Flu Trend Prediction Based on Massive Data Analysis

Tsung-Hau Chen
Department of Computer Science and Engineering
National Taiwan Ocean University
Keelung, Taiwan
e-mail: 00457145@mail.ntou.edu.tw

Yung-Chiao Chen, Jiann-Liang Chen, Fu-Chi Chang
Department of Electrical Engineering,
National Taiwan University of Science & Technology,
Taipei, Taiwan
e-mail: chenyc@ntu.edu.tw, lchen1215@gmail.com

Abstract—**Tracking the trend of infectious diseases, such as influenza, supports public health departments making timely and meaningful decisions and greatly help to stabilize the country and to save people's lives. Traditional systems for monitoring epidemics rely on subsequently reporting confirmed cases with at least one-week delay from the actual epidemic peak, however, there might have already been collective concerns revealed by the active social media network messages. Therefore, by real-time information examination, it is possible to detect and track the spread of epidemics in advance to monitor epidemic activities. This paper aims to study the influenza trend by collecting three datasets including the statistics of the Centers of Disease Control in Taiwan from 2010 to 2016, the Google Trends Web search data, and the King Net national medical diagnosis and consultation records. Linear methods are used to analyze the relationships among those three datasets, while establishing a pattern of interrelationships. In this article, we propose a linear forecasting framework, which is able to capture the major peaks during the interval in those years with greater epidemics. It proves that the huge amount of online social behavior information can be an indirect medium to monitor influenza activities. The prediction model based on the framework provides early response to flu pandemic.**

Keywords-big data; influenza; flu; correlation analysis; linear prediction

I. INTRODUCTION

Information searching is the most frequent type of usage among many people’s Web behavior, and Google is currently the most popular search engine. In the Web search behavior, there was a paper summarizing the highest frequency of occurrence of the search terms, exploring their rules of existence, and then establishing statistical model as the prediction tool [1]. Because the Internet has brought many changes and much dependence into our life and has rapidly developed to provide convenient ways for people accessing information, trending that the government has also set up an online diagnosis platform (King Net) [2]. Mainly it is for the large population of netizens to ask about medical-related issues or questions, including inquiry diagnostic information, medical treatment guidelines, online pharmacopoeia, medical encyclopedia, family health education and other diverse topics, meanwhile setting up a dedicated medical team to help answer and advice, without injections, medication or directly getting face-to-face with the doctors. Just submit demands on King Net and you will receive immediately and correctly response.

But the above mentioned model, established from the rule of the search terms, has been questioned about its accuracy in the latest publication [3]. The reason is that the complexity of the Web behavior itself would lead to statistical model failure. Therefore, the motivation of this study is to explore about the specific pandemic: influenza, whether there is any correlation among those three datasets: statistics from the Department of Disease Control, Department of Health, Welfare [4], number of queries to the Google search engine [5], and the number of inquiries from the Web medical diagnosis and consultation platform. Through our linear and nonlinear methods, we analyze and clarify the mutual relationships, also verify the feasibility of making linear prediction from such open datasets as an indirect medium monitoring influenza activity signals.

In recent years, big data analysis, with the power of Hadoop framework, is being widely applied. The core technology, Hadoop MapReduce, featuring high reliability, fault tolerance, load balancing, etc., is particularly suitable for processing huge amount of data. In this paper, the data acquisition from the Web medical diagnosis and consultation platform were processed by MapReduce.

After pre-processing the required Web datasets: statistics from the Department of Disease Control, Department of Health, Welfare (CDC); the number of queries from the Google search engine; and the medical diagnosis and consultation records from King Net; we first subject to linear analysis and calculate the correlation coefficient (ρ), then perform a linear regression to predict the pandemic trends in the year of 2016.

In this paper, the datasets including the CDC statistics, the number of queries in the Google search engine, and the Web diagnostic and consultation records from King Net are pre-processed, to explain their correlations with each other. Besides, the quantified similarities and differences between the single search words and the compound search terms have been also investigated, to understand the reliability and accuracy of such huge amount of data, and finally figure out which features are suitable to be as the indirect medium to monitor flu activities.

This paper is divided into five chapters, and the outline of each chapter are as follow:

1. Introduction: To address the motivation, the research methods, and the contributions of this paper.

2. Definition and Formulation: To investigate the background knowledge and related work.
3. Research Methods: To introduce the research architecture, algorithms, procedures and so on.
4. Results and Discussion: To explain the quantified results and the meaning of them, while discussing the prediction of the trend.
5. Conclusions and Future Work: To summarize the achievements of this research and possible applications in the future.

II. Definition and Formulation

A. Influenza

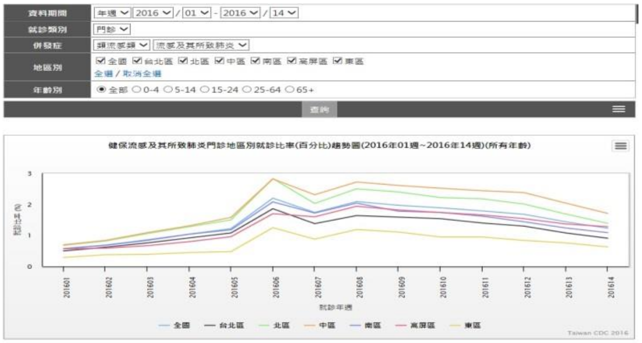
Influenza (hereinafter referred as flu), is an acute viral respiratory disease. The major causal agent is a virus called influenza. Being infected with this virus can cause fever, headache, muscle aches, feeling tired, runny nose, cough, sore throat, and so on. Symptoms vary according to each person's physique, but usually recover within one week. The most terrifying part of flu is that the outbreak could be fast and broadly spreading, and the complexity of the serious complications, especially bacterial and viral pneumonia. At the time of the outbreak, the severe and fatal ethnic groups are mostly in the elderly, and the patients with heart, lung, kidney and other metabolic diseases, anemia or immune dysfunction [4]. In summary, one can tell that flu is fairly a horrible epidemic.

B. Monitoring the Epidemics

Traditional monitoring in the U.S. and Europe relies on subordinated and contracted medical units of the U.S. Centers for Disease and Prevention, U.S. CDC, and European Influenza Surveillance Scheme, EISS. As the sentinel surveillance units on influenza and parainfluenza, they report info according to the virus samples and clinical data. But it is often one or two weeks behind that the U.S. CDC releases the national or regional data weekly [1], which is quite incapable for disease control response and policy making. In Taiwan, surveillance of epidemics is also conducted through various infectious disease surveillance systems, including legal infectious diseases control, infectious disease symptoms control, contracted virus laboratories, densely populated institutions, designated physicians and schools, monitoring systems such as 1922, physicians, medical personnels and school faculties. If a case was found that meets the definition of notifiable communicable diseases, it shall be reported to the responsible authority within the legally appointed time range [4]. The CDC would disclose the statistics on the website as shown in Figure 1, the CDC statistical information inquiry schematics.

The new epidemiological surveillance method does not need to collect samples, nor collecting reports from the medical units at all levels, without being in touch with the virus laboratory, it only needs to utilize and analyze those Web data [6], [7], [8], [9], [10]. But the Web search behavior is complex and often ambiguous, which affects the Web data analysis. For example, the GFT made too few or too many forecasts in both 2011 and 2013, as shown in Figure 2, and Google has so far modified the system at least twice [11].

In this paper, we also take advantage of this concept to quantify the Web data, prove the feasibility, and analyze the correlations. Besides analyzing the number of queries of Google search, together the Web medical diagnosis and consultation platform records are added in quantitative comparison to the CDC data.



資料期間	年週 / 2016 / 01 - 2016 / 14
就診類別	門診
併發症	類流感類　流感及其所致肺炎
地區別	☑全國 ☑台北區 ☑北區 ☑中區 ☑南區 ☑高屏區 ☑東區 全選 / 取消全選
年齡別	● 全部 ○0-4 ○5-14 ○15-24 ○25-64 ○65+

Figure 1. The CDC statistical information inquiry schematics [4]

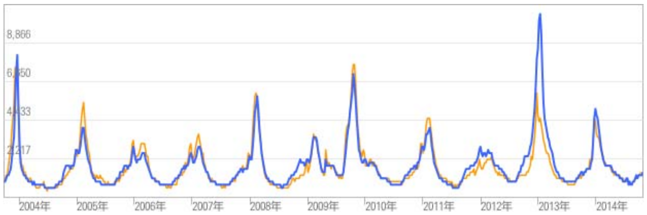


Figure 2. GFT predict trends (blue: Google; yellow: U.S. CDC) [11]

C. MapReduce [12], [13]

MapReduce is a parallel processing concept of software architecture, which is also the core of Hadoop processing, and especially fit to big data analysis. MapReduce can be detach into two steps: (1) Map: Mapping the key-value pair to an intermediate value through the Map function, and the intermediate value of this group will be sent to the Reduce function for processing. (2) Reduce: The intermediate value with the same key - value will be aggregate together to get desired result. MapReduce provides high reliability operation, because if a node error occurs, Master will re-assign the work to other nodes. Master keeps monitoring the status of Slave Worker with fault tolerance mechanism. In addition, when Master allocates work, it gives priority to accessing nodes with a relatively short distance node. Such architecture effectively lowers the bandwidth of network transmission. During processing, if Master find a node with heavier loads, it will allocate work to other idling node, in order to reach load balancing.

D. Correlation Coefficient

The famous statistician, Karl Pearson designed many well-known statistical indicators, one of them would be the correlation coefficient. When we study the relationships between two sets of random variables, we can determine the correlations by this coefficient. The correlation coefficient is calculated by the difference between the sample value and the mean, which is called deviation. Similarly, based on the deviation between two variables and their respective average values, the correlation between the two variables is reflected by multiplying those two deviations. According to different relationships, the names of the statistical indicators are also different. For example, linear relation between two variables is called correlation coefficient, and the statistic indicators that reflect the curvature between two variables are called nonlinear correlation coefficient or nonlinear determination. In this paper mainly use correlation coefficient to study the linear correlation of the targets. Covariance is defined in formula (1), which can be used to describe the case where two random variables X and Y changed along with each other,

$$Cov(X,Y) = E[(X - E(X))(Y - E(Y))] \quad 1)$$

Two random variables X and Y's Correlation Coefficient ρ are defined in formula (2),

$$\rho = \frac{Cov(X,Y)}{\sqrt{Cov(X,X)}\cdot\sqrt{Cov(Y,Y)}} = \frac{\sigma_{xy}^2}{\sqrt{\sigma_x^2}\cdot\sqrt{\sigma_y^2}} = \frac{\sigma_{xy}^2}{\sigma_x\cdot\sigma_y} \quad (2)$$

where$\sigma_x$and $\sigma_y$are the Standard Deviation of two variable X and Y, σ_x^2 and σ_y^2 each represent the Variance of variable X and Y respectively, as defined in formula (3).

$$\sigma_x^2 = Cov(X,X) = E[\left(X - E(X)\right)^2] \quad (3)$$

Assuming two sample variables Xi and Yi are independent and identically distribution, and are randomly selected from the population in normal distribution. Precursor and are the respective mean, as defined in formula (4), wherein n is the number of samples.

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad (4)$$

The covariance of two sample variables is as shown in (5)

$$S_{XY}^2 = \frac{\sum_i^n (X_i - \bar{X})(Y_i - \bar{Y})}{n-1} \quad (5)$$

Thus, we can define the correlation coefficient of two sample variables in formula (6) [14].

$$\rho = \frac{S_{xy}^2}{S_x \cdot S_y} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i-1}^{n}(X_{i\_}\overline{X})^2 \sum_{i=1}^{n}(Y_i - \overline{Y})^2}} \quad (6)$$

In this paper, Xi and Yi refers to the CDC statistics and Google Trends data, or the CDC statistics and King Net diagnosis records, which are used to analyze the correlation between two different groups of sample variables.

E. Regression Analysis

Regression analysis mainly describes causal relationship between two variable sets. One may use the independent variable *x* to predict the dependent variable y. First, by the existing (*x, y*) data to determine *y = f (x)*, the function of *x*, then feed a new *x* into this equation to get the prediction of *y*. Both the independent variable *x* and the dependent variable *y* must be numeric [7].

The following formula (7), is a simple linear regression model,

$$Y_i = \gamma_0 + \gamma_1 X_i + \varepsilon_i \quad (7)$$

Let *Yi* be the *i-th* observation value of the dependent variable, γ_0 and γ_1 be the intercept and slope of the parameters to be estimated, *Xi* be the *i-th* observation value in the independent variable, ε_i be the error, and *Yi*'s estimated value $\widehat{Y_i} = \gamma_0 + \gamma_1 X_i$; Since γ_0 and γ_1 are unknown parameters, they must be estimated first. The most commonly used estimation rule is the least square estimation (LSE). Try to find a set of γ_0 and γ_1, so that the sum of squares of the estimation errors can be minimized. The estimated error is equal to the observation value minus the estimated value, as shown in equation (8),

$$e_i = Y_i - \hat{Y}_i \quad (8)$$

$$\mathrm{Q} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 = \sum_{i=1}^{n}(Y_i - \gamma_0 - \gamma_1 X_i)^2 \quad (9)$$

To find a minimum square error is to minimize the value of Q in equation (9), so first γ_0 and γ_1 must be partially differentiated.

$$\frac{\partial \mathrm{Q}}{\partial \beta_0} = -2\sum_{i=1}^{n}(Y_i - \gamma_0 - \gamma_1 X_i) \quad (10)$$

$$\frac{\partial \mathrm{Q}}{\partial \beta_1} = -2\sum_{i=1}^{n} X_i(Y_i - \gamma_0 - \gamma_1 X_i) \quad (11)$$

Let (10) and (11) be 0,

$$\sum_{i=1}^{n}(Y_i - \gamma_0 - \gamma_1 X_i) = 0 \quad (12)$$

$$\sum_{i=1}^{n} X_i(Y_i - \gamma_0 - \gamma_1 X_i) = 0 \quad (13)$$

(12) and (13) can be organized into

$$\sum_{i=1}^{n} Y_i - n\gamma_0 - \gamma_1 \sum_{i=1}^{n} X_i = 0 \quad (14)$$

$$\sum_{i=1}^{n} X_i Y_i - \gamma_0 \sum_{i=1}^{n} X_i - \gamma_1 \sum_{i=1}^{n} X_i^2 = 0 \quad (15)$$

Solving simultaneous equations (14), (15) results in

$$\gamma_0 = \bar{Y} - \gamma_1 \bar{X} \quad (16)$$

$$\gamma_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \overline{\bar{X}})^2} = \frac{\frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n-1}}{\frac{\sum_{i=1}^{n} X_i - \bar{X}^2}{n-1}} = \frac{S_{XY}^2}{S_X^2} \quad (17)$$

wherein S_{xy}^2 is the covariant of X and Y, and S_X^2 is the variance of the samples.

In addition to the linear regression model, the nonlinear model can be applied to this linear regression model through variable transformation. For example, polynomial or exponential form can be applied after being processed by this method, as shown in equations (18) and (19) [14].

$$\widehat{Y_i} = \gamma_0 + \gamma_1 X_i + \gamma_2 {X_i}^2 + \cdots + \gamma_{n-1} {X_i}^{\mathrm{n}-1} \quad (18)$$

$$\widehat{Y_i} = \gamma_0 e^{\gamma_1 X_i} \quad (19)$$

Here we take exponential model as an example to explain the variable transformation process. First the equation (19) can be rewritten to equation (20)

$$\widehat{Y_i} = ae^{bX_i} \quad (20)$$

In equation (20), apply natural logarithm on both sides of the equal signs and obtain (21).

$$ln\widehat{Y_i} = lna + bX_i \rightarrow \hat{Y}_i^* = a^* + bX_i \quad (21)$$

By the steps of such a transformation, linear regression analysis can be performed.

III. Research Methods

The research framework of this dissertation is shown in Figure 3, which is mainly divided into several parts: data acquisition, linear correlation and regression analysis. In data acquisition, the CDC statistics and Google Trends data can be used directly after a certain organization. The records of the King Net are messier, therefore, we need to use MapReduce architecture to help organize.

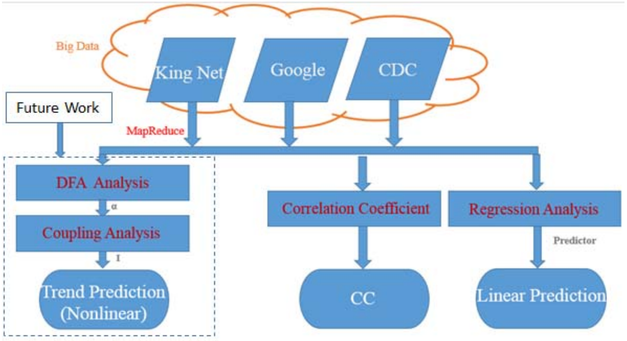


Figure 3. Research flow chart

After obtaining the public data from the Internet, the correlation coefficient ρ can be calculated using equation (6). Figure 4 is based on the CDC statistics and the Google Trends data in the year of 2011 as an example to illustrate processing the correlation coefficient ρ and regression analysis in this paper. First take the CDC statistical data as x, Google Trends data as y, $\bar{x}$ and $\bar{y}$ as their average respectively, and the correlation coefficient ρ between these two datasets can be calculated using the formula (6), as well as the CDC statistics and King Net records.

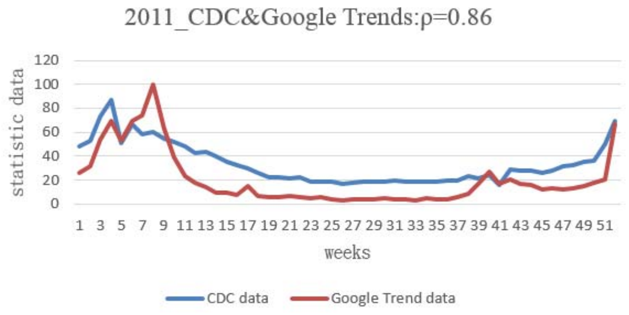


Figure 4. Illustration of calculating the correlation coefficient ρ

Then by using the formula (16) and (17) to find out the weights γ_0 and γ_1 by equation (7) to determine the linear regression line, as shown in Figure 4. Take the CDC statistics as x, and Google Trends data as y, plot the scatter points of x and y, and then use the formula to calculate the regression line as shown in Figure 5.

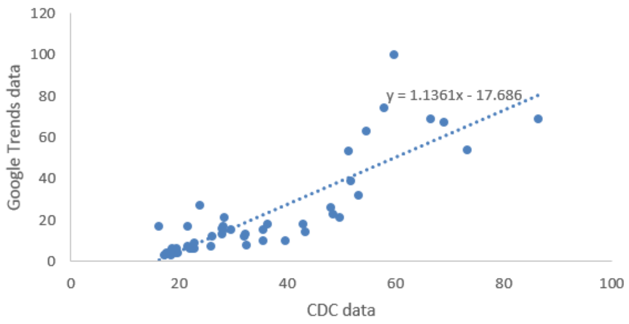


Figure 5. Calculation of linear regression

Demonstrating by the same 2011 data, as shown in Figure 6, the polynomial regression method uses formula (18).

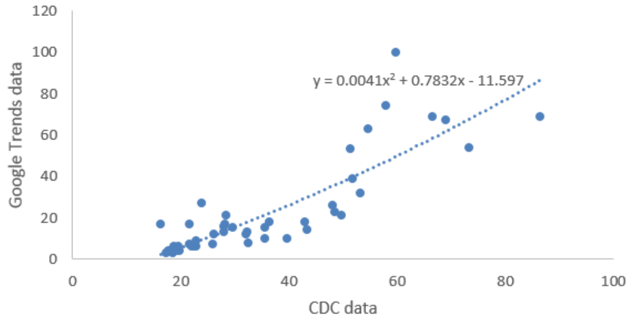


Figure 6. Calculation of polynomial regression

The exponential regression method uses formula (19) as shown as Figure 7.

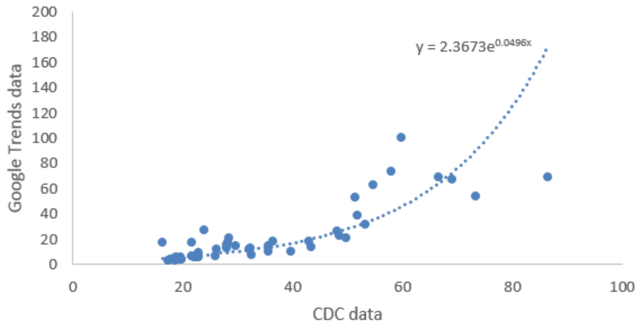


Figure 7. Calculation of exponential regression

IV. Results and Discussion

A. Correlation Analysis

In linear correlation analysis, the correlation of the CDC statistics from 2008 to 2016 with both the Google Trends data and King Net diagnosis records are analyzed respectively by equation (6). For the year 2016, plot the trends as shown in Fig 8, which are the correlation between the CDC statistics and Google Trends data.

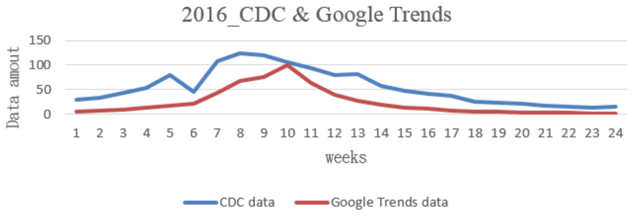


Figure 8. The correlation between the CDC statistics and Google Trends data (2016)

In Figure 9 shows the correlation between the CDC statistics and King Net diagnosis records. There is no significant correlation between the CDC statistics and Kingnet diagnosis records. Therefore, they are not suitable for further regression analysis

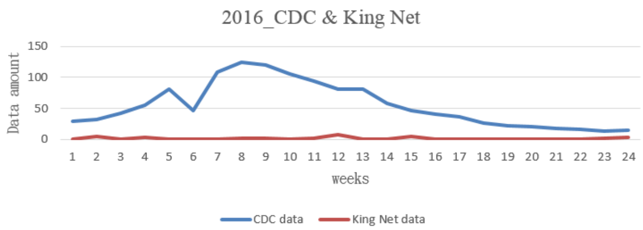


Figure 9. The correlation between the CDC statistics and King Net records (2016)

B. *Regression Analysis*

Using the regression equation (7), (18) and (19) to predict the influenza trend line for the first 24 weeks of 2016. The correlation analysis are conducted between those trend lines and the CDC statistics of the same year, and the linear prediction results are shown in Table I. One can observe that the second-order polynomial prediction is the best result while exponential prediction is the worst. However, The linear regression prediction is simply influenced by the data of 2016 itself, that is, Google Trends data in 2016 have higher linear correlation with the CDC statistics, so the result is better.

TABLE I. THE RESULTS OF LINEAR REGRESSION PREDICTION

regression	linear	second-order polynomial	third-order polynomial	exponential
CC	0.896	0.967	0.946	0.776

V. CONCLUSIONS AND FUTURE WORK

In the beginning of 2016, the flu went pandemic, and people failed to follow the procedure and flocked to the emergency department of the major hospitals. At the time the national health authorities failed to respond effectively, causing the number of beds to be miscalculated, which led to many negative critics in public. The news is still fresh in our memory. The linear influenza prediction framework really capture the trend fluctuations, so the predictions of epidemic trends are helpful in early preparation and in propounding work of prevention. Also we confirm that Google Trends data can be used as an indirect medium monitoring for signs of flu activity. As a result, accurate flu epidemic spikes predictions can help national health units make timely and meaningful decisions, and are of tremendous help in stabilizing and saving the lives of people.

Indirect monitoring of influenza activity signals has been taking place for some years now. According to previous works, most forecasting models accuracy starts to diminish after a few years and the model must be readjusted. Mainly it is because of the complexity of search behavior, and linear prediction can be too much influenced. Therefore, we will introduce a nonlinear analysis method, and to find the epidemic in advance is our main goal.

REFERENCES

[1] J. Ginsberg, M.H. Mohebbi, R.S. Patel, L Brammer, M.S. Smolinski, L. Brilliant, “Detecting influenza epidemics using search engine query data, ” Nature, volume:457(19), pp.1012-1014, 2009.

[2] King Net: http://www.kingnet.com.tw/

[3] S. Yang, M. Santillana, S.C. Kou, “Accurate estimation of influenza epidemics using Google search data via ARGO, ” Proc Natl Acad Sci U S A , volume:112(47), pp.14473-14478, 2015.

[4] Centers for Disease Control, Taiwan: http://www.cdc.gov.tw/

[5] Google Trends: https://www.google.com/trends/

[6] J. Espino, W. Hogan, & M. Wagner, “Telephone triage: A timely data source for surveillance of influenza-like diseases, ” AMIA: Annual Symposium Proceedings, pp. 215–219, 2003.

[7] S. Magruder, “Evaluation of over-the-counter pharmaceutical sales as a possible early warning indicator of human disease, ” Johns Hopkins University APL Technical Digest 24, pp. 349-353, 2003.

[8] A. Hulth, G. Rydevik, & A. Linde, “Web Queries as a Source for Syndromic Surveillance, ” PLoS ONE 4(2): e4378. doi:10.1371/ journal.pone.0004378, 2009.

[9] G. Eysenbach, ”Infodemiology: tracking flu-related searches on the web for syndromic surveillance, ” AMIA: Annual Symposium Proceedings, pp.244-248, 2006.

[10] P. M. Polgreen, Y. Chen, D. M. Pennock, & N. D. Forrest, “Using internet searches for influenza surveillance,” Clinical Infectious Diseases 47, pp.1443-1448, 2008.

[11] Google Flu Trends: https://www.google.org/flutrends/about/how/ html

[12] L. Yang, J. Zhang and Q. Zhang, “PESC:A parallel system for clustering ECG streams based on MapReduce, ” 2013 IEEE Global Communications Conference (GLOBECOM), pp.2604-2609, 2013.

[13] Y. Zhao, J. Wu and C. Liu, “Dache: A data aware caching for big-data applications using the MapReduce framework,” Tsinghua Science and Technology, Volume: 19, Issue: 1, pp.39-50, 2014.

[14] W. C. Wang, “Statistic and Excel Data Analysis,” DrMaster Press Co., Ltd., pp.389-393, 2004.