Seasonal-adjustment Based Feature Selection Method for Predicting Epidemic with Large-scale Search Engine Logs

Thien Q. Tran
thientquang@mdl.cs.tsukuba.ac.jp
University of Tsukuba
Tsukuba, Ibaraki, Japan

Jun Sakuma
jun@cs.tsukuba.ac.jp
University of Tsukuba, Riken AIP
Tsukuba, Ibaraki, Japan

ABSTRACT

Search engine logs have a great potential in tracking and predicting outbreaks of infectious disease. More precisely, one can use the search volume of some search terms to predict the infection rate of an infectious disease in nearly real-time. However, conducting accurate and stable prediction of outbreaks using search engine logs is a challenging task due to the following two-way instability characteristics of the search logs. First, the search volume of a search term may change irregularly in the short-term, for example, due to environmental factors such as the amount of media or news. Second, the search volume may also change in the long-term due to the demographic change of the search engine. That is to say, if a model is trained with such search logs with ignoring such characteristic, the resulting prediction would contain serious mispredictions when these changes occur.

In this work, we proposed a novel feature selection method to overcome this instability problem. In particular, we employ a seasonal-adjustment method that decomposes each time series into three components: seasonal, trend and irregular component and build prediction models for each component individually. We also carefully design a feature selection method to select proper search terms to predict each component. We conducted comprehensive experiments on ten different kinds of infectious diseases. The experimental results show that the proposed method outperforms all comparative methods in prediction accuracy for seven of ten diseases, in both now-casting and forecasting setting. Also, the proposed method is more successful in selecting search terms that are semantically related to target diseases.

KEYWORDS

Search log; feature selection; public health; seasonal adjustment

1 INTRODUCTION AND BACKGROUND

Search engine logs have a great potential in tracking and predicting infectious disease outbreaks [2]. Search logs are accessible in nearly real-time and can be used for detecting epidemic outbreaks more quickly than gathering data from hospitals nationwide, which often takes one or two weeks. A number of studies have been conducted using search logs of different search engines aiming at prediction of epidemic outbreak, such as Google [7][5][13], Yahoo [15] and Baidu [17]. These studies selected a number of search terms and then learned a prediction model in which the search rates of selected search terms were used as the features. Some of these works showed very promising results. For example, [7] reported a prediction accuracy with the correlation of 0.97 in the outbreak prediction problem for influenza.

Conducting accurate prediction using search engine logs is still a challenging task. The first difficulty is the instability of search log data. The search volume of a search term may change in both the short-term and the long-term due to many factors. First, information retrieval activities of users are very unstable and easily affected by environmental factors such as the amount of media or news. Thus, the search volume of a search term may irregularly fluctuate in the short-term [3]. Second, the search volume of a search term may also change gradually in the long run, for example, due to the demographic change of search engine [1]. Without considering such short-term and long-term changes, the resulting prediction would contain serious overestimation or underestimation.

Another difficulty is the computational hardness due to the enormously large number of search terms. For example, on Yahoo! Japan search engine, about 30 millions of different search terms are searched a day, on average. This makes the dimension of the explanatory variable very large. To reduce the learning cost and overfitting, one needs to conduct feature selection to select effective search terms for the prediction model [10].

Moreover, in the epidemic prediction problem, a reliable predictor not only needs to predict outbreaks accurately but also needs to produce the prediction based on meaningful evidence. More precisely, it is necessary to select search terms semantically related to the target infectious disease so that the resulting model is interpretable to experts. Finding such terms is difficult especially for infectious diseases with strong seasonality. Because the outbreak of these seasonal epidemics occurs at around the same time every year, some seasonal search terms not related to the target disease frequently co-occurs with the outbreak. While the total number of the search terms is large, only a small number of search terms are semantically related to the target disease and are buried in non-related search terms. This makes these related search terms extremely hard to find.

Finally, because search logs have a great potential in monitoring various epidemic diseases based on its diversity, methodologies that are applicable to a wide variety of infectious diseases are preferable. However, all existing works only focused on outbreak prediction of one or two major infectious diseases (e.g., influenza and dengue). This left the question of the applicability to other infectious diseases.

In breif, we aim to develop a feature selection method that is:

- **Stable**. The selected search terms must be robust against the instability of search log data, i.e, the changes in both short-term and long-term.
- **Scalable**. It must be implementable and be executable in a reasonable time for large-scale search log data, e.g, up to millions of search terms.
- **Interpretable**. It is desirable that the selected search terms are semantically related to the target infectious disease.
- **Universal**. It must be applicable to a wide range of infectious diseases.

In this work, we introduce a novel feature selection method that takes the instability characteristic of the search logs into account. Our method is built on a decomposition prediction framework. Our first insight is that because the short-term and the long-term changes have different occurrence dynamics, it is better to deal with these changes separately to conduct better prediction. Thus, we decompose these time serieses into three components: seasonal component, trend component and irregular component. Then, we individually train models to predict each component. The final prediction of the infection rate is obtained by combining the prediction of each component. Here, the trend and irregular component can be interpreted as the long-term and short-term change, respectively.

Moreover, we do not only train individual models but also conduct feature selection individually for each component. This is because useful search terms for prediction can vary for each component. Thus, we carefully design a method to select proper search terms that can appropriately predict each component. Furthermore, because we consider the seasonal component separately, our method can avoid selecting non-related seasonal search terms that are highly correlated with the outbreak. Hence, as we show with experimental results later, the resulting models built with our method are also more interpretable.

We note that, in the prediction problem, model selection is an orthogonal task to the feature selection problem. We emphasize that this work focuses on developing a feature selection method, and at some point, a prediction modeling framework for selected search terms. That is, various prediction models can be readily applied to our method. However, because most of the existing studies of search terms selection evaluated their methods with a linear prediction model, we only adopt a simple multivariate linear regression model as the prediction model for the evaluation purpose.

We conducted comprehensive experiments on predicting outbreaks of ten different kinds of infectious diseases to demonstate the universality of the proposed method. From the experimental results, we show that, the prediction model learned with the search terms selected by the proposed methods achieved more accurate and stable prediction compared to the existing methods. In particular, the proposed method achieved better prediction accuracy compared to all comparative methods for seven of ten diseases. For

example, in the case of herpangina, we achieved the correlation of 0.98, which is higher than the correlation coefficient of other methods by at least 12%. We also achieved a prediction error that is at least 2 times smaller than other methods. Moreover, our method outperformed the existing methods in both the now-casting and the forecasting setting. Furthermore, not only in prediction accuracy, our method was also more successful in selecting search terms that are semantically related to the target diseases. Especially, the proposed method was able to select related search terms even in the forecasting setting while the comparative methods almost failed.

2 RELATED WORKS

There are some prediction models have been proposed for the outbreak prediction problem. For example, the GFT method used a univariate logit regression model in which the single aggregate variable obtained by summing up selected search terms were used as the feature [7]. On the other hand, some works employed more representative models, for example, the multivariable regression model, and achieved a better prediction accuracy compared to the GFT model [5][13]. Besides, some studies also proposed using the non-linear frameworks, such as the Gaussian Processes (GPs)[13][14].

Several approaches to select the search terms has been proposed. Some works attempted to leverage prior knowledge about the target disease to select relevant search terms. For example, some studies preliminarily chose keywords related to the target disease and then used these keywords for filtering the search terms [17][15]. There is also an attempt that applies text processing technique to determine meaningful search terms [14]. One common point of these methods is that one has to preliminarily prepare keywords related to the target infectious disease manually. This task is often difficult because of the following two reasons. First, the search terms in search logs are extremely diverse and it is hard to cover all possible keywords (e.g, due to the user's diversity, synonyms, etc.). Secondly, finding related keywords could be hard for some infectious diseases or be costly when conducting for multiple diseases.

There also have been some efforts to automatically select useful search terms without prior knowledge. The Google Flu Trends method (the GFT method) used a correlation based method with the hybrid feature selection framework [7]. Specifically, they first learned a logit regression model for each search term and then ranked all the search terms by the correlation coefficients between the predicted value and the ground-truth. Finally, they select the best feature subset from the search terms with the top-rank scores and obtain a single aggregate variable by summing up the search rates of selected search terms. Despite the fact that the GFT method is scalable and selecting features using correlation coefficient is effective in collecting related search terms, [5] reported that the aggregated feature performs poorly in the prediction task.

Another approach is to use sparse modeling with the regularized regression schemes, e.g, the LASSO [16] method which uses the L1 regularization and the Elastic Net method which combines both the L1 and L2 regularization [18]. Based on the observation that many search terms have search rate time serieses that are strongly correlated to each other, [13] used the Elastic Net method because it is known to perform well in multicollinearity setting. The Elastic Net method can select useful search terms for prediction and can

learn a sparse model which is helpful in understanding the model. However, there is no guarantee that the selected search terms are related to the target disease. Also, learning with the Elastic Net method could be costly for large-scale search logs, e.g, due to the high optimization and hyperparameters tuning cost.

There are also studies focused on the seasonal characteristic of the infectious epidemic. In particular, [6] extended the GFT model by introducing a different ranking rule. To be specific, they first modeled the seasonal variation of the observed time series of the infection rate using the Serfling model and the yearly average model. After that, they ranked the search terms by the correlation of the search rate time serieses with the residual of the seasonal model and the observed infection rate. The final subset of search terms is selected from the ranked list. However, in their experiments because they used only 100 terms obtained from Google Correlate for evaluation, applicability to large-scale search logs is questionable. Our experimental result using their method with our large-scale dataset indicated that their method failed to select proper search terms due to the extreme diversity of the large-scale search logs.

Above all, there was no study that took changes in short-term and long-term in the search logs into account. Moreover, all existing works only focused on one or two specific infectious diseases. This left the question of the applicability to other infectious diseases. In conclusion, none of these methods fully achieved our goal.

3 DATA DESCRIPTION

In this section, we describe the data used in this work.

Epidemic report data: We use the publicly available Infectious Diseases Weekly Report (IDWR) data published by the National Institute of Infectious Diseases (Japan) as the ground-truth data for epidemic breakout [11]. IDWR data contains the number of the infection cases for all infectious diseases with the reporting requirement in Japan. We denote the infection rate of an infectious disease by y_t^d, where d is the disease name and t is a weekly time. Here, the infection rate y_t is the number of the infection cases reported in IDWR data normalized by the population of Japan at time t. While IDWR data is published every week, the data of the epidemic at time t is published one week later, i.e., at time $t + 1$ because it takes time for gathering the data from the hospitals. Among the infectious diseases reported in IDWR data, we focus on ten infectious diseases listed in Table 1. Here we chose ten common diseases which has a relatively large number of infected cases to make sure that desired evidences shows up in our data search logs. Targeting at ten diseases is significantly different from the existing works since they only focused on one or two major infectious diseases.

Search log data: For the search log data, we use the search log data of Yahoo! JAPAN search engine. Currently, Yahoo! JAPAN has approximately 120 million active users quarterly, occupying most of the population of Japan. Therefore, one can use this data to observe the interest of the people in Japan in nearly real-time. Yahoo! JAPAN search log data is collected daily, containing all the search terms in that day and the search volume of each search term. In our work, instead of the search volume, we use the search rate, which is obtained as the search volume of a search term normalized

by the search volume of all search terms at that moment. We denote the search rate of search term q^k at time t as x_t^k and the data of all search terms as X_t^k.

Training and test data: To simplify the notation, we focus on a specific target disease and omit the subscription d - disease name. We define the data set of all search terms and the epidemic data as $D = \{(X_t, y_t)\}$. We then set 80% of D for the training data and the remaining data is used for the test data. We used the training data for the feature selection and model learning. The test data is used only for evaluating the performance of the final model.

4 PROBLEM FORMULATION

In this section, we will formulate the epidemic prediction problem using the search log data.

Objective variable: The value to predict is the infection rate $y_t \in [0, 1]$ at time t. We describe the time series of infection rate (ground-truth) as $Y = \{y_1, y_2, \ldots, y_T\}$ where y_t denotes the infection rate y_t at time t.

Explanatory variable: We use N time serieses $X = \{X^1, X^2, \ldots, X^N\}$ of N search terms $q^1, q^2, \ldots, q^N$ as the explanatory variables. Here, $X^k = \{x_1^k, x_2^k, \ldots, x_T^k\}$ is the time series of search term q^k, and $x_t^k \in [0, 1]$ is the search rate of search term q^k at time t. We also denote the search rate of search term q^k from time 1 to time t as $X_{1:t}^k$ and $X_{1:t} = \{X_{1:t}^1, X_{1:t}^2, \ldots, X_{1:t}^N\}$ for all the search terms.

Prediction problem: The prediction problem is the problem of learning a function f that takes the search log data available until time t, $X_{1:t}$ as the input and returns the infection rate $y_{t+\phi}$ at time $t + \phi$. When $\phi = 0$, we predict the current epidemic at that moment, rather than the future epidemic.

5 PROPOSED METHOD

Here, we discuss the overall framework and the details of our proposed method. First, our overall framework is described as follows:

(1) **Seasonal adjustment:** We decompose the logit time series $logit(Y)$ of the epidemic into three components: trend $\hat{T}$, seasonal $\hat{S}$, and irregular component $\hat{I}$. Similarly, we also decompose the logit time series of 2.5 million candidate search terms. To be specific, for the search term q^k, we decompose $logit(X^k)$ into $\hat{T}^k$, $\hat{S}^k$ and $\hat{I}^k$.

(2) **Feature ranking:** Next, we score the prediction capability of the trend component $\hat{T}^k$ and the irregular component $\hat{I}^k$ of each search term individually. We then sort all the search terms using these scores in descending order.

(3) **Feature subset selection:** From the sorted search term list, we find the optimal combination of search terms for predicting the trend component $\hat{T}$, and irregular component $\hat{I}$ using the wrapper approach for feature subset selection [12]. We apply the forward selection method to choose the candidate search terms and use the multivariate regression model with L2 regularization for the prediction model.

(4) **Prediction model:** Finally, we retrain the prediction model $f_{\hat{T}}$ for the trend component $\hat{T}$, and $f_{\hat{I}}$ for the irregular component $\hat{I}$ with the selected search terms.
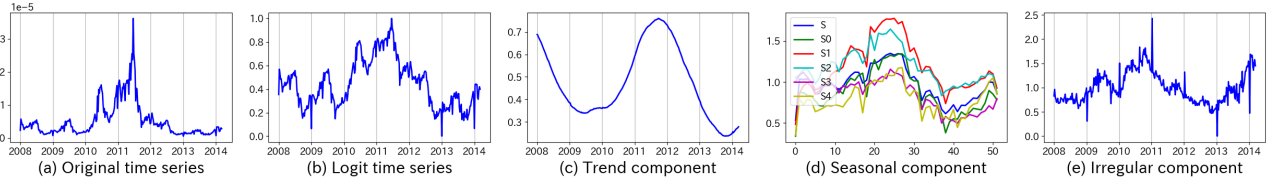
Figure 1: Decomposition result of the infection rate time series for erythema infectiosum

5.1 Seasonal adjustment

Seasonal adjustment is the method for excluding the seasonal component for seasonal time series. Considering that we need to process numerous search terms, we use a basic seasonal adjustment method with the multiplicative model $logit(y_t) = \hat{T}_t * \hat{S}_t * \hat{I}_t$ where the logit time series $logit(Y)$ is decomposed into three components: trend $\hat{T}$, seasonal $\hat{S}$, and irregular component $\hat{I}$. Similarly for the search term q^k, we decompose $logit(X^k)$ into $\hat{T}^k$, $\hat{S}^k$ and $\hat{I}^k$. Next, we describe the interpretations of these components in the epidemic prediction problem and their calculation methods.

As a running example, we use the infection rate of erythema infectiosum. We show the original time series and the logit time series of infection rate of erythema infectiosum in Figure 1(a) and Figure 1(b), respectively.

5.1.1 Trend component $\hat{T}$.

The trend component represents the variation of the outbreak in the entire year, i.e., the long-term change in the time series. The trend component is calculated by the moving average. Because most of the infectious diseases focused in our work have a one-year cycle, we use the moving average MA_{52} with the width of 52 weeks (see Figure 1c). Because we used moving average of the past 52 weeks, data in the first year becomes unusable

$$\hat{T}_t = MA_{52}(logit(y_t)) = \frac{1}{52} \sum_{i=t-52}^{52} logit(y_i). \tag{1}$$

5.1.2 Seasonal component $\hat{S}$.

The seasonal component $\hat{S}$ represents the common pattern of each cycle and is calculated as follows. First, we calculate the de-trended time series $\hat{D}_t = \frac{logit(y_t)}{\hat{T}_t}$ by removing the trend component from the original time series. The seasonal component $\hat{S}$ is calculated as the average of all cycles of $\hat{D}$. Specifically, if we denote the set of indexes of data points which have the same position with respect to the cycle as $I_t = \{\ldots, t-2m, t-m, t, t+m, t+2m, \ldots\}$, the seasonal component $\hat{S}$ is derived as follows (see Figure 1d)

$$\hat{S}_t = \frac{1}{|I_t|} \sum_{i \in I_t} \hat{D}_i. \tag{2}$$

We assume that the seasonal component does not change in a short time. For this reason, we do not build a prediction model for the seasonal component, but instead, use the seasonal pattern $\hat{S}$ calculated by the epidemic data in the training period for prediction.

5.1.3 Irregular component $\hat{I}$.

The irregular component $\hat{I}_t$ represents the component that could not be expressed by the trend component $\hat{T}_t$ and the seasonal component $\hat{S}_t$. In other words, the irregular component represents the short-term change in the time series. The irregular component can be calculated as follows:

$$\hat{I}_t = \frac{logit(y_t)}{\hat{T}_t * \hat{S}_t}. \tag{3}$$

We illustrate the irregular component for the case of erythema infectiosum in Figure 1e.

5.2 Feature ranking

Scoring function: We define the scoring functions of search terms with respect to the seasonal component $\hat{S}^k$, trend component $\hat{T}^k$ and irregular component $\hat{I}^k$ as $Score_{[S]}(q^k)$, $Score_{[T]}(q^k)$ and $Score_{[I]}(q^k)$, respectively. Let $cor(X, Y)$ be the Pearson correlation between X and Y. Our scoring functions are defined as follows:

$$Score_{[S]}(q^k) = max(cor(\hat{S}, \hat{S}^k), 0), \tag{4}$$

$$Score_{[T]}(q^k) = \max_{\varepsilon=1}^{3} cor(\Delta_\varepsilon(\hat{T}), \Delta_\varepsilon(\hat{T}^k)) * Score_{[S]}(q^k), \tag{5}$$

$$Score_{[I]}(q^k) = cor(\hat{I}, \hat{I}^k) * Score_{[S]}(q^k). \tag{6}$$

We score q_k with respect to the seasonal component by the Pearson correlation between $\hat{S}$ of infection rate and $\hat{S}^k$ of the search term q^k. To evaluate the trend component, we use the first-order differential time series $\Delta_\varepsilon(Y) = \{y_{t+\varepsilon} - y_t\}_{t=1}^{T-\varepsilon}$ which is obtained by the difference of two data points. Here ε is the distance between two data points. We calculate the correlation coefficients between the two differential time series $\Delta_\varepsilon(\hat{T})$ and $\Delta_\varepsilon(\hat{T}^k)$ using several different ε, and then adopt the highest correlation coefficient among all ε. The irregular component is evaluated by the correlation coefficient between the irregular component $\hat{I}$ of the infection rate and $\hat{I}^k$ of the search term q^k.

Moreover, we multiply the seasonal score $Score_{[S]}(q^k)$ to the scoring function $Score_{[T]}(q^k)$ and $Score_{[I]}(q^k)$ to exclude non-related search terms that coincidentally have high correlation with $\hat{T}$ and $\hat{I}$. In Figure 2, we use an example about four search terms in the case of erythema infectiosum to show that multiplying the seasonal score works appropriately to select related search terms. To be specific, we consider two related search terms 'fifth disease pregnant' and 'fifth disease adult symptoms'. For non-related search terms, we consider 'Mie prefecture juniours soccer' and 'Mion Kamikawa video' ('Mion Kamikawa' is an actress name). We see that, compared to the related ones, these non-related search terms still achieve a similar or even higher correlation coefficients in the trend and irregular component, i.e., $cor(\Delta_\varepsilon(\hat{T}), \Delta_\varepsilon(\hat{T}^k))$ and $cor(\hat{I}, \hat{I}^k)$. However, because these search terms are non-related to the disease, their seasonal scores are significantly lower than the related ones. Thus, we can use the seasonal score $Score_{[S]}(q^k)$ to diminish the final score $Score_{[T]}(q^k)$ and $Score_{[I]}(q^k)$ of these non-related search terms. In

(a) Season component: $\hat{S}$ and $\hat{S}^k$ (b) Trend component: $\Delta_1(\hat{T})$ and $\Delta_1(\hat{T}^k)$ (c) Irregular component: $\hat{I}$ and $\hat{I}^k$
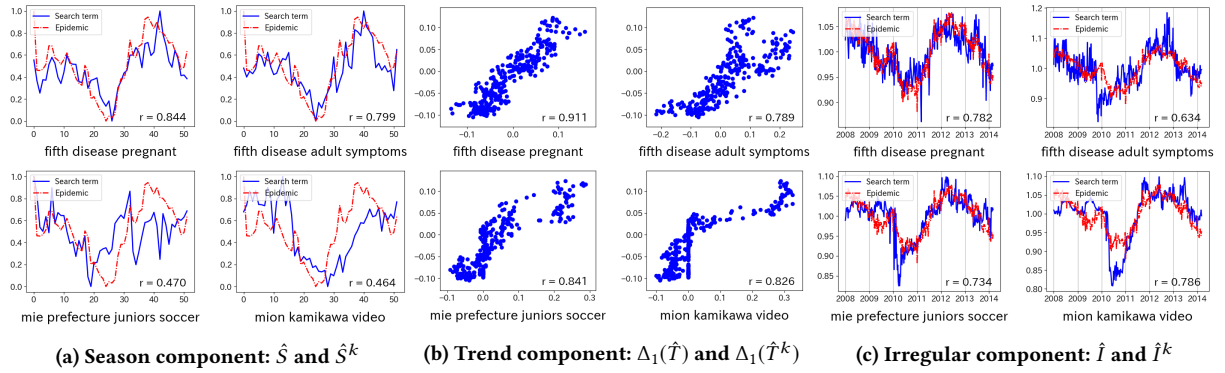
Figure 2: The relationship between the epidemic of erythema infectiosum and four search terms for three components. We also annotate the Pearson correlation $cor(\hat{S}, \hat{S}^k)$, $cor(\Delta_1(\hat{T}), \Delta_1(\hat{T}^k))$ and $cor(\hat{I}, \hat{I}^k)$ in each subfigure. The top row and bottom row present results of search terms semantically related and non related to the target disease, respectively.

conclusion, our scoring functions are expected to guarantee the meaningfulness of selected search terms for the prediction model.

Furthermore, because the search terms with negative correlation are not proper for epidemic prediction apparently, we set the negative scores of the seasonal component to 0 to get rid of the case that a positive score is given by the product of two negatives.

Sorting: For the trend component, we sort the search term list in descending order by the scoring function $Score_{[T]}(q^k)$ and denote the sorted list as $Q_{[T]}$. Similarly, for the irregular component, we sort the search term list in descending order by the scoring function $Score_{[I]}(q^k)$ and denote the sorted list as $Q_{[I]}$.

5.3 Feature subset selection

Next, we find the optimal feature subset from the sorted lists $Q_{[T]}$ and $Q_{[I]}$. Because we used the same feature selection algorithm and prediction model for both the trend and irregular component, here we only describe the proposed method for the case of the trend component, i.e., using $Q_{[T]}$ to predict $\hat{T}$. First, we denote the index set of selected search terms as $\Omega_{[T]} = \{\omega_1, \omega_2, \ldots, \omega_p\}$ and the corresponding input matrix as $\hat{T}^{\Omega_{[T]}} = (\hat{T}^{\omega_1}, \hat{T}^{\omega_2}, \ldots, \hat{T}^{\omega_p})$.

We use the multivariate regression for the prediction model:

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

To overcome the multicollinearity problem, we learn this model with the L2 regularization. The quadratic loss function with L2 regularization is defined as follows:

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

We denote Ω_n as the index set of the selected feature subset when finished evaluating the search term q^{k_n}. We described our selection method as follows:

(1) **Initiallization:** We start with the empty set $\Omega_0 = \emptyset$.
(2) **Searching direction:** We evaluate each search term in the sorted list $Q_{[T]} = \{q^{k_1}, q^{k_2}, \ldots, q^{k_N}\}$ in order and decide whether to add them to the prediction model.

(3) **Feature subset evaluating:** We evaluate the prediction capability of a feature subset with the 5-fold cross-validation. To be specific, we learn a multivariate regression model using the training folds and use it to predict the trend component with the validation fold. The score of the feature subset will be obtained by the averaging the MSE values on five validation folds. We denote the score of the index set Ω as $WrapperScore_{PP}(\Omega)$.
(4) **Feature subset updating:** Consider the case that we have to evaluate the next candidate search term $q^{k_{n+1}}$ to determine Ω_{n+1}. If $WrapperScore_{PP}(\Omega_n \cap k_{n+1})$ is smaller than $WrapperScore_{PP}(\Omega_n)$, we update $\Omega_n = \Omega_{n-1} \cap \{k_n\}$. Otherwise, we maintain $\Omega_n = \Omega_{n-1}$.
(5) **Stop condition:** We stop adding features to Ω_n when the update fails five consecutive times.

5.4 Final prediction model

Let $\hat{T}^{\Omega_{[T]}}$ and $\hat{I}^{\Omega_{[I]}}$ be the matrixes obtained for trend component and irregular component by feature selection. Finally, we learn two functions $f_{[T]} : \mathbb{R}^{|\Omega_{[T]}|} \to \mathbb{R}$ and $f_{[I]} : \mathbb{R}^{|\Omega_{[I]}|} \to \mathbb{R}$. Here $f_{[T]}$ takes $\hat{T}_t^{\Omega_{[T]}}$ as the input and returns $\hat{T}_t$. Similarly, $f_{[I]}$ takes $\hat{I}_t^{\Omega_{[I]}}$ as the input and returns $\hat{I}_t$. In other words, $f_{[T]}$ and $f_{[I]}$ predicts the trend and the irregular component, respectively. Moreover, we use the seasonal component $\hat{S}$ computed by the epidemic data of the training period as the seasonal component for the prediction model. Hence, the prediction value of infectious epidemic at t is given as $y_t = logistic(\hat{S}_t * f_{[T]}(\hat{T}_t^{\Omega_{[T]}}) * f_{[I]}(\hat{I}_t^{\Omega_{[I]}}))$.

5.5 Extension for future prediction

Next, we discuss the extension to forecasting problem when $\phi \neq 0$ where ϕ denotes the time-lag parameter. Here, we learn a function f that use the data $X_{1:t}$ to predict infectious rate at time $t + \phi$, $y_{t+\phi}$.

We use the following scoring functions for the feature selection process of the forecasting problem. For the seasonal component, we use the same scoring function as when $\phi = 0$. For the trend and irregular component, we first compute the time-lag time series for search rate $Y_\phi = \{y_t' | y_{t+\phi}' = y_t\}$ obtained by shifting original time series Y by ϕ weeks. Then, we use this time-lag time series Y_ϕ to

calculate the lagged trend $\hat{T}_\phi$, seasonal $\hat{S}_\phi$ and irregular component $\hat{I}_\phi$. We defined the scoring function as follows:

$$Score_{[S]}(q^k) = max(cor(\hat{S}, \hat{S}^k), 0), \tag{9}$$

$$Score_{[T]}(q^k) = \max_{\varepsilon=1}^{3} cor(\Delta_\varepsilon(\hat{T}_\phi), \Delta_\varepsilon(\hat{T}^k)) * Score_{[S]}(q^k), \tag{10}$$

$$Score_{[I]}(q^k) = cor(\hat{I}_\phi, \hat{I}^k) * Score_{[S]}(q^k). \tag{11}$$

The final prediction model is defined as follows:

$$y_{t+\phi} = logistic\big(\hat{S}_{\phi_t} * f_{[T]}(\hat{T}_t^{\Omega_{[T]}}) * f_{[I]}(\hat{I}_t^{\Omega_{[I]}})\big). \tag{12}$$

6 EXPERIMENT

In this section, we discuss the experiments and the results.

6.1 Experiment data

IDWR data: We consider infectious diseases that the number of patients is relatively large because the clues for predicting the epidemic will hardly appear in the search logs if the number of patients is small. In our experiments, we focus on the ten infectious diseases shown in Table 1.

Search log: For the search log data, we use the most frequently searched 2.5 million search terms for the experiment. nIndeed, the search term ranked 2.5×10^6th from the top is searched roughly 450 times a day, on average.

Data period: In our experiments, we use 9 years of data from 2007 to 2015. In addition, 80% of the dataset, which contains 374 weeks from January 01, 2007 to March 11, 2014, was used for the feature selection and training phase. The other 20% of the data set, which contains 94 weeks from March 12, 2014, to December 31, 2015, was used for evaluating the final model.

The logit function: The search rates and infection rates used in this work are normalized within the range of $[0, 1]$. We use the logit function $logit(x) = ln(\frac{x}{1-x})$ to map the rates in $[0, 1]$ to $[-\infty, \infty]$. Also, because the logit function cannot be calculated when $x = 0$, before applying the logit function, we replace all the data points with value 0 with the smallest non-zero value of that time series.

6.2 Experimental setting

be evaluate the performance of the proposed method and three comparative methods on the prediction of the outbreak of ten different types of infectious diseases, with different time-lag parameters $\phi \in \{0, 1, 2\}$.

To evaluate the prediction accuracy, we use the Symmetric Mean Absolute Percentage Error (sMAPE) [9] and the Pearson correlation.

Additionally, in the field of public health, the prediction is required to be made based on meaningful evidence. To evaluate the relevance of the search terms selection, we manually evaluate the interpretability of selected search terms. Specifically, we determined a search term as a related one if it contains either the target disease's name, symptoms or treatment. Moreover, we count the number of the semantically related health-care search terms listed by feature selection methods and then calculate the ratio of these related search terms to the total number of selected search terms. We call this ratio the relevance ratio of the selected search terms.

6.3 Comparative baseline

We compare our method with three baselines, the GFT method [7], the ElasticNet method [13], and the method of Neils et al. [6]. First, the GFT method is evaluated with our search log data of 2.5 million search terms. For the ElasticNet method, we conducted the learning process on top 1000 search terms ranked by the GFT method. The regularization parameters were chosen using 5-fold cross-validation. Finally, for the method of Neils et al., we first applied their ranking method to rank our 2.5 million search terms and then select the best feature subset from the top 100 search terms. For the forecasting setting, we shift the infection rate time series by ϕ week to obtain the time-lag time series and then apply the same algorithm as the now-casting setting.

6.4 Experiment results

6.4.1 *Was the prediction accuracy improved ?*

In the Table 1, we show the Pearson correlation values and the sMAPE values in the test period for each disease with different time-lag parameters ϕ.

We first explain the results of now-casting, i.e, when $\phi = 0$. We see that our proposed method achieved higher correlation coefficients compared to all the other methods for seven of ten diseases. For the remaining three infectious diseases: chickenpox, strep throat, mycoplasma pneumonia, the proposed method could not achieve the best correlation coefficient while it still attained reasonable prediction with correlation coefficients that are close to the best one of the comparative methods. In contrast, the GFT method and the ElasticNet method only achieved the best correlation coefficient for only two of ten diseases. A similar observation can also be obtained for the sMAPE value.

From Table 1, we also see that the proposed method also achieved lower sMAPE values and higher correlation coefficients for the forecast problem, i.e., when $\phi \neq 0$. When the time lag parameter becomes larger, the accuracy of the comparison methods decreases drastically. Though this decrease of accuracy is still observed for our method, the decreasing is much slower compared to the other methods. In particular, for the case of HFMD, Erythema infectiosum, PCF and Gastroenteritis, our proposed method achieved the highest correlation coefficient for all the time-lag parameters ϕ.

We show the prediction results when $\phi = 0$ for four different diseases: influenza (Figure 3a), PCF disease (Figure 3b), herpangina (Figure 3c). We see that, for all these diseases, the proposed method performed more accurately and stably compared to the other methods. To be specific, for the GFT method, we observed the long-term misprediction in the case of influenza and herpangina. We also observed the short-term misprediction in the case of PCF disease. Compared to the GFT method, the ElasticNet method attained more accurate predictions. However, we observed that ElasticNet still gives the long-term mispredictions for herpangina and PCF disease. In contrast, we did not observe such kinds of mispredictions in the predictions by the proposed method. In other words, the proposed method is significantly more stable than the existing methods. We also showed the trend component and irregular component prediction of these diseases in Figure 3a, 3b, and 3c. We see that our method succeeded in predicting both the trend and irregular component for these diseases.

(a) Influenza(The flu)
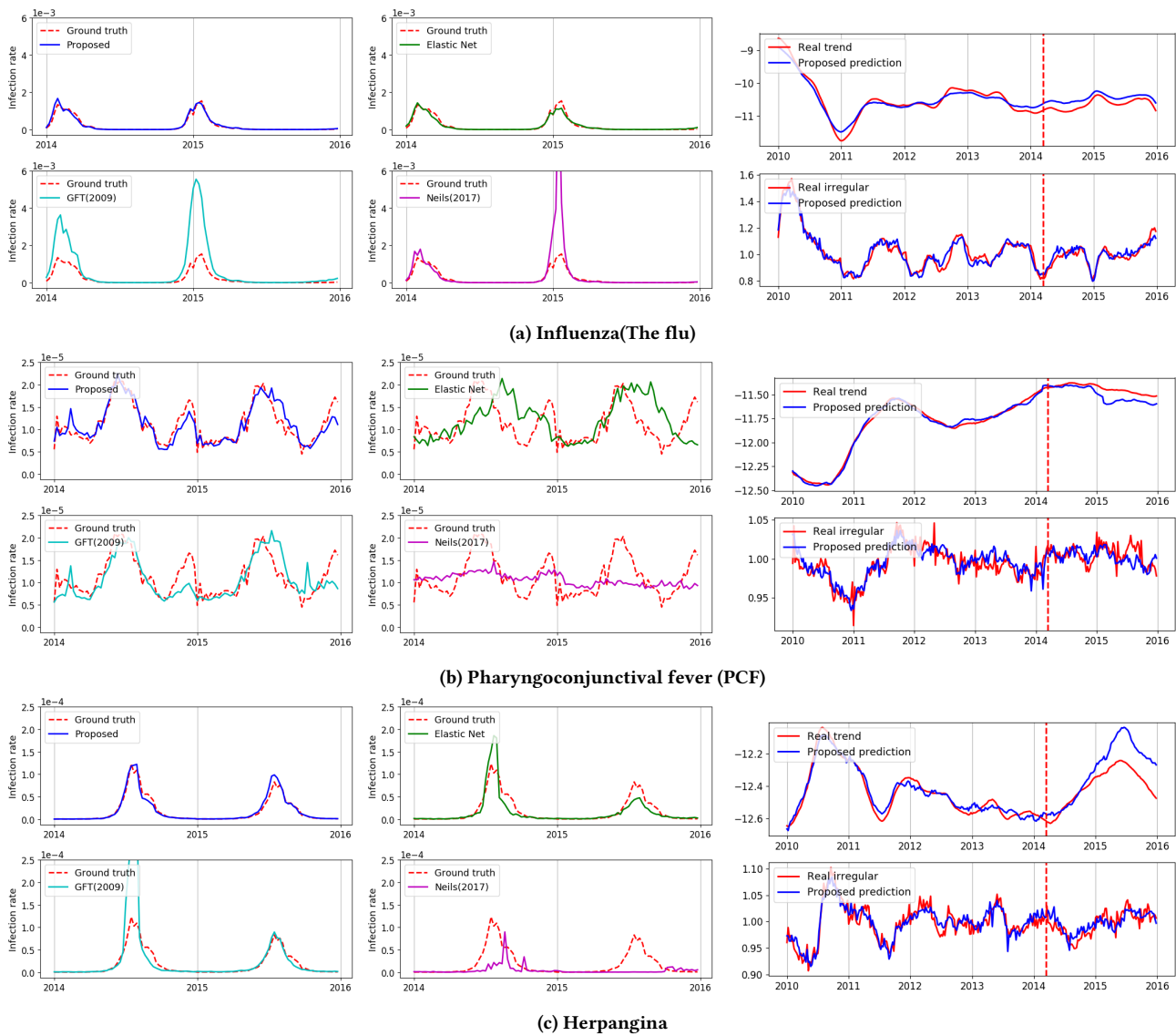
(b) Pharyngoconjunctival fever (PCF)

(c) Herpangina

Figure 3: (Column 1 and 2) The prediction by the proposed method and three comparative methods for four infectious diseases. (Column 3) The prediction for the trend and irregular component by our method. The red vertical line indicates the test period.

6.4.2 *Which component is harder to predict?*

From Figure 3c, we see that the proposed method showed some overestimation and underestimation in predicting the trend of herpangina in the test period. We show Table 2 to discuss this failure. In this table, we show the correlations of the top-5 search terms of the trend component in the case of four diseases, including herpangina. We observed that the correlation coefficients of the trend component $\max_{\varepsilon=1}^{3} cor(\Delta_{\varepsilon}(\hat{T}_{\phi}), \Delta_{\varepsilon}(\hat{T}^{k}))$ in the case of herpangina are significantly weaker than those of the other three diseases. Hence, in the case of herpangina, even for the top-rank search terms, the trend change of the search terms are incapable of reflecting the trend change of the epidemic. Thus, it is difficult to predict the trend of herpangina using the search terms.

Compared to herpangina, we observed that the proposed method can predict more accurately for the trend component of the other infectious diseases. However, predicting the trend component is still much harder than predicting the irregular component. As confirmed in Figure 3c, the prediction error of the trend component tends to increases with time. To be precise, for many infectious diseases, the prediction error of the trend component in the second year of the test period is often greater than that of the first year. In contrast, we did not observe such difference in the prediction accuracy of the irregular component.

6.4.3 *Can we select meaningful search terms?*

In Table 1, we also show the relevance ratio and the actual number of search terms semantically related to the target diseases

	Correlation				sMAPE				Relevant ratio				
Time lag	GFT(2009)	ElasticNet	Neils(2017)	Proposed	GFT(2009)	ElasticNet	Neils(2017)	Proposed	GFT(2009)	ElasticNet	Neils(2017)	Proposed[T]	Proposed[I]
Influenza (The flu)													
0	0.95	**0.99**	0.77	**0.99**	92.39	58.47	44.95	**30.24**	100% (7/7)	33% (9/27)	6% (6/99)	100% (28/28)	100% (12/12)
1	**0.99**	0.98	0.69	0.98	95.37	44.14	42.75	**31.02**	100% (91/91)	51% (51/99)	6% (6/99)	100% (7/7)	100% (10/10)
2	0.93	**0.95**	0.66	0.93	65.29	42.53	70.56	**41.80**	100% (4/4)	38% (229/589)	4% (4/96)	100% (13/13)	100% (4/4)
Hand, foot and mouth disease (HFMD)													
0	**0.98**	0.97	0.85	**0.98**	**30.32**	32.88	65.80	41.69	100% (4/4)	23% (3/13)	0% (0/60)	100% (6/6)	100% (16/16)
1	0.95	0.98	0.83	**0.99**	57.94	**32.28**	61.39	46.77	50% (2/4)	15% (2/13)	0% (0/64)	100% (3/3)	88% (23/26)
2	0.56	0.73	0.63	**0.97**	78.89	65.13	71.74	**40.83**	0% (0/3)	10% (3/30)	0% (0/79)	100% (3/3)	100% (11/11)
Chickenpox (Varicella)													
0	**0.78**	0.58	0.37	0.77	43.07	64.92	40.95	**34.64**	85% (6/7)	10% (4/38)	0% (0/75)	30% (3/10)	50% (4/8)
1	0.40	**0.83**	0.20	0.56	56.07	32.02	44.99	53.04	14% (7/47)	16% (2/12)	0% (0/4)	18% (2/11)	13% (3/22)
2	0.43	0.29	0.24	**0.58**	60.37	79.49	**44.49**	48.87	7% (5/71)	0% (0/93)	0% (0/67)	33% (4/12)	20% (3/15)
Erythema infectiosum (Fifth disease)													
0	0.76	0.75	0.63	**0.93**	38.30	**30.22**	70.79	46.57	62% (10/16)	42% (3/7)	5% (3/51)	100% (2/2)	88% (8/9)
1	0.84	0.83	0.87	**0.93**	29.30	**22.97**	59.23	78.14	47% (10/21)	60% (3/5)	9% (1/11)	71% (5/7)	80% (8/10)
2	-0.25	0.72	0.82	**0.89**	71.40	66.49	66.49	82.79	37% (10/27)	60% (3/5)	9% (1/11)	66% (4/6)	88% (8/9)
Pharyngoconjunctival fever (PCF)													
0	0.76	0.31	0.35	**0.91**	20.06	34.90	29.72	**12.96**	58% (10/17)	4% (3/73)	0% (0/27)	80% (4/5)	44% (4/9)
1	0.71	0.33	0.21	**0.92**	26.60	29.82	34.26	**12.97**	75% (3/4)	15% (3/20)	0% (0/18)	42% (3/7)	41% (5/12)
2	0.67	0.43	0.26	**0.92**	27.27	27.09	30.86	**13.25**	75% (3/4)	8% (2/24)	0% (0/35)	33% (2/6)	26% (8/30)
Herpangina (Mouth blisters)													
0	0.83	0.86	0.34	**0.98**	45.50	50.34	120.96	**21.07**	85% (6/7)	4% (3/70)	0% (0/37)	100% (8/8)	55% (11/20)
1	**0.92**	**0.92**	-0.07	0.86	51.00	48.89	119.56	**39.40**	16% (2/12)	3% (1/31)	0% (0/11)	100% (8/8)	34% (24/70)
2	0.91	0.92	-0.11	**0.97**	60.90	57.13	112.45	**41.98**	0% (0/29)	0% (0/15)	0% (0/1)	100% (6/6)	46% (12/26)
Streptococcal pharyngitis (Strep throat)													
0	0.83	**0.91**	0.28	0.86	21.56	**12.66**	40.22	24.87	100% (2/2)	40% (2/5)	0% (0/33)	41% (7/17)	13% (4/29)
1	0.57	0.24	0.40	**0.80**	60.95	**37.35**	40.43	38.83	3% (2/63)	0% (0/27)	0% (0/66)	22% (5/22)	6% (7/108)
2	0.56	0.30	0.30	**0.65**	53.96	33.74	**32.74**	42.21	2% (2/67)	0% (0/84)	0% (0/66)	20% (5/24)	9% (4/41)
Epidemic keratoconjunctivitis (EKC)													
0	0.55	0.53	0.07	**0.82**	22.82	28.71	24.98	**14.59**	100% (2/2)	33% (2/6)	0% (0/1)	60% (3/5)	40% (8/20)
1	0.33	0.41	-0.49	**0.56**	22.44	43.40	28.04	**19.17**	2% (1/41)	11% (2/18)	0% (0/1)	57% (4/7)	18% (5/27)
2	0.32	0.44	-0.45	**0.73**	21.55	42.57	27.60	**15.86**	11% (4/36)	3% (2/56)	0% (0/1)	100% (3/3)	13% (2/15)
Gastroenteritis (Infection diarrhea)													
0	0.87	0.87	0.33	**0.96**	34.32	17.13	34.24	**12.73**	100% (40/40)	33% (8/24)	0% (0/18)	100% (13/13)	100% (20/20)
1	0.87	0.84	0.25	**0.95**	39.41	18.33	42.37	**10.48**	100% (2/2)	53% (7/13)	0% (0/16)	100% (26/26)	100% (21/21)
2	0.86	0.81	0.02	**0.93**	14.85	19.35	42.13	**12.88**	50% (5/10)	29% (5/17)	0% (0/70)	100% (20/20)	90% (10/11)
Mycoplasma pneumonia (Walking pneumonia)													
0	0.88	**0.89**	-0.18	0.83	41.06	35.90	35.31	32.13	26% (13/50)	40% (4/10)	0% (0/3)	50% (2/4)	12% (1/8)
1	0.49	**0.91**	-0.14	0.70	34.65	35.59	35.66	27.75	16% (13/79)	50% (4/8)	0% (0/3)	36% (4/11)	5% (1/18)
2	0.37	0.75	0.00	**0.77**	37.13	48.76	38.31	35.00	10% (1/10)	27% (3/11)	0% (0/11)	33% (1/3)	3% (1/28)

Table 1: The correlation coefficient (Column 1) , the sMAPE value (Column 2) and the relevance ratio with the actual number of related search terms among the selected ones (Column 3) of the proposed method and three comparative methods.

Rank	Herpangina	Influenza	Fifth disease	HFMD
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

Table 2: The trend correlation $\max_{\varepsilon=1}^{3} cor(\Delta_\varepsilon(\hat{T}_\phi), \Delta_\varepsilon(\hat{T}^k))$ of 5 top-rank search terms for the trend component

among the selected search terms. For example, for the now-casting setting of erythema infectiosum, 17 search terms were selected by the GFT method, in which 10 of them were related to erythema infectiosum. We express these values as '62% (10/16)'. We call the value '58%' the relevance ratio of the selected features. In addition, we also show some of the search terms selected by our method in the appendix.

We see that the proposed method was able to achieve better relevance ratio than the comparative methods. In particular, for the case of the now-casting problem, the proposed and the GFT method

achieved the best relevance ratio among all methods. However, when ϕ is large, the GFT method almost failed to select semantically related search terms. In contrast, the proposed method can select semantically related search terms even when ϕ is large. The ElasticNet method was able to learn a sparse model for most of the cases, whereas many selected search terms are non-related to the target diseases. Furthermore, the method of Neils et al. completely failed in selecting semantically related search terms for all the diseases. The reason is that their selecting rule does not work well when applied to large-scale search logs. To be specific, because of the enormous diversity of the search terms, Neils et al.'s method selected many search terms that are highly correlated to the residual of the seasonal model, but non-related to the target disease. This failure in selection explains the poor prediction accuracy of this method shown in Table 1.

For some diseases, such as chickenpox, streptococcal pharyngitis, EKC and mycoplasma pneumonia, our method selected many non-related search terms for the irregular component. The first reason is
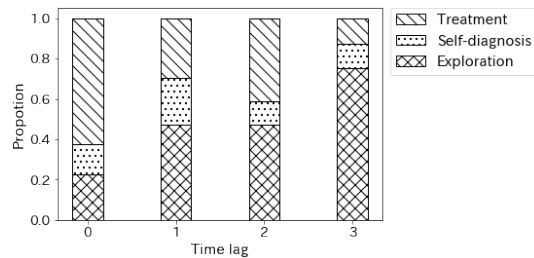
Figure 4: Categories of selected search term for influenza

that the irregular component is often noisy. In addition, the irregular component of these infectious diseases fluctuates tightly around 1.0. As a consequence, it is difficult to distinguish the related search terms from the non-related ones. Fortunately, because the irregular components fluctuate tightly around 1.0, they can merely be treated as noise. Thus, predicting the irregular component becomes relatively less important in the overall epidemic prediction. Even so, the proposed method still performed acceptably in predicting the irregular component of these diseases.

6.4.4 *What kinds of search terms were selected?*

Next, we focus on the case of influenza to analyze the meaning the search terms selected by our method. Borrowing the ideas from existing studies on web searching behavior [4], we categorize selected search terms into three categories: self-diagnosis search term, treatment search term, and exploration search term. To be specific, self-diagnosis search term is the search term that contains symptom keyword, such as 'influenza fever' or 'influenza headache'; treatment search term is the search term about treatments or concerns of infected users, such as 'influenza medicine' or 'influenza going to school'. These two categories are considered to be searched when users actually infected. Finally, exploration search term is the kind of search term that can be searched when the user is whether infected or not, such as 'influenza infectiousness', 'influenza epidemic'. We assume that such self-diagnosis and treatment search terms are more efficient than exploration search terms for predicting the outbreak.

In Figure 4, we show the proportion of these categories for each time-lag ϕ in the case of influenza. Here we summed up the number of selected search terms for trend and irregular components to calculate the proportion. We see that the larger the ϕ, the greater the proportion of exploration search terms becomes. In other words, the larger the ϕ, the smaller the proportion of the search terms that are efficient for predicting the outbreak becomes. This is because the symptoms of influenza typically last only five to seven days for the majority of persons [8]. This decreasing in the number of efficient search terms also well explains the decrease of the prediction accuracy in the case of influenza shown in Table 1.

Due to the randomness of the feature selection process, this observation might not generalize to other infectious diseases. Another insight is that both methods failed to select related search terms for pertussis and mycoplasma pneumonia. We note that these infectious diseases both have comparatively few patients and weak seasonality. Hence, we conclude that a reasonable number of patients and strong seasonality are necessary for the epidemic prediction using search logs.

7 CONCLUSIONS

In this work, we proposed a novel scalable feature selection method and a prediction modeling framework based on seasonal adjustment in the infectious epidemic prediction problem using search engine logs. By considering each component individually, we achieved more accurate and stable outbreak prediction against both the short-term and long-term changes in search engine logs. We showed that our proposed method is significantly outperformed the existing methods by conducting comprehensive experiments on predicting the outbreak of ten kinds of infectious diseases. The experimental results showed that the proposed method can give accurate predictions for both now-casting and forecasting setting. Moreover, the proposed method is more successful in selecting search terms semantically related to target diseases.

REFERENCES

[1] Monica Anderson and Andrew Perrin. 2017. Tech Adoption Climbs Among Older Adults. (2017).
[2] John S Brownstein, Clark C Freifeld, and Lawrence C Madoff. 2009. Digital disease detectionâĂŤharnessing the Web for public health surveillance. *New England Journal of Medicine* 360, 21 (2009), 2153–2157.
[3] Declan Butler. 2013. When Google got flu wrong. *Nature* 494, 7436 (2013), 155.
[4] Marc-Allen Cartright, Ryen W White, and Eric Horvitz. 2011. Intentions and attention in exploratory health search. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*. ACM, 65–74.
[5] Patrick Copeland, Raquel Romano, Tom Zhang, Greg Hecht, Dan Zigmond, and Christian Stefansen. 2013. Google disease trends: an update. *Nature* 457 (2013), 1012–1014.
[6] Niels Dalum Hansen, Kåre Mølbak, Ingemar J Cox, and Christina Lioma. 2017. Seasonal Web Search Query Selection for Influenza-Like Illness (ILI) Estimation. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 1197–1200.
[7] Jeremy Ginsberg, Matthew H Mohebbi, Rajan S Patel, Lynnette Brammer, Mark S Smolinski, and Larry Brilliant. 2009. Detecting influenza epidemics using search engine query data. *Nature* 457, 7232 (2009), 1012.
[8] Heidi Godman. 2016. How long does the flu last? (2016).
[9] Paul Goodwin and Richard Lawton. 1999. On the asymmetry of the symmetric MAPE. *International journal of forecasting* 15, 4 (1999), 405–408.
[10] Isabelle Guyon and André Elisseeff. 2003. An introduction to variable and feature selection. *Journal of machine learning research* 3, Mar (2003), 1157–1182.
[11] National Institute Of Infectious Diseases (JAPAN). [n.d.]. Infectious Diseases Weekly Report (IDWR). https://www.niid.go.jp/niid/en/idwr-e.html.
[12] Ron Kohavi and George H John. 1997. Wrappers for feature subset selection. *Artificial intelligence* 97, 1-2 (1997), 273–324.
[13] Vasileios Lampos, Andrew C Miller, Steve Crossan, and Christian Stefansen. 2015. Advances in nowcasting influenza-like illness rates using search query logs. *Scientific reports* 5 (2015), 12760.
[14] Vasileios Lampos, Bin Zou, and Ingemar Johansson Cox. 2017. Enhancing feature selection using word embeddings: The case of flu surveillance. In *Proceedings of the 26th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 695–704.
[15] Philip M Polgreen, Yiling Chen, David M Pennock, Forrest D Nelson, and Robert A Weinstein. 2008. Using internet searches for influenza surveillance. *Clinical infectious diseases* 47, 11 (2008), 1443–1448.
[16] Robert Tibshirani. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* (1996), 267–288.
[17] Qingyu Yuan, Elaine O Nsoesie, Benfu Lv, Geng Peng, Rumi Chunara, and John S Brownstein. 2013. Monitoring influenza epidemics in china with search query from baidu. *PloS one* 8, 5 (2013), e64323.
[18] Hui Zou and Trevor Hastie. 2005. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67, 2 (2005), 301–320.

A SELECTED SEARCH TERMS

Here we show some of top selected search terms by the proposed method for the now-casting task (originally in Japanese)

Influenza (The flu)

Trend component	Irregular component
influenza incubation period	influenza incubation period
influenza infection period	influenza infection period
after influenza fever	tamiflu side effect
tamiflu side effect	influenza recovery
influenza recovery	influenza school
school of influenza	influenza fever
influenza antipyretic	influenza fever
influenza going out	influenza antipyretics
influenza work	influenza a type

Pharyngoconjunctival fever (PCF)

Trend component	Irregular component
PCF fever	PCF fever
pharyngeal conjunctival fever	riders jacket
pharyngeal conjunctivitis	adeno virus
tamiflu side effect	adeno
riders jacket	hachioji castle
hachioji castle	PCF disease
adeno	kuwabata blog
	epidemic conjunctivitis
	bromidrosis

Hand, foot and mouth disease (HFMD)

Trend component	Irregular component
HFMD pregnant women	HFMD disease
HFMD (Hand Foot Mouth)	HFMD pregnant women
HMFD (Hand Mouth Foot)	HFMD adult
adults HFMD	HFMD treatment
HFMD stomatitis	HFMD infection
HFMD newborn	adults HFMD
	HFMD medicine
	HFMD fever
	high fever HFMD

Hand, foot and mouth disease (HFMD)

Trend component	Irregular component
HFMD pregnant women	HFMD disease
HFMD (Hand Foot Mouth)	HFMD pregnant women
HMFD (Hand Mouth Foot)	HFMD adult
adults HFMD	HFMD treatment
HFMD stomatitis	HFMD infection
HFMD newborn	adults HFMD
	HFMD medicine
	HFMD fever
	high fever HFMD

Chickenpox (Varicella)

Trend component	Irregular component
chickenpox (in Kanji)	chickenpox (in Kanji)
chickenpox (in Hiragana)	chickenpox (in Hiragana)
ikimonogakari thank you lyrics	chickenpox pregnant woman
http://www.kuronekoy	http://www.kuronekoy
first love forecast (Comic)	0471792338
doltz	playstation 3 torne
electric toothbrush	ikimonogakari thank you lyrics
chickenbox (in Romaji)	Liquefied Phenol
gem night deck (Card Game)	

Herpangina (Mouth blisters)

Trend component	Irregular component
helpangina	helpangina
herpangina pregnant woman	HFMD
helpangina incubation period	helpangina incubation period
HFMD disease	helpangina adult
helpengina (typo)	HFMD prevention
HFMD epidemic	HFMD food
helbangina (typo)	helpangina medicine
HFMD adult	helpangina symptoms
	aiko set list (singer show ticket)

Erythema infectiosum (Fifth disease)

Trend component	Irregular component
adult fifth disease	fifth disease adults
fifth disease	fifth disease (in Hiragana)
	erythema infectious
	fifth disaese pregnant women
	fifth disease incubation period
	adults fifth disease
	fifth disease (in Katakana)
	skype bulletin board
	unisia combination tablet

Gastroenteritis (Infection diarrhea)

Trend component	Irregular component
viral gastroenteritis	viral gastroenteritis
infectious disease gastroenteritis	children vomiting
children vomiting	infectious disease gastroenteritis
gastrointestinal cold	infant vomitting
epidemic gastroenteritis	infectious gastroenteritis
gastroenteritis	vomiting diarrhea
baby gastroenteritis	gastroenteritis
diarrhea vomiting	epidemic emesis diarrhea
vomiting diarrhea	child gastroenteritis