Monitoring Epidemic Alert Levels by Analyzing Internet Search Volume

Xichuan Zhou*, *Member, IEEE*, Qin Li, Zhenglin Zhu, Han Zhao, Hao Tang, and Yujie Feng

***Abstract*—The prevention of infectious diseases is a global health priority area. The early detection of possible epidemics is the first and important defense line against infectious diseases. However, conventional surveillance systems, e.g., the Centers for Disease Control and Prevention (CDC), rely on clinical data. The CDC publishes the surveillance results weeks after epidemic outbreaks. To improve the early detection of epidemic outbreaks, we designed a syndromic surveillance system to predict the epidemic trends based on disease-related Google search volume. Specifically, we first represented the epidemic trend with multiple alert levels to reduce the noise level. Then, we predicted the epidemic alert levels using a continuous density HMM, which incorporated the intrinsic characteristic of the disease transmission for alert level estimation. Respective models are built to monitor both national and regional epidemic alert levels of the U.S. The proposed system can provide real-time surveillance results, which are weeks before the CDC's reports. This paper focusses on monitoring the infectious disease in the U.S., however, we believe similar approach may be used to monitor epidemics for the developing countries as well.**

***Index Terms*—Hidden Markov model (HMM), infectious disease, outbreak surveillance, search engine.**

I. Introduction

NOTIFIABLE infectious diseases, such as the hepatitis, cause over a million infections in the U.S. every year [1]. Early detection of disease activity, when followed by a rapid response, can reduce both social and medical impact of the disease. However, conventional surveillance systems, e.g., the centers for disease control and prevention (CDC), rely on the clinical data. Specifically, a network of sentinel laboratories performs disease test, by counting and classifying pathogens collected from patients, while a network of sentinel physicians reports the number of people diagnosed with notifiable infectious diseases. The CDC's reports regarding epidemic activities are no longer current when released to health care professionals. Generally, the CDC publishes the surveillance results weeks after epidemic outbreaks. To improve the early detection of epidemic outbreaks, we designed a syndromic surveillance system based on analyzing disease-related Google search volume. The search volume data is public available and updated by the Google trend service on daily basis. So the search-based system can provide update-to-date surveillance results.

As estimated 113 million people in the U.S. search online for information about medical problems each year [2]. Most people searching for medical information use a search engine [2]. Internet search engine users include patients and their families and health care professionals [3], [4]. Since a large population of people search online for medical information, thus the pattern of how and when people search may provide clues or early indications about future concerns and expectations. Cooper *et al.* found that Internet searches for specific cancers were correlated with their estimated incidence [5]. As an earlier attempt to use search data for epidemic surveillance, Polgreen and Ginsberg proposed to estimate the influenza trend using Yahoo and Google search queries, respectively, [6], [7]. Wilson discussed different Internet-based methods, including search engine based method, for disease outbreaks detection [8].

The key of existing researches is the assumption that the search volume trends of disease-related terms are highly correlated with actual infection trends [6], [7]. This assumption is consistent with the observation for the disease of influenza, which infects a large population and has millions of related searches submitted on search engines. Fig. 1(a) shows the morbidity trend of influenza and the Google search volume for the term of "influenza". As one can see, the morbidity trend and the Google search trend have contemporary high and low points. The positive correlation characteristic allows researchers to use the regression-based methods for disease trend estimation [6], [7], [9]. Lately, Zhou proposed to estimate the number of measle infections in China using the search volume data [9]. Due to the large population of infections in China, the measle morbidity trend was also found to be highly correlated with the search-volume trends of measle-related terms.

However, most of the notifiable infectious diseases, with less infections and searches, may not satisfy the assumption of highly correlation between the disease trends and the related search volume trends. Fig. 1(b) shows the example of the hepatitis disease in America. Note that the influenza trends are smooth, however, the hepatitis morbidity trend and the Google search volume of the term "hepatitis" are more noisy. This is mainly because the hepatitis has much less occurrences than the influenza. Due to the *noisy effect*, the disease morbidity trend and the search volume trend are no longer highly correlated, which violates

Manuscript received July 4, 2012; revised September 17, 2012 and October 29, 2012; accepted October 30, 2012. Date of publication November 20, 2012; date of current version January 16, 2013. This work was supported in part by the National Natural Science Foundation of China under Grant 61103212) and the Natural Science Foundation Project of CQ CSTC under Grant cstcjjA40005). *Asterisk indicates corresponding author.*

*X. Zhou is with the College of Communications Engineering Chongqing University, Chongqing 400044, China. He is also a member of the No. 24 research institute of CETE (e-mail: zxc@cqu.edu.cn).

Q. Li and H. Zhao are with the Centers for Disease Control and Prevention of Chongqing, Chongqin 400038, China.

Z. Zhu and H. Tang are with the Chongqing University, Chongqing 400044, China.

Y. Feng is with the Southwest Hospital, Chongqing 400044, China.

Digital Object Identifier 10.1109/TBME.2012.2228264

0018-9294/$31.00 © 2012 IEEE

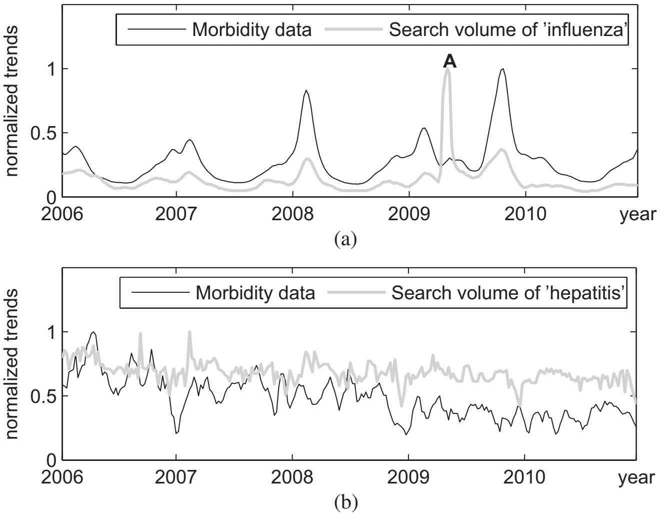


Fig. 1. Normalized morbidity trends published by the CDC (black) and the normalized Google search trends of disease-related terms (gray). Note that the hepatitis trends are more noisy than the influenza trends for the period from January 2006 to December 2010. (a) Influenza. (b) Hepatitis A.

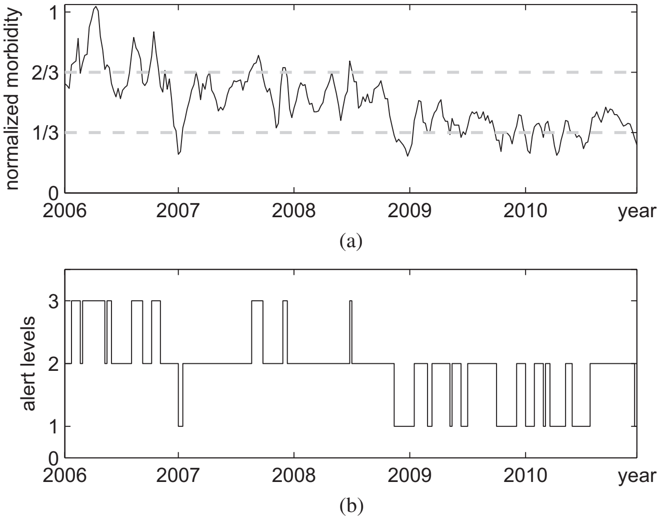


Fig. 2. Normalized morbidity trend of hepatitis A is evenly split and discreetly represented as three epidemic alert levels. (a) Normalized morbidity trend of hepatitis A in America. (b) Epidemic levels of hepatits A disease.

the assumption of earlier researches. The noisy effect is common for the notifiable infectious diseases because the number of infections and search queries submitted are much less than the case of influenza. Moreover, for regional epidemic surveillance, the noisy effect may play a more important role due to the limited number of search queries submitted. To handle the noise, we used a two-step method in our research, i.e., smooth representation and state-space model estimation. Specifically, we discretely represented the morbidity trends using multiple epidemic alert levels, which reduced the noise level in morbidity trend data (see Fig. 2).

Besides the noise in trend data, it has also been observed that the search volume might be affected by factors other than disease infections. For example, the searches of the term "AIDS" submitted on the World AIDS Day increase dramatically due to media reports every year. Fig. 1 shows another example of the media effect for the disease of influenza. As one can see in Fig. 1(a), the search volume of the term "influenza" had a high point in May 2009 (point A). However, the dramatic increase in search volume was not related with contemporary infections in America (black curve). In fact, the increase was caused by the media reports of an H1N1 outbreak in the country of Mexico. Several cases of H1N1 infections were also reported in the U.S., which attracted a lot of attention and caused a dramatic increase in flu-related searches in America [10].

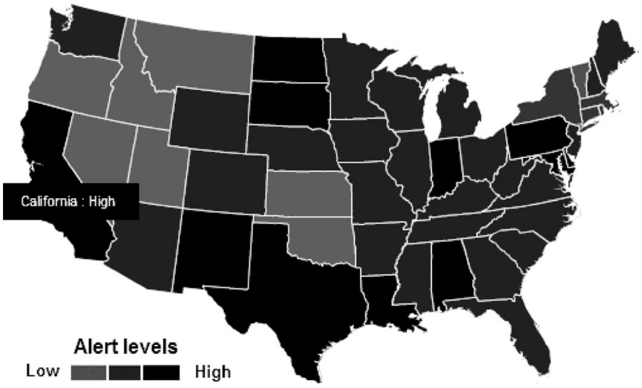


Fig. 3. Illustration of the state-level surveillance system. The gray level of each state indicates the epidemic alert levels in respective regions.

It has been widely accepted that the transmission of infections is highly correlated at continuous time steps [11]. To reduce the effect of irrelevant searches, we proposed to estimate the discrete alert levels using a continuous density hidden Markov model (HMM), instead of direct assessment of search data. Since the Markov method predict the present alert level based on previous alert levels, the proposed system is more robust against the irrelevant searches. Prior to our research, Zhou *et al.* proposed to incorporate the epidemic characteristic using a generalized Kalman filter with periodic parameters [12]. However, their method was based on the assumption of periodic morbidity trends, which may not be satisfied for most notifiable infectious diseases.

Since, the Google trend service provides both the national and regional search volume data, we can build both the national and state-level surveillance models. Fig. 3 illustrates the state-level surveillance system we built. The colors of the states represent the regional epidemic alert levels. Specifically, higher alert levels indicate higher risk of epidemic outbreaks. Furthermore, since the search volume data is updated timely, the resulting weekly estimates are consistently weeks ahead of the CDC surveillance reports. Real-time surveillance results may enable public health officials and health professionals to better respond to epidemics outbreaks. If a particular region experiences an early increase in alert levels, it may be possible to focus additional resources on that region to identify the source of the outbreak, providing extra drug capacity or raising local media awareness as necessary.

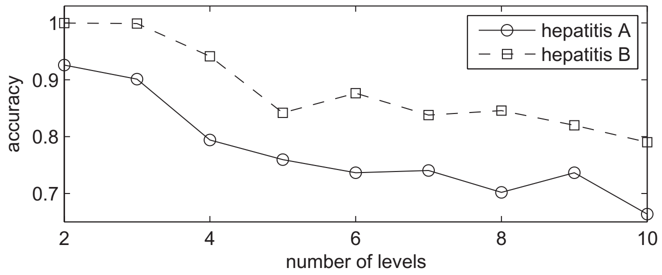


Fig. 4. Change of the average accuracy of the HMM for national-level estimation of hepatitis alert levels. The number of alert levels used for discrete trend representation changed from two to ten.)

II. Data Sources

We used the hepatitis as an example of notifiable infectious diseases to demonstrate our method. We employed two types of data to train and evaluate the model, i.e. the morbidity data published by the CDC and the search volume data published by the Google trend service.

A. Morbidity Data

At each week, State Department of Health Services collects the new cases of hepatitis reported and sends the data to the CDC [14]. The CDC verifies and publishes the provisional new cases periodically. Specifically, the CDC publishes the new hepatitis morbidity data every week, typically with one to four weeks' reporting lag. We obtained the weekly data published from January 2006 to December 2010. It included the hepatitis cases reported in America and each state. Fig. 1(b) shows the weekly updated hepatitis A morbidity data obtained from the CDC.

To calculate the epidemic alert levels, we evenly split the morbidity trend into n levels. For example, suppose n equals three. Suppose m_t denotes the morbidity data of week t, and $M = \max_t m_t$ denotes the maximum number of weekly infections published by the CDC. Let x_t denote the alert levels of the week t. Then we have,

1) if $m_t < \frac{1}{3}M, x_t \Leftarrow 1$, else
2) if $m_t < \frac{2}{3}M, x_t \Leftarrow 2$, else
3) $x_t = 3$.

By definition, higher alert levels indicates larger number of infections reported by the CDC. Generally, larger number of n shows more detailed information but results in lower estimation accuracy in average. The main reason for the accuracy drop is lacking of training data to estimate the parameters. Specifically, the surveillance model has n^2 parameters in the state transition matrix. The larger the n is the less accurate the transition matrix could be estimated with limited training data. Fig. 4 shows the accuracy of national hepatitis alert level estimation with different number of levels n. We chose to use three-level models in our system, since they seemed to have adequate information and estimation accuracy.

B. Search Volume Data

Different from [6] and [7], our research is based on Google trend service [15], which is publicly accessible. We collected the search volume data from January 2006 to the December 2010. The search volume was calculated by aggregating the search queries for the given term (a word or a combination of words) submitted in a selected area. The area involved could be a country or a state. For each term, the volume data was normalized from zero to one. Fig. 1(b) shows the search volume of the term "hepatitis" we obtained (gray curve), which reflects how many searches have been submitted in the U.S.

TABLE I
Brief List of Terms Selected for Hepatitis Surveillance

Categories	Typical terms	Rate
Name	hepatitis, hepatitis A (B, C, D, E)	5 %
Symptom	hepatitis symptoms, splenomegaly, fever, headache, hepatomegaly, appetite loss, abdominal discomfort,liver enlargement, liver inflammation, lymphadenopathy jaundice, decreased blood pressure	34 %
Cause	bacterial, virus, anaplasma, echovirus alcoholic, ethanol, metabolic disorders	21 %
Diagnosis	liver blood test, HAV, HCV, ALT	15 %
Treatment	ribavirin, ganciclovir, paracetamol minocycline, amoxycillin, isoniazid	6 %
Related disease	fatty liver, lassa fever, dengue, cirrhosis influenza, ebola, yellow fever	10 %
Other terms	mushroom, NASH, drink	9 %

One important consideration in our research is the selection of disease-related terms. We first collected a candidate set of terms by domain knowledge. Specifically, we collected 473 terms from the Wikipedia hepatitis articles [17]. We obtained their national-level and state-level search volume for the period from Jan. 2006 to Dec. 2010. Google trend service provides query filters with respect to different knowledge domains. We used the health domain filter in our research. Since the Google only provides the search volume of the terms with adequate data, the terms with not enough queries submitted were removed from the candidate set.

Then, we used an automatic method to further select the best terms from the candidate set. Specifically, the top-related terms were selected using a greedy forward selection method (GFS). The GFS method started with an empty term set, and added the term whose addition provided the best estimation accuracy of the epidemic alert levels. Experiment showed that the national-level accuracy for hepatitis surveillance quickly climbed to 81% by adding about 90 terms. Then the accuracy slowly increased to over 90% until about 300 terms were used to build the model. We also found that removing any single term from the selected term set would not significantly affect the results, mainly because the search volume of most disease-related terms were correlated with one another. Further investigation showed that the selected terms included the names, causes, symptoms, diagnosis method, and related diseases of hepatitis. Intuitively, we intended to estimate the hepatitis alert levels, shown in Fig. 2(b), by multiple contemporary search volume data as in Fig. 1(b).

Table I briefly lists the terms used in our syndromic surveillance system. The selected terms can be roughly categorized in to seven groups. Over 50% of the terms we found were directly

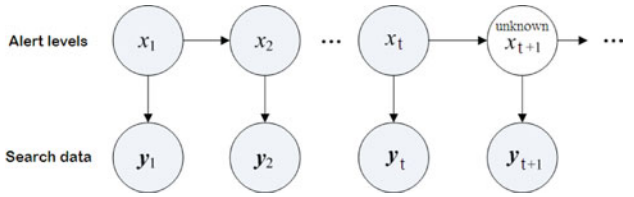


Fig. 5. Illustration of the HMM for surveillance purpose. The states of the system are weekly epidemic alert levels calculated by the CDCs reports. At the week $t+1$, the state-space model estimates the unknown alert level x_{t+1} using the current search data $\mathbf{y}_{t+1}$ and previous alert level x_t.

related to the hepatitis disease, including the symptoms, diagnosis, treatment, and causes. Search quires about hepatitis-related diseases could also help to predict the hepatitis alert levels. The selected terms also included a small proportion of terms, which were not obviously related. For example, the search volume of terms Nash and mushroom are found to be helpful in hepatitis surveillance. We checked and found that the term NASH could be the abbreviation of nonalcoholic steatohepatitis, which on biopsy of the liver resembles alcoholic hepatitis. On the other hand, toxin-containing mushrooms is known to be a cause of hepatitis.

III. Model and Method

In classic epidemic models, the occurrence of infections at continuous time steps are highly related [13]. Therefore, one could use a Markov process to describe the relation between the alert levels at continuous time steps. Suppose $x_t \in \{1,2,\dots,n\}$ is the epidemic alert level at week t. Using the alert level as the states of the Markov process, the state entered at each step depends on an initial probability vector $\boldsymbol{\pi}$ and a state transition matrix $\mathbf{A}$, where

$$\begin{aligned}&\boldsymbol{\pi} = \{\pi_i\},\ \pi_i = \Pr(x_1 = i),\ i = 1,\dots,n\\ &\mathbf{A} = a_{i,j},\ a_{i,j} = \Pr(x_{t+1} = j|x_t = i),\ i,j = 1,\dots,n. \qquad (1)\end{aligned}$$

Suppose the vector $\mathbf{y}_t$ in k dimensional Euclidean space is the search volume vector of k disease-related terms at the week t. We used the continuous density HMM to describe the relation between alert levels and contemporary search volume (see Fig. 5). The continuous density HMM is a special type of HMM with real observations [16]. The probability distribution of the search volume vector $\mathbf{y}_t$ depends on the current alert level x_t. We made the assumption that, given the alert level $x_t = j$, the conditional probability density of $\mathbf{y}_t$ was Gaussian. In this case, a conditional mean vector $\boldsymbol{\mu}_j$ and a conditional covariance matrix $\boldsymbol{\Sigma}_j$ determined the density corresponding to state j . We denoted this density as $\mathcal{G}(\mathbf{y};\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j)$. In summery, a continuous density HMM was defined by the parameters of

$$\boldsymbol{\lambda} = (\mathbf{A},\boldsymbol{\pi},\Theta), \text{ where } \Theta = \{\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j\}_{j=1}^n$$

The goal of our research is to estimate the current alert levels using current search data and earlier published CDC surveillance data. To overcome the CDC reporting lag, the search-based surveillance system was performed in an online fashion. At the week $t+1$, we used the CDC data and the search data before the tth week as the training data. The training process is to update model parameter set $\boldsymbol{\lambda}$ by maximizing the following likelihood function as

$$\begin{aligned}&\max_{\boldsymbol{\lambda}} \mathcal{L}(\boldsymbol{\lambda}|\mathbf{y}_1,\dots,\mathbf{y}_t,\mathbf{x}_1,\dots,\mathbf{x}_t)\\ &= \max_{\boldsymbol{\lambda}} \Pr(\mathbf{y}_1,\dots,\mathbf{y}_t,\mathbf{x}_1,\dots,\mathbf{x}_t|\boldsymbol{\lambda})\\ &= \max_{\boldsymbol{\lambda}} \pi_{x_1}\Pi_{i=2}^t a_{x_{i-1}x_i}\Pi_{i=1}^t \mathcal{G}(\mathbf{y}_i;\boldsymbol{\mu}_{x_i},\boldsymbol{\Sigma}_{x_i}). \qquad (2)\end{aligned}$$

After training, we can use the $(t+1)$th week's search data to predict the $(t+1)$th week's epidemic alert level. The prediction is achieved by choosing the alert level with the maximum a posterior probability as

$$\begin{aligned}&\max_{j\in\{1,\dots,n\}} \Pr(x_{t+1} = j|x_t,\mathbf{y}_{t+1},\boldsymbol{\lambda})\\ &= \max_{j\in\{1,\dots,n\}} \Pr(x_{t+1} = j|x_t,\boldsymbol{\lambda})\Pr(\mathbf{y}_{t+1}|x_{t+1},\boldsymbol{\lambda})\\ &= \max_{j\in\{1,\dots,n\}} a_{x_t,j}\mathcal{G}(\mathbf{y}_{t+1};\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j). \qquad (3)\end{aligned}$$

Since the Google search volume is updated on a daily basis, one can predict the epidemic alert level in real time, which is generally one to four weeks ahead of the CDC's reports. For some diseases, e.g. tuberculosis, similar search based prediction could be over ten weeks ahead of the CDC's reports.

IV. National Results for Hepatitis

By using the CDC data and the search queries submitted in the U.S., we can estimate the national epidemic alert levels. Since the linear regression model was commonly used to estimate the epidemic trends [6], [7], [9]. We compared our models with the linear regression model:

$$\mathbf{y}_t = \boldsymbol{\beta} m_t + \boldsymbol{\epsilon} \qquad (4)$$

where $m_t \in R$ was the normalized morbidity trend for the week $t, \mathbf{y}_t$ was the search volume vector of hepatitis related terms collected at the week t, and $\boldsymbol{\epsilon}$ was the error term. After estimating the morbidity trend m_t using liner regression method, the alert levels $x_t \in \{1,\dots,n\}$ were generated by evenly splitting the trend m_t into n levels.

We also compared our method with the state-of-the-art classification method. Specifically, we used the alert levels as classification labels, and treated the level estimation process as a classification task of the search trend data. We used the classic Bayes classifier in the experiment to assess the search trend data. All the regression model, Bayes classifier and the HMM were trained in an incremental way. Specifically, at each week, all the search data and morbidity data before the last week were used for training. The alert level of the current week were estimated according to equation 3 using the search data of the current week.

Fig. 6 shows the national epidemic-alert-level estimation for hepatitis. The proposed continuous density HMM had 8.1% and 1.8% estimation error rate for hepatitis A and B respectively. On the other hand, the regression based method had 34.1% and 36.5% error rate for hepatitis A and B respectively. And the Bayes method had 12.5% and 14.3% error rate for hepatitis A

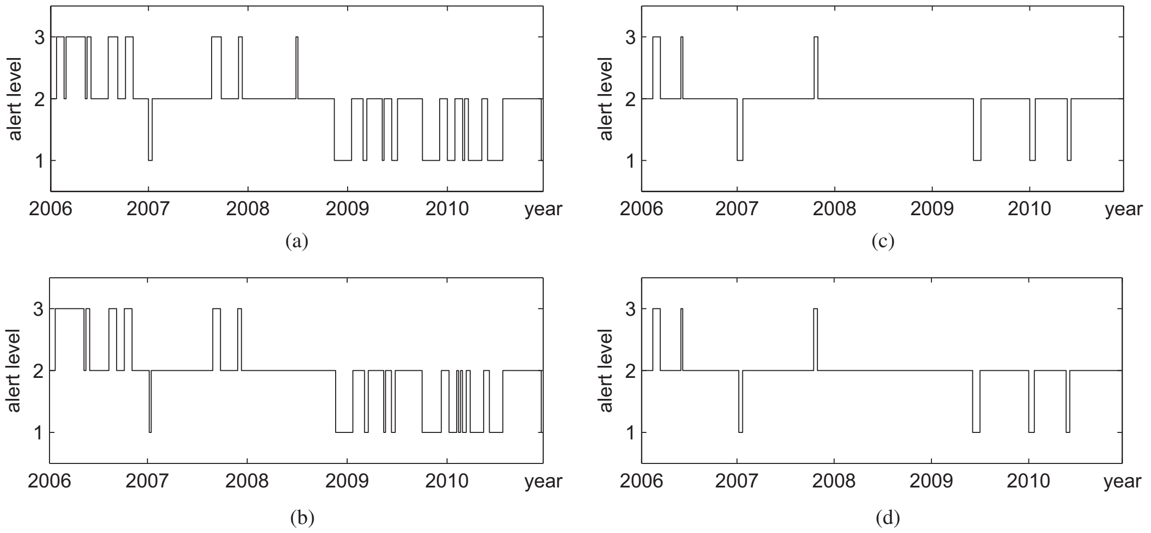


Fig. 6. American hepatitis alert level estimation using the continuous density HMM. (a) Acctual epidemic alert levels of hepatitis A. (b) HMM estimated alert levels of hepatitis A. (c) Actual epidemic alert levels of hepatitis B. (d) HMM estimated alert levels of hepatitis B.

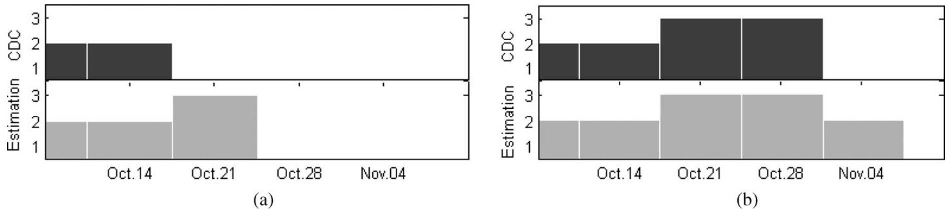


Fig. 7. Epidemic alert level estimation of the hepatitis A disease in real time. For the period from Oct. 2007 to Nov. 207, the search based surveillance results were consistently seven days before the CDC's reports. (a) Surveillance results available until Oct. 21, 2007. (b) Surveillance results available until Nov. 7, 2007.

and B respectively. We analyzed the results and found that the regression method showed worse performance because the search volume data was noisy for hepatitis related terms (Fig. 1(b)), which 'passed' the noise to the estimated alert levels by the linear model. On the other hand, by using discrete representation, the HMM and Bayes classifier could significantly reduce the noise level in the estimation.

Besides the error rate, we also examined the false alarm rate of different models. Suppose the surveillance system started the outbreak alarm if the highest level was reached. For the three-level model estimation, the continuous density HMM had 3.2% false alarm rate for hepatitis A and 0.3% false alarm rate for hepatitis B. Compared to the regression based method, the false alarm rete of the HMM estimation was 12.7% and 17.1% lower for hepatitis A and B respectively. Compared to the Bayes classifier, the false alarm rate of the HMM estimation was 4.7% and 4.5% lower for hepatitis A and B respectively. Further investigation showed that the false alarms were mainly caused by irrelevant searches in Google trend data. By incorporating the transmission characteristic in the Markov model, the HMM significantly reduced the false alarm rate. Lower false alarm rate could reduce the social cost for preparing for a false epidemic outbreak.

The search-based system was designed to monitor the epidemic outbreaks in real time. The surveillance results can be published at the end of each week, which is one to four weeks ahead of the CDC's hepatitis report. To illustrate the temporal lag between the model estimates and the CDC's reports, we show the online estimation from October to November of the year 2007 in Fig. 7. At October 21st, our syndromic system indicated that the alert level of hepatitis increased to level three; similarly, we detected a decrease of alert level in November 4th. Both results were later confirmed by the CDC surveillance data.

V. State Results for Hepatitis

Google also supplies the search volume of given terms for each state of America. Using the state-level search volume, we can estimate the regional epidemic alert levels. Fig. 3 is an illustration of the state-level syndromic surveillance system we built. The gray levels of the regions indicate the epidemic alert levels of an infectious disease.

Similar to the national models, both the continuous density HMM, the Bayes classifier and the regression model were trained in an incremental way. At each week, all the search data and morbidity data before the last week were used for training. The alert level of the current week were estimated using the model and the search data of the current week. We estimated the hepatitis alert levels for all the states where more than one infections occurred in each year. We evaluated the error rate and the false alarm rate for the Bayes classifier, the continuous density HMM and the regression method respectively. The results are summarized in Table II. Similar to the national surveillance, the linear regression method had significantly larger error rate

TABLE II
THIS TABLE LISTS THE HEPATITIS RESULTS FOR THE AMERICAN STATES WHERE OVER ONE INFECTIONS OCCURRED FROM JAN. 2006 TO DEC. 2010

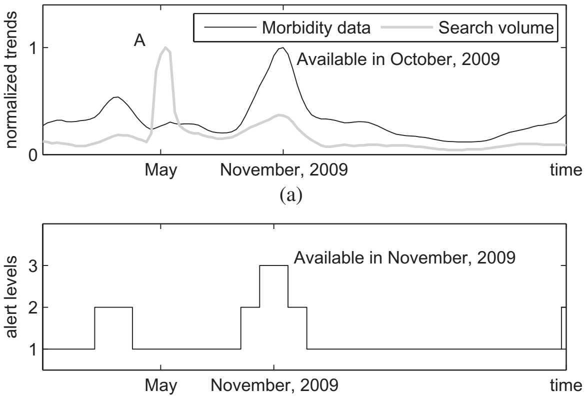
States	**Hepatitis A**			**Hepatitis B**			**States**	**Hepatitis A**			**Hepatitis B**		
	Bayes	HMM	LR	Bayes	HMM	LR		Bayes	HMM	LR	Bayes	HMM	LR
AR	11.3 (5.3)	8.7 (2.4)	19.0 (9.1)	15.1 (7.2)	11.3 (5.7)	21.7 (9.8)	AZ	18.5 (9.5)	14.6 (5.3)	19.4 (11.3)	18.5 (9.1)	15.4 (4.3)	21.2 (11.5)
CA	9.7 (6.5)	9.1 (4.7)	21.1 (2.2)	14.2 (6.4)	13.2 (5.2)	26.2 (14.3)	CO	8.1 (4.2)	7.2 (3.1)	19.3 (10.9)	12.3 (5.2)	10.1 (4.5)	17.3 (8.6)
CT	13.5 (7.1)	9.5 (4.9)	17.8 (8.6)	9.1 (4.3)	6.5 (4.2)	14.0 (5.3)	FL	21.9 (11.2)	18.7 (8.2)	24.2 (11.3)	17.5 (8.9)	13.0 (5.3)	28.2 (10.1)
GA	19.0 (11.6)	15.9 (7.2)	27.9 (13.1)	19.0 (8.8)	17.3 (4.3)	24.2 (9.1)	IL	15.3 (7.4)	7.6 (3.3)	22.9 (9.7)	17.3 (7.2)	13.1 (5.8)	21.3 (9.6)
IN	12.3 (6.2)	7.8 (3.1)	18.6 (8.2)	10.4 (6.9)	9.2 (6.2)	13.0 (9.4)	KS	6.1 (4.0)	4.8 (2.2)	9.7 (4.6)	3.1 (1.3)	2.3 (0.9)	3.8 (1.4)
KY	17.1 (9.5)	13.2 (7.2)	23.7 (13.1)	14.9 (7.7)	14.7 (5.2)	27.2(14.5)	LA	9.6 (4.2)	5.8 (3.3)	15.7 (5.1)	9.3 (5.2)	7.7 (3.4)	12.5 (4.8)
MD	9.3 (5.1)	7.3 (3.5)	15.3 (4.3)	10.3 (4.8)	7.2 (4.1)	24.0 (9.3)	MI	8.1 (4.3)	7.9 (4.1)	17.3 (9.1)	19.4 (7.0)	14.3 (2.1)	21.1 (5.7)
MO	5.3 (2.2)	2.3 (1.4)	12.9 (8.2)	6.2 (4.2)	4.6 (2.1)	7.5 (3.5)	MS	12.4 (5.7)	11.6 (4.9)	19.8 (9.0)	10.0 (4.4)	9.3 (4.3)	13.7 (6.9)
MT	14.7 (8.4)	13.7 (7.2)	21.7 (12.2)	13.7 (6.2)	11.2 (6.3)	17.4 (10.1)	NE	12.6 (6.5)	7.3 (3.9)	15.3 (14.5)	12.3 (6.2)	11.2 (5.3)	14.6 (10.3)
NM	10.4 (4.7)	7.1 (3.9)	17.3 (5.5)	10.8 (6.5)	9.3 (6.2)	25.4 (13.5)	NY	10.6 (4.2)	5.1 (1.5)	17.2 (6.3)	17.3 (7.0)	11.2 (6.1)	19.5 (8.7)
OH	19.6 (8.3)	12.4 (5.5)	28.2 (12.1)	17.4 (8.4)	11.3 (7.7)	19.3 (9.1)	OR	8.4 (4.7)	6.8 (3.3)	11.6 (4.1)	9.4 (5.1)	3.8 (0.9)	13.3 (5.9)
PA	14.8 (7.2)	11.5 (6.5)	16.2 (9.4)	17.4 (8.1)	11.3 (4.4)	21.3 (15.1)	SD	12.3 (5.0)	7.5 (3.5)	14.3 (6.2)	8.7 (4.1)	5.8 (2.7)	11.3 (4.3)
SC	4.6 (2.1)	1.5 (0.1)	10.0 (3.2)	5.3 (2.5)	3.8 (1.2)	6.3 (2.3)	TN	17.5 (7.9)	14.9 (6.5)	19.3 (7.2)	11.3 (5.0)	8.3 (3.9)	15.7 (6.0)
TX	16.3 (9.3)	11.5 (6.2)	19.0 (7.3)	17.2 (8.2)	14.2 (5.9)	21.4 (13.5)	WA	11.5 (5.7)	8.1 (4.1)	21.1 (9.5)	17.3 (8.3)	13.2 (7.2)	14.9(6.5)
WV	10.4 (5.4)	6.9 (2.7)	13.0 (5.3)	15.3 (8.4)	12.3 (5.9)	19.0 (7.2)	WI	4.4 (2.9)	1.5 (0.7)	10.0 (4.3)	5.4 (2.5)	3.8 (1.9)	6.3 (2.7)

than the Bayes classifier and the HMM method, and the HMM method had lower false alarm rate than the regression method and the Bayes classifier.

VI. RESULTS OF OTHER DISEASES

The search engine based system could also be implemented to monitor the alert levels of other infectious diseases. We examined the influenza data and the Lyme disease data from January 2006 to December 2011 in the U.S. With the same approach as the hepatitis disease, we built the surveillance models for the influenza and Lyme disease, respectively. The proposed HMM-based method could predict the alert levels of the influenza and Lyme disease with 91.7% and 84.7% accuracy in average, respectively. The comparative Bayes classification method only achieved 85.3% and 77.6% accuracy in average for the influenza and Lyme disease, respectively. The regression-based method showed less accurate results, with 81.3% and 74.2% accuracy in average for influenza and Lyme disease, respectively.

By using a Markov model to describe the disease transition characteristic, the proposed method becomes more robust against the irrelevant searches. For example, Fig. 8(a) shows that the search volume of the term "influenza" had a high point in May 2009 (point A). Note that the dramatic increase in search volume was not related to contemporary morbidity trend (thin black curve). In fact, the increase in search volume was caused by the media reports of an H1N1 outbreak in the country of Mexico [10]. Several infections in the U.S. were also reported, which caused a lot of attention and an dramatic increase in flu-related searches over the Internet. As one can see in the Fig. 8(b), the alert level of the first week in May 2009 was correctly predicted as level one, which was made not only based on contemporary search volume but also based on previous alert levels. On the other hand, the regression- based methods and classification-based methods wrongfully predicted the alert level as level three in May 2009.

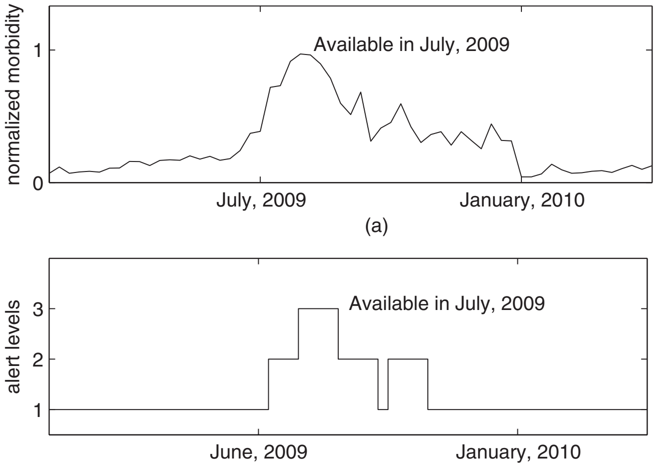


(a)

(b)

Fig. 8. Influenza epidemic detected by the search engine based system. (a) Normalized search volume and morbidity data. (b) Estimated alert levels.

As the motivation of our research, the proposed system could detect a potential epidemic before it was confirmed by the CDC. The term "epidemic" has different definitions under different circumstances. In this paper, we roughly define the period with the highest alert level as an epidemic. By definition, the epidemic period has the largest number of infections reported by the CDC. Figs. 8 and 9 show two examples of the real-time surveillance based on alert level prediction. In both figures, the upper thin black curves are the morbidity data reported by the CDC and the lower stairs are the predicted alert levels. The proposed system correctly predicted the influenza epidemic in November and the Lyme epidemic in late July 2009, which were both later confirmed by the CDC data. It is worth noting that the estimation results was given in real time, which was over a week ahead of the CDC reports.



(a)

(b)

Fig. 9. Lyme disease epidemic detected by the search engine based system. (a) Normalized morbidity data. (b) Estimated alert levels.

VII. Discussion

The timing and length of epidemic outbreaks vary from time to time, complicating the efforts to produce reliable and timely alarms for epidemic outbreaks. However, we were able to estimate the epidemic alert levels using the Google queries related to the infectious diseases. We built a syndromic surveillance system for the U.S. country and each state to provide the surveillance results in real time. The system used a continuous density HMM to incorporate the following two types of relations for disease surveillance purpose,

1) the relation between alert levels of continuous time steps;
2) the relation between alert levels and contemporary searches of disease-related terms.

Traditional disease surveillance networks publish the disease occurrence data on weekly (or monthly) basis, usually with weeks' reporting lag. Since Google trend data are updated on a daily basis, the surveillance results of our method are one to four weeks ahead of the CDC's reports. With weeks of lead time, public health officials could mount a more effective early response.

This article used hepatitis, influenza, and the Lyme disease as examples to explain the effectiveness of the proposed method, but it could be applied to monitor other notifiable infectious diseases too. The early detection provided by this approach may become an important line of defense against future epidemics in the U.S., and perhaps eventually in international settings, including those, which lack the infrastructure required for traditional surveillance. We believe the technology can be especially extended for the developing countries, such as China and India, because of their large numbers of infections and Internet users.

References

[1] D. Adams, M. Kathleen, R. Jajosky, J. Ward, P. Sharp, W. Anderson, J. Abellera, A. Aranas, M. Mayes, M. Wodajo, D. Onweh, and M. Park. (2010). Summary of Notifiable Diseases, United States, 2010. Centers for Disease Control and Prevention (CDC), Atlanta, GA. [Online]. Available: http://www.cdc.gov/mmwr/mmwr_nd/index.html

[2] S. Fox. "Online health search 2006," Pew Research Center's Internet American Life Project, Washington, DC, Oct. 2006.

[3] M. Bundorf, T. Wagner, S. Singer, and L. Baker, "Who searches the internet for health information?," *Health Serv. Res.*, vol. 41, pp. 819–36, Jun. 2006.

[4] J. Diaz, R. Griffith, J. Ng, S. Reinert, P. Friedmann, and A. Moulton, "Patients' use of the internet for medical information," *J. General Internal Med.*, vol. 17, no. 3., pp. 180–185, Mar. 2002.

[5] C. Cooper, K. Mallon, S. Leadbetter, L. Pollack, and L. Peipins, "Cancer internet search activity on a major search engine, united states 2001–2003," *J. Med. Internet Res.*, vol. 7, no. 3, p. e36, Jul. 2005.

[6] J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant, "Detecting influenza epidemics using search engine query data," *Nature*, vol. 457, pp. 1012–1014, 19 Feb. 2009.

[7] P. Polgreen, Y. Chen, and D. Pennock, "Using internet searches for influenza surveillance," *Clin. Infect. Dis.*, vol. 47, no. 11, pp. 1443–1448, Dec. 2008.

[8] K. Wilson and J. Brownstein, "Early detection of disease outbreaks using the internet," *Can. Med. Assoc. J.*, vol. 180, no. 8, Apr. 14, 2009.

[9] X. Zhou and H. Shen, "Notifiable infectious disease surveillance with data collected by search engine," *J. Zhejiang Univ. Sci. C, Comput. Electron.*, vol. 11, no. 4, pp. 241–248, Apr. 2010.

[10] "The CDC morbidity and mortality weekly report," *Novel Influenza A (H1N1) Virus Infection—Mexico, March–May, 2009*, Jun. 5, 2009, vol. 58, no. 21, pp. 585–589.

[11] J. R. Williams, D. J. Nokes, G. F. Medley, and R. M. Anderson, "The transmission dynamics of hepatitis B in the UK: A mathematical model for evaluating costs and effectiveness of immunization programmes," *Epidemiol Infect.*, vol. 116, no. 1, pp. 71–89, Feb. 1996.

[12] X. Zhou, J. Ye, and Y. Feng, "Tuberculosis surveillance by analyzing google trends," *IEEE Trans. Biomed. Eng.*, vol. 58, no. 8, pp. 2247–2254, Aug. 2011.

[13] T. Tassier, SIR Model of Epidemics, Anual report, 2005.

[14] Centers for Disease Control and Prevention(CDC). (2012, May 20) [Online]. Available in http://wonder.cdc.gov/mmwr/mmwrmorb.asp

[15] The Google Trend Service. (2012, May 20). [Online]. Available: http://www.google.com/insights/search/

[16] J. Gauvain and C. Lee, "MAP estimation of continuous density HMM: Theory and applications," in *Proc. Workshop Speech Natural Lang., 1992*, pp. 185–190.

[17] The Hepatitis. (2012, May 20). [Online]. Available: http://en.wikipedia.org/wiki/Hepatitis

Authors' photographs and biographies not available at the time of publication.