Influenza Patients Are Invisible in the Web: Traditional Model Still Improves the State-of-the-Art Web Based Influenza Surveillance

Eiji ARAMAKI*, **, Sachiko MASKAWA*, Mizuki MORITA***
* Center for Knowledge Structuring, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo, Japan.
** Precursory Research for Embryonic Science and Technology (PRESTO), Japan Science and Technology Agency (JST).
*** Department of Fundamental Research, National Institute of Biomedical Innovation (NIBIO), 7-6-8 Saito-Asagi, Ibaraki, Osaka, Japan.
eiji.aramaki@gmail.com

Abstract
Although web-based information extraction systems draw much attention, most of such systems assume that the web directly reflects the real world. For instance, Google flu trend, which is one of the state-of-the-art influenza surveillance systems, relies on the basic idea that the amount of the influenza related search queries directly correlates with the number of the influenza patients. However, the real patients suffering from influenza symptoms are invisible in the web, because they do not use Internet. Considering this gap, this paper employs an infectious model, assuming that a potential patient utilizes Internet at the first sign of flu. The proposed model improves two types of the state-of-the-art systems, Google based system (from 0.837 correlation to 0.928) and Twitter based system (from 0.898 correlation to 0.918). This study demonstrated that a simple model could easily improve the web-based surveillance.

Introduction
Influenza epidemics are a major public health concern, because it causes tens of millions of respiratory illnesses each year [1][1]. To reduce the victims, the early detection of influenza epidemics is a national mission in every country(2, 3), motivating various influenza surveillance organizations: the U.S. has the Centers for Disease Control and Prevention (CDC), the E.U. has its European Influenza Surveillance Network (EISN), and Japan has its Infection Disease Surveillance Center (IDSC). Their surveillance systems fundamentally rely on both virology and clinical data. For example, the IDSC gathers influenza patient data from around 5,000 clinics and releases summary reports once a week. Such manual systems typically have a 1–2 week reporting lag, which has been pointed out as a major flaw.

In an attempt to provide earlier influenza detection, various new approaches are proposed. Espino et al. (4) focuses on a telephone triage service. They reported the number of telephone calls correlates with influenza epidemics. Magruder (5) used the amount of over-the-counter drug sales. Johnson et al (6) utilizes the web access log of a health care site. The last approach has received much attention because the scale and the rapidness outperform the other materials, motivating various follow up studies.

Currently, most of web-based methods are classified into two types: (1) a search query-based (7-10) and (2) a micro blogging based (11-14).
1 Query-based approach: The query-based approach assumes that the number of influenza related queries correlate with the number of influenza patients. Among various web search engines, Google based approach (10) achieved the highest accuracy.
1 Micro blogging based approach: Another approach, micro blogging based, relies on the assumption that the influenza related post reflects the epidemics. Currently, Twitter based approach (14) achieved the highest accuracy.

Although the above two approaches use different materials, they share the same idea to catch an action of influenza patient via the Internet. However, this idea implies an unnatural situation that an influenza-infected person uses a computer when he/she are suffering from influenza symptoms, such as a high fever, nausea, chill and so on.

This study employs more natural assumption that a person uses Twitter or Google at the first sign of flu. To implement this assumption, we utilize a traditional infectious model (15) that was proposed in 1927. The
Model classified people into three categories, (S) people who are still susceptible, (I) people who are infected, and (R) people who have recovered. We basically rely on the model, but we newly assume that new patient (S-to-I transition) could be obtained from the web. Because another transmission ratio (I-to-R transition) is well known in the previous study (16). The both transitions enable to estimate the current number of infection patients.

In the experiments, the proposed model improves two types of the state-of-the-art systems, Google based system (from 0.382 correlation to 0.843) and Twitter based system (from 0.851 correlation to 0.917), demonstrating the validness of our assumption that people use Internet at the first sign of flu.

**Material and Methods**

Google data: Google, a popular web-search engine, has been collecting the queries from all over the world. By using the data, Google also provides various web services. One of such services is Google Flu Trends, which is the state-of-the-art query based influenza surveillance (10). The method utilizes the frequency of the highest k influenza correlated queries. Although the query data is not publicly available, the estimated results of Google Flu Trends are published on the web. For this study, we obtained the data starting from 2008 to 2010.

Twitter data: Twitter, a popular micro-blogging service, increased the user rapidly. Its community estimated as 120 million worldwide, posts more than 5.5 million messages (tweets) every day (17). By using this massive data, several Twitter based influenza surveillance is proposed (11-14). Among them, Aramaki et al. (14) achieved the highest accuracy. Their method regards that the frequency of the influenza related tweets reflects the influenza epidemics. For this study, we collected 300 million tweets, starting from 2008 to 2010 via Twitter API. Then, we applied the Aramaki’s algorithm to them.

Original SIR Model: To produce the number of influenza patients, we utilized the SIR model, which is proposed by Kermack and McKendrick (15). The original model divides population into three distinct groups: susceptible, S(t), infected, I(t), and recovered, R(t).

1. S(t) is the number of individuals not yet infected with the disease at time t.
2. I(t) denotes the number of individuals who have been infected and are capable of spreading the disease to those in the susceptible category.
3. R(t) is the compartment used for those individuals who have been infected and then recovered from the disease. Those in this category are not infected again in the same season.

Using a fixed population, S(t)+I(t)+R(t), derived the following equations:

\[
\begin{align*}
\frac{dS}{dt} &= -\beta IS \\
\frac{dI}{dt} &= \beta IS - \nu I \\
\frac{dR}{dt} &= \nu I 
\end{align*}
\]

where \(\beta\) is the infection rate of the disease (the probability of the S-to-I transition), \(\nu\) is the recover ratio (the probability of the I-to-R transition).

Modified SIR Model: Our approach is basically based on the original SIR model, but three extra assumptions were made for this study (Figure 1).

1. A person who has just caught the influenza (S-to-I transition) is observed by the web-based surveillance systems (\(\beta IS\) is obtained by the systems).
2. The I-to-R transition ratio \(\nu\) is a fixed value. We checked the various values (such as \(\nu=0.20\) and \(0.38\)), which comes from the previous infectious disease study (15).
3. At the start of each season, there is no persons belong to Infected and Recovered (I(t)=0 and R(t)=0)).

Due to these assumptions, we could estimate the number of influenza patients during the entire season. We apply the above model to the following two previous methods, GOOGLE (10) and TWITTER (14). We call new methods respectively GOOGLE+SIR and TWITTER+SIR.

![Figure 1: The proposed infectious model. This model assumes that the S-to-I transition is observed via Twitter, and another transition, the I-to-R transition, is fixed (\(\nu=0.38\) in this figure) [19].](image-url)
Table 1. The Correlation Ratio between the number of influenza patients (gold standard data) and each method prediction (four methods). The SIR model significantly improves the correlation of both the micro blogging based and query based.

<table>
<thead>
<tr>
<th></th>
<th>Micro blogging based</th>
<th>Query based</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>TWITTER</td>
<td>TWITTER+SIR (( \nu =0.20 ))</td>
</tr>
<tr>
<td>2009 Winter</td>
<td>0.809</td>
<td>0.856</td>
</tr>
<tr>
<td></td>
<td>(+0.047)</td>
<td>(+0.027)</td>
</tr>
<tr>
<td>2010 Winter</td>
<td>0.962</td>
<td>0.976</td>
</tr>
<tr>
<td></td>
<td>(+0.014)</td>
<td>(+0.013)</td>
</tr>
<tr>
<td>All</td>
<td>0.898</td>
<td>0.918</td>
</tr>
<tr>
<td></td>
<td>(+0.020)</td>
<td>(+0.016)</td>
</tr>
</tbody>
</table>

* The number in bold indicates the significance correlation \((p=0.05)\). The number with asterisk indicates the significant improvement compared than without model methods.

Figure 2: The Influenza Timeline of 2009 (top) and 2010 (bottom). The gray area indicates the gold standard (the number of patients reported by the real surveillance). The other lines are predicted values. The X-axis shows the date and the Y-axis shows the number of influenza patients (normalized by the max value in the season). Web-based surveillance systems (GOOGLE and TWITTER) assume that web information directly reflects the infectious epidemics. In reality, web information precedes the real epidemics. The infectious model solves this time lag (GOOGLE+SIR and TWITTER+SIR).
Results

We investigated the performance using actual influenza reports provided by the Japanese IDSC (gold standard) mentioned before. We separated the above data into two periods as follows:

- 2009 Winter (from Nov. 9, 2008 to Apr 5, 2009)

The results are presented in Table 1. The SIR model based method (TWITTER+SIR and GOOGLE+SIR) shows higher correlation than methods without the model respectively (TWITTER and GOOGLE). The improvement of GOOGLE+SIR (+0.092 point) is significant (p=0.05), demonstrating the validness of the proposed approach. The improvement of TWITTER+SIR (0.020 point) is not significant. One of the reasons is that the number of sample is not enough.

Conclusions

This paper proposes the web based influenza detection method that considers a time lag of infections onset. The proposed method employs the SIR infection disease model. The experiment results demonstrate the feasibility of the proposed model. The result suggests that people do Internet action, such as tweet or Google search, shortly before catching the flu.

References