

Twitterを用いた感染症発生動向の視覚化

Visualization of the occurrence trend of infectious diseases using Twitter

松本 流星^{1*} 吉田 稔^{1,2} 松本 和幸^{1,2} 北 研二^{1,2}
Matsumoto Ryusei¹ Yoshida Minoru^{1,2} Matumoto Kazuyuki^{1,2} Kita Kenji^{1,2}

¹ 徳島大学工学部知能情報工学科

¹ Faculty of Engineering, Tokushima University

² 徳島大学大学院理工学研究部

² Institute of technology and science, Tokushima University

Abstract: The purpose of this research is to extract comments on infectious diseases from Twitter and to create an infection map. As a method, we check the facts of the extracted utterance, extract the past tweets of the infected person from the result, and estimate the place where the person lives. We used SVMs both for the fact check and place estimation.

1 はじめに

本研究では、Twitter から感染症に関する発言を抽出し、感染症マップを作成することを目標とする。現在、一般ユーザが感染症などの病気に関する情報を入手する場合、テレビのニュース、新聞、インターネットなど限られた情報源しか存在しない。また、世の中に飛び交っている多種多様な情報の中でも病気や感染症といった人間の生命に関する情報は、多くの人々にとって重要であると考えられる。これらを解決する方法として、SNS の中でも特にリアルタイム性が高く、且つ一度に大量の情報を入手できる Twitter を利用することで、感染症に関する情報を入手し、現在どこでどのような感染症がどのくらい流行しているのかということがわかる感染症マップのようなものがあれば、一般ユーザにとって有益な情報となり得る。

本研究では、Twitter 上に投稿されているツイートから感染症の名称が含まれているツイートを抽出する。抽出したツイートの文章の内容から投稿した人物が感染症に感染しているかどうかを判定する事実性解析を行う。事実性解析の結果、感染症に感染していると判定された場合、そのツイートの投稿者の過去のツイートを検出し、そのツイートの内容からその人物の住んでいる位置を推定する。位置推定した結果を用いて感染症マップを作成する。

2 既存研究

Twitter と病気を組み合わせた研究はこれまでにいくつか行われている。荒牧ら [1] の研究では、Twitter 上に投稿されているインフルエンザに関連する発言を抽出し、抽出した発言に対して SVM を用いた事実性解析を行い、その結果よりインフルエンザの流行検出を行うという手法を提案している。また、荒牧が開発した Twitter からインフルエンザに関連する発言を抽出し、日本全国の地図と重ねることでインフルエンザの流行地図を提供するサービスの「インフルくん」[6] がある。しかし、これらはインフルエンザのみを対象としており、また感染している投稿者のツイートに GPS により付与されたジオタグ (位置ラベル) やプロフィールに記載されている現在地を元に感染者の位置を示すため、ジオタグが付与されていないツイートやプロフィールに現在地が記載されていないユーザに対して対応ができないといった問題点がある。また、既存研究ではインフルエンザの流行検出を最終目標としているが、本研究ではインフルエンザ以外の複数の感染症を対象とし、感染症に感染している人物の住んでいる場所を推定し、地図上に示すことで感染症マップを作成することを最終目標とする。

2.1 事実性解析

事実性解析とは、文章中のある事象が実際に起こったことなのか、あるいは起こる可能性を述べただけなのかを判定する手法である。本研究では、以下の二項

*連絡先：徳島大学工学部知能情報工学科
〒 770-8506 徳島県徳島市南常三島町 2 丁目 1 番地
E-mail: c501306048@tokushima-u.ac.jp

目について事実性解析による判定を行う。

① ツイートの内容からその投稿者が感染症に感染しているかどうか

② 位置情報を表す語を含むツイートの内容からその投稿者がその場所に住んでいるかどうか

事実性解析の手法としてSVMを用いた教師あり機械学習を用いて行う。SVMについては2.2項で詳しく述べる。

2.2 SVM

SVM(サポートベクトルマシン)[5]とは、機械学習における教師あり学習のパターン認識のアルゴリズムの一種である。SVMは、線形入力素子を利用して2クラスのパターン識別器を構成する手法である。

線形分離超平面上に正例と負例の2つのクラスが存在し、この2クラスを分類する識別面を求めるとする。識別面が正例と負例それぞれの最も近い事象との距離であるマージンが最大になるように識別面を設定する。この識別面を用いることで未知の事象に対してのクラスを求めることができる。

本研究では、感染症に関しては感染症に感染しているツイートを正例、感染症に感染していないツイートを負例とし、位置推定に関しては位置情報を表す語を含むツイートの中で投稿者が自分の住んでいる場所でツイートしているものを正例、それ以外のものを負例として識別面を求め、クラスが未知の新たなツイートに対して分類を行う。

3 提案手法

3.1 提案手法の概要

本研究で提案する手法は、Twitter上に投稿されている感染症に関するツイートの文章を対象に、クラス分類を行い、その投稿者が感染症に感染しているかどうかを判定するものである。その結果から、感染症に感染していると判定された人物の過去のツイートの文章より、その人物が住んでいる場所を推定し、それらの結果を元に視覚化を行う。以下に手順を示す。

- Twitter上に投稿されている感染症の名称が含まれているツイートを取得する
- 取得したツイートに対して事実性解析を行い、投稿者が感染症に感染しているか否かを判定する
- 感染症に感染していると判定されたユーザの過去のツイートを取得する

- 取得したツイートに対して位置情報データベースを用いて位置情報を表す語を検出し、位置推定を行う
- 位置推定結果より、地図上にプロットした感染症マップを作成する

この提案手法を数式化すると以下ようになる。

- ① ユーザ U から得られた感染症に関する1ツイートを u_i とする
- ② 感染症に関するツイート u_i に対して事実性解析を行い、感染していると判定したツイートを u_{it} とする
- ③ 感染していると判定したツイート u_{it} を投稿したユーザ U の過去のツイートを u_p とする
- ④ 過去のツイート u_p に対して位置情報データベースを用いてマッチングを行い、位置情報を表す語 $x(u_p)$ を検出する
- ⑤ 語 $x(u_p)$ を含んでいる過去のツイート u'_p の前後でつぶやかれているツイート u'_q を検出し、 u'_p との関係性を元にその人物の住んでいる位置を推定する
- ⑥ ツイート u'_p の中から位置を推定したツイート u'_t に対して事実性解析を行う

3.1.1 感染症に関するツイートの事実性解析

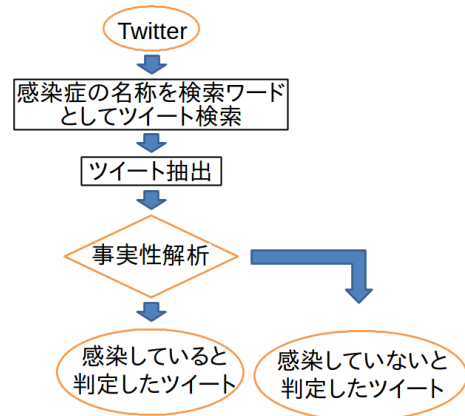


図 1: 感染症に関するツイートの事実性解析のフローチャート

感染症に関するツイートの事実性解析について、簡単な流れを図1に示す。

- ① Twitter上に投稿されているツイートから、感染症の名称を検索ワードとして検索を行い、その名称が含まれているツイートを取得する。
- ② 取得したツイートに対してSVMを用いた事実性解析を行い、その投稿者が感染症に感染しているか否かを判定する。その結果から、感染症に感染していると

判定したツイートと感染していないツイートとに分類する。

3.1.2 位置情報の推定

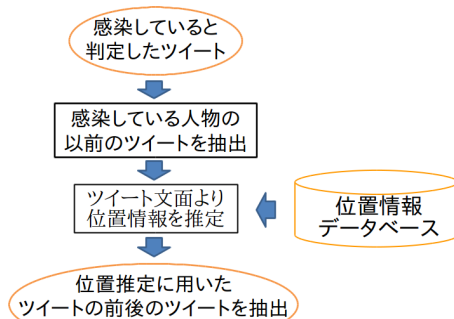


図 2: 位置情報の推定のフローチャート

位置情報の推定について、簡単な流れを図 2 に示す。

- ① 事実性解析の結果、感染症に感染していると判定されたユーザの過去のツイートを取得する。
- ② 取得した過去のツイートの中から地名やスポット名などを含んでいるものを、位置情報データベースを用いて検出する。
- ③ 検出した語を含んでいるツイートの前後でつぶやかれているツイートを抽出し、そのツイートとの関連性を見る。その結果から、位置推定を行う。

3.1.3 位置推定に関するツイートの事実性解析

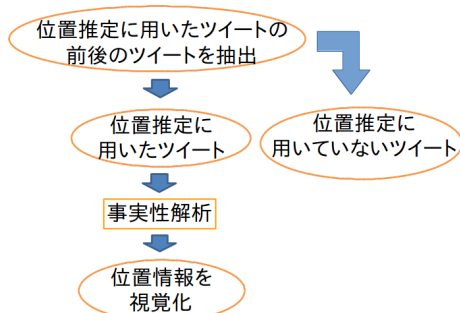


図 3: 位置推定に関するツイートの事実性解析のフローチャート

位置推定に関するツイートの事実性解析について、簡単な流れを図 3 に示す。

- ① 位置推定に用いたツイートと用いていないツイートとに分類する。
- ② 位置推定の精度をより向上させるために、位置推定に用いたツイートに対して感染症に関するツイートと

同様の事実性解析を行う。

③ 事実性解析を行った後に、位置推定に用いたツイートを元に地図上にプロットし、視覚化を行う。

3.2 提案手法のアルゴリズム

3.2.1 感染症に関するツイートの取得

ここでは、Twitter からツイートを取得する方法を説明する。本研究ではツイートを取得するために TwitterAPI を使用する。検索対象として夏に流行している 7 種類の感染症の名称もしくは同義語を用いる。対象とする感染症については、4 節で詳しく述べる。同義語を用いる理由として、本研究で対象にした感染症の多くはその名称よりも同義語がツイート文に含まれていることが多かったためである。例として「流行性耳下腺炎」という感染症を挙げると、この感染症の別名は「おたふく風邪」であり、一般的には後者の方が使われていることが殆どである。このように正式名称よりも通称で用いられることが多い感染症が多いため、同義語を用いる。

ここで例として、「ヘルパンギーナ」という感染症の名称を用いて検索を行う実験方法を説明する。

図 3.4 は取得したツイート例であるが、これを各感染症毎に 400 ツイート取得し、事実性解析を行う。

病院で検査したら、ヘルパンギーナだった
 ヘルパンギーナ疑惑浮上
 やっぱ流行ってるよね、ヘルパンギーナ
 @XXX 今流行してるヘルパンギーナじゃないですか？

図 4: ヘルパンギーナのツイート例

3.2.2 感染症に関する事実性解析

ここでは、取得したツイートに対して事実性解析を行う。事実性解析を行う手法として SVM を用いる。SVM を用いる判定には、学習用データとテストデータが必要となる。ここで使用する学習用データは、予め対象となる各感染症の名称で取得した 1000 ツイートとする。また、1000 ツイートの内訳として正例と負例それぞれが 500 ツイートとなるようにツイートを取得する。テストデータは、3.1.1 項で各感染症毎に取得した 400 ツイートとする。学習用データに対して感染症に感染しているか、感染していないかを目視で判定し、手動でラベル付け (1, -1) を行う。図 5 が学習用データの例である。ラベルの意味として、「1」は感染しているもの、「-1」は感染していないものとなっている。

この学習用データを用いて事実性解析を行い、取得したツイートに対してラベル付けを行う。

病院で調べたらヘルパンギーナだったわ 1
今ヘルパンギーナでダウン中 1
ヘルパンギーナかもしれない -1
今周りでヘルパンギーナが流行ってる -1
ヘルパンギーナ怖いなあ・・・ -1

図 5: 感染症に関する学習用データの例

明日から研修旅行だぜ!
↓
1日目終了乙
↓
2日目オワタ
↓
東京ばなな買い忘れた。買おう。(「東京」を検出)
↓
長いようで短いような研修旅行でしたー!

図 7: 旅行もしくは観光を表しているツイート例

3.2.3 位置情報の推定

前項の感染症に関する事実性解析で感染していると判定されたツイートから、そのツイートを行っているユーザのユーザ ID を用いて、そのユーザの過去につぶやかれているツイートを抽出する。また、過去のツイートの基本的な取得条件として、そのユーザの最新のツイートから 200 ツイートまでを取得する。200 ツイートで取得する理由として、ツイート数が少ないユーザから位置情報を推定するのは非常に難しくなるからである。そのため、取得可能なツイート数が 200 以下のユーザは除外する。また、推定を行う人数は各感染症毎に感染していると判定された中から無作為に選んだ 100 人とする。

過去のツイートを抽出したら、そのツイートに対して位置情報データベースを用いてマッチングを行う。位置情報データベースとは、県名や市名、駅名などのスポット名といったような位置情報を表す語を元に作成したデータベースである。県名と市名に関しては、国土交通省国土政策局国土情報課が提供している位置情報ダウンロードサービス [7] の「大字・町丁目レベル位置参照情報」というものを参照している。図 6 はその情報の一部である。この位置情報データベースと抽出した

1	国道路線名	市区町村名	大字・町丁目名	市区町村番号	道路番号	道路番号	経度	緯度	経度
2	北海道	札幌市中央区	南十二条西五丁目	2	12	106529.3	-75.6611	43.037701	141.321212
3	北海道	札幌市中央区	南十二条西六丁目	2	12	106532.4	-75.6614	43.037501	141.320699
4	北海道	札幌市中央区	南十二条西七丁目	2	12	107388.8	-75.6113	43.029995	141.320831
5	北海道	札幌市中央区	南十二条西八丁目	1	12	107379.7	-75.6485	43.030403	141.320831
6	北海道	札幌市中央区	南十二条西九丁目	1	12	106984.4	-75.6391	43.034403	141.320831
7	北海道	札幌市中央区	南十二条西十丁目	1	12	106945.9	-75.6378	43.033805	141.320831

図 6: 位置情報参照データ

ツイートをマッチングを行うことで、ツイートの文章中に含まれている位置情報を表す語を検出する。位置情報を表す語を検出できたら、その語を含んでいるツイートの前後でつぶやかれているツイートを抽出する。このツイートを利用することで、位置情報を表す語を含むツイートとの関係性を見ることができる。例として旅行もしくは観光中であることを示しているツイートがある場合、その前後でつぶやかれているツイートはその旅行先あるいは観光先に関係しているツイートである傾向がある。取得したツイートに対して旅行あるいは観光でその場所を訪れているかを手動で判別する。そのツイートの前後でつぶやかれているツイートを抽出し、それらのツイートに一貫性があれば旅行あるいは観光中であると判定する。図 7 が例である。

3.2.4 位置推定に関する事実性解析

前項の位置情報の推定精度をより向上させるために、位置情報に関しても感染症に関するツイートと同様に事実性解析を行う。ここで使用する学習用データは、位置情報を表す語が含まれている 1000 ツイートとする。また、1000 ツイートの内訳として正例と負例それぞれ 500 ツイートとなるようにツイートを取得する。ここでの正例は、位置情報を表す語の場所に住んでいると判定できるツイートを意味し、負例はその場所に住んでいない判定できるツイートを意味している。テストデータは、前項の位置推定でユーザの住んでいる位置を推定できるようなツイートとする。学習用データに対して正例あるいは負例であるかどうかを目視で判定し、手動でラベル付け (1, -1) を行う。図 8 が学習用データの例である。ラベルの意味として、「1」は投稿者が自分の住んでいる場所でツイートしているもの、「-1」はそれ以外のものとなっている。

思ったより早く東京に帰れそうだ 1
1ヶ月の出張やっと終わったー!!!ただいま名古屋!!! 1
@XXX 昨日はお疲れ様ー!!また大阪で待ってる! 1
週末は一人で東京に行く -1
買ってから明日学校行こうかな 博多で一泊降るべき? -1

図 8: 位置情報に関する学習用データの例

4 実験

本研究では、以下の 7 つの感染症を対象として、事実性解析の精度評価の結果を示す。

- ① ヘルパンギーナ
- ② おたふく風邪
- ③ 感染性胃腸炎
- ④ 流行性結膜炎
- ⑤ マイコプラズマ
- ⑥ プール熱
- ⑦ 手足口病

本研究では、夏に流行する感染症を対象としているため、7 月～8 月の間で収集したツイートを実験データとして用いる。位置推定に関しては、新たに取得したツイッターアカウントを対象に位置推定を行った結果を示す。

また, 本研究では,SVM を用いた事実性解析を行うために Scikit-learn[4] を用いる.Scikit-learn とは,Python のオープンソース機械学習ライブラリで,SVMをはじめとし,Random Forest,Gradient Boosting,k 近傍法などを含む様々な分類, 回帰, クラスターリングアルゴリズムを備えている. 本研究では,SVM を用いた教師あり機械学習を行う.

4.1 評価方法

感染症に関する評価方法に関しては,4.2.1 項の式 (4.1) を用いた精度評価を行う. 位置情報の推定に関する評価方法に関しては, 新たに評価用アカウントからデータを取得する. 評価用データとして, 新たに Twitter から 100 個のユーザアカウントを取得する. 取得条件は以下のようにしている.

- ① プロフィールに位置情報が記載されている
- ② Twitter クライアント名, アカウント名, 表示名, プロフィールのうち,1 つ以上に「BOT」「Bot」「bot」のいずれかを含む

上記の条件の①に関しては, ツイート文から位置情報を推定した結果が正しいかどうかを簡単に判定することを目的としている. また, プロフィールに位置情報を記載しているユーザは自分の住んでいる地域を表す語を含んでいるツイートが位置情報を記載していないツイートよりも多い傾向にある. ②に関しては, Bot アカウントを除外することを目的としている.

4.2 実験結果

4.2.1 感染症に関する事実性解析

ここでは, 対象となっている 7 つの感染症に対して SVM を用いた事実性解析の精度の評価結果を表 1 に示す. 精度の算出式は式 (4.1) で行う.

事実性解析の精度 = (事実性解析の正解数)/(取得したツイート数) ... (4.1)

表 1: 感染症に関する事実性解析の実験結果

	精度
ヘルパンギーナ	83.75
おたふく風邪	66.25
感染性胃腸炎	79.75
流行性結膜炎	72.00
マイコプラズマ肺炎	84.00
プール熱	68.50
手足口病	67.75

表より, 66~84[%] という結果になった.

4.3 位置推定に関する事実性解析

ここでは, 評価用アカウントから取得したツイートに対して SVM を用いた事実性解析の精度の評価結果を表 2 に示す. 精度の算出式は式 (4.2) を用いる.

事実性解析の精度 = (事実性解析の正解数)/(位置推定を行ったツイート数) ... (4.2)

位置推定を行ったツイートとは, 例として東京という語が含まれているツイートの場合にそのツイートから投稿者が東京に住んでいると推定できたツイートを意味する.

表 2: 位置推定に関する事実性解析の実験結果

	精度
位置推定の事実性解析	78.00

表より,78.00[%] という結果になった.

次に, 3.2.2 項で感染していると判定されたツイートの投稿者の位置推定に関する事実性解析を行った結果を示す.

表 3: 位置推定の事実性解析の実験結果

	精度
ヘルパンギーナ感染者	81.00
おたふく風邪感染者	79.00
感染性胃腸炎感染者	76.00
流行性結膜炎感染者	62.00
マイコプラズマ肺炎感染者	78.00
プール熱感染者	71.00
手足口病感染者	65.00

表より, 62~81[%] という結果になった.

4.4 位置情報の視覚化

ここでは, ヘルパンギーナの感染者を位置推定した結果を元に作成した感染症マップを示す. 地図上に示しているマークは, 以下のような意味をする.

赤色のマーク:感染者が 1 人

青色のマークと数字:感染者が 2 人以上 10 人未満

黄色のマークと数字:感染者が 10 人以上



図 9: 位置情報に関する学習用データの例

5 考察

5.1 感染症に関する事実性解析

本研究では、SVM を用いて事実性解析を行ったが結果より感染症によって誤分類されている割合に変化があることがわかる。この要因として、「おたふく風邪」という感染症は基本的に子供がかかる感染症である。本研究では、ツイート文の投稿者本人ではなくその人物の身内が感染している場合でも正例と扱っている。例として「息子がおたふく風邪にかかった」というツイート文があった場合、このツイート文は正例となる。また、「子供が通っている学校でおたふく風邪が流行っている」、「娘と同じ幼稚園の子がおたふく風邪にかかっている」というように自分の子供が通っている学校などで流行していることを示しているツイートが多い傾向があった。このため、おたふく風邪や手足口病といった子供に流行る感染症に関しては精度が低くなったと考えられる。より精度を向上させるためには、学習用データの増加と偏りを少なくすることが有効であると考えられる。

5.2 位置推定に関する事実性解析

位置推定に関する事実性解析に関して、今回は2種類のデータを用いた評価を行った。誤分類される要因として、今回実験に用いたツイート文の多くは、一つのツイート文に一つの位置情報を表す語を含んでいるものが多い傾向にあった。しかし、一つのツイート文中に複数の位置情報を表す語が含まれているものもあり、誤分類されているツイートの多くは複数の位置情

報を表す語を含んでいるものであった。より精度を向上させるためには、学習用データの増加、ノイズ除去方法の改良あるいは追加、位置情報データベースを増加することが有効であると考えられる。現在行っているノイズ除去の方法は、音符などの特殊な記号のみとしている。これに加えてより多くの特殊な記号や顔文字などを除去することで精度が向上すると考えられる。位置情報データベースは、現在はスポット名に関しては「スカイツリー」や「鳥取砂丘」、イベント名に関しては「東京ゲームショウ」といったような全国的に知られているものが多いため、もう少し知名度の低いものも追加することで精度が向上すると考えられる。

6 結論

本研究では、感染症に関するツイートに感染しているか否かを判定する分類器と位置情報に関するツイートに対してそのツイートが住んでいる場所をつぶやかれているか否かを判定する分類器を作成し、その結果から感染症マップを作成した。結論として、実験よりTwitter ユーザの位置を推定することができたが、その中には正しく推定できなかったものもあった。そのため、事実性解析の際に用いる学習用データの増加や偏りを失くすことが今後の課題となる。

参考文献

- [1] 荒牧 英治, 増川 佐知子, 森田 瑞樹: Twitter Catches the Flu:事実性判定を用いたインフルエンザ流行予測, 情報処理学会研究報告, 2009
- [2] 杉谷 卓哉, 白川 真澄, 原 隆浩, 西尾 章治郎: 教師あり教科学習を用いたツイート投稿時のユーザ位置推定手法, 情報処理学会研究報告, 2013
- [3] 東京都感染症情報センター, <http://idsc.tokyo-eiken.go.jp/>
- [4] python の機械学習ライブラリ scikit-learn の紹介, <http://sucrose.hatenablog.com/entry/2013/05/25/133021>
- [5] サポートベクターマシン入門, <http://home.hiroshima-u.ac.jp/tkurita/lecture/svm.pdf>
- [6] インフルくん, http://mednlp.jp/influ_map/
- [7] 位置情報ダウンロードサービス, <http://nlftp.mlit.go.jp/isj/>