大規模検索ログを用いた感染流行予測における季節調整を考慮した特徴選択手法

TranQuang Thien† 佐久間 淳†,††,†††

† 筑波大学大学院システム情報工学研究科 〒305–8571 茨城県つくば市天王台 1 丁目 1–1
†† JST CREST 〒102–0076 102-0076 東京都千代田区五番町 7 K ’ s 五番町
††† 理化学研究所 革新知能統合研究センター 〒103–0027 東京都中央区日本橋 1–4–1 日本橋一丁目三井ビルディング 15 階
E-mail: †thientquang@mdl.cs.tsukuba.ac.jp, ††jun@cs.tsukuba.ac.jp

あらまし 検索ログは感染症の流行予測に有効なデータの一つである. 各検索語の検索量の時系列は, 感染症の流行をより早く予測するために利用できるだけでなく, 医療データの管理が不完全な地域での流行把握にも有効である. しかし, 検索ログを用いる場合, 正確かつ安定に予測を行うことは容易ではない. なぜなら, ある検索語の検索量が様々な要因により短期的または長期的に変化するためである. 要するに, いい予測モデルを学習するための検索語を選択することが難しいである. 本研究では, この安定性の問題を解決するために, 季節調整をベースにした特徴選択手法と予測モデルを用いる. 具体的に, 我々は時系列を趨勢変動成分, 季節変動成分, 不規則変動成分の 3 つの成分に分解し, れぞれの成分に対して独立に特徴選択と学習を行う. 提案手法が既存手法より優れることを示すために, 10 種類の感染症を用いて実験を行った. 実験結果では, 提案手法が全ての感染症において優位性を示した.

キーワード 検索ログ, 特徴選択, 集団検診, 時系列解析, 季節調整

1. ま え が き

近年, インターネット, 特に検索エンジンは人々の健康管理において重要な情報源になりつつある. 2013 年の調査によると, 米国の成人の 59% がインターネットを用いて健康に関する情報を検索した. さらに, インターネットを用いて自己診断を行ってから医者に相談した人の中で, 自己診断した健康状態医師の診断と一致した人の割り合いは 77% だった [2]. この事実に基づき, すい臓がんを罹患したユーザの特定 [1], 妊婦の関心のパターン発見 [4], 薬の副作用の初期発見 [5] などのように, 大規模な検索ログデータを用いる様々な研究が行われてきた. また, 感染症の流行予測に注目する研究も多数あった [3] [9] [10] [8] [11].

検索ログは感染症の流行予測に有効である. 実際, Google [3] [9] [10], Yahoo [8], Baidu [11] などの検索ログを用いる感染症の流行予測に関する研究が行われた. 診療データを病院から収集し, 感染症の流行予測をする場合, 集計に 1 週間から 2 週間がかかるのに対して, 検索ログはほぼリアルタイムに集計できるため, 検索ログから感染症の流行を予測できれば, より早く流行を把握することができる. また, 検索ログによる流行予測は, 診療データの管理が不完全な発展途上国などの地域では, より有効であると考えられる. このように検索ログは大きな可能性を持っているが, 予測器の構築は慎重に行われる必要がある. なぜなら, 感染流行の予測の間違いは, 科学に対する信頼の喪失を引き起こしかねない. また, 誤った予測は医療資源を無駄に消費させるリスクに繋がる恐れがある.

検索ログから流行予測にはいくつかの困難性が知られている. まず, 各検索語の検索量は様々な要因によって短期的または長期的に変化するため, 精度の高い予測器を学習することは簡単ではない. まず, 人々の情報検索行動は不規則であり, かつニュースや報道によって影響を受けやすいため, 検索量が短期的に変動することがある. 他に, 検索エンジンのユーザ構成などの経時変化のため, 検索語の検索量は長期的にも変化する. つまり, このような傾向を持つ検索ログを利用する場合, これらを考慮した手法を用いる必要がある.

もう一つの困難性は検索エンジンの検索語数が膨大な点である. 例えば, Yahoo!Japan の検索エンジンでは, 平均一日におよそ 3×10^7 個の異なる検索語が検索される. これを予測に用いる場合, 説明変数の次元が非常に高くなる. この場合, 計算コストを減らし, 過学習を防ぐために, 有効な検索語を選択する, いわゆる特徴選択を行う必要がある. さらに, 公衆衛生においては, 予測が正しいだけでなく, 予測の解釈性も求められる. つまり, 予測したい感染症に関連する検索語を選択する必要がある. 実際, 全検索語の膨大な数に対して, 特定の感染症に関連する検索語の数が非常に小さいと考えられる. 従って, 予測モデルに使うための, 有効かつ解釈性を持つ少数の検索語を選択できるスケーラブルな手法が必要である.

この特徴選択問題の最も簡単なアプローチとして, 感染症に直接関連するキーワードを含む検索語 (例 インフルエンザ, タミフル) などを人手で選択する方法が提案された [11] [8]. これらの研究はインフルエンザを対象にしているが, その他の多様な感染症への適用を考えた場合, このような人手での特徴選択は現実的ではない. また, エキスパートによる特徴選択は, 解釈性を保証できるが, 必ずしもそれらを用いた予測モデルの予測精度が高いとも限らない. 従って, 予測精度に基づいて自動的に有効な検索語を選択できる手法が必要である.

事前知識を導入することなく, 検索ログからの特徴量を自動的に選択する研究として, J. Ginsberg らの Google Flu Trend(GFT) の手法 [3] およびそのフォローアップ研究があ

る [9] [15] がある. [3] では, アメリカの CDC 報告を正解データ, 最も検索量の高い 50,000,000 検索語を特徴の候補として用いた. 特徴選択の過程ではまず, 個々の検索語についてロジット回帰モデルを予測モデルとして学習し, 予測値と正解データの CDC 報告との相関を用いて検索語をスコアリングする. 次に, 最も高い相関係数を持つ一部の検索語を特徴として選択した. 最終的な予測モデルは選択された特徴量の総和を特徴量とする 1 次変数ロジット回帰モデルで与えられた. なお, この研究は後に Google Flu Trends サービスとして実用化された.

GFT の最も大きな問題点は予測が不安定なところである. 後に 6.3 節で示される我々の実験結果から, GFT 手法を用いた予測感染率は, 実際の感染率よりも短期的そして長期的に大きく高い値を示す傾向が頻繁に見られた. また, 実際に GFT サービスで公開される予測が実際の流行を上回る誤った予測を与えるという報告がされた [9]. Google は 2013 年のレポートで, GFT 手法がニュースなどの報道に敏感である根拠を示し, 誤った予測を与える理由を説明した. また, 同レポートでは GFT 手法が長期的な変化に対応できないことも示された. このような不安定性の問題以外に, GFT 手法は次のような弱点がある. 第一に, 選択された特徴量の総和を最終的な特徴量とする方法は個々の検索語の変動を考慮できないため適切ではない. 第二に, 相関係数を用いる予測精度の評価は予測精度を過剰評価または過小評価することがある [13]. 第三に, GFT の研究を含む既存研究のほとんどは特定の一つの感染症に着眼することが多く, 他の感染症に対する汎用的な適用性には疑問がある.

本研究は検索ログデータを用いる予測の安定性を解決するために, 季節調整を用いて時系列を季節調整で季節変動成分, 趨勢変動成分, 不規則変動成分の 3 つの成分に分解し, 各成分を独立に扱う. 具体的に, 我々はこの分解をベースにするスケーラブルな特徴選択手法と分割可能な予測モデルを提案する. ここで, 不規則変動成分は短期的な変化, 趨勢変動成分は長期的な変化に対応すると考えられる. 各成分を独立に扱うことによって, 短期的な変化と長期的な変化をより対応できるようになり, より安定な予測モデルが学習可能になる. さらに, 特徴選択では, より適切な評価指標として相関係数の代わりに二乗誤差を用い, より表現力のあるモデルとして L2 正則化付き多変量線形回帰モデルを用いた.

我々はこの検索ログを用いる感染症の流行予測問題において, より正確かつ安定な予測が達成できた. 提案手法が既存手法より優れることを示すために, 10 種類の感染症において実験を行った. 実験結果では, 提案手法が 10 種類の感染症において提案手法より高い予測精度を達成できることが示された. 特に, インフルエンザの予測の場合, 提案手法は予測感染率と実際の感染率の相関は 0.99 であり, 極めて高精度である. また, 提案手法は, 既存手法より 9 倍小さい予測誤差を達成した. また, 提案手法は現在予測だけでなく, 将来予測においても予測精度のおける優位性があることが実験的に確認できた. さらに, 提案手法は, 既存手法より解釈性の高い検索語を選択できることも, 実験的評価により確認できた.

2. デ ー タ

本研究に用いられるデータについて説明する.

感染症報告データ: 我々は, 公的に利用可能な国立感染症研究所の感染症発生動向調査（IDWR）速報データを正解データとして用いる [7]. このデータは報告義務のある全ての感染症を対象として集計されるデータである. 感染症のある時点での感染率を y_t^d と表すことができる. ここで, d (disease) は感染病名, t (time) は週単位の時間を表す. IDWR データは毎週発表されているが, データ収集に時間がかかるため, 時刻 t の流行データがその 1 週間後に公開される. また, IDWR の感染症発生動向調査データのうち, 表 1 に記載した感染症を対象とした.

検索ログデータ: 検索ログについては, Yahoo!Japan 検索エンジンの検索ログデータを用いる. 現在の Yahoo!JAPAN では月間のアクティブユーザー ID 数が 2,700 万 ID 以上であり, 日本人口の 1/5 も占めている. 従って, 国民がどのようなことに関心を持っているかリアルタイムで観察することができる. また, 今回利用する Yahoo!Japan のデータは 2007 年から 2016 年まで, 全ユーザの検索ログである. データは日毎に集計され, その日に検索された全ての検索語と各検索語の検索回数が格納されるデータである. 本研究では検索回数でなく, 検索割合を用い, 時刻 t での検索割合を x_t^k と表す. ここで, 検索割合 x_t^k は, 時刻 t における検索語 q^k における q^k の検索回数を t における全検索語の検索回数で正規化した値である.

学習データとテストデータ: 検索語 q^k と感染症報告データからなるデータセットを改めて $D^k = (x_t^k, y_t)$ とする. このデータを 4 : 1 の割合で学習用データ D_{tr}^k とテストデータ D_{te}^k に分ける. D_{tr}^k は特徴選択と学習に用いられ, D_{te}^k は最終モデルの性能評価のみに用いられる.

3. 検索ログを用いる感染症流行の予測問題

次に, 本研究が扱う感染症の流行予測を定式化する.

目的変数: 予測したい目的変数は時間 t における感染率 $y_t \in [0,1]$ である. そして感染率の時系列を $Y = y_1, y_2, \ldots, y_T$ とする. 感染率 y_t は時刻 t に IDWR データで報告される患者数をその時点の人口で正規化した値である.

説明変数: 予測に使う説明変数として, N 個の検索語 $q^1, q^2, \ldots, q^N$ に関する N 個の時系列 $X = \{X^1, X^2, \ldots, X^N\}$ を用いる. ここで, $X^k = \{x_1^k, x_2^k, \ldots, x_T^k\}$ は検索語 q^k の時系列である. ただし, $x_t^k \in [0,1]$ は時刻 t において, 検索語 q^k の検索割合である. また, 検索語 q^k に関する時刻 t までの時系列を $X_{1:t}^k$ とし, $X_{1:t} = \{X_{1:t}^1, X_{1:t}^2, \ldots, X_{1:t}^N\}$ とする.

予測問題: 予測問題は関する時刻 t までの検索ログ $X_{1:t}$ を入力として時刻 $t+\phi$ の感染率 $y_{t+\phi}$ を出力する関数 f を学習する問題である. $\phi = 0$ のとき, 予測問題は将来の流行ではなく, 現在の流行を予測することになる. 病院から感染症データを集計する場合, およそ 1 週間の遅れが生じるのに対して, 検索ログを用いる場合ほぼリアルタイムな予測が期待できる.

4. 既存手法

次に, J. Ginsberg et al. の手法 [3] を説明する. 彼らの手法は次のような 3 ステップから構成される.

4.1 特徴量の順位づけ

スコアリング関数: 最初に, 彼らは予測問題に対して各検索語の適性を個別に評価した. 具体的に, N 個の検索語について, 次のロジット回帰モデルを用いて N 個の関数 $f^k : X^k_{1:t} \to y_{t+\phi}$ を D^k_{tr} で個別に学習を行う.

$$logit(y_t) = logit(f^k(x^k_t)) = \beta^k_0 + \beta^k_1 * logit(X^k_t) + e_t \quad (1)$$

ここで, β_0 はバイアス項で, β_1 は係数, e_t は誤差項である. そして, ロジット関数 $logit(p)$ はロジスティック関数の逆関数で, $logit(p) = \log(\frac{p}{1-p})$ で表される. 彼らは 5-fold 交差検証を用いて, 各検証 fold で予測した値と正解データとの平均 Pearson 相関係数で各特徴 (つまり検索語) をスコアリングした.

順位づけ: ここで, 検索語 q^k の X^k と感染症の Y で計算されたスコアリング関数を $Score_{GG}(X^k)$ と表記する. 彼らは次に全検索語をスコアリング関数 $S_{GG}(X^k, Y)$ で降順に並べ替える. 並べ替えた検索語リストを $q^{k_1}, q^{k_2}, \ldots, q^{k_N}$ と表記する.

4.2 特徴量の部分集合選択

順位づけられた検索語リストについて, 最適な検索語の部分集合を探索する. 彼らが用いた選択方法は以下の式で表すことができる.

$$\underset{\mathbf{\Omega} \subset \{k_1, k_2, \ldots, k_M\}}{\operatorname{argmax}} \; Score_{GG}(\sum_{k \in \mathbf{\Omega}} X^k) \quad (2)$$

ここで, $\mathbf{\Omega}$ は検索語のインデックスの集合で, $\sum_{k \in \mathbf{\Omega}} X^k$ は, $\mathbf{\Omega}$ に対応する検索語の時系列の総和を値とする時系列である. この最適化問題は検索語数について指数時間を要するために, 以下のヒューリスティックを用いた. 既存手法は順番に $k_1, k_2, \ldots$ を $\mathbf{\Omega}$ に追加し, 選択された検索語の総和で新しい時系列を作り, 最もスコアリング関数 $S_{GG}(\sum_{k \in \mathbf{\Omega}} X^k)$ の値が高い組み合わせを選択する. 最終的に得られた最適な時系列を X^{best} とする.

4.3 最終的な予測モデル

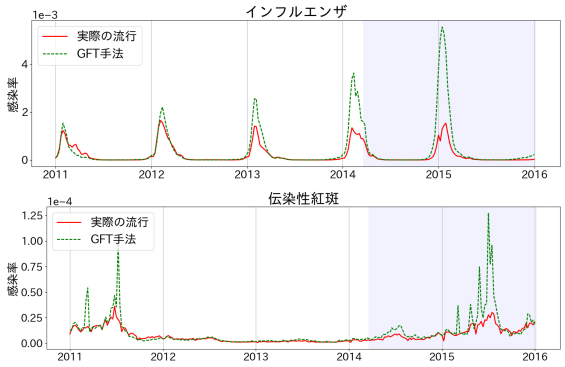
最終的に, 最適な時系列 X^{best} を入力とする関する $f^{final} : \mathbb{R} \to \mathbb{R}$ を学習データを用いて学習する. 時間 t における感染症流行は $y_t = f^{final}(x^{best}_t)$ で予測される.

4.4 GFT 手法の問題点

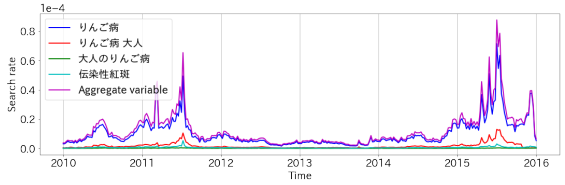
次に, 6.3 節で示される実験結果を用いて GFT 手法の問題点を説明する.

4.4.1 短期的と長期的な変化に対する誤予測

まず, 図 1(a) から, GFT 手法が短期的と長期的な変化に対応できていないことがわかる. 図 1(a) の上図では, 予測が流行の全シーズンにおいて大きく間違う予測を示した. このような間違いの原因は検索エンジンのユーザの変化などの長期的な変化に対応できないからだと考えられる. そして, 図 1(a) の下図から, GFT 手法は短期的な間違いも起こしていることがわかる. この種類の間違いの原因は情報検索行動がニュースなどの環境要因に影響されやすいからだと考えられる.



(a) GFT 手法の種類の誤予測. 青い領域がテスト期間である



(b) ある検索語による支配

図 1: GFT 手法の 2 つの問題点

このような誤予測の原因が異なるため, それぞれ独立に対応することが望ましい. 従って, 我々は季節調整を用いて, 各時系列を趨勢変動成分, 季節変動成分, 不規則変動成分の 3 つの成分に分解する. ここで, 不規則変動成分と趨勢変動成分はそれぞれ短期的と長期的な変化に対応すると考えられる.

4.4.2 1 次変数ロジット回帰モデルの弱点

GFT 手法のもう一つの問題点は選択された特徴量の総和を最終的な特徴量とするところにある. まず, このやり方は各検索語の変動を消すため, 予測に有効な情報が失われることがある. また, 総和をとることで, 検索率が高い検索語が他の検索語の影響を圧倒し, 最終的な特徴量を支配することがある. このような支配的な検索語が誤予測を引き起こす原因となりうる. なぜなら, よく検索される検索語は感染してないユーザでも検索する可能性が高いからである. また, このような支配的な検索語はニュースのような環境要因に影響を受けることで発生する可能性がある.

図 1(b) では, 支配的な検索語が存在し, その検索語が誤予測をもたらす例を示している. 具体的に, 図 1(b) は伝染性紅斑の流行予測において, GFT 手法が選択した 4 つの検索語とこれらの合計時系列を示している. 図から, 「りんご病」という検索語が他の検索語と比べて検索率が圧倒的に高いことがわかる. さらに, 図 1(a) の実際の感染率と比較すると, 「りんご病」そして合計時系列には 2011 年前半と 2015 の夏の期間において異常な短期的な変動が見られる. そして, これらの異常な変動は GFT の予測結果にも反映される. 要するに, 支配的な検索語「りんご病」が誤予測の主な原因と考えられる.

本研究は, 多重線形性の問題を考慮しつつより高い表現力を持つ予測モデルとして, L2 正則化付き多変量線形回帰を用いる.

5. 提案手法

本章では, 提案手法の全体的なフレームワークと詳細を説明す

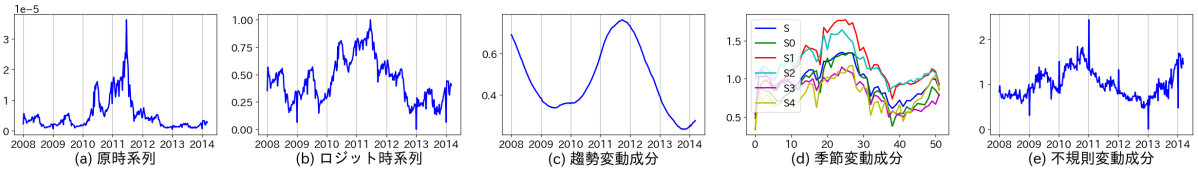


図 2: 手足口病の流行率の時系列における分解結果

る. まず, 我々の手法の全体的なフレームワークは以下である.

（1） **季節調整:** 我々は感染率のロジット時系列 $logit(Y)$ を趨勢変動成分 $\hat{T}$, 季節変動成分 $\hat{S}$, 不規則変動成分 $\hat{I}$ に分解する. そして同様に, 候補の 250 万個の検索語のロジット時系列を分解する. 具体的に, 検索語 q^k の場合, 我々は $logit(X^k)$ を $\hat{T}^k$, $\hat{S}^k$ と $\hat{I}^k$ に分解する.

（2） **特徴量の順位づけ:** 次に, 各検索語の趨勢変動成分 $\hat{T}^k$ と不規則変動成分 $\hat{I}^k$ を独立に評価してスコアリングする. そして, 全検索語をこれらのスコアで降順に並び替える.

（3） **特徴量の部分集合選択:** 並べ替えた検索語リストから, 特徴選択の Wrapper 法を用いて, 趨勢変動成分 $\hat{T}$ と不規則変動成分 $\hat{I}$ の各成分の予測に最適な検索語の組み合わせを探索する. 候補の選択方法では Forward selection 法を用い, 予測モデルでは L2 正則化付き多変量線形回帰を用いる.

（4） **予測モデル:** 最後に, 我々は選択された検索語を用いて, 趨勢変動成分 $\hat{T}$ について $f[T]$, 不規則変動成分 $\hat{I}$ について $f_{[I]}$ を学習する.

5.1 季節調整

5.1.1 時系列分解モデル

季節調整とは季節性のある時系列に対して, その季節的な成分を除く手法である. 時系列の分解手法は多数存在するが, 本研究は膨大な数の検索語を処理する必要があるため, 最も簡単な季節調整手法を用いて乗法モデル $y_t = \hat{T}_t * \hat{S}_t * \hat{E}_t$ で原系列 y を趨勢変動成分 (Trend) $\hat{T}$, 季節変動成分 (Seasonal) $\hat{S}$ と不規則変動成分 (Irregular) $\hat{I}$ に分解する. 次に, 各成分の解釈と計算方法を説明する.

5.1.2 趨勢変動成分 (Trend) $\hat{T}_t$

趨勢変動成分は時系列の長期的な変化を表す. 趨勢変動成分の計算は移動平均を用いて計算することができる. 本研究で扱っている感染症の多くは 1 年間の周期を持つため, 幅が 52 の移動平均 MA_{52} を用いる（図 2c 参照）. ただし, 移動平均を用いる場合, データの最初の 1 年間が利用できなくなることに注意されたい.

$$\hat{T}_t = \mathrm{MA}_{52}(logit(y_t)) = \frac{1}{52}\sum_{i=t-52}^{52} y_i \tag{3}$$

5.1.3 季節変動成分 (Seasonal) $\hat{S}_t$

季節変動成分 $\hat{S}$ は各周期のパターンを表す成分である. 季節変動成分 $\hat{S}$ は次のように計算できる. まず, 趨勢変動成分 $\hat{T}$ を除いた時系列 $\hat{D}_t = \frac{y_t}{\hat{T}_t}$ (Detrended) を計算する. このとき, 季節変動成分 $\hat{S}$ は時系列 $\hat{D}$ の各周期の平均に等しい. 具体的には, 時間 t について, 周期に対する位置が同じデータポイントのインデックス集合を $I_t = \{\ldots, t-2m, t-m, t, t+m, t+2m, \ldots\}$

No.	感染症	季節性の強さ	2015 年の報告数
0	インフルエンザ	0.827	1,169,041
1	手足口病	0.925	381,720
2	水痘	0.943	77,614
3	伝染性紅斑	0.764	98,521
4	咽頭結膜熱	0.888	72,150
5	ヘルパンギーナ	0.984	98,212
6	A群溶血性レンサ球菌咽頭炎	0.954	401,274
7	流行性角結膜炎	0.651	98,212
8	感染性胃腸炎	0.938	987,912
9	マイコプラズマ肺炎	0.339	10,384

表 1: 各感染症の 2015 年の報告数と季節性の強さ

とし, 季節変動成分 $\hat{S}$ は次のように計算できる. (図 2d 参照)

$$\hat{S}_t = \frac{1}{|I_t|}\sum_{i\in I_t} \hat{D}_i \tag{4}$$

我々は, この季節変動成分が短時間では大きく変化しないと仮定する. そのため, 季節変動成分について予測を行うのではなく, 学習期間の感染データを用いて計算した季節変動成分を予測モデルに用いる.

表 1 は本研究が対象とする感染症と各感染症の季節性を示している. 具体的に, 我々は 2007 年から 2013 年までのデータを用いて, 連続する 3 年間の 5 組 {2007,2008,2009},{2008,2009,2010}...{2011,2012,2013}を作り, 各 i 番目の組みに対応する部分時系列を $Y^{[i]}$ の季節変動成分 $\hat{S}^i$ を計算する. 最後に, 各組みの季節変動成分の Pearson 相関の平均で評価する. 平均相関が高いほど, その感染症が強い季節性を持つと考えられる. また, 図 2d では, 伝染性紅斑の流行時系列の各部分時系列の季節変動成分が示される.

5.1.4 不規則変動成分 (Irregular) $\hat{I}_t$

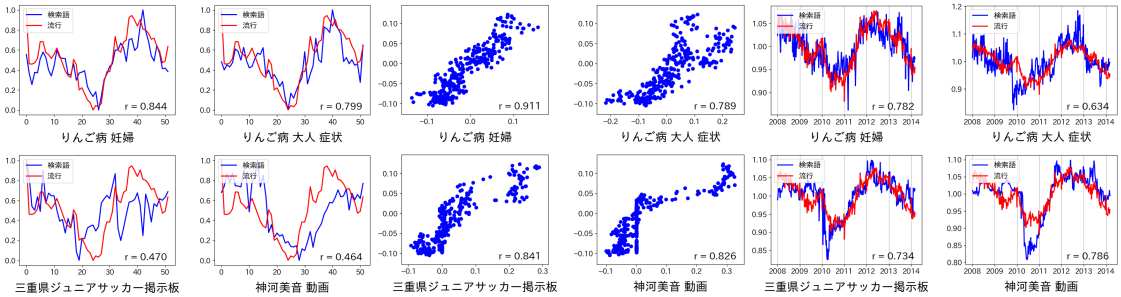
不規則変動成分 $\hat{I}_t$ は趨勢変動成分 $\hat{T}_t$ と季節変動成分 $\hat{S}_t$ で表現できなかった成分を表す. 要するに, 不規則変動成分は時系列の短期的な変動を表す成分である. 不規則変動成分は次のように計算できる.

$$\hat{I}_t = \frac{y_t}{\hat{T}_t * \hat{S}_t} \tag{5}$$

5.1.5 特徴量の順位づけ

スコアリング関数: 検索語 q^k の季節変動成分 $\hat{S}^k$, 趨勢変動成分 $\hat{T}^k$, 不規則変動成分 $\hat{I}^k$ のスコアリング関数 $Score_{[S]}(q^k), Score_{[T]}(q^k), S_{[I]}(q^k)$ を次のように定義する. ただし, $cor(X,Y)$ 時系列 X と Y の Pearson 相関係数を表すとする.

$$\begin{aligned} Score_{[S]}(q^k) &= max(cor(\hat{S}, \hat{S}^k), 0) \\ Score_{[T]}(q^k) &= \max_{\varepsilon=1}^{3} cor(\Delta_\varepsilon(\hat{T}), \Delta_\varepsilon(\hat{T}^k)) * Score_{[S]}(q^k) \\ Score_{[I]}(q^k) &= cor(\hat{I}, \hat{I}^k) * Score_{[S]}(q^k) \end{aligned} \tag{6}$$



(a) 季節変動成分: $\hat{S}$ and $\hat{S}^k$　　(b) 趨勢変動成分: $\Delta_1(\hat{T})$ and $\Delta_1(\hat{T}^k)$　　(c) 不規則変動成分: $\hat{I}$ and $\hat{I}^k$

図 3: 伝染性紅斑と 4 つの検索語の関係性. 各サブ図に相関 $cor(\hat{S}, \hat{S}^k)$, $cor(\Delta_1(\hat{T}), \Delta_1(\hat{T}^k))$ と $cor(\hat{I}, \hat{I}^k)$ が示される

ここで, 季節変動成分では感染率の $\hat{S}$ と検索語 q^k の $\hat{S}^k$ の Pearson 相関係数でスコアリングする. 趨勢変動成分では離れた 2 点の差を表す差分時系列 $\Delta_\varepsilon(Y) = \{y_t - y_{t-\varepsilon} | y_t, y_{t-\varepsilon} \in Y\}$ を用いる. 具体的に, 幾つかの異なる距離 ε の差分時系列 $\Delta_\varepsilon(\hat{T}), \Delta_\varepsilon(\hat{T}^k)$ の相関を計算し, 最も高い相関係数で評価する. そして, 不規則変動成分では感染率の $\hat{I}$ と検索語 q^k の $\hat{I}^k$ の相関で評価する.

また, 関連はないが偶然に趨勢変動成分 $\hat{T}$ または不規則変動成分 $\hat{I}$ の相関が高い検索語を除くため, $S_{[T]}(q^k), S_{[I]}(q^k)$ に $S_{[S]}(q^k)$ を掛けている. これらのスコアリング関数の有効性を示すために, 伝染性紅斑の場合における 4 つの検索語の例を図 3 で示す. 図 3 では関連する 2 つの検索語「りんご病　妊娠」と「りんご病　大人　症状」と, 無関係な 2 つの検索語「三重県ジュニアサッカー掲示板」と「神河美音 動画」について考察を行う. まず, 趨勢変動成分と不規則変動成分において, これらの無関係な検索語は関連する検索語に比べて, 同様またはより高い相関係数 $cor(\Delta_\varepsilon(\hat{T}), \Delta_\varepsilon(\hat{T}^k))$ と $cor(\hat{I}, \hat{I}^k)$ を達成している. しかし, これらの検索語が感染症に無関係であるため, 季節変動成分の相関が非常に低いことが見られる. 従って, 季節変動成分の相関を用いて, 無関係な検索語のスコア $cor(\Delta_\varepsilon(\hat{T}), \Delta_\varepsilon(\hat{T}^k))$ と $cor(\hat{I}, \hat{I}^k)$ を低下させることができる. つまり, これらのスコアリング関数は選択された検索語の関連性の保証を与えてくれる.

さらに, 流行予測問題において負の相関係数を持つ検索語が有効であることは考えにくいため, 負の値同士の積によって高い正のスコアが与えられることを防ぐために, 負の季節変動成分は 0 にする.

順位づけ:スコアリング関数 $S_{[T]}(q^k)$ で降順に並べ替えられた検索語リストを $Q_{[T]}$ とする. また, 同様にスコアリング関数 $S_{[I]}(q^k)$ で降順に並べ替えられた検索語のリストを $Q_{[I]}$ とする.

5.1.6　特徴量の部分集合選択

次に, 順位づけした検索語リスト $Q_{[T]}$ と $Q_{[I]}$ について最適な特徴の組み合わせを探索する. ただし, 本研究は $\hat{T}$ と $\hat{I}$ について同じ予測モデルと同じ特徴選択アルゴリズムを用いるため, ここでは趨勢変動成分だけについて説明をする, 要するに $Q_{[T]}$ を用いて $\hat{T}$ を予測する場合である. まず, 選択された検索語のインデックス集合を $\Omega_{[T]} = \{\omega_1, \omega_2, \ldots, \omega_p\}$ とし, それに対応する行列を $X^{\Omega_{[T]}} = (\hat{T}^{\omega_1}, \hat{T}^{w_2}, \ldots, \hat{T}^{w_p})$ とする.

予測モデルでは, 次の多変量回帰モデルを用いる.

$$\hat{T}_t = f(\hat{T}_t^{\Omega_{[T]}}) = \beta_0 + \sum_{i=1}^{p} \beta_i \hat{T}_t^{\omega_i} + e_t \tag{7}$$

また, 多重共線性を回避するために, L2 正則化を用いる. この時の損失関数は次のようになる.

$$L(\hat{T}^{\Omega_{[T]}}, \hat{T}) = \frac{1}{2} \sum_t \left(\hat{T}_t - f(\hat{T}_t^{\Omega_{[T]}})\right)^2 + \lambda \frac{1}{2} \sum_{j=1}^{p} \beta_j^2 \tag{8}$$

最後の検索語 q^{k_N} を選択し終わった時に得られるインデックス集合を Ω_N とする. 我々の手法は以下で説明される.

（1） **初期化:** 空集合 $\Omega_0 = \emptyset$ から始める.

（2） **探索方向:** 並び替えたリスト $Q_{[T]} = \{q^{k_1}, q^{k_2}, \ldots, q^{k_N}\}$ の検索語を順番に選択するかどうかを評価する.

（3） **特徴量部分集合の評価:** 5-fold 交差検証を用いて各部分集合を評価する. 具体的に, 学習 fold のデータを用いて多変量線形回帰モデルを学習し, 学習したモデルを用いて検証 fold を用いて予測する. ある部分集合のスコアは各検証 fold での平均二乗誤差で与えられる. また, インデックス集合 Ω のスコアを $WrapperScore_{\mathrm{PP}}(\Omega)$ と表記する.

（4） **特徴量部分集合の更新:** 次に検索語 $q^{k_{n+1}}$ を評価して Ω_{n+1} を求める場合を説明する. もし, $WrapperScore_{\mathrm{PP}}(\Omega_n \cap k_{n+1})$ が $WrapperScore_{\mathrm{PP}}(\Omega_n)$ よりも小さければ, $\Omega_n = \Omega_{n-1} \cap \{k_n\}$ と更新する. そうでなければ $\Omega_n = \Omega_{n-1}$ と維持する.

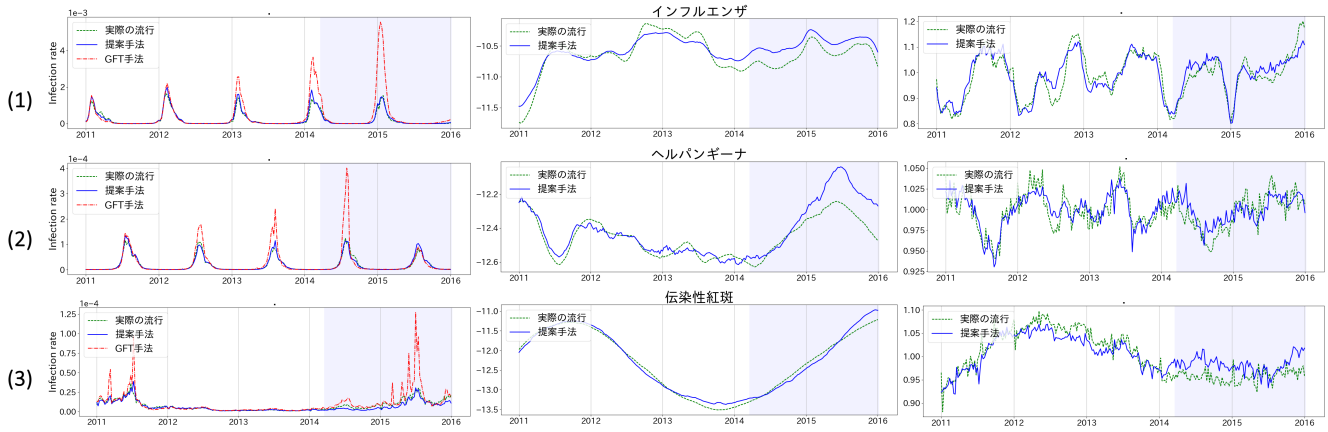
（5） **停止条件:** 更新が 5 回連続失敗すれば終了する.

5.2　提案モデル

得られた最適な特徴量集合をそれぞれ $\hat{T}^{\Omega_{[T]}}$ と $\hat{I}^{\Omega_{[I]}}$ とする. 最後に,$f_{[T]} : \mathbb{R}^{|\Omega[T]|} \to \mathbb{R}$ と $f_{[I]} : \mathbb{R}^{|\Omega[I]|} \to \mathbb{R}$ を学習データセットで再学習する. また, 季節変動成分は学習データ期間での感染症流行の季節変動成分を $\hat{S}$ を用いる. 従って, 時間 t における感染症流行は $y_t = logistic(\hat{S}_t * f_{[T]}(\hat{T}_t^{\Omega_{[T]}}) * f_{[I]}(\hat{I}_t^{\Omega_I}))$ で予測される.

5.3　将来予測への拡張

次に, $\phi \neq 0$ の将来予測問題への拡張を示す. つまり, $X_{1:t}$ を用いて $y_{t+\phi}$ を予測する問題を考える. 提案手法の順位づけでは次のスコアリング関数を用いる. 季節変動成分は $\phi = 0$ のスコアリング関数を用い, 趨勢変動成分と不規則変動成分は $\phi \neq 0$ 週間だけずらした時系列 $Y_\phi = \{y'_t | y'_{t+\phi} = y_t\}$ から計算された $\hat{T}_\phi$ と $\hat{I}_\phi$ でスコアリングする.



(a) 原時系列の予測　(b) 趨勢変動成分の予測　(c) 不規則変動成分の予測

図 4: インフルエンザ, ヘルパンギーナ, 伝染性紅斑における予測結果

$$
\begin{aligned}
Score_{[S]}(q^k) &= max(cor(\hat{S}, \hat{S}^k), 0) \\
Score_{[T]}(q^k) &= \max_{\varepsilon=1}^{3} cor(\Delta_\varepsilon(\hat{T_\phi}), \Delta_\varepsilon(\hat{T}^k)) * Score_{[S]}(q^k) \\
Score_{[I]}(q^k) &= cor(\hat{I_\phi}, \hat{I}^k) * Score_{[S]}(q^k)
\end{aligned}
\tag{9}
$$

また, 最終の予測モデルは次のように定義できる.

$$
y_{t+\phi} = logistic(\hat{S}_{\phi_t} * f_{[T]}(\hat{T_{\phi_t}}^{\Omega_{[T]}}) * f_{[I]}(\hat{I_{\phi_t}}^{\Omega_{[I]}})) \tag{10}
$$

また, GFT 手法のの将来予測への拡張として, 以下の予測モデルが考えられる.

$$
logit(y_{t+\phi}) = logit(f^k(x_t)) = \beta_0^k + \beta_1^k * logit(X_t^k) + e_t \tag{11}
$$

6. 実　　験

6.1 実験データ

IDWR データ: 我々は患者数が比較的に多い感染症を注目する. なぜなら, 患者数が少ないとその感染症の流行予測の手がかりが検索エンジンに現れないからだ. 具体的に, 我々は表 1 で示される 10 個の感染症を実験対象とする.

検索ログ: 検索ログデータについて, 本研究は最も検索される 2.5×10^6 個の検索語を対象として実験を行う. 2.5×10^6 番目の検索語はおよそ一日平均 450 回検索される.

データ期間: この実験では, 全データセットの期間は 2007 年から 2015 年までの 9 年間のデータを用いる. そして, 学習データセットは全データの 80%で, 2007 年 01 月 01 日から 2014 年 3 月 11 日までの 374 週間, である. また, テストデータは全データセットの 20%で 2014 年 03 月 12 日から 2015 年 12 月 31 日の 94 週間を用いる.

ロジット関数: 本研究で用いられたデータは $[0, 1]$ の範囲にある割合データである. このような割合データを扱うモデルとして, ロジスティック関数の逆関数であるロジット関数 $logit(x) = ln(\frac{x}{1-x})$ を用いて, $[0, 1]$ の割合データを実数領域 $[-\infty, \infty]$ のデータに変換する. また, ロジット関数は $x = 0$ において計算不可であるため, ロジット関数を適用する前に, 値が 0 であるデータ点に対して, その時系列の 0 でない最小の値を代用する.

6.2 実 験 設 定

我々は提案手法と GFT 手法を用いて 10 種類の感染症の流行を時間ラグ $\phi = \{0, 1, 2, 3, 4\}$ について実験を行う. 予測精度は次の平均絶対パーセント誤差 (MAPE) と相関係数を用いて評価することにする. ただし, Y は正解データで, Y' は予測の時系列である.

$$
\text{MAPE}(Y, Y') = \frac{100}{n} \sum_{t=1}^{T} \left| \frac{Y_t - Y'_t}{Y_t} \right|
$$

また, 公衆衛生では, 予測の正確性だけでなく, 予測の解釈性も求められるため、我々は二つの手法が選択した検索語の関連性を主観評価する. 具体的には, 我々は選択された検索語の中で, 対象の感染症と関連ある検索語を数え, それらの割合を算出する. また, 提案法の場合, この割合が趨勢変動成分と不規則変動成分の合計で計算される. 我々はこの割合を信頼度割合と呼ぶ.

6.3 実 験 結 果

6.3.1 予測精度は改善されたか?

表 2 は $\phi = 0$ のテスト期間での MAPE 値と相関係数を示している. 提案手法が全ての感染症においてより低い MAPE 値と高い相関を達成できた. 特に, インフルエンザの予測では提案手法が既存手法より 9 倍小さい MAPE 値を達成できた. また, ヘルパンギーナ, 流行性角結膜炎, 感染性胃腸炎でも約 2 倍位以上の差が見られた. 他に, 伝染性紅斑, 咽頭結膜炎, ヘルパンギーナ, 流行性角結膜炎, 感染性胃腸炎において, 相関係数が大幅に改善できた.

図 4a では, $\phi = 0$ の時の, インフルエンザ (1a), ヘルパンギーナ (2a), 伝染性紅斑 (3a) の予測結果を示している. 提案手法が既存手法に比べて, より正確かつ安定に予測ができることがわかる. 具体的にインフルエンザとヘルパンギーナ場合では, 既存手法が長期的な誤予測を起こし, 伝染性紅斑では短期的なが誤予測が見られる. それに対して, 提案手法の予測ではこのような間違いが見られない. また, 図 4b と 4c から, 提案手法が趨勢変動成分と不規則変動成分を予測できていることが分かる.

The MAPE value and correlation coefficient							Ratio of related search terms					
ϕ	0	1	2	3	4	BEST(ϕ)	ϕ	0	1	2	3	4
インフルエンザ												
GFT	345.07(0.95)	294.89(0.99)	129.66(0.93)	156.54(0.82)	251.24(0.70)	129.66(2)	GFT	100% (7/7)	100% (91/91)	100% (4/4)	100% (1/1)	15% (5/32)
Proposed	**38.07**(0.99)	**38.92**(0.98)	**67.06**(0.93)	**121.14**(0.89)	**108.12**(0.82)	**38.07(0)**	Proposed[T]	100% (28/28)	100% (8/8)	100% (13/13)	100% (2/2)	100% (5/5)
							Proposed[I]	100% (12/12)	100% (10/10)	100% (4/4)	100% (6/6)	100% (6/6)
手足口病												
GFT	42.50(0.98)	46.06(0.95)	67.46(0.56)	63.25(0.76)	**56.87**(0.77)	42.50(0)	GFT	100% (4/4)	50% (2/4)	0% (0/3)	0% (0/2)	0% (0/41)
Proposed	**33.25**(0.98)	**36.49**(0.99)	**37.94**(0.97)	**48.23**(0.94)	61.42(0.80)	**33.25(0)**	Proposed[T]	100% (6/6)	100% (3/3)	100% (3/3)	100% (2/2)	100% (6/6)
							Proposed[I]	100% (17/17)	80% (21/26)	100% (16/16)	77% (17/22)	26% (41/155)
水痘												
GFT	67.66(0.78)	96.57(0.40)	108.67(0.43)	136.89(0.44)	121.11(0.46)	67.66(0)	GFT	85% (6/7)	14% (7/47)	7% (5/71)	0% (0/16)	0% (0/45)
Proposed	**46.18**(0.77)	**90.05**(0.56)	**75.22**(0.61)	**82.83**(0.60)	**75.91**(0.63)	**46.18(0)**	Proposed[T]	30% (3/10)	18% (2/11)	33% (4/12)	20% (3/15)	9% (2/22)
							Proposed[I]	50% (4/8)	25% (3/12)	33% (3/9)	44% (4/9)	60% (3/5)
伝染性紅斑												
GFT	58.45(0.76)	**39.27**(0.84)	61.52(-0.25)	61.54(-0.25)	61.56(-0.25)	39.27(1)	GFT	62% (10/16)	47% (10/21)	37% (10/27)	9% (9/91)	11% (10/85)
Proposed	**35.35**(0.93)	55.47(0.93)	**35.05**(0.91)	**45.93**(0.91)	**52.77**(0.88)	**35.05(2)**	Proposed[T]	100% (2/2)	71% (5/7)	100% (2/2)	75% (6/8)	66% (6/9)
							Proposed[I]	88% (8/9)	80% (8/10)	88% (8/9)	88% (8/9)	80% (8/10)
咽頭結膜熱												
GFT	20.05(0.76)	22.58(0.71)	23.13(0.67)	20.54(0.77)	32.90(0.80)	20.05(0)	GFT	58% (10/17)	75% (3/4)	75% (3/4)	0% (0/11)	0% (0/14)
Proposed	**13.80**(0.90)	**12.48**(0.92)	**14.35**(0.91)	**13.62**(0.93)	**18.37**(0.91)	**12.48(1)**	Proposed[T]	66% (4/6)	50% (3/6)	33% (2/6)	66% (2/3)	66% (2/3)
							Proposed[I]	45% (5/11)	33% (4/12)	25% (5/20)	29% (5/17)	17% (10/57)
ヘルパンギーナ												
GFT	64.52(0.83)	58.68(0.92)	117.39(0.91)	340.75(0.87)	71.50(0.89)	58.68(1)	GFT	85% (6/7)	16% (2/12)	0% (0/29)	0% (0/3)	0% (0/67)
Proposed	**23.36**(0.97)	**48.92**(0.86)	**33.21**(0.96)	**26.87**(0.94)	**38.85**(0.83)	**23.36(0)**	Proposed[T]	100% (8/8)	100% (8/8)	100% (6/6)	100% (7/7)	100% (8/8)
							Proposed[I]	52% (11/21)	35% (25/70)	40% (23/57)	36% (16/44)	30% (22/71)
A群溶血性レンサ球菌咽頭炎												
GFT	26.31(0.83)	45.29(0.57)	41.82(0.56)	41.56(0.55)	**32.11**(0.61)	26.31(0)	GFT	100% (2/2)	3% (2/63)	2% (2/67)	4% (2/46)	3% (3/100)
Proposed	**18.36**(0.89)	**20.38**(0.90)	**32.58**(0.73)	**38.94**(0.70)	34.19(0.79)	**18.36(0)**	Proposed[T]	53% (7/13)	54% (6/11)	25% (6/24)	20% (3/15)	12% (3/24)
							Proposed[I]	17% (5/29)	6% (7/108)	14% (4/28)	6% (7/107)	4% (5/109)
流行性角結膜炎												
GFT	28.33(0.55)	20.38(0.33)	20.02(0.32)	20.52(0.32)	20.83(0.33)	20.02(2)	GFT	100% (2/2)	2% (1/41)	11% (4/36)	5% (2/36)	0% (0/25)
Proposed	**14.82**(0.83)	**20.01**(0.58)	**15.54**(0.73)	**18.10**(0.47)	**18.90**(0.40)	**14.82(0)**	Proposed[T]	60% (3/5)	100% (3/3)	100% (3/3)	40% (2/5)	33% (2/6)
							Proposed[I]	42% (8/19)	21% (4/19)	13% (2/15)	14% (1/7)	20% (1/5)
感染性胃腸炎												
GFT	45.60(0.87)	54.62(0.87)	15.55(0.86)	19.63(0.84)	20.45(0.84)	15.55(2)	GFT	100% (40/40)	100% (2/2)	50% (5/10)	0% (0/15)	0% (0/17)
Proposed	**14.41**(0.96)	**11.57**(0.95)	**14.75**(0.93)	**14.35**(0.91)	**15.70**(0.90)	**11.57(1)**	Proposed[T]	100% (13/13)	100% (16/16)	100% (18/18)	100% (9/9)	100% (4/4)
							Proposed[I]	100% (20/20)	100% (21/21)	90% (10/11)	74% (26/35)	60% (3/5)
マイコプラズマ肺炎												
GFT	55.99(0.88)	43.72(0.49)	50.95(0.37)	**43.44**(0.70)	59.12(0.03)	43.44(3)	GFT	26% (13/50)	16% (13/79)	10% (1/10)	0% (0/12)	0% (0/43)
Proposed	**29.76**(0.85)	**28.01**(0.70)	**43.02**(0.77)	43.94(0.76)	**42.23**(0.76)	**28.01(1)**	Proposed[T]	40% (2/5)	36% (4/11)	33% (1/3)	33% (1/3)	25% (1/4)
							Proposed[I]	12% (1/8)	5% (1/18)	3% (1/28)	25% (2/8)	14% (1/7)

表 2: (左) 二つの手法のテスト期間においける MAPE 値と相関係数 (括弧の中で示す). BEST(ϕ) 列は最適な MAPE 値とその時の ϕ を示す. (右) 選択された関連する検索語の数と割合

順位	ヘルパンギナー	インフルエンザ	伝染性紅斑	手足口病
1	0.83	0.95	0.91	0.93
2	0.68	0.94	0.85	0.90
3	0.63	0.94	0.91	0.86
4	0.62	0.93	0.95	0.85
5	0.63	0.93	0.93	0.88

表 3: 上位の 5 つの検索語の趨勢変動成分の相関

また, 表 2 から, 現在予測だけでなく, 将来予測において, 提案手法がより小さい MAPE 値を達成できたこと分かる. 特に, インフルエンザ, 水痘, 咽頭結膜熱, ヘルパンギナー, 流行性角結膜炎, 感染性胃腸炎において, 提案手法が全ての ϕ において MAPE 値が小さい. また, 最適の ϕ での MAPE 値予測精度についても, すべての感染症において提案手法が既存手法より優れる予測精度を示した.

6.3.2 最も予測が難しい成分はどれか?

図 4 (2b) では, ヘルパンギーナの趨勢変動成分の予測において, 提案手法がいくつかの間違いを起こしたことが分かる. その原因は表 3 を用いて説明することができる. この図ではヘルパンギーナと他の 3 つの感染症の趨勢変動成分のランキングにおいて上位 5 つの検索語とその相関係数 $\max_{\varepsilon=1}^{3} cor(\Delta_\varepsilon(\hat{T}_\phi), \Delta_\varepsilon(\hat{T}^k))$ を示している. 表から, 他の感染症と比べて, ヘルパンギーナの上位にある検索語の趨勢変動成分の相関が非常に低いことが分かる. 要するに, これらの検索語の趨勢変動成分の変化は実際の流行の趨勢変動成分の変化を反映していないことが分かる. そのため, これらを用いてヘルパンギーナの趨勢変動成分を予測することは難しいことが分かる.

ヘルパンギーナ以外の感染症の趨勢変動成分の予測においては提案手法がヘルパンギーナの場合より高い精度を達成したが, 不規則変動成分と比べて趨勢変動成分の予測が難しいことが見られた. また, 予測誤差が時間とともに大きくなる傾向もある. 具体的に, ほとんどの感染症において, テスト期間の 2 年目の趨勢変動成分の予測誤差が 1 年目より大きいことが見られた. それに対して, このような誤差の増加は不規則変動成分に見られなかった. また、本研究は性別, 年齢などのユーザ情報を使用していないが, このようなユーザ属性は趨勢変動成分の予測

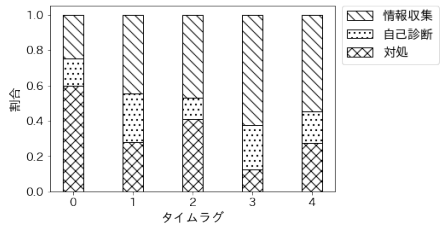


図 5: インフルエンザについて選択された検索語の構成

に有効だと考えられる.

6.3.3 関連する検索語を選択できるか?

表 2 左では, 選択された検索語の中で主観評価において関連すると評価された検索語の数と割合が示される. 例えば, 伝染性紅斑の $\phi = 0$ の時の GFT 手法の結果では, 16 検索語が選択され, その中関連する検索語は 10 個である. それらを「62%(10/16)」と示す. ただし, これの信頼度割合は利用する側がモデルの結果を信用するかどうかを決める際のみに有効であり, 予測制度に対する保証を与えるものではない.

まず全体的に, 提案手法が GFT 手法よりよい信頼度割合を達成できた. 特に, ϕ が大きい場合, 既存手法がほとんど関連する検索語が選択できなくなるのに対して, 提案手法が ϕ が大きい時でも関連する検索語を選択できる. また, 割合だけでなく, 提案手法はより多くの関連する検索語を選択できた.

しかし, 提案手法は水痘, 咽頭結膜炎, 流行性角結膜炎, マイコプラズマ肺炎などの不規則成分において, 多くの無関係な検索語を選択した. その原因は不規則成分が非常にノイジーであるからだと考えられる. また, これらの感染症では不規則成分が 1 の周辺をタイトに振動するため, 関連する検索語と無関係な検索語を区別することが難しい. ただし, このような不規則成分は単なるノイズと見なすことができ, 全体の予測問題にほとんど影響を及ぼさない. そのため, 提案手法による予測の誤差が許容される範囲に長あっていると考えられる.

6.3.4 どのような検索語が選択されたか?

最後に, 選択された検索語の分類について, インフルエンザに注目して考察を行う. 関連する検索語を自己診断検索語, 対処検索語, 情報収集検索語の 3 のカテゴリーに分類する. 具体的に, 自己診断検索語は「インフルエンザ 微熱」, 「インフルエンザ　頭痛」のような, 感染症な症状に関する検索語である. 次に, 対処検索語は「インフルエンザ 登校」や「インフルエンザ　解熱」のような治療や休暇に関する検索語である. このような検索語は, ユーザが実際に感染したときに検索されると考えれらる. 最後に, 情報収集検索語は「インフルエンザ 潜伏期間」や「インフルエンザ 感染力」のように, 一般の人が感染症について調べるときによく検索される検索語である. 流行予測では自己診断検索語と対処検索語のほうが有効だと考えられる. 図 5 は異なる ϕ において, 提案手法で選択された検索語の 3 カテゴリーの割合を示している. ただし, 各カテゴリーの検索語数は趨勢変動成分と不規則変動成分の合計で計算する. ϕ が大きいほど, 情報収集検索語の割合が大きくなることが分かる. 要するに, 予測に有効な検索語の割合が小さくなると考えられる. そして, 予測精度に関してもこの傾向に一致する傾向が見られた. しかし, 特徴選択プロセスのランダム性のため, このような考察が他の感染症にも成り立つとは限らない. また, 全体的にも提案手法のほうが対処検索語をより多く選択できた.

7. 結　　論

本研究は感染症の流行予測において, 季節調整を考慮するスケーラブルな特徴選択手法と分割可能な予測モデルを提案した. 各成分を独立に扱うことによって, より正確かつ検索ログの短期的な変化と長期的な変化に対して安定な予測モデルを達成できた. また, 提案手法は現在予測だけでなく, 将来予測にも有効である. そして, 10 種類の感染症について実験を行い, 提案手法が既存手法より優れることを示した. しかし, このようなタスクは検索ログデータやタスクに強く依存すると考えられる. 従って, 他の検索ログや別のタスクへの適用は微調整が必要だと考えられる.

謝　　辞

本研究は JST CREST および科学研究費 16H02864 の助成を受けました. また, 本研究はヤフー株式会社から提供頂いた検索ログデータを利用しました. ここに記して謝意を表します.

文　　献

[1] J. Paparrizos, R.W. White, E. Hovitz Detecting Devastating Diseases in Search Logs. *KDD*, 2016.

[2] S. Fox and M. Duggan Heath Online 2013. *Pew Internet and American Life Project*, 2013.

[3] J. Ginsberg et al. Detecting influenza epidemics using search engine query data. *Nature*, Vol. 457, 2009.

[4] J. Paparrizos, R. W. White, E. Hovitz Exploring time-dependent concerns about pregnancy and childbirth from search logs. *SIGCHI*, 737–746, 2015.

[5] R.W. White et al. Early identification of adverse drug reactions from search log data. *Journal of Biomedical Informatics*, Vol. 59, 42–48, 2016.

[6] E.Vayena et al. Ethical Challenges of Big Data in Public Health. *PLoS Computational Biology* 2015.

[7] National Institute Of Infectious Diseases (JAPAN) Infectious Diseases Weekly Report (IDWR) https://www.niid.go.jp/niid/en/idwr-e.html

[8] P.M. Polgreen et al. Using Internet Searches for Influenza Surveillance. *Clinical Infectious Diseases*, Vol. 47, Issue. 11, 1443–1448, 2008,

[9] P. Copelane et al. Google disease trends: An update *International Society of Neglected Tropical Diseases*, 2013.

[10] A.J. Ocampo et al. Using search queries for malaria surveillance, Thailand *Malaria Journal*, 2013.

[11] Y. Qingyu et al. Monitoring Influenza Epidemics in China with Search Query from Baidu *PLoS One*, 2013.

[12] D. Butler et al. When Google got flu wrong *Nature 494*, 155–156, 2013

[13] Olson D. et al. Reassessing Google Flu Trends Data for Detection of Seasonal and Pandemic Influenza: A Comparative Epidemiological Study at Three Geographic Scales *PLoS Computational Biology*, 2013

[14] Godman H. How long does the flu last? *Harvard Health Publishing*, 2016

[15] M. Santillana et al. What Can Digital Disease Detection Learn from (an External Revision to) Google Flu Trends? *American Journal of Preventive Medicine*, 2014

[16] R. Kohavi and G. John. Wrappers for feature selection *Aritificial Intelligence*, 97(1-2), 273-324, 1997