

Twitter のつぶやき発信地推定法に関する提案 ～大規模災害発生時における Twitter の活用に向けて～

A proposal of how to estimate sending area of tweet data:

Practical use of twitter in large scale disaster

坂巻 英一十

Yoshikazu Sakamaki

1. はじめに

2011 年 3 月 11 日に発生した東日本大震災が東北 3 県を始めとした太平洋沿岸地域に甚大な被害をもたらしたことは記憶に新しい。地震の直後に発生した停電の影響により通信網は寸断され、自治体や政府は情報がないまま被災地の救援活動を余儀なくされたのである。

こうした中、注目されたのが Twitter を始めとした SNS である。通信手段が寸断された中、阪神淡路大震災発生当時はまだ生まれていなかった SNS を活用することにより、我々は被災者の安否確認や救援物資の需要把握を行うことができたのである。

ところが、現在、Twitter はつぶやきの発信地を特定できる仕組みをサービスとして提供していない。そのため、情報をリアルタイムに発信できる、といったメリットがある反面、その情報がどこから発信されたものなのか、を把握することはできないのである。つまり、いざ、安否確認や物資の需要把握を行おうとした場合に、どこで誰が救助を求めているのか、どこで何がどのくらい必要なのか、といった情報まで特定することができない場合が多いのが現実である。

東日本大震災発生直後につぶやかれたツイートを分析することにより、将来、同様の大規模災害が発生した際の防災対策に Twitter を活用しようといったワークショップが 2012 年 10 月に開催された。このワークショップでは、震災発生直後の人の流れ、どこで助けを求めている人達がいるか、といったことをツイートデータから把握する試みが多く報告された。ところが、実際にはつぶやきの発信地が不明なケースが大半であり、そもそもこれらのつぶやきがどこから発信されたものなのか、を把握することができなければ、ツイートデータを防災対策に役立てることは難しいのではなかろうか。

そこで、本稿では Twitter の防災対策への活用に焦点を当て、発信地が不明なつぶやきの発信元を、単純ベイズ分類器を用いて推測するための仕組みを提案することを目的とした研究を報告する。

2. 本研究における提案モデルの概要

本節では、単純ベイズ分類器を用いて、発信地が不明なツイートデータの発信元を推測する仕組みの構築方法について説明する。

2.1 単純ベイズ分類器とは

単純ベイズ分類器とは、事象間における極めて強い独立性を仮定した上で、ベイズの定理を用いて事象を特徴の類似したもの同士に分類することを目的とした確率的

分類手法である。

単純ベイズ分類器に関する初期の頃の研究としては、Domingos ら[1]が挙げられる。Domingos らは単純ベイズ分類器を用いた分類に関する効率性に理論的な理由を示している。また、McCallum ら[2]は単純ベイズ分類器において広く利用されている多項モデルとベルヌーイモデルについてモデルの予測精度に関する比較を行っている。

単純ベイズ分類器が最も身近に応用されている例としては、マイクロソフト社の Outlook に代表されるメールソフトに搭載された迷惑メールフィルターではなかろうか。今日、単純ベイズ分類器はメールフィルタリング機能を始め、文書分類を始めとした様々な分野で応用されているのである。

この単純ベイズ分類器をツイートデータの発信元の推測に応用することはできないだろうか。先述したように、Twitter 上ではリアルタイムに様々な情報がつぶやかれている反面、ツイートデータの大半は発信地が不明であり、ツイートデータを大規模災害発生時における災害対策に活用する場合、まずは、個々のつぶやきの発信地を特定することが必要になると考えられる。

そこで、本研究では、緯度経度が記録されており、発信地が特定できるツイートデータを教師情報として使用し、単純ベイズ分類器につぶやきを学習させることで、発信地が不明なつぶやきの発信元を推測する仕組みを構築することを試みる。

2.2 つぶやきの発信地推定法

ここで本研究において提案する単純ベイズ分類器を用いて、発信地が不明なつぶやきの発信元を推測する方法について説明する。今、

i ツイート番号($i=1,2,\dots,I$)

m 発信地エリア番号($m=1,2,\dots,M$)

$TWIT_i$ i 番目に発信されたツイート

$AREA_m$ m 番目のエリア

とした時、 i 番目のツイート $TWIT_i$ が発信された時、それがエリア $AREA_m$ から発信される確率を

$$p(AREA_m | TWIT_i) \quad (1)$$

で表すことにする。今、(1)式に対してベイズの定理を適用すると、(1)式は

$$p(AREA_m | TWIT_i) = \frac{p(AREA_m)p(TWIT_i | AREA_m)}{p(TWIT_i)} \quad (2)$$

$$\propto p(AREA_m)p(TWIT_i | AREA_m)$$

と書くことができる。

ここで、 i 番目のツイート $TWIT_i$ に含まれる k 番目の単語を $WORD_{ik}(k=1,2,\dots,K_i)$ と書くことにする。この時、ツイートに出現する単語間に完全な独立性があると仮定すると(2)式は

$$p(TWIT_i | AREA_m) = \prod_{k=1}^{K_i} p(WORD_{ik} | AREA_m) \quad (3)$$

のように書き換えることができる。

(3)式を(2)式へ代入すると、

$$p(AREA_m | TWIT_i) \propto p(AREA_m) \prod_{k=1}^{K_i} p(WORD_{ik} | AREA_m) \quad (4)$$

となる。

ここで、 $p(WORD_{ik} | AREA_m)$ を定式化する必要がある。今、 $T(WORD_{ik}, AREA_m)$ をエリア $AREA_m$ から発信されたつぶやきにおいて、 $TWIT_i$ の中にある k 番目の単語が出現する回数とする。この時、 $p(WORD_{ik} | AREA_m)$ は

$$p(WORD_{ik} | AREA_m) = \frac{T(WORD_{ik}, AREA_m)}{\sum_{i=1}^I \sum_{k=1}^{K_i} T(WORD_{ik}, AREA_m)} \quad (5)$$

によって定式化することができる。

更に、ツイートデータの中には一般に多くの単語が含まれているので、 $\prod_{k=1}^{K_i} p(WORD_{ik} | AREA_m)$ は非常に小さな値になるこ

とが多い。そのため、(4)式を計算すると、計算機がアンダーフローを起こす可能性があり、計算上のアンダーフローを回避するために、(4)式の両辺に対して自然対数をとることとする。(4)式の両辺において自然対数をとると(4)式は(6)式のように書くことができる。

$$\begin{aligned} \log p(AREA_m | TWIT_i) \\ \propto \log p(AREA_m) + \sum_{k=1}^{K_i} \log p(WORD_{ik} | AREA_m) \end{aligned} \quad (6)$$

この結果、ツイート $TWIT_i$ がつぶやかれたときに、そのツイートが発信されたエリアは、関数 $f(x)$ が最大になるような x を $\operatorname{argmax}(x)$ と書くことにすると、

$$\begin{aligned} \operatorname{argmax}(\log p(AREA_m | TWIT_i)) \\ = \operatorname{argmax} \left(\log p(AREA_m) + \sum_{k=1}^{K_i} \log p(WORD_{ik} | AREA_m) \right) \end{aligned} \quad (7)$$

によって与えられることになる。本研究では、(7)式を満たす $AREA_m$ を、 $TWIT_i$ が発信されたエリアとみなし分析を行う。

2.3 モデルの妥当性に関する検証

(7)式を利用すると、発信地が不明なツイートに対して発信地を推測することが可能になる。ここで、第 3.2 項においてモデル構築に利用したデータを学習用データ (in-sample-data)、構築されたモデルの予測精度を検証するために使用するデータを検証用データ (out-of-sample-data) と呼ぶことにする。

今、検証用データにおいて、

i' ツイート番号 ($i'=1, 2, \dots, I'$)

m エリア番号 ($m=1, 2, \dots, M$)

$TWIT_{i'}$ i' 番目に発信されたツイート

$AREA_m$ m 番目のエリア

$WORD_{i'k}$ i' 番目のツイートに含まれる k 番目の単語 ($k=1, 2, \dots, K_{i'}$)

とした時、 i' 番目のツイート $TWIT_{i'}$ が発信された時にそれがエリア $AREA_m$ から発信された確率は

$$\begin{aligned} p(AREA_m | TWIT_{i'}) \\ \propto p(AREA_m) \prod_{k=1}^{K_{i'}} p(WORD_{i'k} | AREA_m) \end{aligned} \quad (8)$$

となる。

ここで、 $p(WORD_{i'k} | AREA_m)$ は(5)式と同様にして、

$$p(WORD_{i'k} | AREA_m) = \frac{T(WORD_{i'k}, AREA_m)}{\sum_{i=1}^I \sum_{k=1}^{K_i} T(WORD_{ik}, AREA_m)} \quad (9)$$

によって定式化することができる。

ところが、つぶやきに含まれる全ての単語が学習用データに含まれていれば(8)式を計算することは可能であるが、検証用データに学習用データに含まれていない単語が含まれている場合、(9)式を計算することが困難になる。こうした問題はゼロ頻度問題と呼ばれており、単語に対するスムージングを行うことで、影響を緩和することが可能になる。一般的に利用されているスムージングの方法として、全ての単語について単語の出現回数に 1 を加えるラプラススムージング (Laplace Smoothing) と呼ばれる方法がある。ラプラススムージングを利用すると(9)式は(10)式のように書き換えることができる。

$$p(WORD_{i'k} | AREA_m) = \frac{T(WORD_{i'k}, AREA_m) + 1}{\sum_{i=1}^I \sum_{k=1}^{K_i} (T(WORD_{ik}, AREA_m) + 1)} \quad (10)$$

ラプラススムージングを利用すると、学習用データに出現しなかった単語が、検証用データに現れた場合でも、その単語の出現確率がゼロにならないようにすることが可能になる。そこで、本研究では $p(WORD_{i'k} | AREA_m)$ の計算に(10)式を利用することにする。

3. 実データへのモデルの適用

本研究では株式会社ホットリンク社(本社: 東京都千代田区、代表取締役社長: 内山幸樹氏)が、ソーシャルメディア分析ツール「クチコミ@係長」によるデータ収集技術を応用して収集し、全国の研究機関向けに公開した東日本大震災に関連したツイートデータを使用した。本研究における検証実験で使用したデータの概要は以下の通りである。

[収集した実験データ]

#tsunami, #jishin 等、震災に関連するハッシュタグまたはキーワードが含まれるツイート。

[収集項目]

ツイートの投稿日時、ツイート本文、プロフィール情報、発信場所の緯度経度に関する GPS 情報、等

[収集期間]

2011 年 3 月 11 日(震災当日)～2011 年 4 月 4 日

[データ量]

66,088,205 ツイート

以上

参考文献

- [1] Domingos, P. and Michael, P., "On the optimality of the simple Bayesian classifier under zero-one loss", Machine Learning, p.103-137, Kluwer Academic Publishers Hingham (1997)
- [2] McCallum, A. and Nigam, K., A Comparison of Event Models for Naive Bayes Text Classification, AAAI-98 Workshop on Learning for Text Categorization(1998)