: Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment, in *ICWSM* (2010)