Association for Information Systems
AIS Electronic Library (AISeL)

PACIS 2010 Proceedings

Pacific Asia Conference on Information Systems (PACIS)

2010

A Study of Machine Learning Models in Epidemic Surveillance: Using the Query Logs of Search Engines

Ze-Han Fang
National Taiwan University, r97725031@ntu.edu.tw

Jian-Shuin Tzeng
National Taiwan University, d97725003@ntu.edu.tw

Chien Chin Chen
National Taiwan University, paton@im.ntu.edu.tw

Tzu-Chuan Chou
National Chengchi University, tzuchuan@nccu.edu.tw

A STUDY OF MACHINE LEARNING MODELS IN EPIDEMIC SURVEILLANCE: USING THE QUERY LOGS OF SEARCH ENGINES

Ze-Han Fang, Department of Information Management, National Taiwan University, Taipei, Taiwan, R.O.C., r97725031@ntu.edu.tw

Jian-Shuin Tzeng, Department of Information Management, National Taiwan University, Taipei, Taiwan, R.O.C., d97725003@ntu.edu.tw

Chien Chin Chen, Department of Information Management, National Taiwan University, Taipei, Taiwan, R.O.C., paton@im.ntu.edu.tw

Tzu-Chuan Chou, Graduate Institute of National Development, National Chengchi University, Taipei, Taiwan, R.O.C., tzuchuan@nccu.edu.tw

Abstract

Epidemics inevitably result in a large number of deaths and always cause considerable social and economic damage. Epidemic surveillance has thus become an important healthcare research issue. In 2009, Ginsberg et al. observed that the query logs of search engines can be used to estimate the status of epidemics in a timely manner. In this paper, we model epidemic surveillance as a classification problem and employ query statistics from Google to classify the status of a dengue fever epidemic. The query logs of twenty-three dengue-related keywords serve as observations for machine learning and testing, and a number of machine learning models are investigated to evaluate their surveillance performance. Evaluations based on a 5-year real world dataset demonstrate that search engine query logs can be used to construct accurate epidemic status classifiers. Moreover, the learned classifiers generally outperform conventional regression approaches. We also apply various machine learning models, including generative, discriminative, sequential, and non-sequential classification models, to demonstrate their applicability to epidemic surveillance.

Keywords: Text Mining, Classification, Query Log Analysis.

1 INTRODUCTION

In the last decade, the world has witnessed several serious epidemics, such as SARS, H5N1, and H1N1. Epidemics are characterized by a large number of deaths and widespread transmission; and they always cause considerable social and economic damage. Epidemic surveillance is thus important in every country to prevent the spread of deadly diseases. Normally, epidemic surveillance is performed by a country's center for disease control (CDC), which collects various epidemic-related data, such as the number of infections and over-the-counter drug sales, from different medical organizations. The data are aggregated to determine the status or level of the epidemic to help governments take the necessary precautions to prevent the disease spreading further. However, a major problem with the surveillance mechanism is that the data aggregation process is time-consuming (Eysenbach 2006). For example, in Taiwan, the CDC usually takes about a week to aggregate data, so the resulting precautions and prevention measures may not be appropriate responses to the current status of the epidemic.

Eysenbach (2006) observed that people generally use search engines to seek health information or assess an epidemic's status. Indeed, the query logs of epidemic-related keywords are consistent with the development of epidemics. Figure 1 shows the number of dengue fever infections reported by Taiwan's CDC and the query frequency of the keyword "dengue" on Google Insight[1] on a weekly basis from 2005 to 2008. The number of infections, an important indicator of an epidemic's status, is used by many CDCs. A large number of infections means that the intensity of the epidemic is severe. The results in Figure 1 demonstrate that the query frequency of epidemic-related keywords and an epidemic's status are highly correlated; that is, when people search with epidemic-related keywords frequently, the epidemic is usually severe, and vice versa.

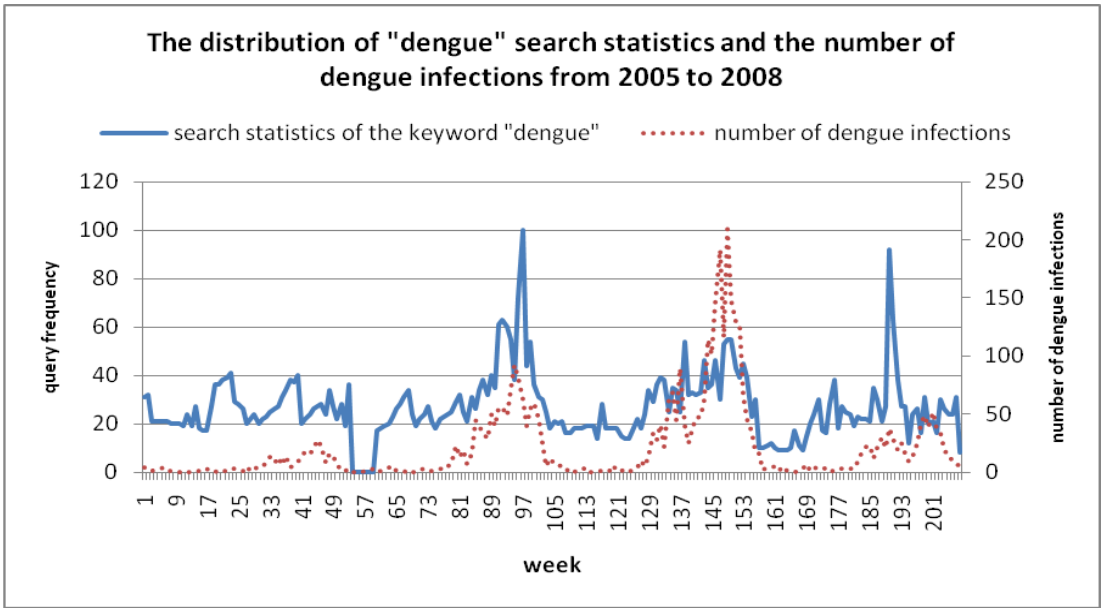

Figure 1. The query frequency of the keyword "dengue" and the number of dengue infections

A number of researchers have proposed using query logs for epidemic surveillance (e.g., Eysenbach 2006; Ginsberg et al. 2009; Polgreen et al. 2008). The query frequencies of epidemic-related keywords are modeled by regression equations to predict the number of infections or the status of an epidemic. In this paper, we model epidemic surveillance as a classification problem and assess the surveillance performance of two types of machine learning models, namely, generative and discriminative classification models. To model the correlation between the query frequencies of epidemic-related keywords and the evolution of an epidemic, we employ the following state-of-the-art sequential and non-sequential classification models: naïve Bayes (NB) (Manning et al. 2008),

[1] http://www.google.com/insights/search/

maximum entropy (ME) (Berger et al. 1996), support vector machine (SVM) (Steinwart and Christmann 2008), hidden Markov models (HMM) (Markov 1913), maximum entropy Markov models (MEMM) (MaCallum et al. 2000), and conditional random fields (CRFs) (Lafferty 2001). Evaluations based on a 5-year real world dataset show that machine learning models together with query logs generally outperform conventional regression-based surveillance approaches.

The remainder of this paper is organized as follows. Section 2 contains a review of related works on epidemic surveillance. In Section 3, we model epidemic surveillance as a classification problem and resolve it with a number of machine learning models. We evaluate the surveillance performance of various approaches in Section 4. Then, in Section 5, we summarize our conclusions and consider avenues for future research.

2 RELATED WORK

The most widely used surveillance method was developed by Serfling (Serfling 1963). The method, which is used by many CDCs, is based on a natural phenomenon whereby the development of epidemics normally follows a seasonal pattern. For instance, in the northern hemisphere, influenza usually starts to affect people in the fall. Serfling thus estimated epidemic-related statistics like the proportion (or number) of deaths from pneumonia and influenza (P&I) or the proportion (or number) of physician visits for influenza-like illnesses (ILIs). Both are important indicators of an influenza outbreak's severity, and are derived by the following periodic regression model:

$$y_t = \mu_0 + \theta t + \alpha \sin(2\pi t/52) + \beta \cos(2\pi t/52) + \varepsilon_t, \qquad (1)$$

where y_t is the epidemic-related statistic at timestamp t; the parameter μ_0 is a weekly base number of epidemic-related statistics; ε_t is a random noise factor whose mean is zero and variance is σ^2; θ describes a secular trend of epidemics; and the sine-wave component $\alpha\sin(2\pi t/52)+\beta\cos(2\pi t/52)$ models the annual recurrence of epidemics. Basically, the equation formulates a sinusoid, which has a 52-week cycle, to capture the seasonal regularity of epidemics. Given a sufficient amount of epidemic-related statistics and the corresponding timestamps, the values of the parameters μ_0, θ, α, β, and σ^2 can be acquired by minimizing the squared difference between the estimated and real statistics (Serfling 1963). Then, the derived parameters and the regression model are used to predict the statistics and trend of epidemics for future timestamps.

Rath et al. (2003) proposed using a hidden Markov model to represent a time series of ILIs. The model characterizes ILI distributions during epidemic and non-epidemic periods by Gaussian and exponential distributions respectively. The reported evaluation results show that the proposed model can reduce the number of cases of false epidemic detection, and that machine learning-based surveillance approaches are promising.

In recent years, the accessibility of query information through many search engine services (e.g., the Google Insight service) has enabled researchers to utilize the query logs of web searches for epidemic surveillance (Eysenbach 2002; Eysenbach 2006; Ginsberg et al. 2009; Polgreen 2008). The idea of using query logs for epidemic surveillance was proposed by Eysenbach (2002). Subsequently, Eysenbach (2006) investigated the correlation between the usage of epidemic-related terms queried on Google and the intensity of epidemics, and found that search engine query logs are effective surveillance indicators. Polgreen et al. (2008) modeled the relationship between searches on Yahoo! and the intensity of influenza outbreaks in the United States. They counted the frequency of queries containing influenza-related terms on a daily basis and divided the count by the total number of searches. The normalized frequencies of the queries in one week were then averaged to construct the following linear model:

$$C_t = \beta_0 + \beta_1 s_{t-x} + \beta_2 t + \varepsilon_p, \qquad (2)$$

where C_t is the epidemic-related indicator (e.g., ILI or P&I) at timestamp t (measured in weeks); s_{t-x} is the average frequency of a query at timestamp t-x; and x is a lagging factor measured in weeks. To

determine an appropriate lag, Polgreen et al. examined 11 possible values for x (ranging from 0 to 10) and compared the R^2 value for each model. More recently, Ginsberg et al. (2009) searched Google databases and selected 45 query terms out of 50 million common queries as indicators for an influenza surveillance model. The authors developed a linear regression model that accurately correlates the log-odds of ILI-related physician visits with the log-odds of the selected query terms.

While the above studies demonstrate that query logs are promising sources of data for epidemic surveillance, to the best of our knowledge, no work has demonstrated their effectiveness in sophisticated machine learning approaches. In this paper, we define epidemic surveillance as a classification problem, and investigate the effectiveness of query logs for epidemic surveillance under two types of machine learning classification models, namely, generative models and discriminative models.

3 CLASSIFICATION MODELS FOR EPIDEMIC SURVEILLANCE

3.1 Problem Definition

As mentioned earlier, we define epidemic surveillance as a classification problem. Let $C = \{c_1, c_2, \ldots, c_K\}$ be a set of status classes of an epidemic defined by the CDC; $Q = \{q_1, q_2, \ldots, q_M\}$ be a set of epidemic-related query terms we are interested in; and random variable $s_t \in C$ represent the status of the epidemic at timestamp t. Given an observation vector $\underline{O}_t = \langle o_{1,t}, o_{2,t}, \ldots, o_{M,t} \rangle$ in which $o_{i,t}$ is the query frequency of q_i at t downloaded from Google Insight, the most likely status of the epidemic at time t would be:

$$\hat{c} = \arg\max_{1 \le k \le K} P(s_t = c_k \mid \underline{O}_t). \quad (3)$$

To categorize the status of the epidemic, we must first model the distribution of $P(c_k|\underline{O}_t)$; when there is no ambiguity, $P(c_k|\underline{O}_t)$ is used instead of $P(s_t=c_k|\underline{O}_t)$. Normally, discriminative models or generative models are used to compute the distribution (Nallapati 2004). Discriminative approaches model the posterior probability $P(c_k|\underline{O}_t)$ directly, or they construct a confidence function $g(c_k|\underline{O}_t)$ that scores the confidence to derive the observation vector $\underline{O}_t$ from the given class. By contrast, generative approaches model the conditional probability $P(\underline{O}_t|c_k)$ and the prior probability $P(c_k)$ indirectly, and estimate the posterior probability $P(c_k|\underline{O}_t)$ in terms of Bayes' theorem (Hogg and Tanis 2005).

In the following sub-sections, we introduce a number of state-of-the-art generative and discriminative classification models for machine learning-based epidemic surveillance. We also investigate, sequential classification models, i.e., HMM, MEMM, and CRFs, to determine whether the trend of epidemics can be used to categorize an epidemic's status. For each classification model, we compile a training dataset comprised of a sequence of epidemic statuses $\langle \mathcal{s}_1, \mathcal{s}_2, \ldots, \mathcal{s}_N \rangle$[2], where $\mathcal{s}_i \in C$, and the corresponding observation vectors $\langle \underline{\mathcal{O}}_1, \underline{\mathcal{O}}_2, \ldots, \underline{\mathcal{O}}_N \rangle$ to acquire the model parameters.

3.2 Generative Models

- **Naïve Bayes** (NB): NB resorts the assumptions of positional independence and conditional independence (Manning et al. 2008) to Bayes' theorem and expands $P(c_k|\underline{O}_t)$ as follows:

$$\begin{aligned} \hat{c} &= \arg\max_{1 \le k \le K} P(c_k \mid \underline{O}_t) \\ &= \arg\max_{1 \le k \le K} P(c_k) P(\underline{O}_t \mid c_k) \\ &= \arg\max_{1 \le k \le K} P(c_k) \prod_{m=1}^{M} P(o_{m,t} \mid c_k), \end{aligned} \quad (4)$$

[2] Released by the CDC, Taiwan

where the model parameters $P(c_k)$ and $P(o_{m,t}|c_k)$ can be acquired from the training dataset by using maximum likelihood estimation (MLE) (Manning et al. 2008).

- **Hidden Markov Model** (HMM): HMM is a classic sequential classification model. It classifies a sequence of observations $<\underline{O}_1, \underline{O}_2, \ldots, \underline{O}_T>$ by computing the following probability:

$$P(< s_1, s_2, \ldots, s_T >|< \underline{O}_1, \underline{O}_2, \ldots, \underline{O}_T >) = P(s_1)\prod_{t=1}^{T-1} P(s_{t+1} \mid s_t)P(\underline{O}_{t+1} \mid s_{t+1}), \quad (5)$$

where $s_t \in C$ denotes the status of an epidemic at time t; and $P(s_1)$, $P(s_{t+1}|s_t)$, and $P(\underline{O}_{t+1}|s_{t+1})$ are model parameters that represent the initial state probability, the state transaction probability, and the observation probability respectively. As the epidemic statuses in the training dataset are available in this research, the model parameters can be acquired directly by using MLE (Chen et al. 2005; Chen et al. 2009). HMM classifies a sequence of observations by searching for a status series $<s_1, s_2, \ldots, s_T>$ that maximizes the probability. Usually, Viterbi's algorithm (Viterbi 1967) is implemented to improve the search efficiency.

Unlike HMM, the classification of NB only depends on the given observation vector, as shown in Equation 4. The NB model ignores the trend of an epidemic because it is assumed that epidemic statuses are independent of each other. HMM, by contrast, considers the trend of an epidemic by incorporating the probabilities of the state transitions (i.e., $P(s_{t+1}|s_t)$) and the initial state (i.e., $P(s_1)$) into the classification.

3.3 Discriminative Models

- **Support Vector Machine** (SVM): SVM is a state-of-the-art discriminative model that has achieved superior classification performances in various application domains (Steinwart and Christmann 2008). Theoretically, SVM is a binary classification model that classifies an observation $\underline{O}_t$ into a positive class or a negative class based on the following equation:

$$f(\underline{O}_t) = sign(\underline{W}^T \Phi(\underline{O}_t) + b), \quad (6)$$

where Φ is a kernel function that maps $\underline{O}_t$ to a high-dimensional kernel space, $\underline{W}$ is a weight vector in the kernel space, and b is the intercept of a decision hyperplane. When $f(\underline{O}_t) = 1$, $\underline{O}_t$ is assigned to the positive class; otherwise, $\underline{O}_t$ belongs to the negative class. Given a training dataset, the model parameters W and b can be derived by solving the following constrained optimization problem:

$$\min \frac{1}{2}\underline{W}^T\underline{W} + C\sum_{t=1}^{N}\varepsilon_t, \quad (7)$$

such that

$$s_t(W^T\Phi(\underline{O}_t) + b) \geq 1 - \varepsilon_t \text{ and } \varepsilon_t \geq 0, \quad t \in \{1,\ldots,N\}, \quad (8)$$

where C is a regularization term that controls overfitting; and ε_t is a slack variable of mis-classification for a training example.

To apply SVM to our multi-class epidemic surveillance problem, we adopt a popular and efficient multi-class classification strategy called *one-against-one* (Hsu and Lin 2002). The strategy first constructs $K(K\text{-}1)/2$ SVM models, each of which is a binary classifier against a pair of the status classes. To classify an observation $\underline{O}_t$, each classifier performs a binary classification on the observation, which is then assigned to the majority class selected by the classifiers.

- **Maximum Entropy** (ME): ME, also known as multi-class logistic regression, computes the probability of a status c_k, given an observation $\underline{O}_t$, as follows:

$$P(c_k \mid \underline{O}_t) = \frac{1}{Z(\underline{O}_t)}\exp(\sum_j w_j * f_j(c_k, \underline{O}_t)), \quad (9)$$

where f_j is a binary feature function defined by domain experts, and w_j is its weight. A binary function returns 1 (or true) when a specific condition exists between c_k and $\underline{O}_t$. For example, a feature function of epidemic surveillance returns true when $o_{1,t} = 15$ and $s_t = c_1$. One advantage of ME is that domain experts can specify arbitrary feature functions to examine any potential relationships between status classes and observations. We introduce the feature functions used for evaluations in Section 4. $Z(\underline{O}_t) = \sum_{i=1}^{K}\exp(\sum_j w_j * f_j(c_i, \underline{O}_t))$ is a smoothing factor used to normalize the output within the range [0,1]. Given a training dataset, the weights of the feature functions can be derived by conditional maximum likelihood estimation (Jurafsky and Martin 2008).

- **Maximum Entropy Markov Model** (MEMM): MEMM is a sequential classification model that augments ME by considering the status transitions in a classification. It classifies a sequence of observations by computing the following probability:

$$P(< s_1, s_2, ..., s_T > | < \underline{O}_1, \underline{O}_2, ..., \underline{O}_T >) = \prod_{i=1}^{T} P(s_t \mid s_{t-1}, \underline{O}_t), \qquad (10)$$

where $s_t \in C$ denotes the status of an epidemic at time t. The probability $P(s_t|s_{t-1}, \underline{O}_t)$ is modeled by ME as follows:

$$P(s_t \mid s_{t-1}, \underline{O}_t) = \frac{1}{Z(s_{t-1}, \underline{O}_t)} \exp(\sum_j w_j * f_j(s_t, \underline{O}_t, s_{t-1})). \qquad (11)$$

Unlike ME, the feature functions of MEMM incorporate the previous status s_{t-1} into classification. For instance, a feature function in MEMM is true when $o_{1,t} = 15$ and $s_t = c_1$, given that $s_{t-1} = c_2$. The corresponding normalization factor $Z(s_{t-1}, \underline{O}_t)$ is represented as follows:

$$Z(s_{t-1}, \underline{O}_t) = \sum_{i=1}^{k} \exp(\sum_j w_j * f_j(c_i, \underline{O}_t, s_{t-1})). \qquad (12)$$

The weights of feature functions can be acquired by using the same training approach as that used for ME. Once again, the Viterbi algorithm is used to classify the best status sequence efficiently (McCallum et al. 2000).

- **Conditional Random Fields** (CRFs): In this paper, we adopt linear-chain CRFs, which compute the probability of a status sequence given a series of observations as follows:

$$P(< s_1, s_2, ..., s_T > | < \underline{O}_1, \underline{O}_2, ..., \underline{O}_T >) = \frac{1}{Z} \exp(\sum_{t=1}^{T} \sum_j w_j * f_j(s_{t-1}, s_t, < \underline{O}_1, \underline{O}_2, ..., \underline{O}_T >, t)), \qquad (13)$$

where w_j is the weight of the feature function f_j acquired from the training data, and Z is a smoothing factor that normalizes the output within the range [0,1]. Unlike MEMM, the feature functions of CRFs consider the whole sequence of observations $<\underline{O}_1, \underline{O}_2, \ldots, \underline{O}_T>$. This resolves a limitation of MEMM, i.e., some statuses in a status sequence are determined completely by their previous statuses if the latter have distinct observation degrees (Bottou 1991). CRFs solve this problem by using a single distribution definition to consider all the observations in a status sequence (Lafferty 2001). In CRFs, domain experts can define two types of feature functions, namely, transition feature functions and status feature functions. The outputs of transition feature functions depend on the previous status s_{t-1} and the current status s_t, whereas status feature functions only consider the current status. In the next section, we introduce the selected feature functions. Once again, the Viterbi algorithm is employed to search for the best status sequence.

4 PERFORMANCE EVALUATIONS

4.1 Evaluation Dataset and Performance Metrics

The prevalence of dengue fever in Taiwan causes numerous deaths as well as considerable social and economic damage. In this paper, we assess the performance of machine learning models in the surveillance of dengue epidemics. Taiwan's CDC classifies dengue fever's status into five levels: minimal, low, moderate, high, and intense, based on the number of dengue infections reported in one week. Table 1 shows the definition of each level.

Number of infections	Level
# of instances = 0	minimal
1 <= # of instances <= 2	low
3 <= # of instances <= 7	moderate
8 <= # of instances <= 26	high
27 <= # of instances	intense

Table 1. Classification of dengue fever infections

We collected information about dengue infections from Taiwan's CDC for a 5-year period (2004 to 2008) and identified their weekly statuses for evaluation. Twenty-three Chinese keywords related to *dengue*, *aedes aegypti*, *fever*, *symptoms*, *disease*, *prevention*, *viruses*, *hemorrhagic fever*, *vaccine*, *mosquito*, *lamp for killing mosquitos*, and *mosquito incense coils* were selected manually from a Taiwanese dengue prevention brochure as observations for the classification models. Then, we downloaded the query statistics of the keywords searched in Taiwan during the evaluation period from Google Insight for machine learning and testing. Google discretizes the frequency of queries into 101 levels (i.e., from 0 to 100). However, the range is too broad, so many levels do not have any training observations and produce zero probability in a classification. To alleviate this data sparseness problem, we further normalize the range into 11 levels (i.e., from 0 to 10) by partitioning the statistics provided by Google into equal deciles. The resulting evaluation dataset contains 260 epidemic statuses, each of which is associated with an observation vector of the normalized statistics of the 23 keywords.

To determine whether the query logs of search engines are effective indicators for epidemic surveillance, we compare the classification models described in Section 3 with Serfling's method and Polgreen et al.'s linear regression approach. As we have 23 observation keywords, we redefine the latter model as follows:

$$C_t = \beta_0 + \sum_{q=1}^{23} (\beta_q s_{t-x_q}) + \beta_{24} t + \varepsilon_p, \qquad (14)$$

where β_q is the coefficient (or weight) of a query term q, and q's lagging factor (i.e., x_q) is determined by using Polgreen et al.'s method. For each compared method, we use 5-fold cross-validation (Manning et al. 2008) to derive credible results. That is, we evaluate the performance of each method over five runs. In each evaluation run, the statuses and the observations for one year are selected for testing, and the remaining data is used for training the model parameters. Then, the results of the 5 evaluation runs are averaged for comparison. The evaluation metrics are macro/micro-average accuracy and one-step accuracy. The former estimates the fraction of status classifications that are correct, while the latter disregards one-step errors and treats them as correct classifications. As the distributions of the status classes are very non-uniform in the evaluation runs, a simple misclassification in a small class could cause a huge variation in the macro-average performance. In contrast, the micro-average is insensitive to class sizes and is thus appropriate for evaluating the overall performance of each method.

To evaluate ME and CRFs, we use the open source tools MaxEnt[3] and CRF++[4] respectively. An advantage of these tools is that feature functions can be specified simply through feature template definitions. For ME, the set of feature functions defined by our template forms the Cartesian product of current epidemic statuses and normalized keyword statistics. The set of functions thus considers every possible relationship between status classes and observation keywords. For MEMM and CRFs, the functions are further extended by considering all possible previous epidemic statuses. For SVM, LIBSVM[5] is employed and the polynomial kernel is selected because of its superior classification performance.

4.2 Classification Accuracy Evaluations

Table 2 shows the micro-average classification accuracy and one-step accuracy scores derived by the compared models. For each machine learning model, one-tail paired t-tests are applied to determine whether the model together with the query logs outperforms Serfling's method and Polgreen et al.'s method significantly. Tables 3 and 4 list the t-values of the statistical tests. As shown in Table 2, all the machine learning models outperform Serfling's method and Polgreen et al.'s method. However, for ME and MEMM, the improvement over Serfling's method is not statistically significant with a 90% confidence level, as shown in Table 3. The significant improvements of the machine learning models over Polgreen et al.'s method indicate that sophisticated machine learning models are more appropriate than conventional linear regression approaches for epidemic surveillance. Among the models, HMM achieves the best classification performance; its micro-average accuracy rate and one-step accuracy rate reach 48.46% and 88.85% respectively. Figure 2 shows that the HMM classification results align with the true dengue statuses. As the HMM model classifies the status of dengue accurately, it captures the development of dengue correctly. The accuracy of Serfling's method is 26.54%. The inferior performance is due to the effect of global warming because the temperature and humidity appropriate for dengue epidemics change year by year (Tsai and Liu 2005). This climate variability makes the periodic model ineffective.

Model	Accuracy	One-step Accuracy
Serfling's method (1963)	26.54%	75%
Polgreen et al.'s method (2008)	24.23%	70.38%
Naïve Bayes	46.15%	83.46%
HMM	48.46%	88.85%
SVM	37.69%	88.08%
ME	31.54%	67.62%
MEMM	30%	67.31%
CRFs	40%	80.77%

Table 2. The classification accuracy and one-step accuracy rates of the compared models

It is interesting to note that, although a number of researchers (e.g., Goutte et al. 2004; Nallapati 2004) have demonstrated that discriminative models are superior to generative models in terms of data classification, the generative models (i.e., NB and HMM) outperformed the discriminative models in our experiment. After investigating this result, we concluded that the relatively inferior performance of discriminative models is due to the criterion applied by our feature selection function. In this paper, to accord with the conditional independence assumption of generative models and ensure that comparisons are fair, the selected feature functions do not examine the correlations between the observation keywords. This restricts the power of discriminative models because domain experts are encouraged to consider various relationships between observations to derive representative feature functions for classification. Additionally, Ng and Jordan (2001) suggest that discriminative models are inferior to generative models if the size of training data is not large enough. To assess the validity of this conjecture, we are designing sophisticated feature functions and performing further experiments on a large evaluation dataset.

[3] http://maxent.sourceforge.net/
[4] http://crfpp.sourceforge.net/
[5] http://www.csie.ntu.edu.tw/~cjlin/libsvm/

Model	t-value
Naïve Bayes	3.64***
HMM	2.92***
SVM	2.00*
ME	0.66
MEMM	0.36
CRFs	3.65***

*, **, ***, ****, and ***** represent one-tail paired t-tests with α=0.1, 0.05, 0.025, 0.01, and 0.005 respectively

Table 3. *One-tailed paired t-test analyses of the machine learning methods' improvement over Serfling's method*

Model	t-value
Naïve Bayes	3.61***
HMM	3.73***
SVM	4.55****
ME	2.33**
MEMM	2.72**
CRFs	4.65*****

*, **, ***, ****, and ***** represent one-tail paired t-tests with α=0.1, 0.05, 0.025, 0.01, and 0.005 respectively

Table 4. *One-tailed paired t-test analyses of the machine learning methods' improvement over Polgreen et al.'s method*

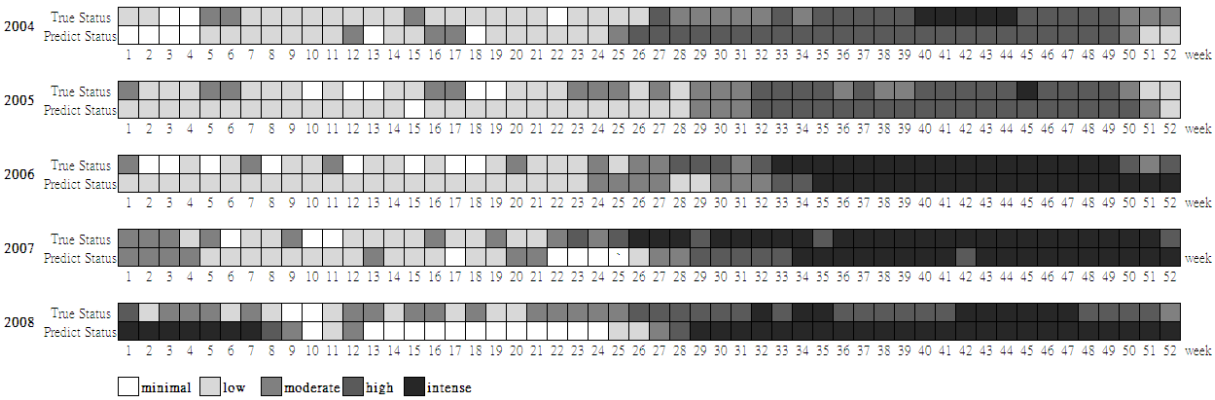

Figure 2. *The classification results of HMM aligned with the true dengue statuses*

Model	Accuracy	One-step Accuracy
Serfling's method (1963)	46.85%	100%
Polgreen et al.'s method (2008)	38.74%	84.68%
Naïve Bayes	56.76%	88.29%
HMM	61.26%	92.79%
SVM	39.64%	90.09%
ME	31.53%	63.06%
MEMM	35.14%	66.67%
CRFs	37.84%	71.17%

Table 5. *The classification accuracy and one-step accuracy of the compared models when a dengue epidemic becomes serious*

A surveillance model should be able to make correct status predictions when a monitored epidemic becomes serious. In our 5-year dataset, more than half (57.3%) of the dengue statuses are minimal, low, or moderate. To evaluate the surveillance performance of the compared models when a dengue

fever outbreak becomes serious, we removed the minimal, low, and moderate statuses from the dataset. Table 5 shows the classification accuracy of the compared models. Once again, HMM achieves the highest accuracy (61.26%); and the generative models outperform the discriminative models and the regression approaches.

To summarize, the superior performance of the machine learning models demonstrates that the query logs of search engines can be used to construct accurate epidemic surveillance systems. The learned models generally outperform the conventional regression approaches. In practice, as search engine query logs are readily available through online Web services, surveillance methods that employ machine learning models and query logs can provide timely and accurate epidemic information earlier than CDCs. For instance, in this study, the estimated statuses of a dengue epidemic are consistently one week ahead of Taiwan's CDC surveillance reports.

4.3 Effectiveness of Epidemic Trend in Classification

The sequential models HMM and MEMM augment NB and ME by considering previous epidemic statuses in the classification. We compare the performances of the four models to determine whether an epidemic's trend is an effective factor in epidemic surveillance. As shown in Table 2, HMM gains from the epidemic trend as it outperforms NB, but MEMM is the less accurate than ME. Nevertheless, the t-test statistics shown in Table 6 indicate that the performance differences between the sequential and non-sequential models are not statistically significant. This indicates that the trend of an epidemic is of little help in machine learning-based epidemic surveillance. Sequential machine learning models are only effective when time series (or sequential) data implies sequential patterns. For instance, MEMM is effective in information extraction because the syntactic structure of sentences provides valuable patterns for named entity identification (McCallum et al. 2000). However, as shown in Figure 3, epidemics usually spread rapidly, so information about the previous status is of little help in predicting the current status of an epidemic.

Model	t-value
HMM vs. NB	0.98
MEMM vs. ME	0.78

*, **, ***, ****, and ***** represent one-tail paired t-tests with α=0.1, 0.05, 0.025, 0.01, and 0.005 respectively

Table 6. The paired t-test analyses of sequential and non-sequential models

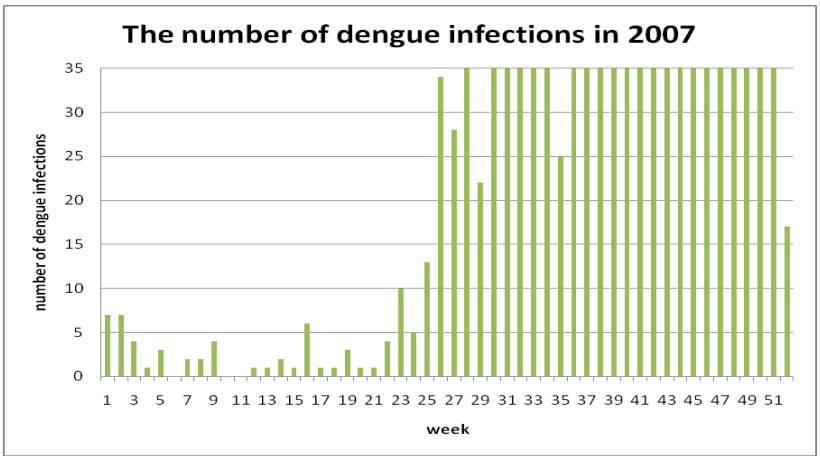

Figure 3. The number of dengue infections in 2007

5 CONCLUSION AND FUTURE WORK

The query logs of search engines can be used as effective indicators in epidemic surveillance. As query logs are readily available through online Web services, they can provide timely and accurate information about an epidemic. Previous works have modeled the statistics of query logs linearly to predict the status of epidemics; however, to the best of our knowledge, no work has evaluated their effectiveness using machine learning models. In this paper, we model epidemic surveillance as a classification problem and assess the effectiveness of query logs using state-of-the-art machine learning models. Evaluations based on a 5-year real world dataset demonstrate that query logs can be used to construct accurate epidemic status classifiers. Moreover, the learned classification models outperform the classic regression-based methods of Serfling and Polgreen et al. Contrary to expectations, our experiment results show that generative models are superior to discriminative models. Interestingly, incorporating the trend of an epidemic into machine learning models is of little help in epidemic surveillance.

Broadly speaking, query logs can be considered as a kind of Web 2.0 data, because Internet users can freely access query information or contribute their opinions when performing Web searches. In the future, we will employ query logs and machine learning models to investigate social and economic phenomena. For example, we will analyze the query logs of economic-related terms to predict business cycles or forecast the sales of a new product. The analyses would help us realize the capability of this kind of collective intelligence for predicting the outcomes of the above phenomena. In addition, we will develop a keyword selection mechanism to investigate the correlation between query terms and the evolution of a social issue to identify representative observation keywords automatically. We will also consider various feature functions to examine the performance of discriminative models further.

Acknowledgements

The authors would like to thank the anonymous reviewers for their valuable comments and suggestions. This work was supported in part by NSC 97-2221-E-002-225-MY2.

References

Bottou, L. (1991). Une approche th´eorique de l'apprentissage connexionniste: Applications `a la reconnaissance de la parole. PhD thesis, Universit´e de Paris XI.

Berger, A.L., Della Pietra, V.J., and Della Pietra, S.A. (1996). A Maximum Entropy Approach to Natural Language Processing. Computational Linguistics, 22 (1):39-71.

Chen, C.C., Chen, M.C., and Chen, M.S. 2005. LIPED: Hmm-based life profiles for adaptive event detection. In Proceedings of KDD '05, 556–561.

Chen, C.C., Chen, M.C., and Chen, M.S. 2009. An Adaptive Framework for Event Detection Using HMM-based Life Profiles. ACM Trans. on Information Systems (TOIS), Vol 27, Issue 2, Article No. 9.

Eysenbach, G. (2002). Infodemiology: The epidemiology of (mis)information. Am J Med, 113:763-765.

Eysenbach, G. (2006). Infodemiology: tracking flu-related searches on the web for syndromic surveillance. AMIA Annu Symp Proc, 244-248.

Ginsberg, J., Mohebbi, M.H., Patel, R.S., Brammer, L., Smolinski, M.S., and Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. Nature, 457:1012-1015.

Goutte, C., Cancedda, N., Gaussier, E., and De´jean, H. (2004). Generative vs Discriminative Approaches to Entity Extraction from Label Deficient Data. In proceedings of JADT 2004, 10–12.

Hsu, C.W. and Lin, C.J. (2002). A comparison of methods for multi-class support vector machines. IEEE Transactions on Neural Networks, 13(2):415-425.

Hogg, R.V. and Tanis, E.A. (2005). Probability and Statistical Inference (7ed Edition). Prentice Hall.

Jurafsky, D., and Martin, J.H. (2008). Speech and Language Processing (2nd Edition). Prentice Hall.

Lafferty, J., McCallum, A., and Pereira, F. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Proceedings of ICML 2001, 282-289.
McCallum, A., Freitag, D., and Pereira, F. (2000). Maximum entropy Markov models for information extraction and segmentation. In Proceedings of ICML 2000, 591–598.
Morbidity and Mortality Weekly Report (MMWR). http://www.cdc.gov/mmwr/
Manning, C.D., Raghavan, P., and Sch¨utze, H. (2008). Introduction to Information Retrieval. Cambridge University Press. Online 17/08/2007. http://nlp.stanford.edu/IR-book/information-retrieval-book.html.
Markov, A.A. (1913). An example of statistical investigation in the text of 'Eugene Onyegin' illustrating coupling of 'tests' in chains. In Proceedings of the Academy of Sciences 7, 153–162.
Nallapati, R. (2004). Discriminative models for information retrieval. In Proceedings of SIGIR 2004, 64–71.
Ng, A.Y., and Jordan, M. (2001). On discriminative vs. generative classifiers: A comparison of logistic regression and Naive Bayes. NIPS, 14:841-848.
Polgreen, P.M., Chen, Y., Pennock, D.M., and et al. (2008). Using Internet Searches for Influenza Surveillance. Clinical Infectious Diseases, 47: 1443–1448.
Rath, T.M., Carreras, M., and Sebastiani, P. (2003). Automated detection of infuenza epidemics with hidden Markov models. Advances in intelligent data analysis V. Springer-Verlag, 521–531.
Serfing, R.E. (1963). Methods for current statistical analysis of excess pneumonia-infuenza deaths. Public Health Reports, 78:494–506.
Steinwart, I., and Christmann, A. (2008). A. Support Vector Machines. Springer, New York.
Tsai, H.T., and Liu, T.M. (2005). Effects of global climate change on disease epidemics and social instability around the world. Human Security and Climate Change, 21-23.
Viterbi, A. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm, IEEE Transactions on information Theory, 13(2): 260- 269.