Tutorial

Healthc Inform Res. 2017 October;23(4):343-348.
https://doi.org/10.4258/hir.2017.23.4.343
pISSN 2093-3681 • eISSN 2093-369X

HIR
Healthcare Informatics Research

Methods Using Social Media and Search Queries to Predict Infectious Disease Outbreaks

Dong-Woo Seo, MD, PhD[1], Soo-Yong Shin, PhD[2]

[1]Department of Emergency Medicine, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Korea; [2]Department of Computer Science and Engineering, Kyung Hee University, Yongin, Korea

Objectives: For earlier detection of infectious disease outbreaks, a digital syndromic surveillance system based on search queries or social media should be utilized. By using real-time data sources, a digital syndromic surveillance system can overcome the limitation of time-delay in traditional surveillance systems. Here, we introduce an approach to develop such a digital surveillance system. **Methods:** We first explain how the statistics data of infectious diseases, such as influenza and Middle East Respiratory Syndrome (MERS) in Korea, can be collected for reference data. Then we also explain how search engine queries can be retrieved from Google Trends. Finally, we describe the implementation of the prediction model using lagged correlation, which can be calculated by the statistical packages, i.e., SPSS (Statistical Package for the Social Sciences). **Results:** Lag correlation analyses demonstrated that search engine data/Twitter have a significant temporal relationship with influenza and MERS data. Therefore, the proposed digital surveillance system can be used to predict infectious disease outbreaks earlier. **Conclusions:** This prediction method could be the core engine for implementing a (near-) real-time digital surveillance system. A digital surveillance system that uses Internet resources has enormous potential to monitor disease outbreaks in the early phase.

Keywords: Digital Syndromic Surveillance System, Disease Outbreak, Social Media, Search Engine

I. Introduction

Emerging infectious diseases, such as Severe Acute Respiratory Syndrome (SARS) in 2002, the H1N1 pandemic in 2009, and Middle East Respiratory Syndrome (MERS) in 2015, have highlighted the necessity for a syndromic surveillance system, which can play a significant role in detecting the beginning of an infectious disease outbreak [1-3]. However, traditional surveillance systems mainly depend on case reports, such as influenza-like illness (ILI) reports, which have time-delays in reporting and case confirmation. To enable the earlier detection of infectious disease outbreaks, a syndromic surveillance system should utilize real-time or near-real-time data, i.e., school or work absenteeism, over the counter medication sales, or the volume of Internet-based health inquiries [4-7]. Among diverse alternative data sources, search queries, social media, and website visits have proven potential for digital surveillance systems [3,5,7-10]. In Korea, there is a national health alert system operated by the National Health Insurance Service which uses data from social media and blogs [11].

Submitted: July 11, 2017
Revised: 1st, August 7, 2017; 2nd, August 24, 2017
Accepted: September 10, 2017

Corresponding Author
Soo-Yong Shin, PhD
Department of Computer Science and Engineering, Kyung Hee University, 1732 Deogyeong-daero, Giheung-gu, Yongin 17104, Korea.
Tel: +82-31-201-2543, E-mail: sooyong.shin@khu.ac.kr

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

ⓒ 2017 The Korean Society of Medical Informatics

In this paper, we briefly introduce how social media and search queries can be used to predict infectious disease outbreaks. This prediction method could be the core engine for implementing a (near-) real-time digital surveillance system.

II. Case Description

1. Data Collection

To implement a prediction model for the digital surveillance system, the statistics of an infectious disease and digital data, such as search engine queries or social media data, should be collected. Disease statistics are used as target data, and digital data is used as input for the prediction model.

1) Collection of disease statistics

ILI data in Korea can be gathered from Korea Centers for Disease Control and Prevention (KCDC) [12]. Since the ILI reports of the KCDC are published in word processor formats, such as HWP files in Korean and DOC files in English, the data should be manually curated. Figure 1 shows an example ILI report published by the KCDC. The ratio of ILI in week 28 was 6.0.

The MERS statistics of Korea in 2015 were released via the official governmental MERS statistics site. However, since that site was not accessible at the time of this study, the statistics could be alternately collected from Wikipedia [13]. When curating disease statistics, researchers should be cautious about time intervals. ILI data is weekly data, whereas MERS statistics are reported as daily data.

2) Choosing keywords

To collect digital data from search engine queries or Twitter, the keywords that relate to influenza or MERS need to be determined first. This step is the most important step for the performance of disease prediction model. To choose the keywords, both laypersons' opinions and experts' opinion are taken into consideration, since search queries and social media data are generated by laypersons. In our previous works regarding influenza prediction, we conducted a survey by quota sampling based on sex and age to choose keywords [8,9]. Moreover, we also included keywords that were chosen by physicians, such as fever, cough, and sore throat for influenza prediction, since these keywords are related to the definitions of ILI.

For MERS prediction, MERS and 메르스 (MERS in Korean) were chosen first. Then the top two related combined queries in Korean were added [3]. Moreover, selected keywords were translated into English or Korean since people can perform search queries in both languages. The chosen keywords are shown in Table 1.

◉ Influenza-like illness surveillance (Week 28: July 9, 2017 ~ July 15, 2017)

○ In week 28, Influenza-like illness (ILI) rate was 6.0/1,000 outpatients. The rate (6.0) was increased compared to the previous week (5.8)

※ The influenza warning threshold of 2016-2017 season was 8.9/1,000 outpatients

※ The influenza warning in 2016-2017 season: December 8, 2016 ~ June 2, 2017

Weekly ILI rate

Week	19	20	21	22	23	24	25	26	27	28
ILI rate (/1,000)	6.8	7.6	6.7	4.9	5.1	5.6	5.7	5.3	5.8	6.0

Figure 1. ILI report example from week 28, 2017 (July 9, 2017–July 15, 2017). The ratio is the number of outpatients divided by 1,000.

Table 1. The chosen keywords for influenza and MERS

	Keywords (Korean in parenthesis)
Influenza	Flu (플루), New flu (신종 플루), Abbreviated New flu (신플)[a], Influenza (인플루엔자), New influenza (신종 인플루엔자), Bad cold (독감), New bad cold (신종 독감), Epidemic bad cold (유행성 독감), H1N1[b], Bird flu (조류독감), Swine flu (돼지독감), Tamiflu (타미플루), Vaccine (백신), Prevention (예방), Mask (마스크), Symptom (증상), Sign (증세), Cough (기침), Fever (발열), Neck pain (목아픔), Sore throat (인후통), Throat pain (목통증), PCR[b], Treatment (치료), Complication (합병증), Decease (사망)
MERS	MERS (메르스), MERS symptom (메르스 증상), MERS hospital (메르스 병원)

MERS: Middle East Respiratory Syndrome, PCR: polymerase chain reaction.

[a]Only Korean keyword was used, [b]only English keywords were used.

https://doi.org/10.4258/hir.2017.23.4.343

3) Collection of data from search engines

The daily or weekly trends of the keywords for web search queries can be obtained from Google Trends [14] or Naver DataLab [15]. The newly renovated Naver DataLab only offers weekly trends, and it requires several steps to download the output. Therefore, we only explain how to use Google Trends. Google Trends offers various data patterns starting from the previous minute's data to monthly data based on the selected time range. For example, Figure 2 shows the weekly trend of searches for "influenza" of Google Trends between September 9, 2007, and September 8, 2012. Trend data associated with the predefined keywords were retrieved by setting the location parameter to "South Korea" and setting the time parameters. The time parameters were based on previous disease outbreak seasons, for example, "May 2015 to Jun 2015" for MERS in Korea. May 11, 2015 was the symptom onset day of the first laboratory-confirmed patient. The results can be downloaded in a CSV format by clicking the arrow located in upper right in the graph. The downloaded CSV file consists of two columns. The first column represents the week by indicating the starting date of the week and the second column is the trend data. The researcher should collect the necessary trend data for each keyword. Table 2 shows example influenza statistics from ILI reports (Figure 1) with Google Trends data for the following five keywords: "New influenza," "New flu," "Fever," "Tamiflu," and "Flu" (Figure 2). As shown in Table 2, there is a time-delay pattern between the influenza statistics and Google Trends data.

The important fact is that the current trend data obtained from Google Trends or Naver DataLab are available as normalized values, not the absolute numbers of searches. According to Google Trends, "Numbers represent search interest relative to the highest point on the chart for the given region and time. A value of 100 is the peak popularity for the term. A value of 50 means that the term is half as popular. Likewise, a score of 0 means the term was less than 1% as popular as the peak."

The previous site of Naver DataLab, called Naver Trends offered the separate data trends for mobile and desktop searches [16]. However, Naver DataLab does not distinguish between mobile and desktop trends as Google Trends does. In our previous work, we collaborated with Daum to collect curated data [8,9].

People search web pages using one or multiple words at a time. To reflect this behavior, combinations of the chosen keywords should be considered, for example, "H1N1", "H1N1 Treatment", "H1N1 Symptom", "Influenza", "Influenza Treatment", and "Influenza Symptom".

4) Collection of data from Twitter

Among diverse social media platforms, Twitter has been used as a data sources for digital surveillance systems [17]. In our previous work, the number of tweets containing one of the predefined keywords was collected through Topsy, which is a certified partner of Twitter that offers social searching and social analytics [3]. However, Topsy closed a few years ago. There are alternative services, such as GNIP [18] or Talkwalker [19]; however, researchers would be required to purchase the relevant data from them. Therefore, there is currently a barrier to using Twitter data.

2. Data Analysis

Spearman correlation analyses can be used to examine the correlations between search engine data and disease data. Lag correlation analyses can be used to assess the tempo-

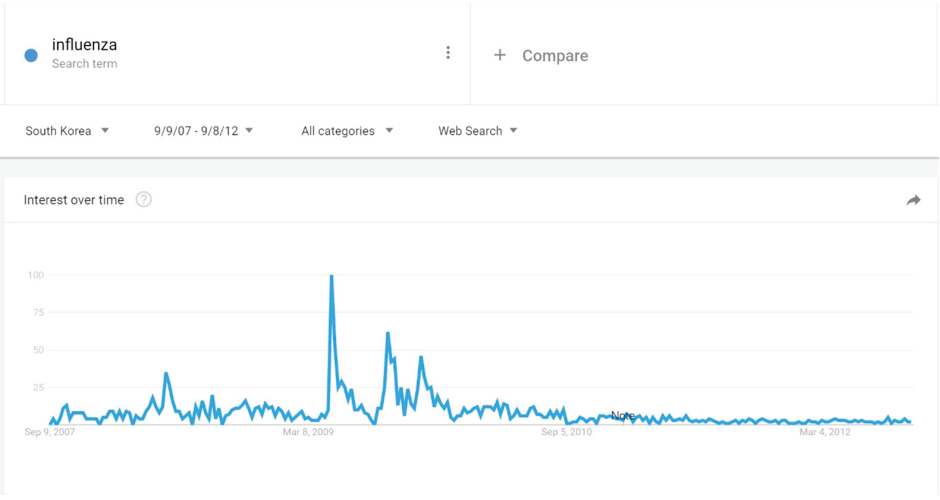

Figure 2. Trends of influenza search queries, according to Google Trends, between September 9, 2007 and September 8, 2012.

Table 2. Example of influenza data and Google Trends data

Date	KCDC data		Search engine query				
	Virological	ILI	New influenza	New flu	Fever	Tamiflu	Flu
06/07/2009	3.6	2.48	0	3	59	0	4
06/17/2009	2.7	2.43	0	4	39	0	5
06/21/2009	4.7	2.40	0	2	49	0	4
06/28/2009	0.0	2.13	0	3	40	0	2
07/05/2009	0.0	1.67	0	4	63	0	2
07/12/2009	1.5	1.69	0	5	59	0	6
07/19/2009	2.2	1.85	0	6	47	0	7
17/26/2009	1.5	2.01	0	3	34	0	5
08/02/2009	1.4	2.23	0	12	32	0	4
08/09/2009	9.4	1.81	0	22	61	0	4
08/16/2009	3.3	1.80	22	31	46	43	33
08/23/2009	3.5	2.76	48	60	54	100	63
08/30/2009	3.1	4.33	100	82	39	89	92
09/06/2009	5.1	5.37	65	53	41	67	57
09/13/2009	12.1	6.32	52	37	75	28	40
09/20/2009	9.1	6.47	16	18	50	18	21
09/27/2009	12.3	7.17	13	15	54	26	15
10/04/2009	16.0	7.26	13	11	33	34	11
10/11/2009	21.7	5.69	18	12	66	42	15
10/18/2009	39.3	9.26	23	20	34	50	23
10/25/2009	57.9	20.29	24	62	59	58	64
11/01/2009	63.9	41.73	15	50	84	64	60
11/08/2009	51.2	44.96	15	47	70	59	60
11/15/2009	46.3	37.71	17	50	61	38	55
11/22/2009	51.1	27.52	19	26	67	28	30
11/29/2009	57.2	28.32	0	19	60	16	20
12/06/2009	55.2	22.42	0	15	73	18	18
12/13/2009	46.7	18.62	0	13	54	20	16
12/20/2009	44.2	12.30	0	14	70	0	15
12/27/2009	42.9	13.15	0	18	40	0	17

Virological and ILI data were extracted manually from ILI reports. The trend data of representative keywords were extracted from Google Trends. Data column represents the week data with starting date.

KCDC: Korea Centers for Disease Control & Prevention, ILI: influenza-like illness.

ral relationships between these sets of data for up to user-defined days or weeks. Significance was set at $p < 0.05$. In our work, we used the SPSS package to obtain statistical values, and the proposed method is summarized as follows. The KCDC data values are placed tidily in a column (Table 2). The search engine query data values to be compare are placed next to it. The data must be arranged in rows so that each row represents a specific time period. The correlation value obtained in this state means present (0 week lag) [8]. To get the correlation coefficient of preceding or lagging week, the column containing a search engine query data value is moved in the desired direction and correlation analysis is performed. In this case, the data that differ by the number of weeks to be compare is placed in the same row.

 https://doi.org/10.4258/hir.2017.23.4.343

To see the changes in correlation coefficients over time, correlation coefficients in subsequent epidemiological intervals are calculated. These correlation coefficients can be used to validate the developed prediction model. If necessary, subgroup analyses for the period are conducted along the same lines. For example, we performed a subgroup analysis focusing on the acceleration and deceleration period of MERS (June 3, 2015–June 26, 2015) adopting the CDC interval [3].

III. Discussion

The proposed digital surveillance system which uses Internet resources has enormous potential to monitor disease outbreaks in the early phase; however, this approach has some limitations as well. First, in our work, it was difficult to choose keywords although they have a considerable effect on the performance of a prediction model. More importantly, keywords should be changed periodically. For example, before 2015 only experts were aware of MERS. However, most Koreans know about MERS nowadays. Since people continuously learn new terminology and change the search keywords they use, keywords should be updated regularly to maintain prediction performance [9]. Second, as in the case of Google Flu, this system can fail to predict disease outbreaks correctly [20]. Therefore, the proposed digital surveillance system should be used with caution or as a complementary method.

Conflict of Interest

No potential conflict of interest relevant to this article was reported.

References

1. Peiris JS, Guan Y, Yuen KY. Severe acute respiratory syndrome. Nat Med 2004;10(12 Suppl):S88-97.
2. Novel Swine-Origin Influenza A (H1N1) Virus Investigation Team, Dawood FS, Jain S, Finelli L, Shaw MW, Lindstrom S, et al. Emergence of a novel swine-origin influenza A (H1N1) virus in humans. N Engl J Med 2009;360(25):2605-15.
3. Shin SY, Seo DW, An J, Kwak H, Kim SH, Gwack J, et al. High correlation of Middle East respiratory syndrome spread with Google search and Twitter trends in Korea. Sci Rep 2016;6:32920.
4. Henning KJ. What is syndromic surveillance? MMWR Suppl 2004;53:5-11.
5. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L. Detecting influenza epidemics using search engine query data. Nature 2009;457(7232):1012-4.
6. Triple S Project. Assessment of syndromic surveillance in Europe. Lancet 2011;378(9806):1833-4.
7. Eysenbach G. Infodemiology: tracking flu-related searches on the web for syndromic surveillance. AMIA Annu Symp Proc 2006;2006:244-8.
8. Cho S, Sohn CH, Jo MW, Shin SY, Lee JH, Ryoo SM, et al. Correlation between national influenza surveillance data and google trends in South Korea. PLoS One 2013;8(12):e81422.
9. Seo DW, Jo MW, Sohn CH, Shin SY, Lee J, Yu M, et al. Cumulative query method for influenza surveillance using search engine data. J Med Internet Res 2014;16(12):e289.
10. Hulth A, Rydevik G, Linde A. Web queries as a source for syndromic surveillance. PLoS One 2009;4(2):e4378.
11. National health alert system [Internet]. Cheongju: National Health Insurance Service; c2017 [cited at 2017 Jul 8]. Available from: http://forecast.nhis.or.kr/menu.do.
12. Korea Centers for Disease Control & Prevention. KCDC ILI reports [Internet]. Cheongju: Korea Centers for Disease Control & Prevention; c2017 [cited 8 Jul 2017]. Available from: http://www.cdc.go.kr/CDC/info/CdcKrInfo0502.jsp?menuIds=HOME001-MNU1175-MNU0048-MNU0050.
13. Wikipedia. 2015 Middle East respiratory syndrome outbreak in South Korea [Internet]. [place unknown]: Wikipedia; c2017 [cited at 2017 Jul 8]. Available from: https://en.wikipedia.org/wiki/2015_Middle_East_respiratory_syndrome_outbreak_in_South_Korea.
14. Google Trends [Internet]. Mountain View (CA): Google; c2017 [cited at 2017 Jul 8]. Available from: https://trends.google.com/trends/.
15. Naver DataLab [Internet]. Seoul: Naver; c2017 [cited at 2017 Jul 8]. Available from: http://datalab.naver.com/.
16. Shin SY, Kim T, Seo DW, Sohn CH, Kim SH, Ryoo SM, et al. Correlation between National Influenza Surveillance Data and Search Queries from Mobile Devices and Desktops in South Korea. PLoS One 2016;11(7):e0158539.
17. Signorini A, Segre AM, Polgreen PM. The use of Twitter to track levels of disease activity and public concern in the U.S. during the influenza A H1N1 pandemic. PLoS One 2011;6(5):e19467.
18. GNIP 2.0 [Internet]. San Francisco (CA): Twitter Inc.;

c2017 [cited at 2017 Jul 8]. Available from: https://gnip.com.
19. Talkwalker [Internet]. New York (NY): Talkwalker Inc.; c2017 [cited at 2017 Jul 8]. Available from: https://www.talkwalker.com.
20. Lazer D, Kennedy R, King G, Vespignani A. The parable of Google Flu: traps in big data analysis. Science 2014; 343(6176):1203-5.

 https://doi.org/10.4258/hir.2017.23.4.343