International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:6, No:12, 2012

Influenza Pattern Analysis System through Mining Weblogs

Pei Lin Khoo, and Yunli Lee

Abstract**—**Weblogs are resource of social structure to discover and track the various type of information written by blogger. In this paper, we proposed to use mining weblogs technique for identifying the trends of influenza where blogger had disseminated their opinion for the anomaly disease. In order to identify the trends, web crawler is applied to perform a search and generated a list of visited links based on a set of influenza keywords. This information is used to implement the analytics report system for monitoring and analyzing the pattern and trends of influenza (H1N1). Statistical and graphical analysis reports are generated. Both types of the report have shown satisfactory reports that reflect the awareness of Malaysian on the issue of influenza outbreak through blogs**.**

Keywords**—**H1N1, Weblogs, Web Crawler, Analytics Report System.

I. Introduction

THE growing of Internet activity has increased the number of informal channel known as social media such as weblogs, facebook, twitter and others. These kinds of channels allow anyone to disseminate any information freely. Due to the influenza spreads around the world, people began to post information about this disease on Internet. Weblogs are one of the channels that contains of disease and outbreak related news shared by among bloggers. The rapid response of social media usage has track the disease activity based on the information posted by Internet users. The interaction of Internet user response allows us to detect the awareness of people for this particular disease. These sources of information provide the different view of global health information instead of traditional communication channels has reporting by government. Early efforts in this area, World Health Organization (WHO) by using the usage of Global Outbreak Alert and Response Network to capture public responses for disease and emerging the diseases to reduce the time of government suppressing the disease information [1].

On another side, the early detection to monitor the human behavior based on search query is proposed by Google Inc [2]. The intelligent of search queries, estimate the current level of influenza activity in certain areas with a large population of web search users. Weblogs are new phenomena among researchers who collect the real-time processing of web documents instead of traditional media content [3]. The mining of social media through weblogs for monitoring the trends of influenza has been researched by Courtney et al. Courtney stated the use of Web and Social Media (WSM) able to evaluate the response rate of people towards influenza outbreaks by detecting theirs flu-posting in weblogs and these crowded of WSM committees could disseminate an important information on infection outbreak for responding the impact of the disease [4]. The analysis on infection outbreak could be analyze by using data mining tool and perform the statistical analysis process.

In this paper, we are using SAS tools one of statistical software to analysis collection data on influenza. SAS data mining tool has been adopted among researchers to develop an analytics report system to analyze their demographic data consists of clinical data. In 2008, Johnson & Johnson Pharmaceutical Research and Development, L.L.C introduced an analytics report system by combine SAS tools and Visual Basic Application (VBA) to develop a tool for a data analytics report system with using clinical demographic data as their research analytics data [5]. This successful effort has been adopted in this proposed system for detecting the affected states of influenza diseases within Malaysia. Fig. 1 illustrates the proposed framework with the work flow and the following paragraph briefly introduces the history of influenza.

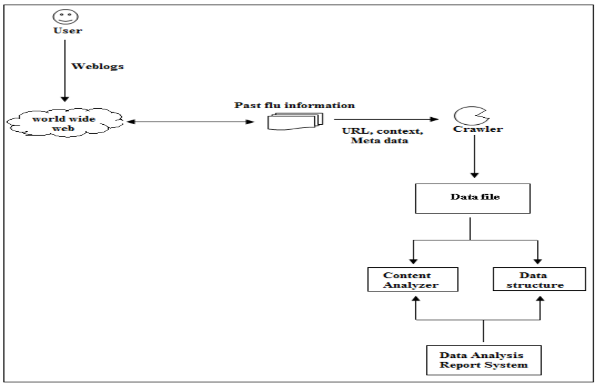


Fig. 1 Proposed framework to identify the possible web process to analyses the content

In the late of April 2009, World Health Organization (WHO) announced the emergence of a novel influenza disease (Virus A) had been reported to 74 countries [6]. This is entirely new virus has not been catching by human in previously. The virus has highly infection by spreading one person to another and the symptoms similar to normal fever, sore throat, headache, muscle or joint pain. Hence, during on

Pei Lin Khoo, is with the Faculty of Science and Technology, Sunway University, Selangor, 46150 Malaysia (e-mail: 09066093@imail.sunway.edu.my).

Yunli Lee, is with the Faculty of Science and Technology, Sunway University, Selangor, 46150 Malaysia (e-mail: yunlil@sunway.edu.my).

the period of time, those who above get symptoms need to do the specific laboratory testing such as blood test to examine the presence of symptoms. In May 2009, U.S. Centers for Disease Control and Prevention found children had no preexisting immunity to the strain of influenza virus, but for those above 60 had some degree of immunity. Hence, those under the age of 25 years (children and teenage) has categorize as high risk group infection [14]. The total of 14,286 confirmed deaths occurred in worldwide has been reported by European Centre for Disease Prevention and Control (ECDC) with weekly updates from regional hospital and 18,036 confirmed deaths has been reported by WHO with gathering the multiple sources of information form WHO Regional Offices and Member States [15].

We briefly consider blogger as a dissemination of the influenza news through weblogs, the content of blogs is stated in web pages. A crawler has been used to retrieve particular topic there are “influenza” (past flu information) specific web pages within Malaysia social media. Once crawler had been reached the maximum of URLs is visited. The data file is the output for crawling tool and contains the collection of visited link. The visited link either extract out in .CSV file or .HTML file format. The ready data file is analyzed on SAS tools to further identify the approach of corresponded weblogs to detect the trends of influenza within Malaysian

II. Related Work

Mining analysis for detecting influenza epidemics has previously conducted a number of researchers. We adopted the method introduced by Courtney et al. for identifying the trends of influenza in public society via social media [4]. The differences of principal are correlation analysis to evaluate the frequency of flu related post and different flu types of post. As concern, health care industry considered using SAS data mining capabilities to detect statistical patterns in pervious through predictive modeling. This model consists of fraudulent pattern by using the methods of rule-only approaches to investigator identify the most important cases [12].

In the meantime, mining technique of correlation analysis is applied for other researchers to detect the human influenza cases. For example, Nicola et al. [7] grouping the data of International Classification of Diseases (ICD-9) to detect the Influenza likes illness (ILI) by using set code of ICD-9. While this type of codes improved evaluate respiratory illness surveillance for public. Additionally, Aron [8] predict the rate of ILI by analyze the message posted in twitter and make the compare of regression models to correlation with Central for Disease Control and Prevention (CDC). Following this, analyzing twitter message to demonstrate the potential Influenza disease applied from Quincey and Kostova [9]. Ritterman et al. [10] showed the early detection of influenza outbreak by using twitter message has been improved the market forecasting models to disseminate the external events such as H1N1 outbreak.

On other side, mining the web information could be able to emerging the trends of behavior from public which use on find the interest in new products. The instant updates of blogging might reveal out the exactly ideas of trends through the information [11].

III. Proposed Methodology

The research conducts two different methodologies to manipulate the analysis of Influenza trends. Section A describes the data and methodology used in the research and section B describes the analysis application tools used in the research.

A. Data Methodology

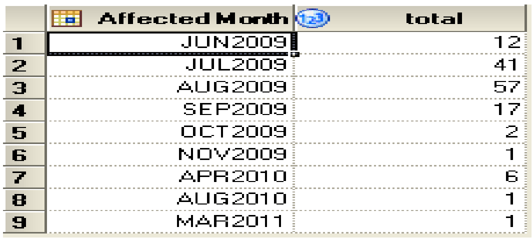
Data collection in social media had used Visual Web Spider (VWS) [13] as a web crawler to crawl the contents of weblogs. VWS is the licensed web crawler, but in this research we are using the trial version of this crawler to test the technicality. As a result, we find the feature of VWS is able to be a dual functional tool such as a crawler and an extractor to extract the content from the URLs link. However, the trial version of the crawling only allows retrieve 100 pages for given URLs, we use several phrases to wide spread crawling such as “H1N1 Malaysia” and “H1N1 weblogs Malaysia”. Hence, the final weblogs we had collected contains 500 related influenza data had posted in the blogs and show in Table I.

TABLE I
Total weblogs from different domain website

Weblogs	Crawled URL
http://my-h1n1.blogspot.com	477
http://malaysianblogs.hitsmojo.com	1
http://blog.malaysiatotal.com	1
http://wordpress.com	7
http://asia.cnet.com/blogs	2
http://blogspot/com	6
http://blogs.straitstimes.com	1
Various domain blog	5

We manually checked through the contents of weblogs to identify the post are written and shared by Malaysian about influenza disease, it to constraint the trends of influenza disease toward Malaysian. We have determined that the peak period between 2009 and 2010 is the period of influenza disease spread widely in according the data collected between the beginnings of June 2009 until the end of August 2010, as shown in Fig. 2. The result of analysis has proved on Director-General of the World Health Organization, Margaret Chan, she announced the end of the H1N1 pandemic on 10 August 2010 and H1N1 influenza event has considered in post pandemic period [15]. A total of 138 confirmed related contents were filtered according from the crawled 500 URLs had listed in Table I.

We identified the increase of posted rate from beginning of peak period for the trends of influenza as shown in Fig. 2. The tool of VWS is successful attempted and pulled the same analysis starting of 2009 where the influenza pandemic zone was posted in world health organization (WHO) statement health report on 11 June 2009 [14, 15].



	Affected Month	total
1	JUN2009	12
2	JUL2009	41
3	AUG2009	57
4	SEP2009	17
5	OCT2009	2
6	NOV2009	1
7	APR2010	6
8	AUG2010	1
9	MAR2011	1

Fig. 2 Affected Month period within Influenza disease

B. Analysis Application Tool

We have developed an analytics report system as analysis application tool to generate combination of interfacing SAS tools to VBA to generate the data processing report. The integration of data mining technologies platform is a common in complement to deliver information processing in the runtime environment. VBA platform has an ability to develop a powerful graphical user interface and SAS platform has an ability to deliver the batch processing on information [5] as showed in Fig. 3. Hence, the combination of VBA and SAS tools will be offered a program to perform the custom task in reporting system. We have introduced a system which produces of statistical analyze and graphical analyze. Statistical analysis consists of statistic mathematical calculation method such as mean, median, mode, standard deviation, cumulative, and frequency which are used to measure the number for affected area in different locations. Graphical analysis consists of bar chart, pie chart, radar, box and whisker plot, histogram and map chart which are used to measure the group of affected area in different location either based on percentage or frequency.

Microsoft office application has executed the communication for SAS tools to generate the final analysis report. The supported file format for the output is RTF file which is the extension integrates with Microsoft Word. The feature of the integration has advantage that allows users to customize the output format report according to their need.

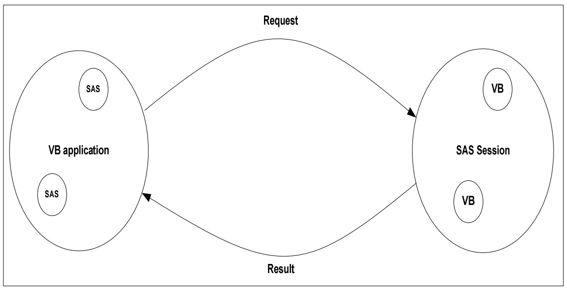


Fig. 3 Integrate concept of Visual Basis Application (VBA) and SAS platform environment

We describe the integration concept of the proposed work as below:

1. VBA as a development for graphical interface for SAS session. The process goes through VBA by sending request to SAS session to process the analysis.
2. SAS session received the request, and depend the requested from VBA to generate the type of statistical report or graphical report. Final result would display by VBA to open the supported file format.

The proposed analytics report system consists of three main sections to perform the analysis output for detecting the trends of influenza data. The working procedure is preceded in the following Fig. 4 and Fig. 5 to show the framework of the system.

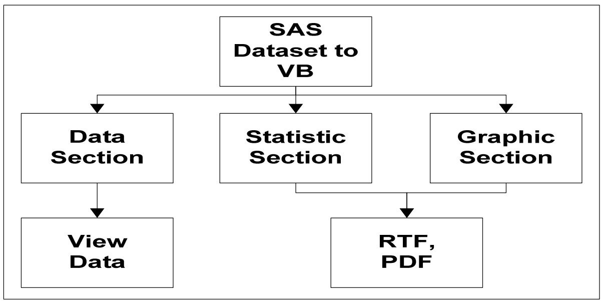
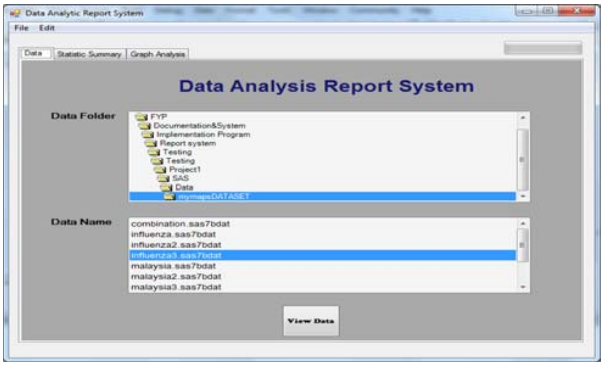


Fig. 4 Process Flow for Analysis Report System

The three main sections of the system are described as below:

1. Data Section: The main module which allows user to select the dataset and used in Statistic module and Graphic module as the analysis data.
2. Statistic Section: The module that to measure the numeric calculation to view in various calculation methods. Such as Mean, Median, Mode, Standard Deviation, Maximum, Minimum, Range, N (number of value), Lower 95% CL for Mean and Upper 95% CL for Mean (confidence interval (CL) for the mean specifies a range of values to within each the unknown cases parameter).
3. Graphic Section: The module allows us to view the entire performance in graphical calculation method. Such as Pie chart, Bar chart, Radar, BoxPlots, Histogram and Map chart.



Data Folder

FYP
Documentation&System
Implementation Program
Report system
Testing
Testing
Project1
SAS
Data
mymapsDATASET

Data Name

combination.sas7bdat
influenza.sas7bdat
influenza2.sas7bdat
influenza3.sas7bdat
malaysia.sas7bdat
malaysia2.sas7bdat
malaysia3.sas7bdat

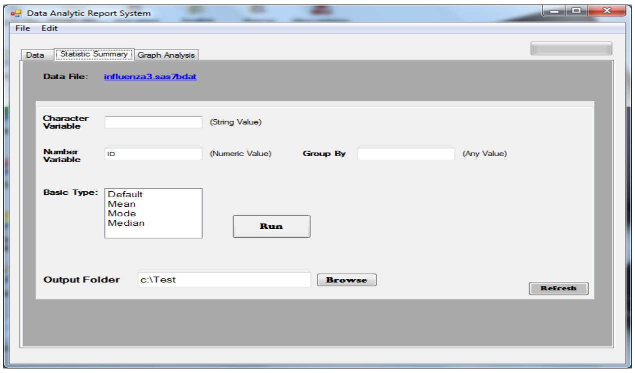
Fig. 5 Framework Data Analysis Report System

IV. Experimental Results

The experiment output is illustrated in below section. We present two different types of statistics output in Section A and two different types of graphics output in Section B to evaluate our analysis.

A. Statistic Output

The working interface for *Statistical Analysis* is shown as Fig. 6 and the function is described as below:



Data File: influenza3.sas7bdat

Character Variable: (String Value)

Number Variable: ID (Numeric Value)

Group By: (Any Value)

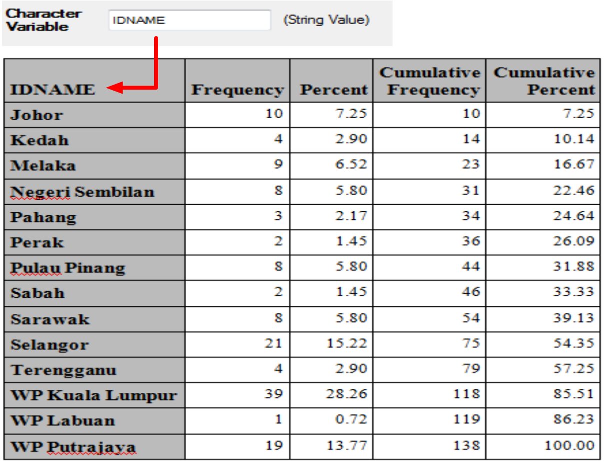
Basic Type: Default, Mean, Mode, Median

Output Folder: c:\Test

Fig. 6 Statistical Analysis framework

1) Data File: Working dataset to perform the analysis.
2) Character Variables: To count the frequency performance on the variable.
3) Number Variables: To perform the Basic Type function.
4) Group By: To cluster the variable in the group.
5) Basic Type: Statistic calculation consists of Default, Mean, Mode and Median.
6) Output Folder: Specify the final output in which local directory.

Under statistic output consists of two different approaches to measure the frequency of affected area. Fig. 7 focuses on count the percentage of frequency for each state by plot in the table. Meanwhile, Fig. 8 focus on numeric measurement with count on row total of value in the table.



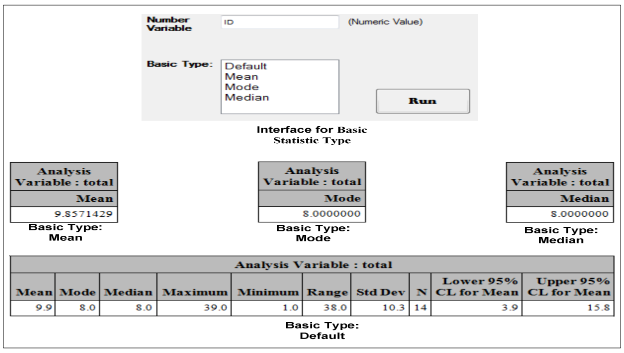
Character Variable: IDNAME (String Value)

IDNAME	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Johor	10	7.25	10	7.25
Kedah	4	2.90	14	10.14
Melaka	9	6.52	23	16.67
Negeri Sembilan	8	5.80	31	22.46
Pahang	3	2.17	34	24.64
Perak	2	1.45	36	26.09
Pulau Pinang	8	5.80	44	31.88
Sabah	2	1.45	46	33.33
Sarawak	8	5.80	54	39.13
Selangor	21	15.22	75	54.35
Terengganu	4	2.90	79	57.25
WP Kuala Lumpur	39	28.26	118	85.51
WP Labuan	1	0.72	119	86.23
WP Putrajaya	19	13.77	138	100.00

Fig. 7 Frequency of affected area in Malaysia

In order to study the trends of influenza spread throughout the Malaysia area, we are using the function of character variable to count the frequency of each value. The state of Malaysia is categorized under “IDNAME” to measure affected area of Influenza disease for that state. Fig. 7 lists the state of in affected influenza case. We found that WP Kuala Lumpur has the most affected influenza cases in among state and categorize as dangerous disease zone. We also noticed the state of Perlis and Kelantan were not in the listed as affected area which is disease free zone.

Fig. 8 shows the type of several statistic methods to analyze the variable. The function of number variable is used to define the different statistic methods. We use “total” as ours analysis variable data to measure total influenza case happen in each state. Based on the analysis output, we can conclude the average number for total case in Malaysia by using mean type. The middle value for total case in median type and the most frequency number of case in each state in mode type. The highest influenza cases happen in maximum value and minimum case in certain state.



Number Variable: ID (Numeric Value)

Basic Type: Default, Mean, Mode, Median

Interface for Basic Statistic Type

Analysis Variable : total
Mean
9.8571429

Basic Type: Mean

Analysis Variable : total
Mode
8.0000000

Basic Type: Mode

Analysis Variable : total
Median
8.0000000

Basic Type: Median

Analysis Variable : total									
Mean	Mode	Median	Maximum	Minimum	Range	Std Dev	N	Lower 95% CL for Mean	Upper 95% CL for Mean
9.9	8.0	8.0	39.0	1.0	38.0	10.3	14	3.9	15.8

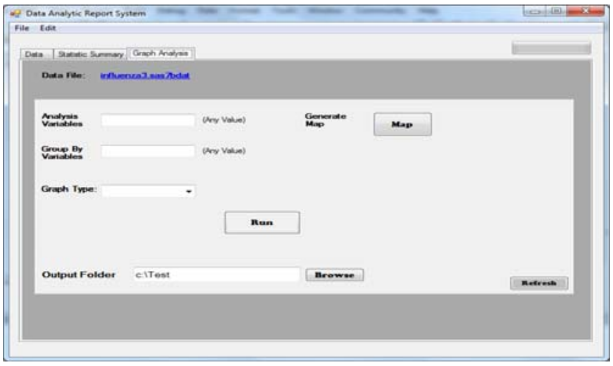
Basic Type: Default

Fig. 8 Measurement in several calculation methods

B. Graphic Output

The working interface for *Graph Analysis* is shown as Fig. 9 and the function is described as below:

1) Data File: Working dataset to perform the analysis.
2) Analysis Variables: To measure the overall performance on the analysis variable.
3) Group By Variables: To cluster the analysis variable in the group.
4) Graph Type: Consists of Pie, Bar, Boxplot, Histogram and Radar.
5) Generate Map: Illustrate the analysis variable in geographical type.
6) Output Folder: Specify the final output in which local directory.



Data File: influenza3.sas7bdat

Output Folder c:\Test

Fig. 9 Graph Analysis section framework

Under graphic output consists of two different representations to illustrate the frequency of affected area by displayed in different colour. Fig. 10 clusters the frequency for each state with plot in the independent bar. Fig. 11 and Fig. 12 visualize the data in map chart to present the views of covered affected area in Malaysia.

Using the graphical chart we verify the group of state by comparing the affected area in heights. We added the function of analysis variables to measures the variable value and group by variables is used on diversity the variable in a group as shown in Fig. 10. We define "IDNAME" in analysis variables with categorized the "Place" of each state to illustrate in independent bar. Over here, different colour for the independent bar has been clustered. The uniform cluster group enables us to inspect which area of the state has the most affected zone in particular state.

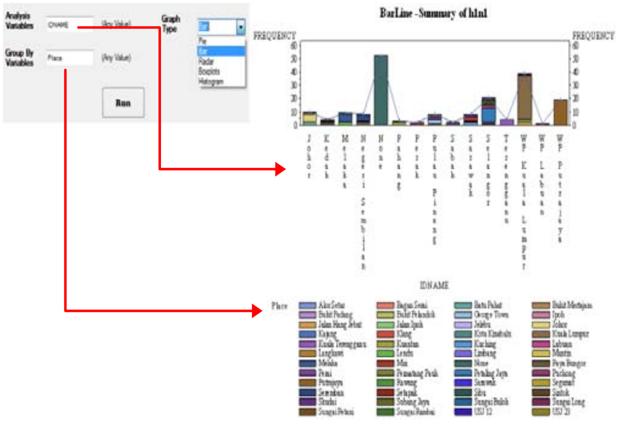
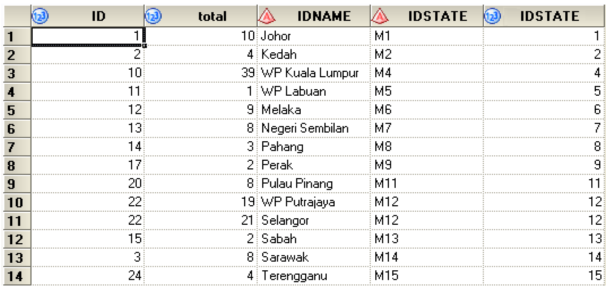


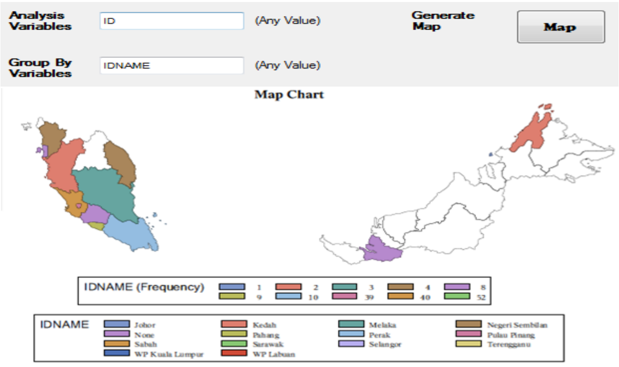
Fig. 10 Bar Line chart shows independent bar for affected area

Map effective for data presentation, data analysis and visualization in coloured regions to show multiple independent variables [16]. Map provides a dimensional picture varies as geographically for the data to allow user to easily identify the clusters of the data in concentration way [17]. In the research, we are introducing SAS/GRAPH GMAP procedure which is a mapping tool from SAS tools to produce the geographic spatial data in analysis influenza zone. The effectiveness of GMAP procedure is introduced by Virginia Health Quality Center (VHQC) in U.S. VHQC in effort used the features of GMAP procedure to develop a series of maps in analyses the pattern of diabetes wellness campaign in the state of Virginia [18]. In Fig. 11, the analysis variables in "ID" as a part of longitude and latitude in SAS/GRAPH for display geographic spatial picture. We define the group by variables using "IDNAME" to visual colour each of the state. The result of map had shown in Fig. 12 as below.



	ID	total	IDNAME	IDSTATE	IDSTATE
1	1	10	Johor	M1	1
2	2	4	Kedah	M2	2
3	10	39	WP Kuala Lumpur	M4	4
4	11	1	WP Labuan	M5	5
5	12	9	Melaka	M6	6
6	13	8	Negeri Sembilan	M7	7
7	14	3	Pahang	M8	8
8	17	2	Perak	M9	9
9	20	8	Pulau Pinang	M11	11
10	22	19	WP Putrajaya	M12	12
11	22	21	Selangor	M12	12
12	15	2	Sabah	M13	13
13	3	8	Sarawak	M14	14
14	24	4	Terengganu	M15	15

Fig. 11 The total of state had affected influenza with ID and IDSTATE to generate the maps



Analysis Variables: ID (Any Value)

Group By Variables: IDNAME (Any Value)

Fig. 12 Malaysia Map shows frequency of affected Influenza area

V. Conclusion

In this paper, the most significant finding is responses of bloggers on the influenza outbreak within Malaysia. We are able to extract the relevant sources of blogs to analyze the pattern of influenza outbreak within Malaysian and bloggers response, based on the evaluation of sources from a posted time period. In addition, this proposed system assists in indentifying the pandemic of disease within Malaysia. We also demonstrated that blogs as one of social media platform could offer as a channel for people to voice and response on the influenza outbreak. The influenza analysis system is used to monitor the possibility of information intervention by measure the frequency of affected area of H1N1 disease outbreak. The system is a useful attempt tool to perform the various type of analysis report such as statistical and graphical report to process the regular analysis data in demographic disease report.

Acknowledgment

The authors would like to thank to Richard Zhou (Johnson

& Johnson Pharmaceutical Research and Development, L.L.C) for valuable guidance and feedback in this project.

REFERENCES

[1] John S. Brownstein, Ph.D., Clark C. Freifeld, B.S., and Lawrence C. Madoff, M.D. (2009) Digital Disease Detection — Harnessing the Web for Public Health Surveillance. The New England Journal of Medicine; med 360;21, published at NEJM.org on May 7, 2009.

[2] 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, no. 7232, pp. 1012–4, Feb 2009.

[3] Il-Chul Moon, Young-Min Kim, Hyun-Jong Lee, Alice H. Oh (2009) Temporal Issue Trend Identifications in Blogs, published in Proceedings of the CSE' 09 International Conference on Computational Science and Engineering Volume-04.

[4] Courtney D Corley, Armin R Mikler, Karan P Singh and Diane J Cook (2010) Monitoring Influenza Trends through Mining Social Media. Int.J.Environ.Res.Public Health published in 22 February 2010.

[5] Richard Zhou (2008). Develop A Demographic Data Analysis Report System By SAS® and Visual Basic, proceedings of the NESUG 2008.

[6] http://www.who.int/mediacentre/news/statements/2009/h1n1_pandemic_phase6_20090611/en/

[7] Nicola Marsden-Haug, Virginia B. Foster, Philip L. Gould, Eugene Elbert Hailiang Wang, and Julie A. Pavlin (2007) Code-based Syndromic Surveillance for Influenzalike Illness by International Classification of Diseases, Ninth Revision published in Vol. 13, No. 2, February 2007.

[8] Aron Culotta (2010) Towards detecting influenza epidemics by analyzing Twitter messages. In 1st Workshop on social media Analytics (SOMA'10), July 25, 2010, Washington, DC, USA, published in ACM 978-1-14503-0217-3.

[9] E. de Quincey and P. Kostkova (2009). Early warning and outbreak detection using social networking websites: the potential of twitter, electronic healthcare. In eHeath 2nd International Conference, Instanbul, Tirkey, September 2009.

[10] J. Ritterman, M.Osborne, and E. Klein. Using prediction markets and Twitter to predict a swine flu pandemic. In 1st International Workshop on Mining Social Media, 2009.

[11] http://www.economist.com/node/15660874

[12] Julie Malida "The Changing Face of Health Care Fraud Detection—Predictive Analytics" BNA's Health Care Fraud Report, Vol. 14, No. 4, Feb. 23, 2011.

[13] http://www.newprosoft.com/web-spider.htm

[14] http://www.who.int/mediacentre/news/statements/2009/h1n1_pandemic_phase6_20090611/en/index.ht ml

[15] http://en.wikipedia.org/wiki/2009_flu_pandemic

[16] Michael Eberhart, (2008) MPH, Philadelphia Department of Public Health. Make the Map You Want with PROC GMAP and the Annotate Facility, proceedings of the NESUG 2008.

[17] Darrell Massengill and Jeff Phillips (2010), SAS Institute Inc., Cary, NC. Tips and Tricks: More SAS/GRAPH® Map Secrets, proceedings of the Paper 134-2010.

[18] Barbara B. Okerson, Virginia Health Quality Center, Glen Allen, VA. Using SAS/GRAPH® GMAP to Enhance a Diabetes Wellness Campaign, proceedings of the Paper 171-30.