Detecting Social Signals of Flu Symptoms

Bumsuk Lee Jinyoung Yoon Seokjung Kim Byung-Yeon Hwang

Dept. of Computer Science and Engineering
The Catholic University of Korea
{bumsuk, my_sk_test, typicalkorean, byhwang}@catholic.ac.kr

Abstract—**A cold and the flu are both respiratory illnesses and they are very common to us. Vaccination is the most effective way to prevent infection of the flu, but there is no way for a cold. Thus, the best strategy for individuals is to stay away from the flu or cold carriers and to wash their hands often. Early detection of flu epidemics and a quick response to that can minimize the impact of the flu. We observed tweets as social signals of flu symptoms to detect the flu epidemics in early stage. We compared a tweet corpus from nine cities in Korea to the weather factors, flu forecast, and Influenza-like Illness datasets. The results show the possibility of using social signals to detect epidemic diseases.**

Keywords-Social Signals, Flu Epidemics, Event Detection

I. Introduction

A cold and the flu are both respiratory illnesses and they are very common to us. Although many researchers have been trying to develop a medicine for a cold and the flu, most medications are not for cure in general, but they just ease the symptoms. Currently, vaccination is the most effective way to prevent infection of the flu, but there is no way for a cold. Thus, the best strategy for individuals is to stay away from the flu or cold carriers and to wash hands often. Early detection of flu epidemics and a quick response to that can minimize the impact of the flu. For such reasons, there are attempts to track flu epidemics. Jeremy Ginsberg et al. proposed a method to analyze Google search engine queries in the United States [1] and visualized the flu trends on Google map [2]. Vasileios Lampos et al. developed a system that monitored and analyzed the data stream of Twitter in the United Kingdom [3, 4]. The system computed and presented a flu-score on a daily basis.

In this paper, we observed tweets as social signals of flu symptoms to detect the flu epidemics in early stage and compared the tweets to the weather factors, flu forecast, and Influenza-like Illness (ILI) datasets. In order to do this experiment, we made a list of keywords that explicitly express the flu symptoms such as sneezing, cough, runny nose, etc. and collected tweets which contain the keywords from nine cities in Korea. Since the sizes of the cities are not the same, we observed and compared the trends instead of the exact numbers of the tweets. Daily four-level flu forecast from Korea Meteorological Administration (KMA) and weekly ILI proportion from Korea Center for Disease Control and Prevention (KCDC) were collected for the comparison. Since November 2011, KMA started furnishing the four-level flu forecast based on the temperature range, the minimum temperature, and the humidity of the day. The four levels are very high, high, normal, and low. KCDC reports every week about the influenza and the reports include the overall proportion of patients, who visited sentinel physicians for ILI in per millage.

This paper is organized as follows. In the next section, we explain the previous studies on tracking flu epidemics that gave us a hint for our research. We introduce the dataset which was used in this paper in Section 3, and our analysis results are presented in Section 4. Finally, we discuss about our results and conclude the paper in Section 5.

II. Related Work

Observation of the flu epidemics from the Internet have been tried before. Polgreen et al. analyzed internet search query logs on Yahoo! search engine to detect changes in disease activities [5]. They investigated the correlations between flu-related searches and actual flu occurrence. Their models predicted an increase in positive flu-cultures and in mortality from flu one to three weeks and up to five weeks in advance respectively. Ginsberg et al. from Google designed a method to estimate the current level of weekly flu activity in each region of the United States by utilizing search queries [1, 2]. Lampos et al. introduced an automated tool that tracks ILI in the United Kingdom from the contents of Twitter in a paper [3] and proposed a method for detecting the flu pandemic by monitoring the social web [4]. They expand the concepts to the general events. They presented a method for inferring the occurrence of an event by exploring the rich amount of unstructured textual data on the social web [6].

III. Data Collection

A. Collecting Data from Twitter

We made a list of keywords related to flu symptoms and collected tweets with the keywords from December 1, 2011 to April 30, 2012. KMA furnished the flu forecast for nine cities in Korea, so the tweets were gathered in the nine cities. We pointed the center of each city and collected tweets with different radius according to the size of the city.

B. Data Refinement

In order to compare the collected data, we had to normalize the data because each data has different scale. Every single data value was normalized into new scale range from 0.0 to 1.0 with a simple equation: v' = (v-Min(V))/(Max(V)-Min(V)) where V is a set of values, v is a single value, and v' is a normalized

COLLABORATECOM 2012, October 14-17, Pittsburgh, United States

DOI 10.4108/icst.collaboratecom.2012.250355

value. Ordinal values, the four-level flu forecast, were converted into numeric values from 1 (lowest) to 4 (highest).

IV. Analysis Result

We present the analysis results in this section. First of all, Figure 1 shows changes of the weather factors over time and the social signals in regard to flu on Twitter in Seoul Metropolitan Region (radius=50km). We considered the temperature range, the minimum temperature, and the humidity of each city as KMA uses these factors for the flu forecast. According to the research papers, the efficiency of influenza virus transmission is dependent on relative humidity and inversely correlated with temperature [7,8]. We can observe this fact on our graph. The minimum temperature and humidity are relatively low until mid February 2012, and the number of tweets is higher during this time than the last half. The flu signals are getting weaker from March 2012 as these two factors are increasing. However, we could not find a reasonable correlation between the temperature range and the flu signals.

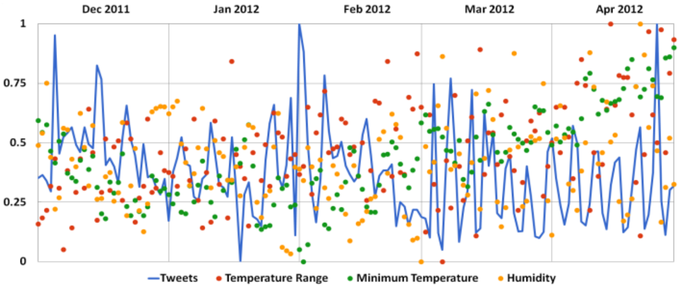

Figure 1. Flu Signals on Twitter and the Weather Factors

Figure 2 describes the comparison result between tweets and flu forecast of Seoul. The pattern of tweets seems very similar to the pattern of the flu forecast. This fact means that the flu forecast may be acceptable or our analysis system may work well.

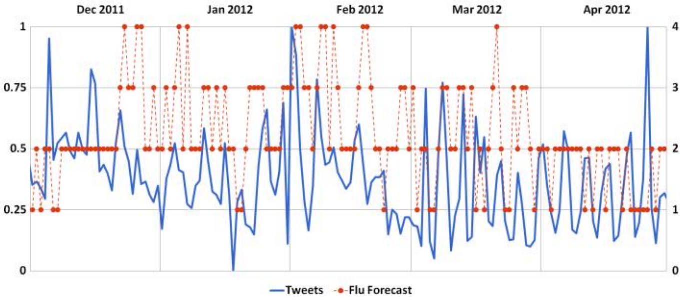

Figure 2. Flu Signals on Twitter and the Flu Forecast

For the last experiment, we compared the tweets with the weekly ILI proportion from KCDC. As shown in Figure 3, the flu signals on Twitter did not match well against the ILI graph. There are possible reasons for this result. As Ginsberg et al. claimed in their paper [1], the traditional reports require 1 or 2 weeks to gather and process data, and consequently there might be a delay between two datasets. The period of time would be another reason. It might be too short to get the statistically significant result, and we may have to observe the trends for a longer period of time.

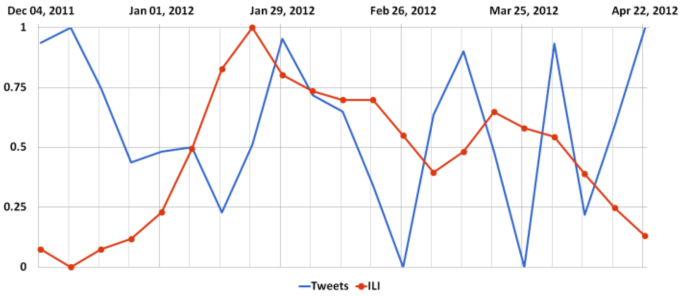

Figure 3. Weekly Flu Signals on Twitter and ILI

V. Conclusion

We observed the social signals of flu symptoms on Twitter and compared the tweet corpus to the weather factors, the flu forecast, and the ILI dataset. We depicted a set of graphs only with the data from Seoul Metropolitan Region in Korea. However, we collected tweet data from nine cities in Korea, and they show the similar aspects with the figures in this paper. Based on the results, we can build a system that warns the flu epidemics in real-time using the Streaming API of Twitter. Approach of our research is similar to the Google's flu trends, but using Twitter would be faster than the analyzing queries on Google. For the future work, we will track the social signals of various topics that can contribute to the life.

VI. Acknowledgements

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (No. 2011-0009407).

References

[1] Jeremy Ginsberg, Matthew H. Mohebbi, Rajan S. Patel, Lynnette Brammer, Mark S. Smolinski, and Larry Brilliant, "Detecting Influenza Epidemics using Search Engine Query Data," Nature, Vol. 457, 2009.

[2] http://www.google.org/flutrends/

[3] Vasileios Lampos, Tijl De Bie, and Nello Cristanini, "Flu Detector - Tracking Epidemics on Twitter," In Proc. of the European Conference on Machine Learning and Principles and Practice on Knowledge Discovery in Database (ECML PKDD 2010), pp. 599-602, 2010.

[4] Vasileios Lampos and Nello Cristianini, "Tracking the Flu Pandemic by Monitoring the Social Web," In Proc. of the 2nd International Workshop on Cognitive Information Processing, pp. 411-416, 2010.

[5] Philip M. Polgreen, Yiling Chen, David M. Pennock, Forrest D. Nelson, and Robert A. Weinstein, "Using Internet Searches for Influenza Surveillance," Clinical Infectious Diseases, Vol. 47, No. 11, pp. 1443-1448, 2008.

[6] Vasileios Lampos and Nello Cristianini, "Nowcasting Events from the Social Web with Statistical Learning," ACM Transactions on Intelligent Systems and Technology, Vol. 3, No. 3, Article 60, 2011.

[7] Brian P. Hanley and Birthe Borup, "Aerosol Influenza Transmission Risk Contours: A Study of Humid Tropics versus Winter Temperate Zone," Virology Journal 7: 98, 2010.

[8] Anice C. Lowen, Samira Mubareka, John Steel, and Peter Palese, "Influenza Virus Transmission is Dependent on Relative Humidity and Temperature," PLoS Pathogens, Vol. 3, Issue 10, pp. 1470-1476, 2007.