A Geospatial Analysis on the Potential Value of News Comments in Infectious Disease Surveillance

Kainan Cui[1,2], Zhidong Cao[2], Xiaolong Zheng[2], Daniel Zeng[2],
Ke Zeng[1,2], and Min Zheng[1,3]

[1] The School of Electronic and Information Engineering,
Xi'an Jiaotong University, China
[2] The Key Lab of Complex Systems and Intelligence Science, Institute of Automation,
Chinese Academy of Sciences, China
[3] Department of Communication, Engineering College of Armed Police Force, China
kainan.cui@live.cn

Abstract. With the development of Internet, widely kind of web data have been applied in influenza surveillance and epidemic early warning. However there were less works focusing on the estimation of geospatial distribution of influenza. In order to evaluate the potential power of news comments for geospatial distribution estimation, we choose the H1N1 pandemic in the mainland of China in 2009 as case. After collecting 75878 comments of H1N1 related news from www.sina.com(a famous news site in the mainland of China), we compared the geospatial distribution of comments against surveillance data. The result shows that the comments data share a similar geospatial distribution with the epidemic data(a correlation of 0.848 p<0.01), especially with a larger data volume(a correlation of 0.902 p<0.01). It suggests that extracting geospatial distribution from comments data for estimation could be an important supplementary method when the surveillance data are incomplete and unreliable.

Keywords: H1N1, infectious diseases, surveillance, geospatial analysis, open source information.

1 Introduction

The prevalence of internet and the threat of emerging infectious disease have driven growing interest in web-based public health surveillance[1, 2]. There are already lots of attempts and applications using different web data sources such as search engine logs[3, 4], news[5], blogs[6, 7], micro-blog[8, 9], wiki[10] and so on. Unfortunately considerable part of the existing works ignored the geographic information contained in the web data and focused on the analysis the temporal dynamic of influenza[11].

The geospatial distribution of infectious diseases is critically important for disease monitoring and control. Although geographic information system (GIS) has already been widely used on visualization and analysis for both epidemic data and web data [12-17], there were little work about evaluating the correlation of web data against

M. Chau et al. (Eds.): PAISI 2011, LNCS 6749, pp. 85–93, 2011.

epidemic data. The uncertain of the correlation weaken the significance of GIS based on web data. In another word, quantitative the correlation of geospatial distribution of web data against epidemic data is the first step for geospatial distribution estimation, and is crucial for the GIS of public health surveillance based on web data.

Fortunately, the development of open source information provides us increasing amount of data with geographic information, which enable us analyze those data from a spatial perspective. News comment is an example of web data with geographic information. The available of geographical information enable us study news comments from a spatial perspective. One of the advantages of news comments is that news comments are straightforward to retrieve. Compared to other web data such as blog and query logs， the news comments are more relevant to certain topic, which means we do not have to apply nature language processing technology for data filtering. When big event happens, the news site will offer a special report containing all related news, which will serve as comments aggregator. In another word, the web sites have done the classification and clustering of comments before we retrieve them. In this paper, we collected comments from www.sina.com and compared the geospatial distribution of news comments against the epidemic data of H1N1 outbreaks in the mainland of China in 2009. The result shows a high positive correlation, which revealed the potential power of news comments for geospatial distribution estimation.

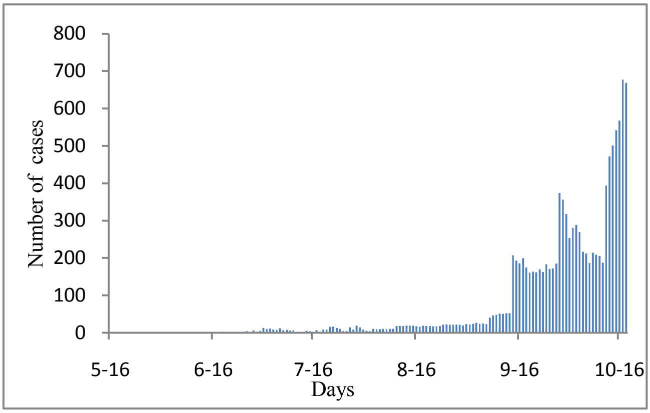
The remainder of this paper is structured as follows. In Section 2, we introduce the epidemiological data, comments data and the analysis method used in our study respectively. In Section 3 we show the result of our geospatial correlation analysis. Based on the result we discuss the possible usages and further work in Section 4. At last we present a conclusion of this paper in Section 5.

2 Data and Methods

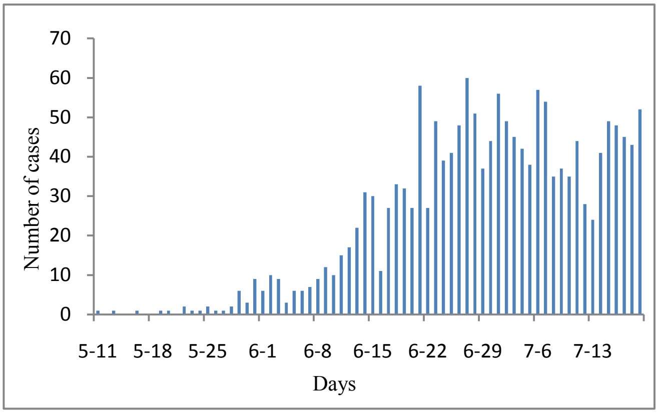
2.1 Epidemic Data

We chose the H1N1 influenza outbreak in the mainland of China in 2009 as case to evaluate the correlation of the geospatial distribution of comments data against epidemic data. The epidemic data used in our study is constituted by two parts. One dataset is obtained through an authority website, which serves 31 province-level regions in the mainland of China (we refer that dataset as CH-ED). The other dataset came from the Beijing Center for Disease Control and Prevention (CDC), which contains the number of reported H1N1 cases in Beijing over 15 geographic divisions (referred as BJ-ED). There are 14 districts and 2 countries in Beijing. For the consideration of reducing the space size difference among regions, we merged Dongcheng District and Xicheng District as city center region in this study.

The CH-ED covers the period from May 11, 2009 to July 9, 2009, and the BJ-ED covers the period from May 16, 2009 to October 18, 2009. Figure 1 show the series plots of the two datasets. Figure 1 (a) is for CH-ED and Figure 1 (b) is for BJ-ED. There is another part of data after October 18, 2009 from CDC, which was estimated based on sample survey. We do not use that part of data with the consideration for accuracy.



(a)



(b)

Fig. 1. Time series plots of number of H1N1 cases, (a) is the reported case number in the mainland of China covering the period from May 11, 2009 to July 9, 2009; (b) is the reported case number in Beijing from May 16, 2009 to October 18, 2009

2.2 Comments Data

We obtained comments dataset from www.sina.com (referred as SINA-CD). www.sina.com is a famous news service offering a full array of Chinese-language news. A special report is a universal entrance for all news related to certain topic in a site, which means we can retrieve all comments of one topic by traversal. The special report for H1N1 is available at http://news.sina.com.cn/z/zhuliugan/index.shtml. We developed a customized crawler for collecting data. After removing duplicated data, there are 75878 comments covering the period from May 27, 2009 to August 11, 2010. Figure 2 shows the time series plot of SINA-CD. Each record in comment dataset contains the following elements: Comment ID, Post time, News title, News URL, Content and Location. Prior to analyze the data, we normalized the comments for duplication. The comments ID were checked and those records whose id has already appeared were removed. The post time was used for time series plot. After a simple analysis of the comments number at different time interval on a day, we found that users like to replay news in morning. The maximum value appears in the period between 9 am and 10 am and the minimum value appears in 3 am and 4 am. In this study we mainly explore the location information. Other elements such as the content can be used for further study.

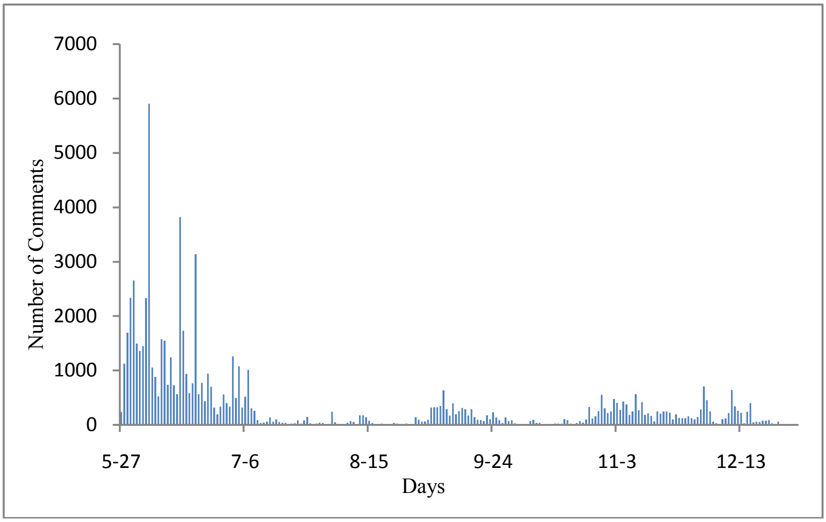


Fig. 2. Time series plots of number of H1N1 related news comments from www.sina.com covering period is from May 27, 2009 to August 11, 2010

The form of location information in comments dataset is string. For most comments the geographical accuracy is in prefecture-level, however for some comments from Beijing, the geographical accuracy is in country-level. We count the comments number in different geographical level by string matching. In particular, we get the comments number for 31 province-level regions in the mainland of China and the comments number for 14 districts and 2 countries in Beijing. After adding the comments number in Dongcheng District and Xicheng District together as comments number in city center region, we get comments number for 15 geographic divisions in Beijing.

2.3 Correlation Analysis

Two geospatial correlation analysis were performed in different geographical levels. We chose the Pearson's Correlation Coefficient to measure the difference between comments and epidemic data in both analyses. In particular, we computed the correlation value of daily counts of comments in SINA-CD against influenza surveillance data in CH-ED and BJ-ED respectively. We defined $P < 0.01$ as statistically significant. After the correlation analyses, several geospatial distribution plots were generated for visual comparison on geospatial patterns of comment data and epidemic data.

The comparison between CH-ED and SINA-CD were referred as SP-C1. This analysis is performed in province-level for the reason that the CH-ED is in province-level. Although the SINA-CD can reach to prefecture-level, we cannot perform further analyses unless the epidemic data with higher accuracy is available.

The comparison between BJ-ED and SINA-CD were referred as SP-C2. This analysis is performed in prefecture-level. In this case the geographical accuracy of BJ-ED is higher than SINA-CD, which means if we could get comments data in township-level, a more detailed comparison is available. For the inconsistent of data period, we align both start date and end date to the earlier one respectively, which means for SP-C1 the data period is form May 11, 2009 to July 9, 2009, for SP-C2 the data period is form May 16, 2009 to October 18, 2009.

3 Result

3.1 SP-C1: Spatial Correlation of CH-ED against SINA-CD

Figure 3 depicts the geospatial distribution of the number of reported H1N1 case before July 19, 2009 and the number of comments of H1N1 news from

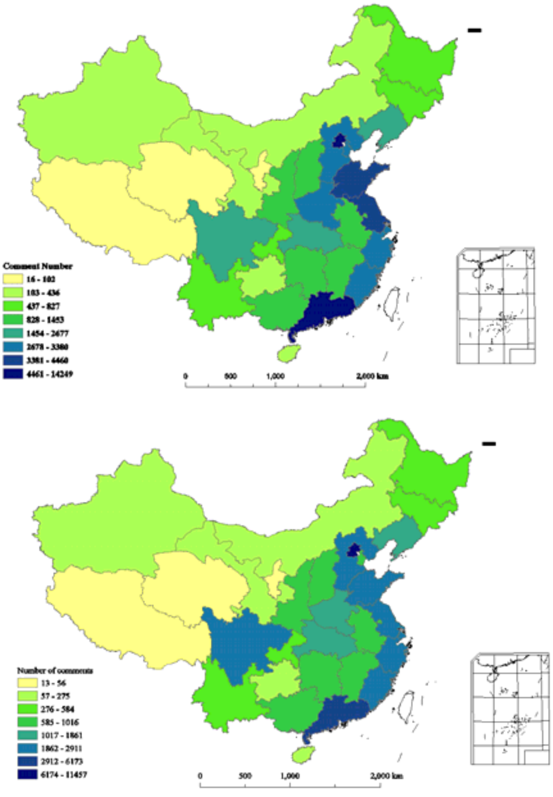


Fig. 3. A comparison of the number of reported H1N1 cases before July 19, 2009 against the number of comments that reply to H1N1 news from www.sina.com. A correlation of 0.848 (*p<0.01*) was obtained over 31 province-level regions.

www.sina.com, which shows that the two datasets have a similar geospatial distribution. From a qualitative perspective, the two datasets have a high correlation of 0.848 (p<0.01) which was obtained by correlation analysis over 31 province-level regions.

3.2 SP-C2: Spatial Correlation of BJ-ED against SINA-CD

Figure 4 depicts the geospatial distribution of the number of reported H1N1 cases before October 18, 2009 and the number of comments of H1N1 news from www.sina.com in Beijing, which shows that the two datasets have a similar geospatial distribution too. From a qualitative perspective, the two datasets have a higher correlation of 0.902 (p<0.01).

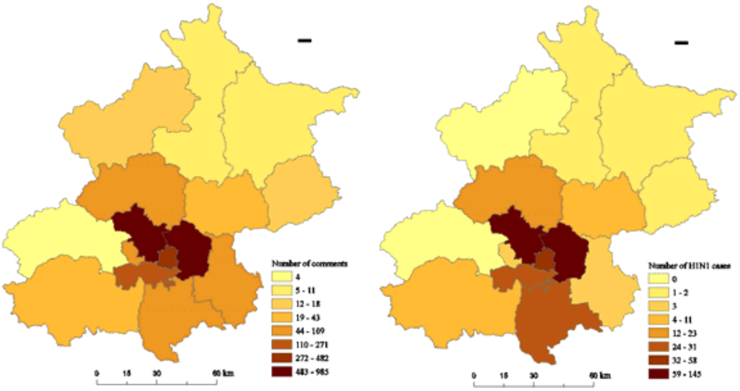


Fig. 4. A comparison of the number of reported H1N1 cases before October 18, 2009 against the number of comments that reply to H1N1 news from www.sina.com. A correlation of 0.902 (*p<0.01)* was obtained over 15 geographic divisions.

Table 1. A comparison for the two spatial correlation analyses

ID	Epidemic data	Number of days covered by epidemic data	Spatial correlation
SP-C1	CH-ED	59	0.848
SP-C2	BJ-ED	155	0.902

3.3 A Comparison of SP-C1 against SP-C2

Table 1 shows the comparison for SP-C1 against SP-C2, the BJ-ED contains data of 155 days has a correlation of 0.902 (p<0.01) with SINA-CD over 15 districts, and the CH-ED contains data of 59 days has a correlation of 0.848 (p<0.01) with SINA-CD over 31 provinces.

4 Discussion

In this paper we attempt to show the correlation of comments of public health related news with epidemic data in a spatial perspective, we used 2009 H1N1 outbreak in the mainland of China as case and perform a correlation analysis to qualitative the relationship. In this section we will first discuss the features and possible usage of comments, and then we will discuss the limitation and future work.

4.1 Features and Possible Usage of News Comments

As a kind of web data, news comments have features affect the way we use it for public health surveillance. Take comments from H1N1 news in www.sina.com as example, there are two main features

- Topic related
- Anonymity

Contrary to other web data like blogs, advertisement click records and so on, the most notable feature of news comments is the topic related. Firstly the special reports were arranged by editors. Those editors will ensure all news in a special report was surrounded to certain topic. Secondly the administrator will review the comments and delete those unrelated posts. This feature means we do not need a keyword matching for retrieving data. General speaking, keyword matching result may includes unrelated noisy record that will lead error in subsequent analysis. In order to attract more comments, the www.sina.com allows user post comment without registration, which means we cannot analyze the relationship among users by applying complex network theory.

The correlation analysis result shows that the comments data have a similar geospatial distribution with epidemic data. This is the main finding in this paper. It is worth to note and that the correlation value is higher when we applied the correlation analysis to the epidemic dataset with larger data volume. In particular, the BJ-ED contains data of 155 days has a correlation of 0.902 ($p<0.01$) with SINA-CD over 15 districts, and the CH-ED contains data of 59 days has a correlation of 0.848 ($p<0.01$) with SINA-CD over 31 provinces. Because of the high geospatial correlation, we may use the geospatial patterns extracted from comments data to estimate the epidemic situation in certain area with survey data in other regions, which could reduce the cost for epidemiological investigations, especially when the surveillance data are incomplete and unreliable for target area.

4.2 Limitations and Future Work

Contrary to other web data the main limitation of news comments for public health surveillance is the timeliness. In H1N1 outbreak in the mainland of China, the first comment appears several days after the report case. Although the development of smart phones and cloud computing will allow user post comment more easily and increase the response speed indirectly, the comments are still disadvantage as data source for early warning systems of public health.

Another limitation is the media interest and public interest for a certain infectious disease. H1N1 is a brand-new disease affecting the globe which leads news site’s special report and thousands of comments by web users. However for other diseases the web users concerned not very much, there may be not enough comments in the news site for research.

There are two main directions for the future work. The first one is to confirm the result that news comments have a high geospatial correlation against epidemic data with more data, we plan to collect more comments data and build geospatial model for prediction. Through comparing the output of model with epidemic data, we will quantitative the prediction power of news comments. Another direction is to mine more information in the comments, such as applying natural language processing technology to analyze the content of comments. The semantic information contained in comments such as the emotion or opinion about certain disease control measure may help us archiving a more comprehensive understand of influenza outbreak.

5 Conclusion

The high spatial correlation against epidemic data shows the news comments data is a potential data source for influenza surveillance. Besides the containing of geographic information, another important feature of comments is topic related, which decrease the difficulty for data retrieve. When epidemiological investigation is too costly or unable to perform in certain area, we may collect comments data and then perform an estimation based on the spatial patterns extracted from the comments as an alternative.

Acknowledgments. This work was supported by Major National Science and Technology Project (Grant No. 2009ZX10004-315 and 2008ZX10005-013), Chinese Academy of Sciences (Grant No. 2F070C01 and 2F08N03), National Natural Science Foundation of China (Grant No. 60921061, 71025001, 91024030, 90924302, 60621001 and 40901219), and U.S. DHS through Grant 2008-ST-061-BS0002.

References

1. Chen, H., Zeng, D.: AI for Global Disease Surveillance. IEEE Intelligent Systems 24, 66–82 (2009)
2. Brownstein, J.S., Freifeld, C.C., Madoff, L.C.: Digital disease detection–harnessing the Web for public health surveillance. New England Journal of Medicine (2009)
3. Ginsberg, J., Mohebbi, M.H., Patel, R.S., Brammer, L., Smolinski, M.S., Brilliant, L.: Detecting influenza epidemics using search engine query data. Nature 457, 1012–1014 (2008)
4. Wilson, K., Brownstein, J.S.: Early detection of disease outbreaks using the Internet. Canadian Medical Association Journal 180, 829 (2009)
5. Collier, N., Doan, S., Kawazoe, A., Goodwin, R.M., Conway, M., Tateno, Y., Ngo, Q.H., Dien, D., Kawtrakul, A., Takeuchi, K.: BioCaster: detecting public health rumors with a Web-based text mining system. Bioinformatics 24, 2940 (2008)
6. Corley, C.D., Cook, D.J., Mikler, A.R., Singh, K.P.: Text and structural data mining of influenza mentions in web and social media. International Journal of Environmental Research and Public Health 7, 596 (2010)
7. Corley, C.D., Mikler, A.R., Singh, K.P., Cook, D.J.: Monitoring influenza trends through mining social media. In: Conference Monitoring Influenza Trends Through Mining Social Media (Year)

8. Culotta, A.: Detecting influenza outbreaks by analyzing Twitter messages. Arxiv preprint arXiv:1007.4748 (2010)
9. Signorini, A.: Social Web Information Monitoring for Health (2009)
10. Laurent, M.R., Vickers, T.J.: Seeking health information online: does Wikipedia matter? Journal of the American Medical Informatics Association 16, 471–479 (2009)
11. Dailey, L., Watkins, R.E., Plant, A.J.: Timeliness of data sources used for influenza surveillance. Journal of the American Medical Informatics Association 14, 626 (2007)
12. Tatem, A.J., Campiz, N., Gething, P.W., Snow, R.W., Linard, C.: The effects of spatial population dataset choice on estimates of population at risk of disease. Popul. Health Metr. 9, 4 (2011)
13. Zhang, Z., Chen, D., Chen, Y., Liu, W., Wang, L., Zhao, F., Yao, B.: Spatio-temporal data comparisons for global highly pathogenic avian influenza (HPAI) H5N1 outbreaks. Plos One 5, e15314 (2010)
14. Balcan, D., Colizza, V., Gonçalves, B., Hu, H., Ramasco, J.J., Vespignani, A.: Multiscale mobility networks and the spatial spreading of infectious diseases. Proceedings of the National Academy of Sciences 106, 21484 (2009)
15. Freifeld, C.C., Mandl, K.D., Reis, B.Y., Brownstein, J.S.: HealthMap: global infectious disease monitoring through automated classification and visualization of Internet media reports. Journal of the American Medical Informatics Association 15, 150 (2008)
16. Cao, Z.D., Zeng, D.J., Wang, Q.Y., Zheng, X.L., Wang, F.Y.: An epidemiological analysis of the Beijing 2008 Hand-Foot-Mouth epidemic. Chinese Science Bulletin 55, 1142–1149 (2010)
17. Cao, Z.D., Zeng, D.J., Zheng, X.L., Wang, Q.Y., Wang, F.Y., Wang, J.F., Wang, X.L.: Spatio-temporal evolution of Beijing 2003 SARS epidemic. Science China Earth Sciences, 1–12 (2010)