Influenza Surveillance Using Volunteered Geographic Information (VGI): A GIS-based Hidden Markov Modeling Approach

Wei Chen[1], Daniel Sui[2]

[1,2] Center for Urban & Regional Analysis (CURA), Ohio State University, Columbus OH 43210 U.S.A
[1,2] Department of Geography, Ohio State University, Columbus OH 43210 U.S.A

[1]Email: chen.1381@osu.edu
[2]Email:sui.10@osu.edu

1. Introduction

Due to the rapid development of a series of Web 2.0 technologies, the past five years have witnessed an explosive growth of geocoded user-generated contents on the Web – popularly known as the Volunteered Geographic Information (VGI) (Goodchild 2007; Elwood 2008). The wikification of GIScience in general and the growth of VGI in particular via cloud computing, open-source, and crowd-sourcing will transform GIScience in many fundamental ways in the years ahead (Sui 2008).

This paper develops a new approach of using VGI for flu surveillance by integrating GIS with a Hidden Markov Model (HMM). The particular type of VGI in this paper is Internet search. Disease related keyword search comprises both location information (in the form of IP address or geotags) and attribute information (in terms of disease related terms). Previous studies have exploited HMM in terms of monitoring disease surveillance data (Le Strat and Carrat 1999; Watkins, Eagleson et al. 2009) but data from Web is not discussed. Other studies explored the possibility of incorporating search data but data spikes problem is not solved (Carneiro and Mylonakis, 2009). The novelty of this paper lies in: first, word similarity analysis is introduced which was widely used by linguistics to process natural language but rarely borrowed by geographers; second, we implemented a HMM based on continuous real valued data and demonstrated its advantages and limitations.

Researchers from Google asserted that the search keywords they chose can predict ILI (Influenza-Like Illness) two weeks in advance (Ginsberg, Mohebbi et al. 2009). However, we do not want to simply accept this conclusion. Instead, we seek to extract new measures from sensor data that give different predictive powers. To achieve this goal, we analyzed the correlation between variables by testing on different steps of lagging. Initial results showed that a semantic combination of key words, rather than using key words directly, could extract sensors that correlate with ILI more significantly.

2. Data

2.1 Sensor data

Google keyword search is used to help us find out all flu related keywords. In total, 800 flu related search phrases are found and 190 significant key words are extracted. NLTK (Natural Language Toolkit), a python library is used to calculate the similarity between keywords (Loper and Bird 2002). Words with high semantic similarity (0.9)

are grouped into meaningful categories. Finally, four categories (*Prevention, Symptom, Treatment, Duration*) are chosen to build our model. The search volume of each category is used as one type of sensor data. An example of retrieving sensor *Symptom* from Google trend using a customized query expression is shown in Table 1.

Table1. Query expression for retrieving sensor *Symptom*.

Keyword Category	**Query** expression
Symptom	(flu|influenza|cold|colds cough|symptom|symptoms|diarrhea) - (swine|bird|avian|dog|dog|dogs)

2.2 Flu status data

Flu status data were obtained from the U.S. Centers for Disease Control & Prevention (CDC). It has 137 records spanning from week 40, 2004 to week 20, 2008.

3. Method

We have argued that search volumes can be "geocoded" to real world locations based on the IP address. We aim to build prediction model for each county that uses Google search engine but in this abstract only the US case is studied for domostration purpose. HMM is implemented as a probabilistic reasoning model to calculate conditional probability of flu status based on sensor data. Flu status is hidden in the sense that the number of infected population at the time of observation is always unknown; however, the consequence of the hidden status can be observed through the sensors, the volume of keyword search.

The simplest Markov chain assumes that current hidden status depends only on one previous status. This is called a first order Hidden Markov Model which is illustrated in Figure 1. X is flu status and E is sensor evidence.

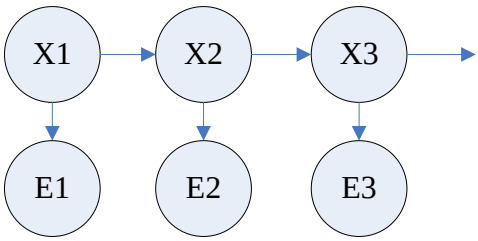

Figure 1. First Order Hidden Markov Model

3.1 Transition Model

The transition model in this first order HMM case is formulated as equation (1).

$$P(X_t|X_{0:t-1}) = P(X_t|X_{t-1}) \quad (1)$$

3.2 Sensor Model

The sensor model based on a single sensor and flu status are given as equation (2).

$$P(E_t|X_{0:t}, E_{0:t-1}) = P(E_t|X_t) \quad (2)$$

3.3 Probabilistic Reasoning

Before we build prediction model, it is necessary to learn following probability distributions: Prior distribution $P(x_t)$ and $P(e_t)$, transitional distribution $P(x_t|x_{t-1})$ and sensor distribution $P(e_t|x_t)$.

3.3.1 Prior distribution:

$$P(x_t) = \left(\frac{1}{\sigma_{x_t\sqrt{2\pi}}} e^{-\frac{(x_t-\mu)^2}{2\sigma {x_t}^2}} \right) \quad (3)$$

$$P(e_t) = \left(\frac{1}{\sigma_{e_t\sqrt{2\pi}}} e^{-\frac{(e_t-\mu)^2}{2\sigma {e_t}^2}} \right) \quad (4)$$

These two prior distributions belong to Gaussian distribution. In equation (3) and (4), μ is the mean of the variable, σ is the standard deviation, P is the probability density function.

For real valued data like ILI and search volumes, we introduce linear Gaussian model to calculate probability distribution.

3.3.2 Transitional distribution:

$$P(x_t|x_{t-1}) = N\left(ax_{t-1} + b, \sigma_{x_t}\right) \left(\frac{1}{\sigma_{x_t\sqrt{2\pi}}} e^{-\frac{\left(x_t-(ax_{t-1}+b)\right)^2}{2\sigma {x_t}^2}} \right) \quad (5)$$

In equation (5), the mean of current flu status x_t can be presented as a linear function of its previous status x_{t-1}.

3.3.3 Sensor distribution:

$$P(e_t|x_t) = N(ax_t + b, \sigma_{e_t}) \left(\frac{1}{\sigma_{e_t\sqrt{2\pi}}} e^{-\frac{(e_t-(ax_t+b))^2}{2\sigma {e_t}^2}} \right) \quad (6)$$

Similar to equation (5), in equation (6) the mean of e_t is presented as a linear function of x_t.

3.3.4 Prediction model

Two prediction models are implemented. They are based on one evidence (equation (7)) and two consecutive evidences (equation (8)) respectively.

$$P(x_t|e_t) = \int_{x_t=x_0}^{+\infty} \frac{P(e_t|x_t)P(x_t)}{P(e_t)} \quad (7)$$

$$P(x_t|e_t, e_{t-1}) = \int_{x_t=x_0}^{+\infty} \int_{x_{t-1}=-\infty}^{+\infty} \frac{P(x_t|x_{t-1})P(x_{t-1})P(e_t|x_t)P(e_{t-1}|x_{t-1})}{P(e_t,e_{t-1})} \quad (8)$$

4. Preliminary Results

Our preliminary results show that one step lagging is optimal for describing the autocorrelation of flu status (See Table 2). It was also found that different steps of lagging of flu status correlate differently with sensor variables. Although one step lagging is optimal for most sensors except *prevention*, it also shows some sensor variables correlated significantly with *ILI* even with two steps of lagging. This makes it possible to predict flu at an earlier stage based on these sensor variables. The results of lagging analysis are summarized in Table 3.

Table 2. Autocorrelation of flu status variable.

Lagged flu status	Flu status	Lag	R	P
ILI	ILI	1	0.97	0.00

Table 3. Comparison of different steps of lagging between flu status and search sensors.

Lagged flu status	Sensor	Lag	R	P
ILI	prevention	16	0.77	0.00
ILI	symptom	1	0.94	0.00
ILI	symptom	2	0.90	0.00
ILI	treatment	1	0.95	0.00
ILI	treatment	2	0.90	0.00
ILI	duration	1	0.85	0.00

We test our model by using *symptom* sensor from week 1, 2008 to week 20, 2008. ILI baseline 2 is used as the threshold of alarming for high flu activity. Probability of ILI exceeding the threshold is calculated based on prediction models (7) and (8).

Table 4 shows the prediction result. It can be seen that probability of ILI exceeding the threshold increases with the increase of sensor value. We achieved a 100% correct alarming rate on our test data. Given more sensors, prediction accuracy can be improved as illustrated by the probability values in Table 4. Figure 2 is a graphical representation of results in Table 4.

Table 4. Prediction model results.

Week	Symptom	ILI	P(ILI>2|e1)	P(ILI>2|e1,e2)	Alarm?
1	0.88	2.447	0.631443		
2	0.9	2.307	0.555164	0.664520898	Yes
3	1.06	2.654	0.570682	0.685182294	Yes
4	1.34	3.971	0.68908	0.826774195	Yes
5	1.76	5.031	0.851925	0.959524029	Yes
6	1.84	5.743	0.969393	0.998414888	Yes
7	2.12	5.964	0.978797	0.999264536	Yes
8	1.82	5.623	0.995074	0.999966125	Yes
9	1.4	4.499	0.976711	0.999104765	Yes
10	1.16	3.828	0.877565	0.972407038	Yes
11	0.96	3.219	0.754966	0.890375623	Yes
12	0.82	2.538	0.616494	0.743592386	Yes
13	0.68	2.073	0.508196	0.599716343	Yes
14	0.58	1.673	0.399286	0.440684304	No
15	0.54	1.313	0.325579	0.331321011	No
16	0.46	1.135	0.297731	0.290788755	No
17	0.44	0.981	0.245568	0.217486819	No
18	0.4	0.87	0.233339	0.200982828	No
19	0.38	0.824	0.209926	0.170312926	No
20	0.36	0.802	0.198757	0.156167131	No

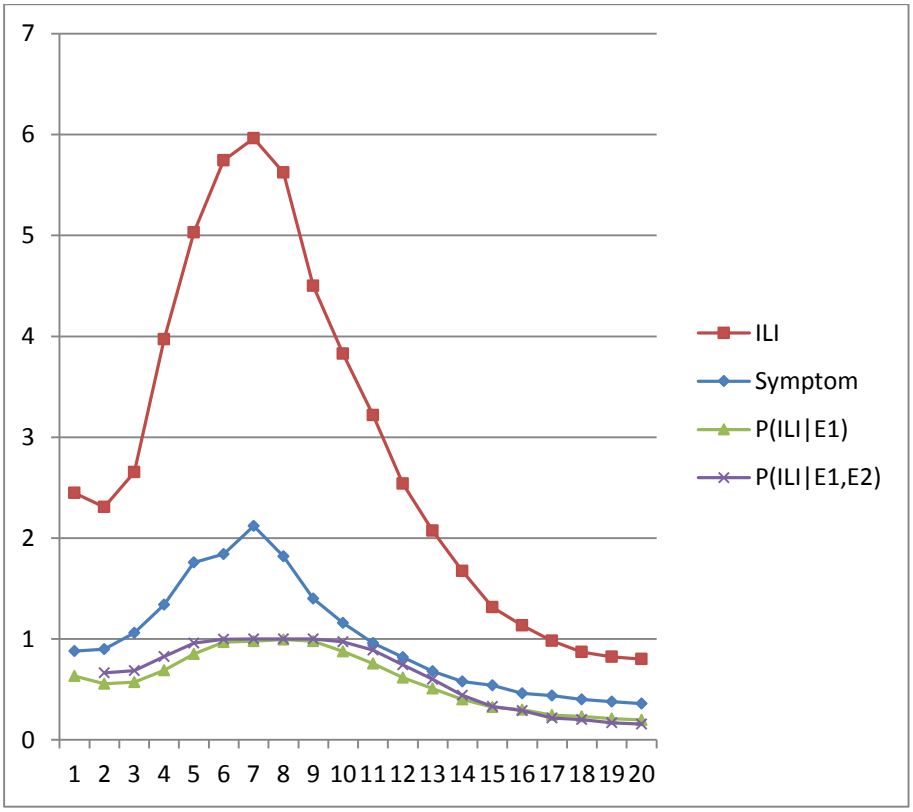

Figure 2. Prediction model results

5. Conclusion

This paper examines the effects of combining search keywords by semantic similarity. Preliminary results show that different steps of lagging can be discovered between flu and sensor variables. HMM can alarm flu activity accurately through probabilistic reasoning. We aim to incorporate GIS in our next step of work to discover the spreading pattern of flu activities geographically. Our research also demonstrated VGI is highly potential data source that may be used to reveal hidden patterns in the population.

References

Elwood, S. 2008. Volunteered geographic information: future research directions motivated by critical, participatory, and feminist GIS. *GeoJournal* 72(3): 173-183.

Ginsberg, J., M. H. Mohebbi, et al. 2009. Detecting influenza epidemics using search engine query data. *Nature* 457(7232): 1012-U1014.

Goodchild, M. 2007. Citizens as sensors: the world of volunteered geography. *GeoJournal* 69(4): 211-221.

Le Strat, Y. and F. Carrat. 1999. Monitoring epidemiologic surveillance data using hidden Markov models. *Statistics in Medicine* 18(24): 3463-3478.

Loper, E. and S. Bird 2002. NLTK: The natural language toolkit, Association for Computational Linguistics.

Sui, D. 2008. The wikification of GIS and its consequences: Or Angelina Jolie's new tattoo and the future of GIS. *Computers, Environment and Urban Systems* 32(1): 1-5.

Watkins, R. E., S. Eagleson, et al. 2009. Disease surveillance using a hidden Markov model. *Bmc Medical Informatics and Decision Making* 9(-.