2012 Fifth International Symposium on Parallel Architectures, Algorithms and Programming

Prediction of Infectious Disease Spread using Twitter: A Case of Influenza

Hideo Hirose, Liangliang Wang
School of Computer Science and Systems Engineering
Kyushu Institute of Technology
Fukuoka, 820-8502 Japan
Email: hirose@ces.kyutech.ac.jp

***Abstract*—Nowadays, detecting the disaster phenomena and predicting the final stage become very important in the risk analysis view-point. The statistical methods provide accurate estimates of parameters when the data are completely given. However, when the data are incomplete, the accuracy of the estimates becomes poor. Therefore, statistical methods are weak in predicting the future trends. The SIR methods, for infectious disease spread prediction, using the differential equations can sometimes provide accurate estimates for the final stage. These methods, however, require some inspection time, which means the delay of analysis at least one week or so when we want to predict the future trends. To detect the disasters and to predict the future trends much earlier, we can use the social network system (SNS).**

In this paper, we have proposed a method to predict the future trend of influenza by using Twitter. We have analyzed the possibility of building a regression model by combining Twitter messages and CDC's Influenza-Like Illness (ILI) data, and we have found that the multiple linear regression model with ridge regularization outperforms the single linear regression model and other un-regularized least squared methods. The model of multiple linear regression with ridge can notably improve the prediction accuracy.

***Index Terms*—Twitter; early detection; influenza; infectious disease; logistic regression; ridge; ILI; AIC; SNS; truncated data.**

I. Introduction

In these days, we feel that large-scale epidemic occur more frequently than those days (e.g. [21]); one reason, among many others, is due to the human behaviors such as the large amount of energy consumption. The climate change and accidents at nuclear-power generation plants are examples of such things. We are becoming nervous to possible big disasters in the future. Therefore, in a risk analysis sense, early detection of disasters becomes much more important nowadays. See Figure 1.

After disasters ceased, analyses using the statistical methods are often performed, and rather accurate estimates for parameters in the underlying probability distributions were obtained. However, when we want to obtain the estimates in the middle stage of a disaster, the data becomes incomplete, resulting in inaccurate estimates. In the risk analysis viewpoint, it would be pity that the scale of disaster is estimated to be smaller than that of the actual one. This kind of phenomena is shown in [13], [14]. According to [11], [12], for example, when the censoring time is before the half time of the final steady stage,

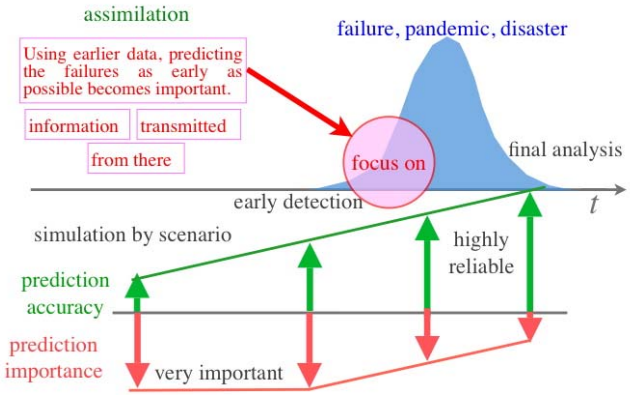

Fig. 1. Importance of early detection of disasters.

the maximum likelihood estimate of the parameter for fatality ratio in SARS is not stably obtained. This indicates that the statistical estimation procedure is very sensitive to the early censoring time in the truncated model or trunsored model; see also [10], [19], [20] for estimability of the maximum likelihood estimates for parameters.

On the contrary, the SIR models [6], [8], [9], [17], [18] and its stochastic models [5] having structures described by the differential equations can provide rather accurate values for the cumulative number of infectious persons even though the censoring time is early. This kind of methods is often used in simulations by scenario because of low computing cost. In Figure 1, the simulations performed in very early stages are illustrated. Another approach is to use the agent methods, in which artificial human beings are moving randomly but following some probability distributions in a computer; see [23], [25]. Such kinds of simulations are also used for simulations by scenario. In these two methods, a consistency between the SIR model and the agent model is observed [24].

The fourth method is to use the internet [1], [2], [3]. In this paper, we have used the social network systems (SNS) to make much earlier disaster detection and prediction. Why it is applicable for early detection is that people in the SNS immediately and promptly respond to other people. In contrast to this, in the cases using reports by medical examination, it

would take at least one week to analyze the observed data.

In this paper, we collected Twitter data, and compared the data with Centers for Disease Control (CDC)'s data in the United States. From building a single linear regression model to multiple linear regression model, we have investigated the methods to predict the future trend of influenza infection spread at early stage accurately. These methods include the logistic regression with the regularization method. The novel approach appears in this paper is to use the regularized method, the ridge regression to improve the prediction accuracy.

II. Data Collection

A. Influenza-Like Illness Data

America's Centers for Disease Control and Prevention (CDC) is keeping the whole nation's influenza epidemic trend under surveillance weekly. The hospitals, clinics and other medical institutes all over the America are reporting the number of patients who have influenza-like symptoms to CDC every week. This influenza surveillance program is called Influenza-Like Illness Surveillance Network (ILINet) [7]. The definition of probability of ILI is shown as follows:

$$P = \frac{\text{number of infected persons}}{\text{number of reported persons}}. \quad (1)$$

In Figure 2, we can see the ILI Rates between the end of 2011 and April 2012 [7]. CDC publishes two kinds of ILI rate; one is unweighted and the other is weighted. Weighted ILI Rate means that it was weighted over the state's population; unweighted, the simple mean. The red line with squared dots is for weighted ILI rate, while the blue line with diamond dots is for unweighted ILI rate.

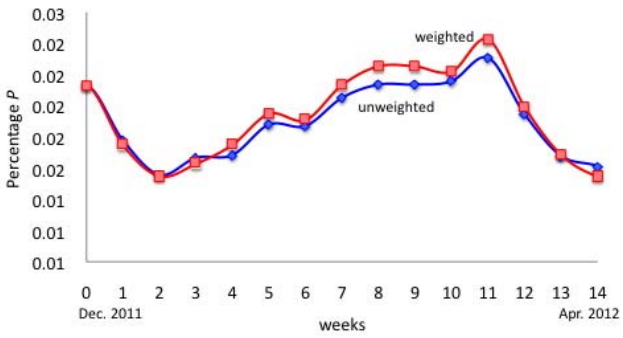

Fig. 2. CDC's ILI Rates: red line with squared dots is for weighted, blue line with diamond dots is for unweighted [7].

B. Influenza Surveillance on Internet

Recently, there has been an increasing tendency of monitoring epidemic trend based on data from the Internet. Most of these researches are monitoring the influenza trend by analyzing news articles, blogs, search engine queries and other data from the Internet. Google Inc. analyzed its query log data which are considered to be connected to influenza, and then published its influenza surveillance program called Google Flu Trend [28]. Google Flu Trend achieved considerably high prediction accuracy for Google's heavy use countries such as the United States and Europe. In Figure 3, we can see the prediction of Google Flu for the United States. The orange line is the real influenza trend, and the blue one is Google's prediction. We can see that Google's prediction fits very well.

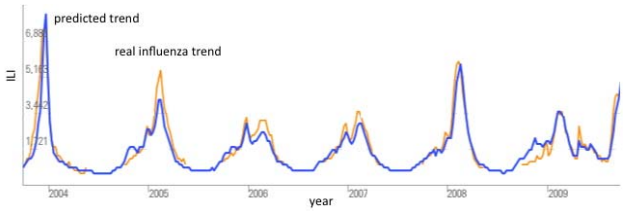

Fig. 3. Google Flu Trend's Influenza ILI trend and its prediction [28].

Meanwhile, the social network society is becoming to play a very important role in our daily life now, people who use Facebook, Twitter and other SNS services are creating tremendous data every day. Research works aimed on stock market prediction [16] and influenza trend prediction [1], [3] has already appeared. Although Google can use its query log data to predict the influenza trend easily, we cannot do that. Instead, we consider the use of Twitter's data as an alternative.

C. Twitter Data

Twitter is a SNS service which people can send and read messages of 140 letters [31]. The message of Twitter is called tweet. Since its foundation from July 2006, Twitter has become a very important and influential information source. In Figure 4, we can see that during the beginning of 2010, Twitter experienced a rapid grow phase. In January 2010, the daily tweet amount had already breakthrough 45 million [32]. Semiocast's latest study reveals Twitter reached the half billion accounts mark in June 2012, including more than 140 million in the U.S. alone [30].

While America's Congress Library is taking the backup of daily tweets, considering the enormous amount, it is very difficult for a single college lab to collect the whole portion of these tweets, so we decided to do samplings. We collected tweets with specific keywords in it. These six keywords are "I", "is", "my", "the", "to", and "you."

We used open source Twitter api application and MySQL database to build our data collection server. The application called 140dev Twitter API Framework was gotten from Adam Green's website [26], and we used Xampp [33] as our web server application. The data collection continued 1 minute per hour and 24 hours a day from the end of 2011 to April 12th 2012. As shown in Figure 5, the whole data collection process lasted 18 weeks, and at last we got a dataset of 8,635,624 tweets. Considering the server crash problem during the beginning and the end of data collection, we used tweets collected from Dec 25th 2011 to April 7th in this paper.

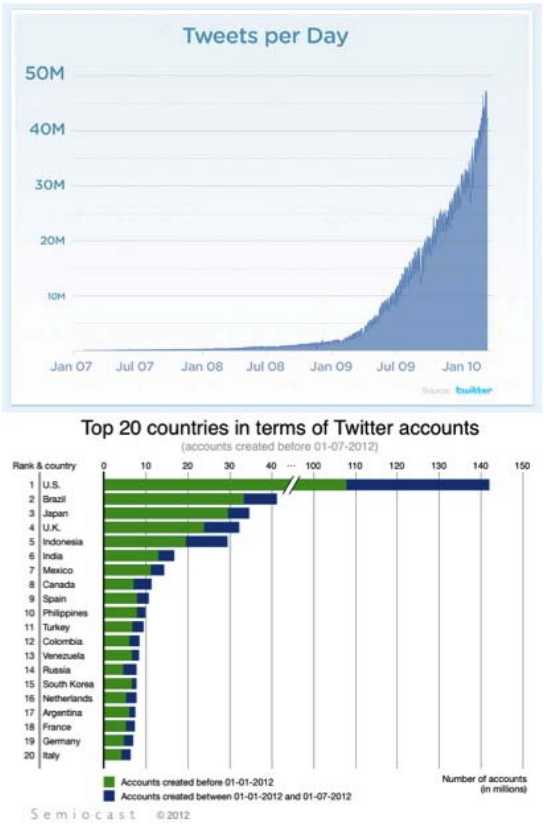

Fig. 4. The growth of Twitter [31] and number of tweets [30].

Meanwhile, CDC publishes the ILI rate on its web site every week. We used the ILI rate over the same period as Twitter data we collected.

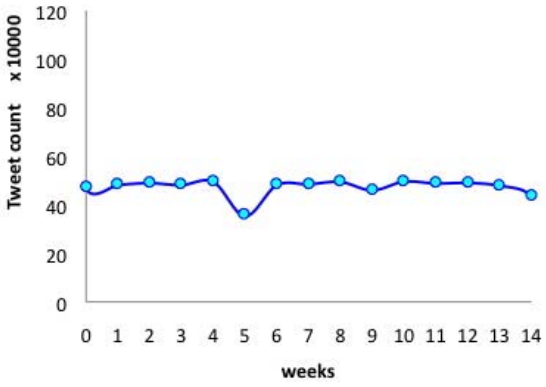

Fig. 5. Twitter data collection process.

We also considered only the tweets in English communicated in the United States rather than the unspecific tweets, because the American users send 80% of the English tweets. We can learn the location information from tweets sent from the devices with location detecting function. We counted the number of these tweets and found that its portion is at most 1.1%. Thus, we did not consider the location of tweets being sent, although it has the valuable information.

III. Simple Linear Regression Model

First, we followed Google Inc.'s Google Flu Trends Project [3] and tried to build a simple linear regression model. The model is shown as,

$$\text{logit}(P) = \beta_0 + \beta_1 \text{logit}(Q(W, D)), \quad (2)$$

$$\text{logit}(P) = \log(\tfrac{P}{1-P}),\ Q(W, D) = \tfrac{|W|}{|D|}. \quad (3)$$

In this formula, P is the ILI rate we obtained from CDC's web site and it is defined as the percentage of people who really suffered from influenza over people who was reported to the ILI Net; here, "$|\cdot|$" means the counted number. We use $Q(W, D)$ to represent the ratio of tweets containing keyword W over the total collected tweet amount of D. Here, W is one of a set of keywords: "flu", "cough", "headache", "fever", "sore throat".

IV. Document Filtering

A document being spurious means that this document is not describing symptoms of people who got influenza but implying other feelings. In our data collection, there are a lot of spurious messages. In Table 1, we can see that spurious messages are not talking about influenza symptoms but people's embarrassments.

In our data collection, the tweets containing these keywords may not be really connected with influenza symptoms. We call these tweets as *negative* samples. On the contrary, the tweets really connected with influenza symptoms is called as *positive*.

We show tweets in Table 1 as negative tweets, and in Table 2 as positive ones.

TABLE I
Negative Message Samples

RT @TheAwkwardTweet: The awkward moment when you start having a random cough attack in the middle of class.
Had been a belieber - 739 days , 63,849,600 seconds , 1,064,160 minutes, 17,736 hours & 105 weeks @Justinbieber . Bieber Fever never ends
i love it when you give me headddddddd! I just hate it when you give me headaches!
Dad wants me to upload songs onto his MP3 for work. Secretly putting Justin Bieber songs on it... Hope your coworkers have Bieber Fever!

TABLE II
Positive Message Samples

Headache, stuffy nose, wishing I had a hot
Woke Up With A Massive Headache , The Chills & The Worst Sore Throat EVER ..... Yep Im Dying :((
Officially not feeling well and fighting a sore throat :(
Went to work fine this morning. Was there for an hour and had to lie down because I thought I was going to throw up. Damn you flu!

We only want to count the number of positive messages; in order to achieve this, we have to classify messages containing influenza-related keywords into positive or negative. Here, the

ratio of one message being positive is defined as the following equation:

$$p(y_i = 1|x_i; \theta) = \frac{1}{1 + e^{(-x_i\theta)}}. \quad (4)$$

In this equation, $x_i = \{x_{ij}\}$ is a vector, and x_{ij} means the frequency of keyword j appearing in document i. Here, θ is a parameter we can learn by using natural language processing toolkit. We can compute the parameter θ using Mallet Machine Learning Toolkit [29].

We computed the ratio of being positive message for each tweet, then we can count the real influenza-related messages. After the computation, we redefine the meaning of being positive as the ratio of being positive message is bigger than 0.5 as follows:

$$Q(W, D) = \frac{\sum_{d_i \in D_W} 1(p(y_i = 1|x_i; \theta) \geq 0.5)}{|D|}. \quad (5)$$

In order to screen out these negative messages, we used Mallet Machine Learning Toolkit to do the document classification work. As shown in Table 3, we can find out that the correlation keywords and weighted ILI rate is considerably improved, and we obtained the largest correlation of 0.5562 between cough and the weighted ILI Rate. We can see that there are a lot of rooms for improvements on the prediction accuracy comparing to the single linear regression model.

TABLE III
CORRELATION BETWEEN KEYWORDS AND WEIGHTED ILI RATE BEFORE AND AFTER FILTERING

keywords	before filtering	after filtering
flu	0.1696	0.2431
cough	0.4346	0.5562
sore throat	0.2179	0.1796
headache	−0.0872	−0.1389
fever	−0.0076	0.2804

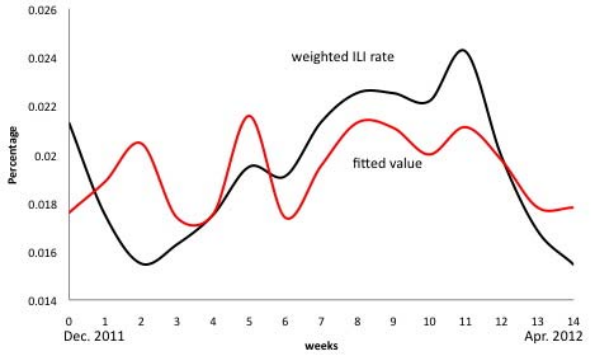

Fig. 6. Fitting result of keyword "cough" with weighted ILI rate

In Figure 6, we can see the fitting result of keyword "cough" with weighted ILI rate, which shows a disappointed result.

V. MULTIPLE LINEAR REGRESSION MODEL

Considering the combination of keywords, it is natural to expand the single linear regression model to a multiple linear regression model [4]. The regression model is shown as,

$$\text{logit}(P) = \beta_0 + \beta_1 \text{logit}(Q(W_1, D)) + \beta_2 \text{logit}(Q(W_2, D)) + \cdots. \quad (6)$$

Here, we used all five keywords as our parameters and the fitting result of the multiple regression model is shown as Figure 7. We split the whole collected data set into two data sets. The first 10 weeks' data is seen as training data while the following five weeks' data is treated as testing data. The dotted line means the prediction using the test data.

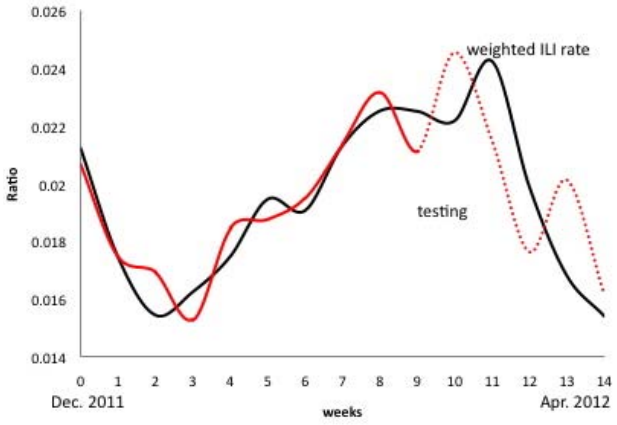

Fig. 7. Fitting result of multiple regression model

VI. MODEL SELECTION BASED ON AIC

By continuously adding parameters, we can build a complicated model to reach the perfect fitting on the training data, but it cannot be applied on the unknown future. To avoid overfitting problem, we applied the AIC (Akaike's Information Criterion) as our model selection standard. By using the AIC model selection, we finally obtained a multiple regression model which has three keywords in it. These are "fever", "flu", and "cough."

VII. RIDGE REGRESSION

Since we have obtained a desirable regression model, we considered strengthening the prediction accuracy. Here, we proposed the ridge regression shrinkage method [4]. Comparing to the traditional solution for coefficients of the linear regression model, the ridge regression adds a diagonal matrix as penalty and its calculation method is shown as,

$$\beta_{ridge} = (X^{'}X + \lambda I)X^{'}Y. \quad (7)$$

We call λ the tuning parameter. By changing λ we can find the proper value to minimize prediction error. The ridge regression is proved to be more effective than the un-regularized

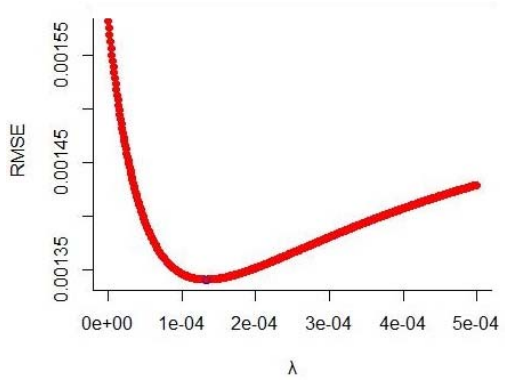

Fig. 8. How λ influence the prediction accuracy

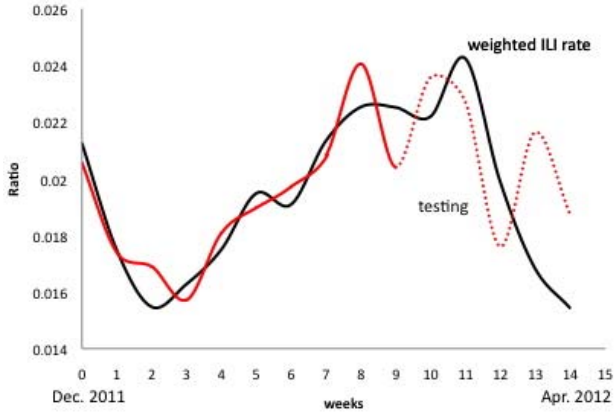

Fig. 9. Ridge regression on weighted ILI rate

least squared method. Figure 8 displayed how λ effected the prediction for root mean square error (RMSE).

Figure 9 shows the fitted value both on training data test data. The computational results in multiple regression analyses are summarized as shown in Table 4. The ridge regression obtains the lowest error on the test data.

TABLE IV
THE SUMMARY OF MULTIPLE LINEAR REGRESSION ANALYSES.

Model	$RMSE$	$RMSE_{ridge}$	Improvement
Weighted	0.002434	0.002234	8.22%
Weighted(AIC)	0.003903	0.002950	24.40%
Unweighted	0.002237	0.001897	15.22%
Unweighted (AIC)	0.003151	0.002356	25.23%

VIII. CONCLUSION

In this paper, we analyzed the possibility of building a regression model by combining Twitter messages and CDC's ILI data. We found that the multiple linear regression model with ridge regularization outperforms the single linear regression model and other un-regularized least squared methods. The model of the multiple linear regression with ridge can notably improve the prediction accuracy.

REFERENCES

[1] A. Culotta, "Detecting influenza outbreaks by analyzing Twitter messages," Science, 16, Issue: May, 1-11, 2010.
[2] A. Culotta, "Towards detecting influenza epidemics by analyzing Twitter messages," Proceedings of the First Workshop on Social Media Analytics (SOMA'10), pp. 115-122. Pages 115-122, 2010.
[3] J. Ginsberg, et.al., "Detecting influenza epidemics using search engine query data," Nature 457, 1012-1014, 2009.
[4] A. E. Hoerl, R. W. Kennard, (1970). "Ridge Regression: Biased Estimation for Nonorthogonal Problems," Technometrics, 12, pp. 55-67.
[5] H. Hirose, and Y. Maki, "Pandemic analysis using the SIR model with stochastic terms, Proceedings of the Annual Conference of JSIAM, 14p-G3-3, 121-122, 2011.
[6] R. Anderson and R. May, *Infectious diseases of humans: Dynamics and control*, Oxford University Press, 1991.
[7] http://www.cdc.gov/flu/weekly/, CDC Weekly Report.
[8] L.R. Elveback, J.P. Fox, E. Ackerman, et al. "An influenza simulation model for immunization studies," *American Journal of Epidemiology*. vol. 103, pp. 152-65, 1976.
[9] O. Diekmann and J.A.P. Heesterbeek *Mathematical epidemiology of infectious diseases: model building, analysis and interpretation*, New York: Wiley, 2000.
[10] W. L. Deemer Jr., D. F. Votaw, Jr., "Estimation of Parameters of Truncated or Censored Exponential Distributions," *Ann. Math. Statist.*, 26, 498-504, 1955.
[11] H. Hirose, "The mixed trunsored model with applications to SARS," *Mathematics and Computers in Simulation*, vol. 74, pp. 443-453, 2007.
[12] H. Hirose, "Estimation for the size of fragile population in the trunsored and truncated models with application to the confidence interval for the case fatality ratio of SARS," *Information*, vol. 12, pp. 33-50, 2009.
[13] H. Hirose, Estimation of the number of failures in the Weibull model using the ordinary differential equation, European Journal of Operational Research, to appear.
[14] H. Hirose, Parameter estimation for the truncated Weibull model using the ordinary differential equation, International Conference on Computers, Networks, Systems, and Industrial Engineering (CNSI 2011), pp.396-399, May 23-25, 2011.
[15] C.Y. Huang, C.T. Sun, J.L. Hsieh, Y.M.A. Chen and H.L. Lin, "A novel small-world model: Using social mirror identities for epidemic simulations," *Simulation*, vol. 81, pp. 671-699, 2005.
[16] Johan Bollen,Huina Mao,Xiao-Jun Zeng, "Twitter mood predicts the stock market", Journal of Computational Science, 2, pp. 1-8, 2011
[17] W. O. Kermack, and A. G. McKendrick. "Contributions to the mathematical theory of epidemics - I," Proceedings of the Royal Society Series A, 115, 700-721, 1927.
[18] W. O. Kermack, and A. G. McKendrick, "Contributions to the mathematical theory of epidemics-III. Further studies of the problem of endemicity," *Proceedings of the Royal Society*, vol. 141A, pp. 94-122. 1933.
[19] Y. Komori, H. Hirose, "Parameter estimation based on grouped or continuous data fo truncated exponential distributions," *Comm. Stat. - Theory and Method*, 31, pp. 889-900, 2002.
[20] M. M. Mittal, R. C. Dahiya, "Estimating the parameters of a truncated Weibull distribution," *Comm. Statist. - Theory Method*, 18 (6), 2027-2042, 1989.
[21] J. Matson and J. Pavlus, Laying Odds on the Apocalypse, Scientific American 303, 82-83 2010.
[22] Y. Toyosaka and H. Hirose, "The consistency between the two kinds of pandemic simulations of the SEIR model and the MAS model," *the 9th International Conference on Computers, Communications and Systems (ICCCS 2008)*, Nov. 7, 2008.
[23] Y. Toyosaka and H. Hirose, "Pandemic simulations by MADE: a combination of multi-agent and differential equations," *The 2009 International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA2009)*, pp. 331-335, 2009.
[24] Y. Toyosaka and H. Hirose, "The consistency of the pandemic simulations between the SEIR model and the MAS model," *IEICE Transactions on Fundamentals*, Vol.E92-A, No.7, pp. 1558-1562, 2009.

[25] Y. Toyosaka and H. Hirose, “Pandemic Simulations by MADE: the Hybrid Method of Multi-Agent and Differential Equations,” *Asia Simulation Conference 2009 (JSST2009)*, October, 2009.
[26] Adam Green, Twitter Database Server http://140dev.com/free-Twitter-api-source-code-library/Twitter-database-server/
[27] Audrey Watters, How the Library of Congress is building the Twitter archive. http://radar.oreilly.com/2011/06/library-of-congress-Twitter-archive.html, 2011.
[28] http://www.google.org/flutrends/ Google Flu Trends Project.
[29] http://mallet.cs.umass.edu/
[30] http://semiocast.com/
[31] Twitter http://www.Twitter.com/
[32] Twitter (2011) #numbers http://blog.Twitter.com/2011/03/numbers.html
[33] http://www.apachefriends.org/jp/xampp-windows.html web server application
[34] WHO, http://www.who.int/csr/don/en/
[35] WHO, http://www.who.int/csr/sars/country/en/