Poster Abstract: Worldwide emerging disease-related information extraction system from news data

Myeonghwi Kim, Inhwan Kim, Miran Lee, Beakcheol Jang
Department of Computer Science, Sangmyung University, Seoul
Corresponding author: Beakcheol Jang (e-mail: bjang@smu.ac.kr)

ABSTRACT

Although there have been many researches on the disease information system with the increased interest in disease, the existing systems have limitations in terms of emerging disease monitoring and internationalization. The purpose of this study is to develop a worldwide emerging disease-related information extraction system from news data, which provides nation-specific disease related information, disease-related topic ranking, map-based number of news articles per region, and various charts showing top disease regions and diseases. Our system is available on the web through http://www.epidemic.co.kr/worldwide.

CCS CONCEPTS

• Applied computing → Wellness and Healthcare information systems

KEYWORDS

Disease-related information, worldwide, emerging, topic, ranking, statistics, chart

ACM Reference format:

Myeonghwi Kim, Inhwan Kim, Miran Lee and Beakcheol Jang. 2018. Poster Abstract: Worldwide emerging disease-related information extraction system from news data. In *Proceedings of 16th ACM Conference on Embedded Networked Sensor Systems (SenSys'18). ACM, Shenzhen, China, 2 pages.* https://doi.org/10.1145/1234567890

1 INTRODUCTION

With the development of the traffic system and the globalization, the infectious disease rapidly and widely spreads over the world. Continuous monitoring of diseases and analysis of disease information have become very important issues. In order to solve this problem, there have been many studies to collect and analyze various disease–related information through the Internet data [1-6].

Freifeld, Clark C., et al. gathered data from Internet news articles and users' reports of the world and created a service, called to Healthmap, which displayed the disease-related information to users [1]. Although it provided useful map-based disease-related information, it could not detect emerging disease-related topic because it only provides news and reports per disease. Collier et al. developed BioCaster [2], which collected disease-related data through Internet-based data such as Google, MMWR, WHO and BioCaster ontology developed by them. Similarly, Canadian government and WHO gathered media data from web-based systems to establish GPHIN [3]. Based on online news, GPHIN provided the early detection of infectious diseases and other public health events. However, BioCaster and GPHIN provided services to government rather than to individual users, which limited the accessibility of individuals. Epispider [4] visualized Promed's data and improved the users' convenience.

In this paper, we present a novel worldwide emerging disease-related information extraction system from news data. It collects news data using API of newsapi.org, where it extracts worldwide disease-related information. Our system provides following four kinds of disease-related information to users. Firstly, it presents disease-related information per nation. Users can choose a nation, and our system provides nation-specific disease-related information. Secondly, it presents disease-related topic ranking [5]. It ranks emerging disease-related topics using Natural Language Processing (NLP) and effective ranking algorithm [6, 7]. Users can grasp emerging disease-related trend at a glance and check news related to the topics. Thirdly, it shows the number of news articles per region in the map. The more news articles, the more events. Lastly, it provides three charts, which show top five regions with most articles, top five infectious diseases with most articles, and monthly top infectious disease with most articles. Our system is now available through http://www.epidemic.co.kr/worldwide.

2 SYSTEM ARICHITECTURE

Figure 1 shows the system architecture of our system. It consists of the data crawler, classification engine, database, web front-end, and web back-end. The data crawler collects worldwide news articles. The classification engine consists of content extractor and term manager modules. The term extractor removes the nonsensical attributes and redundant data in the collected data and sends them to the database. The term manager retrieves the data from the database and generates the term extractor and the term object. The web back-end stores the collected data in the database and converts them into the form of the web front-end. The web front-end is responsible for managing view pages.

The data crawler collects 100 articles per hour based on the search

SenSys'18, November 4-7, 2018, Shenzhen, China

DOI: 10.1145/3274783.3275168

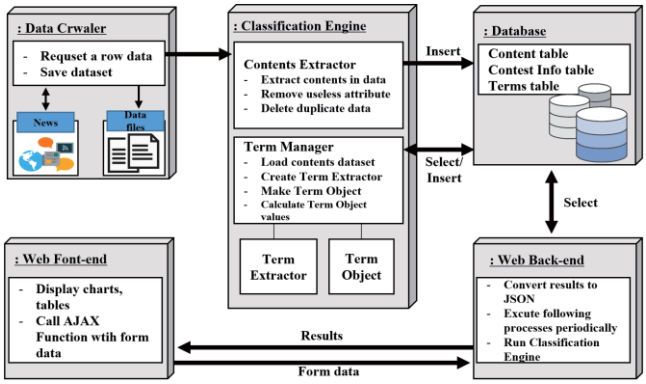

Figure 1: System Architecture

API provided by NewsAPI (https://newsapi.org) and uses the Python-based NLP open source nltk program. The classification engine consists of the content extractor and the term manager. The content extractor calls data collected from the data crawler to remove unnecessary elements. In addition, we check the similarity of duplicate and similar sentences through Sift4 [5]. Sift4 sets the unique sentences in the database's content table. The term manager calls the data stored in the database content table and creates the term extractor and the term object. The term extractor performs generalization and inversion of sentences separated into words. The term object identifies the objects of extracted words. After collecting the words of all sentences, it performs the ranking algorithm. The web back-end switches the data requested by the front end to the JSON type and executes the classification engine module according to the job scheduler time. The web front-end performs visualization works for users based on the collected data. We used dataTable.js to create the disease topic-ranking table and produced a variety of charts using chart.js.

Figure 2: User Interface

3 RESULTS

Figure 2 shows the interface of our system in the mobile environment. The system interface consists of four major parts. First, In Figure 2-A users select the country they are interested in. In Figure 2, users chose China as an example. Figure 2-B shows the map of the country selected by users and the number of disease-related news articles per region. To equalize the meaning of the region of the world, we made the international airport to represent the region. The map is updated monthly. Figure 2-C presents the ranking of emerging disease-related topics in the country. It shows relevant articles when you select any keyword. For the ranking algorithm, we employed CCA [8]. Users can check disease-related trends and news of each country. Figure 2-D provides various charts based on disease-related news statistics. Chart 1 shows top five regions with most articles. Chart 2 shows top five infectious disease with most articles. Lastly, Chart 3 shows monthly top infectious disease with most articles.

4 CONCLUSION

In this paper, we developed the worldwide emerging disease-related information system from news data. With our system, people can identify emerging disease-related topics of the country they are interested in. They can also check the diseased-related news statistics at a glance through the map and charts. This system not only helps to promote the health of people all around the world, but also serves as a good research resource for researchers. Expanding the scope of the country around the world remains as a future work. Our system is now available on the web through http://www.epidemic.co.kr/worldwide.

ACKNOWLEDGMENTS

This work was supported by the national research foundation of Korea grant funded by the Korea government (NRF-2018R1E1A2A02058292 and NRF-2016R1D1A1B03930815).

REFERENCES

[1] Freifeld CC, Mandl KD, Reis BY, Brownstein JS. HealthMap: global infectious disease monitoring through automated classification and visualization of Internet media reports. J Am Med Inform Assoc. 2008; 15: 150–157. https://doi.org/10.1197/jamia.M2544
[2] Collier N, Doan S, Kawazoe A, Goodwin RM, Conway M, Tateno Y, et al. BioCaster: detecting public health rumors with a Web-based text mining system. Bioinformatics. 2008; 24: 2940–2941. https://doi. org/10.1093/bioinformatics/btn534
[3] M. Dion, P. AbdelMalik, and A. Mawudeku. 2015. Big Data: Big Data and the Global Public Health Intelligence Network (GPHIN). Canada Communicable Disease Report 41, 9 (2015), 209.
[4] Keller M, Blench M, Tolentino H, Freifeld CC, Mandl KD, Mawudeku A, et al. Use of unstructured eventbased reports for global infectious disease surveillance. Emerg Infect Dis. 2009; 15: 689. https://doi.org/ 10.3201/eid1505.081114
[5] Yoon, Jungwon, Jong Wook Kim, and Beakcheol Jang. "DiTeX: Disease-related topic extraction system through internet-based sources." PloS one 13.8 (2018): e0201933. e0201933.
[6] Lee Miran, Jong Wook Kim, and Beakcheol Jang. "DOVE: An Infectious Disease Outbreak Statistics Visualization System." IEEE Access (2018).
[7] Jang, Beakcheol, and Jungwon Yoon. "Characteristics Analysis of Data From News and Social Network Services." IEEE Access 6 (2018): 18061-18073.
[8] de Almeida HM, Gonc¸alves MA, Cristo M, Calado P. A combined component approach for finding collection-adapted ranking functions based on genetic programming. Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval. ACM; 2007. pp. 399–406.