ISSN 2409-2665
Journal of Logistics, Informatics and Service Sciences
Vol. 5 (2018) No. 2, pp. 55-67

Research on Internet Search Data in China's Social Problems under the Background of Big Data

Lili Fu[1], Yinhong Dong[2]

[1]Business College, Beijing Union University, Beijing, China.

[2]School of Management, South-Central University for Nationalities, Wuhan, China.

E-mail: lili.fu@buu.edu.cn

Abstract. With the rapid development of the Internet and big data, the research of Internet search data in the social field has become a hot topic. This paper studies and reviews the research of Internet search data in fields including China’s medical, cultural, public opinion, environment, science and technology policy, agriculture, etc. It summarizes current status of research from the perspectives of research methodology, data selection, model construction, etc., so as to point out the direction for future research.

Keywords: Social field, Internet search data, current status of research

1. Introduction

With the rapid development of the Internet, the massive amount of data generated by the Internet has received increasing attention. Internet search data is a kind of special data in big data. It is the data generated by netizens' active search behavior on the Internet. It reflects the information needs of netizens and the psychological state of netizens. It is a mapping of Internet behavior and attitudes on the Internet. According to the “43th Statistical Report on Internet Development in China” released by China Internet Network Information Center (CINIC), as of December 2018, the number of search engine users in China reached 681 million, with a utilization rate of 82.2%, and the number of mobile search users was 654 million, with a usage rate of 80.0%. This indicates that the Internet search data is more generated by the search behavior of search engine users, and the search engine is an important source of Internet search data.

With the rapid development of the Internet and big data, relevant research at home and abroad is increasingly rich in social issues. Abroad research includes epidemiological surveillance (Ginsberg et al., 2009), national elections (Spyros E et al., 2013), etc. Domestic research closely follows abroad research, and some research results have been achieved. Based on the above background, this paper reviews and sorts out the research in the past ten years in which Chinese scholars used Internet research data to analyze social issues in China, focusing on the research fields, research methodology and research effects of domestic scholars, in order to make full use of big data brought by research engines to provide reference for practical research.

2. Concept definition and literature overview

The search data referred to in this article refers to the data generated when netizens search for information in search engines by using words, words and sentences as keywords, then search engines will search in the database to find the website links that match the requirements of netizens. In this process, search engine will record netizens' search behavior, and counts the frequency of the keywords with certain rules to form search data. Baidu's search data is called "Baidu Index", while Google's search data is called "Google Trend" and its upgraded version "Google Insight". The search data of non-search engine websites accounts for a very small proportion in academic research, and is not the mainstream of scholars' research. Therefore, the search data of this paper is mainly from the search data of search engines.

This article searches for thousands of documents in CNKI (China National Knowledge Infrastructure) with the theme of "search data", "Internet search", "Internet search data" and "Baidu index". Excluding some weakly related documents, there are 190 papers that research Internet search data's prediction to the real world. Generally speaking, the research of these papers is mostly concentrated in the economic field and social field (all non-economic field are classified as social fields), and there are almost twenty papers reaching the prediction in the social field.

3. Literature Review

Judging from the results of the literature review, the existing literature research in the social field mainly includes medical research, cultural research, social sentiment research, environmental research, science and technology policy, agriculture and other sub-divisions, but the number of papers in each field is relatively small, compared to research in medical and cultural field.

3.1 Medical Research

That the research on Internet search data has attracted more and more attention is inseparable from foreign scholars Ginsberg et al. (2009) who used Google data to monitor the US flu. Their research makes the research of the big data represented by the Internet search data quickly become a hot topic. Therefore, Chinese scholars have relatively more research in medical field, mainly involving influenza, avian influenza, hand, foot and mouth disease, etc. The research content includes early warning research and correlation research.

The earlier medical research was the correlation research between the search data in Google China and the H1N1 flu in Guangdong Province by Kang Min et al. (2011). In that paper, Pearson correlation analysis method was used to analyze the correlation between the two using data in 2009 and the correlation index reached 0.914. Then descriptive statistics was adopted to compare the search data with the trend chart of the flu. The article believes that “the flu-related Internet search situation well reflects the level of flu activity”, and agrees that “the Internet search data can be used as data source to help the surveillance of infectious diseases such as influenza”. Although early studies did not use complex models for influenza prediction, the role and correlation of Internet search data in influenza surveillance was demonstrated, laying the groundwork for subsequent research. Soon after Kang Min’s research, Li Xiuting et al. (2012) not only studied the correlation between online search information and influenza, but also focused on the ability of Internet search data to predict influenza. They used correlation analysis and principal component analysis to verify the correlation and Internet search synthesis index of the two. By designing three models to compare the accuracy of the prediction, the research found that the model that combines Internet search data which can timely predict changes and historical data which can reflect influenza trend can get better prediction results. Huang Dacang (2015) also obtained similar conclusions in the article on the use of search engine data to monitor hand, foot and mouth disease. After comparing the prediction models of the three cases, the article thinks that the model with the best predictive effect is considered to be combined with historical case data and Internet search data. The research of Wang Ruojia and Li Pei (2016) not only proves that the comprehensive use of influenza historical data and search data can obtain the best monitoring effect, but also finds out that the model combining the multiple linear regression and artificial neural Internet has better performance in prediction, compared to other prediction models.

From the above literature review, it can be concluded that the Internet search data is relatively effective in predicting some diseases. Scholars generally agree that the comprehensive use of data such as Internet search data and actual disease incidence data will make the prediction accuracy higher.

3.2 Culture research

The research of Internet search data in the field of culture is mainly reflected in the prediction of the box office of the movie and the prediction of the art market.

Wang Lian and Jia Jianmin (2014) established a movie box office prediction model based on Internet search data. The film screening time was divided into the premiere week and the follow-up week. The results showed that “the

Internet search data and the weekly box office of the premiere week and the follow-up week, the total box office and the life cycle of the film are significantly related”. This shows that the model is able to predict the movie box office. Tang Zhongjun et al. (2018) based on the theory of consumer decision-making process, use Internet search data and film feature information as independent variables, and construct a joint prediction model including improved Bass model and parametric regression model by using the improved Bass model parameters as dependent variables. The model has a good effect on predicting the demand amount on film day.

Yin Xi (2017) focuses on the relationship between the Internet search index and the art auction confidence index. The article uses the Baidu search index in seven years as a sample, and classifies Internet search index into modern famous artists search index and artwork search index, GDP search index and stock search index. Through building regression models, it studies their impact on the art market confidence index. The focus of Zhu Jiawei's (2019) research is the investment of artwork. The article uses Baidu search index related to painting and calligraphy to synthesize three kinds of comprehensive indexes. Using the time series research method, the market volume of painting and calligraphy can be predicted half a year ahead of time. The exponential model of modern famous artists has a fit of 0.89.

The research in the field of culture can clearly see the degree of familiarity of scholars in their respective fields. The selected keywords are very representative in the industry. The selected methods are mostly regression analysis and time series methods.

3.3 Social sentiment research

There are relatively few literatures on public opinion research based on Internet search data, which is contrary to the frequent use of social Internet information such as Weibo to study public opinion.

Professor Yu Guoming from Renmin University of China published a series of papers to study the Chinese social sentiment index, among which in the paper “Chinese social sentiment under the analysis of big data: overall situation and structural characteristics – modeling based on Baidu hot search keywords (2009-2012)”, Yu Guoming (2013) takes the Baidu search term database from 2009 to 2012 as the research object, and conducts data mining and analysis by selecting the search data of TOP1000 and the fastest rising TOP1000 hot search term, and measure several typical China’s social sentiment indexes including “social warmth”, “social well-being” and “social pressure”.

Zou Wei (2015) studied the spatio-temporal characteristics of the “China-Philippines Huangyan Island Dispute” incident in April 2012 using Sina News data and Baidu Index data. Based on the six-stage model theory of Internet event propagation, natural breakpoint classification , “sequence-scale” method, ESDA technology, correlation analysis and other methods, the influencing factors of the six stages of the dispute over Huangyan Island dispute, the aggregation characteristics of the spatial distribution of Huangyan Island's public opinion search data, the degree of the fluence that public sentiment news data have on public sentiment search data at different time periods, and the distribution of related features, are recognized.

3.4 Environmental research

Environmental research mainly uses Internet search data to build the evaluation system.

Shi Yadong (2018) studies how to use the Internet search data to quantitatively evaluate Public Environmental Concern, thus changing the traditional methods of using questionnaires frequently. The article proposes four levels of progressive relationships by understanding the meaning of public environmental concerns, and then screen out the corresponding environmental keywords and search for the corresponding Baidu index data. It uses the analytic hierarchy process to construct the public environmental concern index, and empirically analyze the Beijing public environmental concern index and its influencing factors. Lu Xing (2017) also studies the public environmental concern index system. He divides the indicator system into four categories including 42 indicators. Based on the keyword data searched by these indicators, the environmental interest index is constructed by using the comprehensive weighting method.

3.5 Science and technology policy research

Research on science and technology policy is relatively rare, but Zhang Yongan et al (2018) use the Internet search data to study the effectiveness of science and technology innovation policy. The research is relatively innovative. Base on rational expectations and information search theory, the article collects keywords according to five scientific and technological innovation policy tools including talent incentives, research and development subsidies, tax preference, government procurement, and service outsourcing, and obtains the keyword search data from January 2011 to June 2017 plus the number of patent applications N periods in advance. The regression model is constructed by using them as independent variables and the number of patent applications as dependent variable. The research results show that the Internet search data and the effectiveness of the science and technology innovation policy (represented by the number of patent applications) have strong dynamic relevance; different science and technology innovation policy tools produce different effects of policy regulation, among which R&D subsidies and talent incentives are important scientific and technological policies, and have the greatest impact on the number of patent applications. Government procurement and service outsourcing have the least impact on the validity of patent applications; the model can predict the number of patent applications.

3.6 Agriculture research

There are also few studies on agriculture, among which agricultural safety warnings are relatively meaningful.

Xue Wenlong and Su Wanyi (2017) used Internet search index to study agricultural safety warnings based on locust disasters. They searched for keywords related to locusts, precipitation, and drought, and these keywords are also categorized into common keywords, related keywords, and other keywords. Using Baidu search engine as sample data, the three categories of keywords and one type of synthetic index, five kinds of early warning results as output, the article builds a search behavior early warning model based on BP neural

Internet. By using MATLAB to carry out simulation test based on locust disaster data, the test results show that the early warning model is reasonable and effective, and it can warn the agricultural safety incidents three to fifteen days in advance.

4. Conclusions

Generally speaking, although the research of Chinese scholars in the social field is not very rich, the entry points of the research are relatively new. The selection of keywords and the characteristics of the research objects are very appropriate. The research can use correlation analysis to study the correlation between search index and traditional data. keyword search data generally uses principal component analysis, classification method, machine learning, etc. to form a composite index. The selection of variables can take the time lag of the search data and the influence of traditional data into account. The choice of the models includes multiple linear regression, time series models, artificial neural Internets, etc. It can be considered that the research paradigm using the Internet search data for forecasting or constructing the evaluation index system is relatively fixed and can be extended to a wider field. But the foundation on which the paper theoretically demonstrates Internet search data's predictability to actual index is not profound, that in the future in-depth research can be done in this area. In addition, the selection of keywords is also a deep point of research. Since the Internet environment is complex and changeable and users' search behaviors are diverse, it's a challenge to build a relatively stable keyword library for some typical traditional indexes in different fields. In the aspect of model construction, a good deal of research is required to done to find out under what kind of environment or condition and which model can have higher prediction accuracy.

Acknowledgements

This research was supported and funded by “Forecast Research of Beijing CPI Based on Network Search Data” (No. SM201311417008), a scientific research program of Beijing Municipal Education Commission. This research was also supported by Fund program: The Ministry of Education of Humanities and Social Science project “Research on Design and Optimization of Increasing Block Pricing Mechanism for Resource Goods” (Approval No: 14YJC630024), China.

References

Ginsberg Jeremy, Mohebbi Matthew, Patel Rajan S, Brammer Lynnette, Smolinski Mark S, Brilliant Larry. Detecting influenza epidemics using search engine query data. Nature. 2009，Vol.457, pp.1012-1014

Huang Dacang. Monitoring of hand, foot and mouth disease based on search engine data [D]. Northeast Normal University, 2015.

Kang Min, Zhong Haojie, Yang Fen, He Jianfeng, Zhang Yurun. Analysis of correlation between the H1N1 flu in 2009 in Guangdong Province and Internet search data. Journal of Tropical Medicine. 2011,11(06):629-632.

Li Xiuting, Liu Fan, Dong Jichang, Lv Benfu. China Influenza Surveillance Based on Internet Search Data[J]. Systems Engineering Theory & Practice,2013,33(12):3028-3034.

Lu Xing. Construction of Public Environmental Concern Indicator System——Based on Internet Search Data[J]. Research World, 2017(06): 35-38.

Shi Yadong. The compilation of public environmental concern index and its influencing factors——Taking Beijing as an example[J]. Journal of Beijing Institute of Technology (Social Science Edition), 2018, 20(05): 46-53.

Tang Zhongjun, Wang Meiyue, Yu Haibo, Wu Fan. The film daily demand forecast combined with improved Bass model and parametric regression model[J]. Industrial Engineering and Management, 2018, 23(04): 16-22+29.

Wang Ruojia, Li Pei. Comparison and Optimization of Influenza Surveillance Model Based on Internet Search Data[J]. Library and Information Service,2016,60(18):122-132.

Wang Lian, Jia Jianmin. Box Office Prediction Model Based on Internet Search——Evidence from Chinese Film Market[J]. Systems Engineering Theory & Practice, 2014, 34(12): 3079-3090.

Xue Wenlong, Su Wanyi. Agricultural Security Early Warning Based on User Search Data[J]. Jiangsu Agricultural Sciences, 2017, 45 (12) :188-191

Yin Xi. Research on Internet Search Index of Chinese Artwork Market [D]. Central Academy of Fine Arts, 2017.

Yu Guoming. Chinese Social Public Opinion under Big Data Analysis: Overall Situation and Structural Features——Public Opinion modeling Based on Baidu Hot Search (2009-2012) [J]. Journal of Renmin University of China,2013,27(05): 2-9.

Zhang Yongan, Song Chenchen, Wang Yanni, Qi Yuan. Research on the Control Effect of Science and Technology Innovation Policy Based on Internet Search[J].Soft Science,2018,32(09):24-29.

Zhu Jiawei. Research on the Prediction of Calligraphy Market Index Based on Internet Search Data [D]. Zhejiang University of Finance and Economics, 2019.

Zou Wei. Research on the Temporal and Spatial Characteristics of Public Sentiment in Huangyan Island Based on ESDA[D]. Nanjing University, 2015.