様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　２８　年　　５　月　２３　日現在

機関番号：　１５１０１
研究種目：　基盤研究(C)（一般）
研究期間：　2013～2015
課題番号：　２５４６０８０１

研究課題名（和文）Webマイニングによる新たな感染症流行予測手法の開発と危機管理支援システムの構築

研究課題名（英文）Detecting and predicting influenza epidemics using Internet-based data.

研究代表者
井上　仁（INOUE, Masashi）

鳥取大学・総合メディア基盤センター・教授

研究者番号：００１７６４３９

交付決定額（研究期間全体）：（直接経費）　　3,000,000円

研究成果の概要（和文）：2000年から現在までのインフルエンザ発生数約4万件を収集して、時系列分析により将来予測の可能性について検討した。３つの時系列分析モデル（ARIMAモデル、指数平滑法、最近隣法）を適応して、その有効性を検討した。推定値と観測値を比較したところ、最近隣法が多くの都道府県で他のモデルよりも高精度の予測値が得られた。
インターネット上のビッグデータを用いた流行予測についても検討した。Twitter及び地方新聞社のHP上のインフルエンザに関連する記事約90万件を抽出し、流行予測への有効性について調べた。その結果、これらの情報はインフルエンザが流行し始める時期の推定に有効な指標であることが示唆された。

研究成果の概要（英文）：Having forecasts for infectious diseases can help support risk management and effective intervention against the outbreak of disease. Recently, Internet data from, for example, Google Trends and Twitter have been highlighted as valuable for faster detection of disease morbidity. The possibility to forecast the future incidence of influenza was analyzed using time series analysis based on retrospective data from the Japanese infectious disease surveillance. And also we investigated whether Internet data are promising data sources for monitoring influenza incidence. Several models were evaluated to find a model yielding the most accurate prediction. A nearest neighbor method was regarded as the best-fir model for the prediction. A comparison between the weekly number of Twitter messages containing the term of influenza and the national surveillance data revealed a strong relationship between both. Twitter messages seemed to be a useful data source to survey the influenza epidemic.

研究分野：　医療情報学

キーワード：　インフルエンザ　流行予測　ビッグデータ

１．研究開始当初の背景

私達人類にとって、感染症は日常に潜んでいる最も身近な脅威である。記憶に新しいところでは２００９年の新型インフルエンザの流行は社会に多大な混乱と被害をもたらし、鳥インフルエンザを筆頭に新たな人獣共通感染症の大流行も懸念されている。感染症の流行は、単に人が罹患するというだけでなく、社会経済的にも国家に大きなダメージを与えるものである。感染症の拡大防止と適切な対応のための感染症危機管理は、国の重要施策としても位置付けられている。情報化が進んだ現在、より有効な感染症危機管理手法が求められている。

２．研究の目的

感染症の発生については、国立感染症研究所が主導する感染症発生動向調査（以後発生動向調査と記す）に基づいて全国各地でデータが収集されており、感染症危機管理の重要な情報源となっている。感染症危機管理の基本は、平時から流行の状況を監視し、的確な流行予測に基づいて、感染拡大を未然に防止することである。発生動向調査では、情報の収集から公表まで約2週間の遅れがあることから、流行の立ち上がりが早い疾患では間に合わない危険性があるとの指摘がなされている。しかしながら、現在の発生動向調査は発生数が提供されるだけであり、流行予測の機能は無い。我々は発生動向調査の更なる有効活用を企図して発生動向調査データを用いた流行予測の試みを行った。

最近インターネット上のビッグデータを用いていち早く感染症流行の兆候を見つけ出そうとする試みが注目されている。インターネットが普及し多くの情報が飛び交っており、インターネット上の情報を閲覧することでリアルタイムに実社会の動向を俯瞰することができる。

今回我々はTwitterに投稿された記事から「インフルエンザ」という単語を含むTwitter投稿記事（以下ツイートと記す）を抽出して、インフルエンザの流行との関係を調べた。更にその情報を用いた流行予測についても検討を行った。またローカル新聞社のホームページで提供されている情報と当地でのインフルエンザ流行との関係についても調査を行った。

３．研究の方法

国立感染症研究所が行う発生動向調査の結果は、感染症疫学センターのホームページで毎週提供されている。我々は、47都道府県ごとの2000年から現在までの週別インフルエンザ報告件数4万件以上を収集してデータベースとして整備した。

流行予測については、SPSS Ver.19時系列分析パッケージを用いて分析を行った。またSPSSとは別に、最近隣法を用いての分析も行った。最近隣法は直近数期の連続するデータのパターンに注目し、これを蓄積された過去のデータと照らしあわせた上で、実測値に適当な重みづけをおこない予測値を導き出す手法である。最近隣法については、計算アルゴリズムに基づいて自作のプログラムを作成して処理を行った。先にも述べたように、発生動向調査では情報の収集から提供まで約2週間を要している。そのため提供された週ごとのデータを用いて翌週の値を予測することは意味が無い。そこで、発生動向調査のデータを用いて流行予測を行う際には、1年間52週を4周ごとにまとめて1年間を13週として計算を行った。以後流行予測については、1年は第1週から第13周の13週として取り扱う。推定モデルの適合度は以下の式で表される平均絶対パーセント誤差（MAPE）を用いて評価した。

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\frac{|A_t - F_t|}{A_t}$$

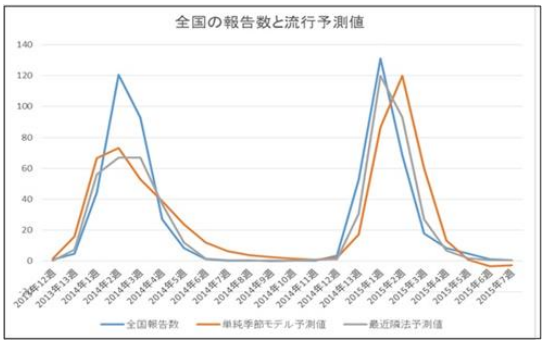
At：実測値

Ft：予測値

また2013年11月から2015年4月の間に投稿されたツイートについて「インフルエンザ」という単語を含んだ記事約200万件を抽出して資料とした。更に鳥取県のローカル新聞社のホームページについて、2012年1月から2015年4月の期間において、毎月曜日に過去1週間分のホームページに掲載された情報から、「インフルエンザ」という単語を含む記事を抽出して資料とした。抽出された記事数は652件である。

４．研究成果

インフルエンザ報告数の全国平均数を用いて流行予測の試みを行った。全国平均の報告数データに対してSPSSの時系列分析機能であるエキスパートモデラーを用いて分析を行ったところ、年間を通して時系列データが周期的なサイクルを呈すると仮定できる単純季節モデルが最も予測に適したモデルであると判定された。単純季節モデルおよび最近隣法を用いて2013年12週から2015年6週の21週において、該当する週より過去のデータを用いて当該週の発生数を上記二つの方法で予測した。予測結果と実際の報告値との関係を下図に示す。



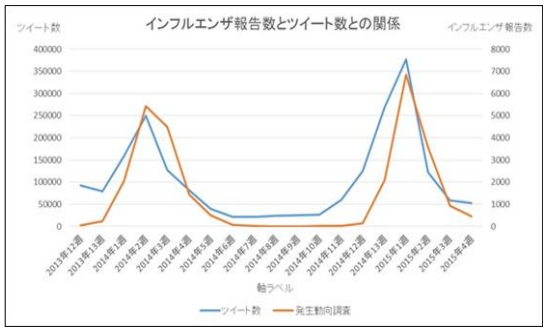
それぞれの適合度の指標である平均絶対パーセント誤差は、単純季節モデルの場合が5.1で、最近隣法の場合が0.67で、最近隣法

が単純季節モデルよりも予測精度が高いという結果が得られた。

長期間にわたるトレンドはほとんどの都道府県で同様である。しかしながら、好発期を含む 11 月から 4 月までを拡大してみると、例えば 2013 年 12 月から 2014 年 4 月までの期間では都道府県ごとにピークを示す時期と高さに違いがあることが分かる。それゆえ、より詳細な流行予測を行うには都道府県ごとに行うことが好ましい。北海道、東京都、愛知県、大阪府、福岡県の５つの地域の時系列データを用いて流行予測を行った結果について記す。以下は５つの地域における単純季節モデルと最近隣法の予測結果である。どの地域においても最近隣法が精度よく予測できた。

	絶対パーセント誤差	
	単純季節モデル	最近隣法
北海道	52.3	0.85
東京都	28.8	0.61
愛知県	1.5	0.56
大阪府	0.93	0.8
福岡県	20.8	2.5

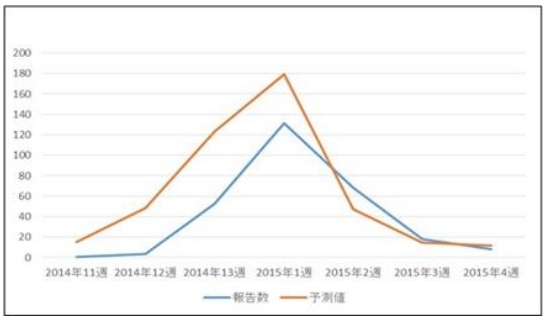
下図は 2013 年 11 月から 2015 年 4 月の間に投稿されたツイートについて「インフルエンザ」という単語を含んだ記事数と発生動向調査による報告数との関係を示したものである。相関係数は 0.84 であり、強い相関関係が認められた。



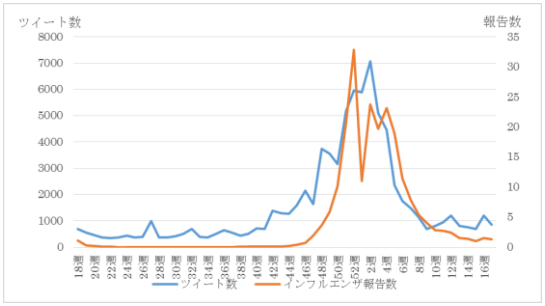
2013 年 11 月から 2014 年 11 月までの期間のデータを用いて、ツイート数を独立変数に、インフルエンザの報告数を従属変数にして回帰分析を行ったところ、ツイート数からインフルエンザ報告数を求める下記の回帰式が求まった。

インフルエンザ報告数＝0.00052×ツイート数－16

下図に、2014 年 12 月から 2015 年 4 月までの期間において、求まった回帰式から推定した値とインフルエンザ報告数との関係を示す。平均絶対パーセント誤差は 9.6 であった。

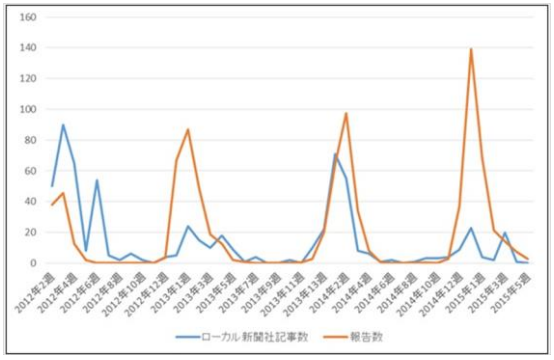


インフルエンザは全国一律に同じタイミングで流行するものではなく、都道府県ごとに異なる様相を呈して伝搬していくため、より詳細な流行状況を把握するには都道府県ごとのデータが必要である。ツイートから都道府県ごとのデータを抽出するには投稿者の所在情報が必要であるが、所在地情報が記録されたものは多くはない。特に地方の県から発信されたと分かるものは少ない。それでも東京からの発信と分かるものは比較的数が多い。下図は東京からの発信について 2014 年 18 週から 2015 年 17 週までのツイート数と東京都のインフルエンザ報告数との関係を示したものである。相関係数は 0.86 であり、非常に強い相関関係が認められた。また特徴的なこととして、実際の流行が立ち上がる前にツイート数の増加が見られる。発生動向調査では、情報の収集から約 2 週間遅れで報告されるが、ツイート数はリアルタイムで把握することができるため、インフルエンザの流行を早期に検出する指標となりうる可能性が示唆された。



ツイートは全国的な流行予測あるいは大都市での早期検出の資料とはなりえても投稿が少ない地方では流行予測の資料として使い難い。そこで我々は、ローカル新聞の記事に注目した。ローカル新聞社のホームページでは地元の情報をタイムリーに提供していると思われる。そこで、鳥取県のローカル新聞社のホームページについて、2012 年 1 月から 2015 年 4 月の期間において、毎月曜日に過去 1 週間分のホームページに掲載された情報から、「インフルエンザ」という単語を含む記事を抽出して資料とした。抽出された記事数は 652 件である。下図は 2012 年第 2 週から 2015 年第 5 週までの鳥取県のインフルエンザ報告数と新聞記事数を示す。相関係数は 0.45 であった。2012 年初頭および 2013

年末から 2014 年初頭においては、両者は同様に大きなピークを示し、ローカル新聞の情報はインフルエンザ流行の兆候を捉える有効な指標となりえた。しかしながら 2012 年末から 1013 年初頭および 1014 年末から 2015 年初頭では両者の類似は大きくなかった。ローカル新聞社の情報はインフルエンザ流行予測には不十分であるが、ローカル新聞社のインフルエンザに関する記事が急激に増える場合は、流行の前兆と捉えることができると示唆される。



５．主な発表論文等

〔雑誌論文〕（計　１件）

Masashi Inoue, Shinsaku Hasegawa, Akihiko Suyama, Masayuki Kakehashi
Development of a Web-based Data Visualization System for Comprehensible Ascertainment of the Spatiotemporal Extent of Infectious Diseases.
Japan Journal of Medical Informatics, 33, 27-32, 2013. 査読有

〔学会発表〕（計　３件）

M. Inoue, S. Hasegawa, M. Kakehashi
DETECTING AND PREDICTING INFLUENZA EPIDEMICS IN JAPAN USING INTERNET-BASED DATA.
IADIS Multi Conference on Computer Science and Information Systems 2014 Proceedings, 414-416,
発表年月日（2014 年 7 月 16 日）
発表場所（リスボン市、ポルトガル）

M. Inoue, S. Hasegawa
WEB-BASED INFORMATION SYSTEM TO SUPPORT RISK MANAGEMENT FOR THE PREVENTION OF INFECTIOUS DISEASE OUTBREAKS.
WWW/INTERNET 2015 Proceedings, 215-217,
発表年月日（2015 年 10 月 25 日）
発表場所（ダブリン市、アイルランド）

井上　仁、長谷川　伸作
ビッグデータを用いた感染症流行予測の試み
第 35 回日本医療情報学連合大会講演論文集, 382-385, 2015.
発表日（2015 年 11 月 1 日）
発表場所（沖縄県宜野湾市）

〔図書〕（計　０件）

〔産業財産権〕
○出願状況（計　０件）

名称：
発明者：
権利者：
種類：
番号：
出願年月日：
国内外の別：

○取得状況（計　０件）

名称：
発明者：
権利者：
種類：
番号：
取得年月日：
国内外の別：

〔その他〕
ホームページ等

６．研究組織
(1)研究代表者
井上　仁（INOUE Masashi）
鳥取大学・総合メディア基盤センター・教授
研究者番号：00176439

(2)研究分担者
（　　）

研究者番号：

(3)連携研究者
岡本　幹三（OKAMOTO Mikizo）
鳥取大学・医学部・講師
研究者番号：40032205