Commentary

Internet Data for Improving Prevention and Control of Global Infectious Diseases

Zhongwei Jia[1,2,3,#]; Xiangyu Yan[1]; Yongjie Li[1]; Jiaqi Ma[4]

By the end of 2018, there were about 5.1 billion internet users worldwide who accounted for 67% of the total population (*1*). Health problems have emerged following widespread use of the internet as a new model of social activity, such as anonymous personal attacks or abuse and information leakages. A more potent threat is the unverified information being provided by the internet or electronic media that can lead to unsubstantiated judgments and improper responses to information regarding an individual's health status. A real example was the Infodemic that occurred at the beginning of the coronavirus disease 2019 (COVID-19) pandemic (*2*).

In 2002, Gunther Eysenbach brought up a creative concept named "infodemiology" to investigate the impact of various information, especially that stemming from the web or electronic media, on people's health (*3*). The information sources of infodemiology cover almost all forms of electronic media, such as browser search histories, social media, mobile phone apps, and some reprocessed internet data (e.g. Google Trends) (*4*). Infectious diseases, such as influenza and Zika virus were the most concerned topics in infodemiology. However, chronic diseases, such as diabetes, cancers, and health-related behaviors, such as drug use, suicide, smoking, and diet were also followed and covered (*4*). Infodemiology is primarily used for monitoring "trends of information" through distribution and change of health-related information rising from web or electronic media in order to provide alarm or prediction. How to identify reasons related with information, further implement interventions, evaluate the effect of interventions, and optimize the scheme have not yet been solved in infodemiology and are still open. Therefore, the theoretical research system of infodemiology needs to be expanded.

Because information is an outcome of internet users and each internet user is related with an actual individual in physical space, it is possible to detect the reasons and motivations of people who produce a kind of internet information. Based on different reasons, health interventions can be devised based on users' characteristics and behaviors, which will be beneficial for those who regard internet and other electronic media as an important resource for health information. These terms and objectives are all covered by epidemiology, but by only focusing on internet users, new informatics methods are needed to analyze data of internet users.

To strengthen the usage and analysis of data rising from the internet, it is necessary to study the characteristics, distribution, and impact factors of internet events related with health and diseases (including the behavior, distribution, and influencing factors of internet users, laws associated with epidemics, and trends and redundancy benefits of internet information); explore the correlation between events on the internet and in real life; optimize the prevention and control strategies of internet information; and serve public health and social governance (*5*). Integration of traditional data and real-time internet data of individuals and subpopulations is the object of further internet studies, and interdisciplinary research methods, such as natural language processing, knowledge graph and machine learning techniques have become the basic methods in this discipline. The most challenging question for further studies is linking an internet event and a physical event. In addition, protecting personal privacy of internet users' health and behavior information is also a rising concern, and methods such as anonymous processing of heterogeneous data from multiple sources must be further studied. Great impact and reform will be also confronted by traditional ethics in internet era.

An example of optimizing use of internet data is the ability to improve HIV/AIDS prevention and control for men who have sex with men (MSM). MSM accounted for 23.0% of new HIV infections in China, and homosexual transmission was the second major HIV transmission route after heterosexual transmission in 2019 (*6*). The internet socializing platforms, such as some geosocial networking applications (GSN apps), have been blamed for this problem. Recent studies indicated that about 41%–63% of Chinese MSM had experienced looking for the casual sexual partners thought GSN apps, among which HIV incidence was about 4 times higher than that of non-users (8.5/100 person-years *vs.* 2.0/100 person-years) (*7*). The anonymity and convenience of dating online was the main reason for GSN becoming popular among MSM, but it was also the primary culprit of increasing HIV infection rates among MSM because these casual sexual partners dating online did not know each other before, let alone their HIV statuses (*8*). In order to avoid these risk behaviors, HIV-related knowledge was recommended to be publicized and delivered through the internet, but the study showed that only about 50% of MSM in China were willing to browse HIV-related knowledge through the internet and electronic media (*9*).

Knowledge graph-based frameworks can be effectively used to provide this knowledge and will be the base for interventions and control efforts online. Knowledge graphs are constructed using graph model based on internet data sources that include two parts, one dataset is the user data from GSN apps, which is used to construct label-based user persona, and another dataset is the data from a website [e.g. World Health Organization (WHO) or Joint United Nations Programme on HIV and AIDS] and literature (e.g. PubMed or Embase), which provides high-quality and evidence-based health information that can be used to build health knowledge graphs and give targeted and timely interventions for users of GSN apps.

Data from GSN apps are composed of structured and semi-structured data, which include demographic information of MSM, self-introduction information, and dating requirements. In demographic information, there are age, height, and weight. The self-introduction information covers MSM users reporting their sex role, physical, and personality characteristics. Dating requirements include their criteria for whom they want to meet. Integrating these data, a personalized user persona for each MSM user will be created by two steps. First, 20% of the user data will be annotated from 5 dimensions by professional staff, which include demographic attributes, social attributes (social position, social relations), behavior habits (lifestyle, sexual behaviors, dating history), interest preferences (shopping, games), and disease statuses (HIV and other sexual transmitted diseases, chronic diseases, and psychological statuses). Second, a support vector machine model will be trained to classify the remaining users and add the other relevant labels (Figure 1).

A health knowledge graph of MSM can then be established by professional information extracted from the unconstructed text of WHO and the literature databases (e.g. PubMed), which includes common diseases and symptoms, professional intervention guidance materials of MSM, disease-specific medical guidelines, and treatment measures for infectious diseases such as HIV, syphilis, etc. The whole process was divided into 4 steps: data cleaning, information extraction, knowledge fusion, and knowledge reasoning (Figure 1) (*10*). First, *data cleaning*–the unstructured texts usually contain noise information, so we removed all the special characteristics (such as @, emoji) and stop words. Second, *information extraction*–for unstructured data, we performed entity extraction, relationship extraction, and attribute extraction on these text data. We used the Long Short–Term Memory Neural Network (LSTM) for named entity recognition (*11*). We converted the filtered text into word vectors and input them into LSTM, and the output was the named entities labeled in the sentences. We also introduced an attention mechanism to improve name recognition. For example, the extracted entities were sexual behavior, HIV, and condoms. For relation extraction, we choosed the weakly supervised learning model to extract relationships for data lacking adequate labels. The OpenNRE was carried out to extract the relationship between diseases and symptoms, diseases and treatment measures, diseases and medical institutions, high-risk behavior and intervention guidance, infectious diseases and detection methods, and so on. We used a rule-based method to extract attributes, including symptoms of common diseases, intervention methods, and the outcome of taking antiviral drugs, etc. For example, an early symptom of HIV is herpes zoster, which has a high-risk sexual infection rate. Third, *knowledge fusion* – we merged different expressions of the same entity extracted from different literature or websites, including the merging of entities and the merging of entity relationships, such as hepatitis B virus and HBV. Hierarchical clustering was used to calculate the similarity between different entities and to partition the entities at different levels and finally form a tree-like clustering structure. Entities with higher similarity were more likely to be merged into the same one, and the fusion results will be reviewed by experts. The final constructed health knowledge graph was represented as an "entity-relationship-entity" triple; for example, anal sex-interventions-condoms, hepatitis B-clinical

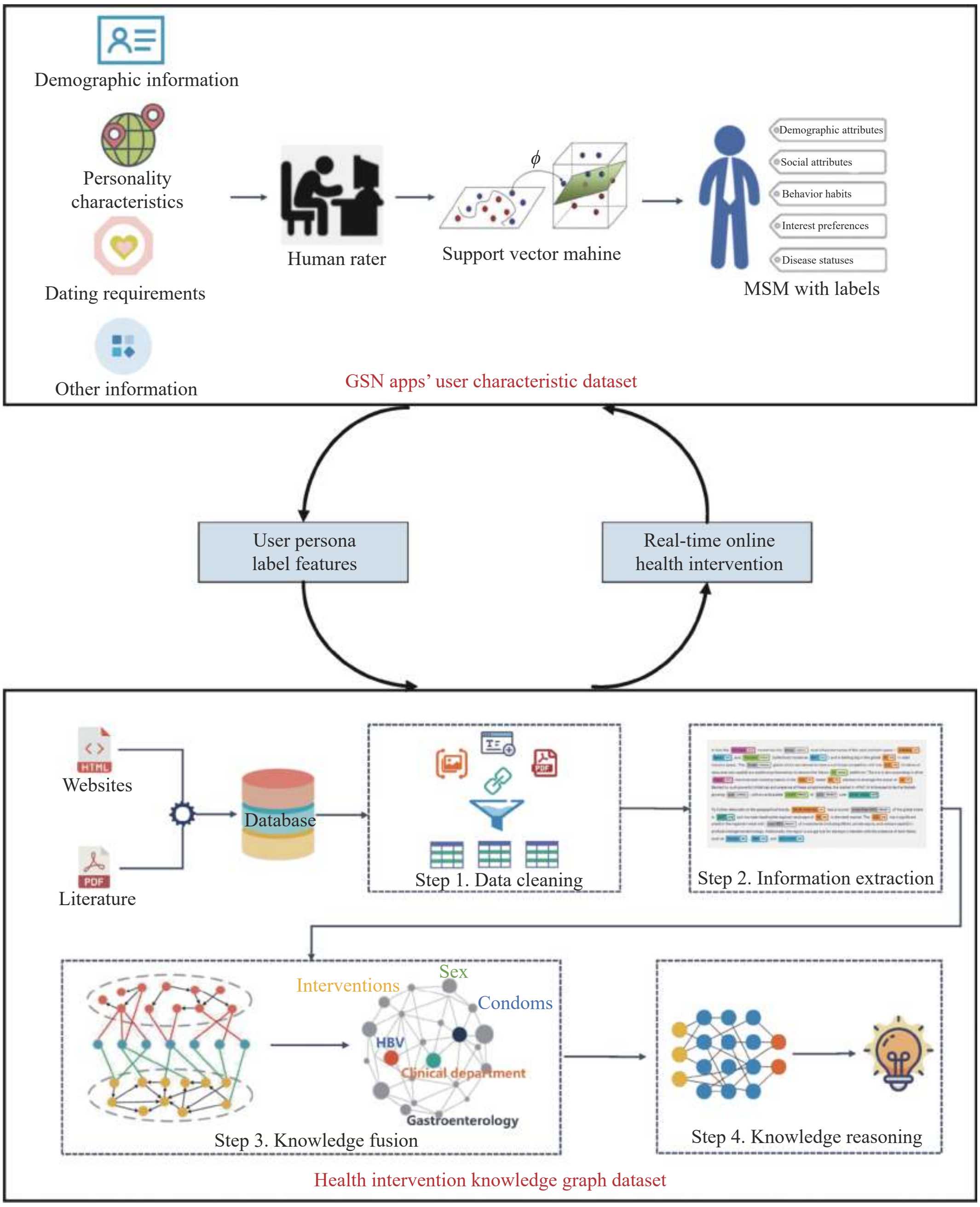

FIGURE 1. Improve HIV/AIDS prevention and control for MSM based on knowledge graph

department-gastroenterology. Fourth, *knowledge reasoning*–based on the completed user persona of MSM, we used the health knowledge graph to push health intervention information for MSM and predict high-risk behaviors. For example, we can push information about the correct use of condoms for people with suspicious high-risk behavior labels. The entire reasoning process will be handled by the convolutional neural network (*12*). Finally, as the data continues to increase, our health knowledge graph will be updated to improve reasoning ability and efficiency.

CONCLUSION

It will be not comprehensive and objective to investigate a specific health problem without integrating the network and electronic media information today. Problems originating from the internet should be solved based on the internet, and it is suitable for the health problems. Using internet-based big data to improve disease prevention and control is still a novel subject which is aiming to narrow the rising gap between health issues caused by internet and the intervention methods in the physical space. With the increasing expansion of network data, it is meaningful in the monitoring and intervention of specific populations on the internet. In this report, we tried to deliver a general idea of public health by a practical case of intervention and control online on MSM. More new approaches deserve to be explored and applied in health problems rising from internet users and enrich the research system of traditional epidemiology. New privacy protection mechanisms need further exploration.

Funding: Key Joint Project for Data Center of the National Natural Science Foundation of China and Guangdong Provincial Government (U1611264); and the National Natural Science Foundation of China (91546203, 91846302).

doi: 10.46234/ccdcw2020.269

# Corresponding author: Zhongwei Jia, urchinjj@163.com.

[1] School of Public Health, Peking University, Beijing, China; [2] Center for Intelligent Public Health, Institute for Artificial Intelligence, Peking University, Beijing, China; [3] Center for Drug Abuse Control and Prevention, National Institute of Health Data Science, Peking University, Beijing, China; [4] Chinese Center for Disease Control and Prevention, Beijing, China.

Submitted: December 03, 2020; Accepted: December 22, 2020

REFERENCES

1. Global System for Mobile Communications Association (GSMA). The mobile economy 2020. 2020. https://www.gsma.com/mobileeconomy/. [2020-11-25].
2. Zarocostas J. How to fight an infodemic. Lancet 2020;395(10225):676. http://dx.doi.org/10.1016/S0140-6736(20)30461-X.
3. Eysenbach G. Infodemiology: the epidemiology of (mis)information. Am J Med 2002;113(9):763 – 5. http://dx.doi.org/10.1016/s0002-9343(02)01473-0.
4. Mavragani A. Infodemiology and infoveillance: scoping review. J Med internet Res 2020;22(4):e16206. http://dx.doi.org/10.2196/16206.
5. Jia ZW. Discussion on construction of network epidemiology from infodemic of COVID-19. Chin J Med Sci Res Manage 2020;33(5): 368-71. http://rs.yiigle.com/CN113565202005/1302784.htm. (In Chinese).
6. National Health Commission of the People's Republic of China. The AIDS epidemic in China remains at a low epidemic level. 2019. http://www.gov.cn/xinwen/2019-12/01/content_5457304.htm. [2020-11-25]. (In Chinese).
7. Xu JJ, Yu H, Tang WM, Leuba SI, Zhang J, Mao X, et al. The effect of using geosocial networking apps on the HIV incidence rate among men who have sex with men: eighteen-month prospective cohort study in Shenyang, China. J Med internet Res 2018;20(12):e11303. http://dx.doi.org/10.2196/11303.
8. Hong H, Xu J, McGoogan J, Dong HJ, Xu GZ, Wu ZY. Relationship between the use of gay mobile phone applications and HIV infection among men who have sex with men in Ningbo, China: a cross-sectional study. Int J STD AIDS 2018;29(5):491 – 7. http://dx.doi.org/10.1177/0956462417738468.
9. Liu SY, Wang KL, Yao SP, Guo XT, Liu YC, Wang BY. Knowledge and risk behaviors related to HIV/AIDS, and their association with information resource among men who have sex with men in Heilongjiang province, China. BMC Public Health 2010;10(1):250. http://dx.doi.org/10.1186/1471-2458-10-250.
10. Liu Q, Li Y, Liu Y, Duan H. Knowledge graph construction techniques. J Comput Res Dev 2016;53(3): 582-600. http://crad.ict.ac.cn/EN/10.7544/issn1000-1239.2016.20148228. (In Chinese).
11. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput 1997;9(8):1735 – 80. http://dx.doi.org/10.1162/neco.1997.9.8.1735.
12. Wang HW, Zhao M, Xie X, Li WJ, Guo MY. Knowledge graph convolutional networks for recommender systems. In: The world wide web conference. San Francisco, CA, USA: ACM. 2019. http://dx.doi.org/10.1145/3308558.3313417.

Zhongwei Jia, PhD, Professor
School of Public Health, Institute for Artificial Intelligence, Peking University