Comparing Social Media and Search Activity as Social Sensors for the Detection of Influenza

Mizuki Morita
National Institute of Advanced Industrial Science and Technology (AIST)
mori-ta.mizuki@aist.go.jp

Sachiko Maskawa
Photonic System Solutions
sachiko.maskawa@gmail.com

Eiji Aramaki
Kyoto University/PRESTO
ei-ji.aramaki@gmail.com

Abstract

Detecting the incidence and prevalence of infectious diseases is important because these diseases affect many people, raise the cost of healthcare, and, in some cases, can lead to a great many deaths. Recently, it has been shown that people's online activity can be applied to detecting the prevalence of influenza. In this study, we compare the characteristics of two kinds of online activities, social media and search activity, as the social sensors capable of supplying information on seasonal epidemics and an unexpected pandemic of influenza. Although both approaches showed quite high performance for the seasonal epidemics, they showed poor performance for the unexpected influenza pandemic. The social media-based approach particularly over-responded to the influenza pandemic.

1 Introduction

Influenza has persisted as a major worldwide public health concern. Although it was discovered early in the last century that a virus causes influenza (Yamanouchi et al., 1919; Smith et al., 1933), influenza has persisted in threatening people and has led to the deaths of a huge number of people around the world (World Health Organization, 2003). Seasonal influenza epidemics typically occur in winter across temperate regions of the world. Although unexpected influenza pandemics rarely occur, we experienced them three times in the 20th century: the Spanish flu, Asian flu, and Hong Kong flu (World Health Organization, 2009).

The routes of transmission of influenza are known. Therefore, effective ways exist to prevent transmission of influenza, such as vaccination; keeping hands clean; avoiding touching the mouth, nose, and eyes; and using a face mask (World Health Organization, 2010). For early detection of the onsets of influenza epidemics and pandemics, some countries have adopted public surveillance agencies such as the Center for Disease Control and Prevention (CDC) in the US, the European Centre for Disease Prevention and Control (ECDC) in the EU, and the Infectious Disease Surveillance Center (IDSC) in Japan (note that the early detection of infectious diseases is an important task, but it is not the organizations' sole reason for existence). A major concern is that reports from these agencies typically have a time lag of 1–2 weeks.

Recently, by aiming at earlier detection of influenza onset, internet-based approaches have adopted access logs of health-related websites (Johnson et al., 2004), search queries to the popular search engine Yahoo! (Polgreen et al., 2008), search queries to a medical website (Hulth et al., 2009), search queries to Google (Ginsberg et al., 2009), and posts on the microblogging site Twitter (Aramaki et al., 2011; Signorini et al., 2011). People's behavior on the internet can be divided into two camps: communicating with others, such as through posting on social media (e.g., Twitter, Facebook, and LinkedIn), and searching for themselves, such as querying search sites (e.g., Google, Yahoo! and Microsoft Bing).

In this study, we have attempted to characterize both social media-based and search activity-based approaches to detecting influenza in respective case of seasonal epidemic and unexpected pandemic influenza prevalence.

2 Methods

We implemented an approach to estimate the number of influenza patients based on Twitter posts in Japanese (Aramaki et al., 2011). We also used results from a Google search query-based approach (Ginsberg et al., 2009). We evaluated the performance of these approaches against public surveillance data in Japan.

Here, the annual influenza season in Japan typically starts from November through December and subsides sometime in April or May, although the trends of the number of patients and the extent of epidemics vary from year to year. In 2009, an extremely important public health issue facing the world was influenza A (H1N1), also known as "swine flu," the first influenza pandemic of the 21st century. In Japan, the influenza A(H1N1)pdm09 viruses were first detected in three returning travelers from abroad on May 9, 2009. After several hundred patients with influenza A (H1N1) had been found, the Japanese Ministry of Health, Labour and Welfare (MHLW) presented a perspective that the number of newly infected patients was decreasing in late May.

Twitter data

We collected Twitter posts in Japanese during November 2008 – July 2009 and November 2012 - June 2013 via the Twitter API (`https://dev.twitter.com/`). We extracted influenza-related Twitter posts, which contained the words "influenza" or "flu" (corresponding words in Japanese, "インフルエンザ" and "インフル," were actually used). The period between November 2008 and April 2009 was defined as the "epidemic period of 2008–2009," that between November 2012 and June 2013 as the "epidemic period of 2012-2013," and that between April 2009 and July 2009 as the "pandemic period of 2009." Although there were actually two waves of the pandemic influenza in 2009 (spring and winter), we only used the first wave for our analyses because the second wave overlapped with the epidemic influenza season of 2009–2010.

2.1 Twitter post-based approach

The total number of influenza patients was calculated according to Twitter posts by influenza patients. Discriminating positive influenza posts from noise posts was conducted as a sentence classification task using natural language processing (NLP) similar to that used to filter spam e-mail. For this task, we implemented a straightforward influenza positive/noise discriminator for Twitter posts based on a machine learning approach, the Support Vector Machine (SVM) (Cortes and Vapnik, 1995). Sentences in each Twitter post were initially divided into words with a morphological analyzer JUMAN (`http://nlp.ist.i.kyoto-u.ac.jp/EN/index.php?JUMAN`) to separate words. Six words both immediately before and after an influenza-related word in a Twitter post (12 words maximum) were selected as input for the SVM discriminator. The details of parameter tuning were described before (Aramaki et al., 2011). Training and performance evaluation was done through a 10-fold cross validation using 922 influenza-related Twitter posts in November 2009 (Twitter posts in this period were not used in the remainder of analyses in this paper), which were annotated manually by Japanese native speakers as either positive or noise influenza Twitter posts (caused 454 positive and 468 negative posts). The performance (F-measure, which is the harmonic mean of precision and recall) of this approach with ten-fold cross validation was 0.76. The number of positive Twitter posts divided by the total number of Twitter posts in Japanese in the same term was defined as the estimated relative number of influenza patients in Japan.

2.2 Google search query-based approach

The relative number of influenza patients estimated by Google was obtained from the Google Flu Trends (Japan Edition) website (`http://www.google.org/flutrends/jp/`). Ginsberg et al. described the algorithm (Ginsberg et al., 2009). In short, Google Flu Trends estimated the number of influenza patients based on the frequency of influenza-related Google search queries, which they found to have a high correlation with the number of patients who consulted physicians.

2.3 Observed number of influenza patients

We obtained the observed number of influenza patients in the epidemic periods of 2008-2009 and 2012-2013, and the pandemic period in 2009 in Japan from the official sentinel survey by the Infectious Disease Surveillance Center (IDSC; `http://idsc.nih.go.jp/`) at the National Institute of Infectious Disease (NIID) of Japan. The numbers of observed patients are based on weekly reports from around 5,000 fixed-point

medical facilities dispersed throughout Japan (about 2,000 internal medicine and 3,000 pediatric departments were selected) to the IDSC under the Law concerning the Prevention of Infections and Medical Care for Patients of Infections. Categorizing influenza-related Twitter posts, we manually classified 200 randomly selected influenza-related Twitter posts from the epidemic period of 2008-2009 and the pandemic period of 2009 into five categories: "Influenza positive" (Twitter posts from influenza patients), "Influenza negative" (includes negations, influenza-positive in the past but already recovered, and receiving vaccination), "Mention or joke about influenza," "News," and "Others."

3 Results

3.1 Performance of the social sensors for epidemic and pandemic influenza

During the seasonal epidemic periods of 2008–2009 and 2012-2013, the number of influenza patients estimated by both the Twitter post-based and the Google search query-based approaches showed good correlations (r=0.82 and r=0.93; r=0.82 and r=0.89) with the number of patients by the public surveillance system (Fig. 1).

However, during the pandemic period of 2009, both Twitter post-based and Google search query-based approaches showed extremely poor performance (r=-0.02 and r=0.23) as predictors of the number of influenza patients (Fig. 2). Although both approaches were able to react to the influenza pandemic, their responses were irrelevant. The Twitter post-based approach particularly over-responded to the pandemic.

3.2 Characterization of Twitter posts in the influenza epidemic and pandemic periods

To identify the cause of the failure in predicting the number of patients during the pandemic period, we sought to identify the differences between the properties of Twitter posts in the epidemic and pandemic periods.

The breakdown of 200 randomly selected Twitter posts that occurred during the two influenza outbreaks were examined (Fig. 3). The performances (F-measures) of the positive/noise discriminator against 200 randomly selected Twitter posts were 0.64 for the epidemic and 0.05 for the pandemic periods. During the pandemic period, the proportions of "Mention or joke about influenza," "News," and "Others" were much higher than those in the epidemic pe-

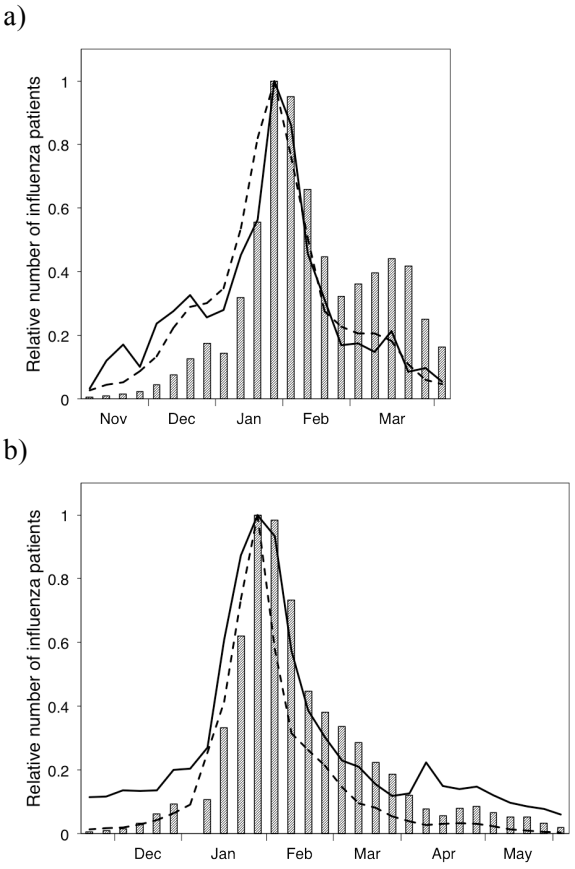

Figure 1. Trends in the relative number of influenza patients per week in the epidemic periods of a) 2008-2009 and b) 2012-2013 in Japan. The values were normalized with those at the peak of each season. The relative quantities of influenza patients per hospital are shown as bars. The estimated relative numbers by Twitter post-based approach are depicted as a solid line. Those by a Google search query-based approach are shown by the dotted line.

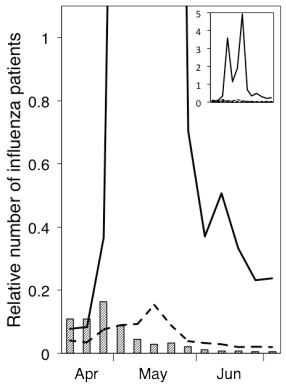

Figure 2. Trends in the relative number of influenza patients per week in the pandemic period of 2009 in Japan. The values were normalized with the same one in Fig. 1a. Data are presented in the same way as in Fig. 1.

riod, and 93% (185/200) of all randomly selected influenza-related Twitter posts fell into one of these three categories. Furthermore, the performances of the positive/noise discriminator were low for these three categories, especially for "Mention or joke about influenza."

4 Discussion

Both Twitter and Google are applicable to the early detection and survey of seasonal epidemic influenza because the correlation of the numbers of patients estimated by the Twitter post-based and Google search query-based approaches with those by the public surveillance system was high (Fig. 1). The Twitter-based and Google-based systems perform in real time, which can reduce the time lag of the current public surveillance systems, making it an actual feasible alternative that is one step ahead of the current official sentinel surveillance system. Because the numbers of Twitter and Google users are notably higher than the number of monitoring spots for the public surveillance system, Twitter and Google were able to monitor epidemics in higher geographic resolution, especially in densely populated areas, where infectious diseases can spread easily among people.

However, the Twitter post-based sensor caused a panic during the influenza pandemic. Moreover, even though it functioned better as a sensor than not responding at all to outbreaks, it completely failed to follow the trends of the outbreak (Fig. 2). If the performance of the positive/noise discriminator was sufficiently high, then such false-positive Twitter posts should have been removed properly. The actual performance was, however, very low in the pandemic period, with an F-measure of 0.05. A major cause of this catastrophic failure appears to be the types of Twitter posts observed frequently during this period. Of all influenza-related Twitter posts, 93% were classified as noise ("Mention and joke about influenza," "News," and "Others") (Fig. 3). The discriminator performance for these categories was poor, perhaps because the posts rarely contained frequently appearing expressions. It was difficult for a supervised machine learning approach when using bag-of-words as features to discriminate between positive data and noise.

Although the Google search query-based approach also got confused, it behaved much more moderately than the Twitter post-based approach did. An important difference between Twitter and Google is that Twitter is a kind of communication tool. Its users have the intention to have their posts read by other users. Studies in psychology have revealed that the following three conditions are related to rumor transmission: *personal anxiety*, *general uncertainty*, and *credulity* (trust in the rumor) (Rosnow, 1991). The environment was ripe for rumors to spread during the influenza pandemic period of 2009. A pandemic influenza in general is able to cause the death of millions of people. Also, it was initially reported that the mortality rate was extremely high in Mexico, and that the supply of vaccine was behind production schedule (*personal anxiety*). Infectability and mortality rates in developed countries were undetermined (*uncertainty*). The government continuously issued official announcements (*credulity*), though they were both unclear and several steps behind. Although some gap might exist in separating the number of Twitter posts and the rate of rumor transmission, the knowledge clarified in the traditional studies of social conversations is apparently applicable to studies of communication through Twitter. Microblogging is not the same as traditional social conversation, but it is apparently related to it. Therefore, we can assume that the amount of Twitter posts will explode under unusual situations such as life-threatening pandemics and bioterrorist attacks.

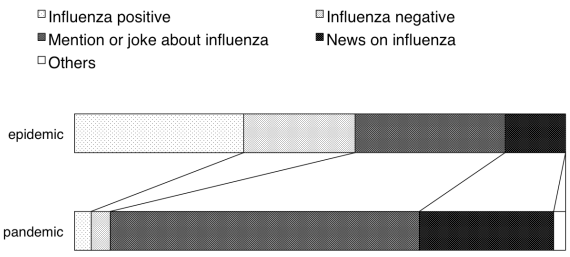

Figure 3. Comparison of types of influenza-related Twitter posts between the epidemic (upper) and the pandemic (lower) influenza periods. "Influenza positive" means Twitter posts from influenza patients. "Influenza negative" means Twitter posts about negation, positive in past but already recovered, and vaccination. The widths are proportional to the ratio of each type of post out of 200 randomly selected Twitter posts.

5 Conclusion

In this study, we have characterized two types of influenza sensors that were based on people's online behavior. Both the social media-based and

the search activity-based approaches could detect seasonal epidemic periods of influenza with fairly good performances, whereas the social media-based approach over-responded to the unexpected pandemic influenza. The number and type of posts on social media are likely to be affected by the condition for rumors to spread, which makes the social media-based approach less effective under such conditions.

Acknowledgements

The authors thank Kazutaka Baba for supporting software maintenance and Genta Kaneyama for providing us with Twitter data. This work was partially supported by the Precursory Research for Embryonic Science and Technology (PRESTO) program of the Japan Science and Technology Agency (JST).

References

Eiji Aramaki, Sachiko Maskawa, and Mizuki Morita. 2011. Twitter Catches The Flu: Detecting Influenza Epidemics using Twitter. EMNLP: 1568-1576.

Corinna Cortes and Vladimir N. Vapnik. 1995. Support-vector networks. Machine Learning 20: 273-297.

Jeremy Ginsberg, Matthew H. Mohebbi, Rajan S. Patel, Lynnette Brammer, Mark S. Smolinski, and Larry Brilliant. 2009. Detecting influenza epidemics using search engine query data. Nature 457: 1012-U1014.

Anette Hulth, Gustaf Rydevik, and Annika Linde. 2009. Web Queries as a Source for Syndromic Surveillance. PLoS ONE 4: e4378.

Heather A. Johnson, Michael M. Wagner, William R. Hogan, Wendy Chapman, Robert T. Olszewski, John Dowling, and Gary Barnas. 2004. Analysis of Web access logs for surveillance of influenza. Stud Health Technol Inform 107: 1202-1206.

Philip M. Polgreen, Yiling Chen, David M. Pennock, and Forrest D. Nelson. 2008. Using Internet Searches for Influenza Surveillance. Clinical Infectious Diseases 47: 1443-1448.

Ralph L. Rosnow. 1991. Inside rumor: A personal journey. American Psychologist 46: 484-496.

Alessio Signorini, Alberto M. Segre, and Philip M. Polgreen. 2011. 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 6: e19467.

Wilson Smith, C. H. Andrewes, and P. P. Laidlaw. 1933. A virus obtained from influenza patients. *Lancet* 2: 66-68.

World Health Organization. 2003. Influenza Fact sheet N°211. Available: http://www.who.int/mediacentre/factsheets/2003/fs211/en/. Accessed June 12, 2012.

World Health Organization. 2009. Influenza Fact sheet N°211. Available: http://www.who.int/mediacentre/factsheets/fs211/en/. Accessed June 12, 2012.

World Health Organization. 2010. What can I do? Available: http://www.who.int/csr/disease/swineflu/frequently_asked_questions/what/en/index.html. Accessed June 12, 2012.

T. Yamanouchi, K. Sakakami, and S. Iwashima. 1919. The infecting agent in influenza: An experimental research. *Lancet* 1: 971-971.