Computer Science and Application 计算机科学与应用, 2021, 11(1), 8-18
Published Online January 2021 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2021.111002

Hans汉斯

谷歌趋势主题热度的地理分布

陆成刚[1]，王丽君[2]，王庆月[3]

[1]浙江工业大学理学院，浙江 杭州
[2]浙江医院，浙江 杭州
[3]宁夏理工学院计算机学院，宁夏 石嘴山
Email: luedge@163.com, luchenggang168@gmail.com

收稿日期：2020年12月9日；录用日期：2021年1月1日；发布日期：2021年1月8日

摘　要

在谷歌趋势中挖掘某个主题搜索热度的地理分布往往具有重要的战略价值，但谷歌趋势只提供单个的关键词(组)的地理分布，对于某个主题涉及到若干个关键词(组)构成的词族是没有合成的搜索热度地理分布的功能。作者从入口检索词出发依托谷歌趋势的相关查询生成主题词族，并利用多关键词(组)的趋势比较功能和单个关键词(组)趋势的地理分布，推导出词族的搜索热度地理分布公式。在应用到中美新冠疫情的地理分布研究中，发现了两国疫情分布与疫情主题的搜索热度的皮尔逊相关性具有相近的规律，即平均疫情主题的搜索热度与各地新冠病例确诊数呈显著的负相关性。另外，中国的平均疫情主题搜索热度与各省区GDP有显著的负相关性，但美国的平均主题搜索热度和各州人均GDP有显著的正相关性。

关键词

谷歌趋势，搜索热度，新冠疫情，地理分布，皮尔逊相关性

Geographical Distribution of Google Trend for the Words Family

Chenggang Lu[1], Lijun Wang[2], Qingyue Wang[3]

[1]Department of Applied Math, Zhejiang University of Technology, Hangzhou Zhejiang
[2]Zhejiang Hospital, Hangzhou Zhejiang
[3]Institute of Computer, Ningxia Institute of Technology, Shizuishan Ningxia
Email: luedge@163.com, luchenggang168@gmail.com

Received: Dec. 9th, 2020; accepted: Jan. 1st, 2021; published: Jan. 8th, 2021

文章引用: 陆成刚, 王丽君, 王庆月. 谷歌趋势主题热度的地理分布[J]. 计算机科学与应用, 2021, 11(1): 8-18.
DOI: 10.12677/csa.2021.111002

Abstract

Exploring the geographic distribution of search interest in a certain topic in Google Trends is often of important strategic value, but Google Trends only provides the geographic distribution of a single keyword (group); for a topic involving word family composed of several keywords (groups), has no synthetic function of searching popularity of each geographic distribution. Starting from the entry search term, the author relies on the relevant queries of Google Trends to generate topic word families, and uses the trend comparison function of multiple keywords (groups) and the geographical distribution of single keyword (group) trends to derive the geographical distribution of search popularity of the words family formula. In the study of the geographic distribution of the COVID-19 epidemic in China and the United States, it was found that the Pearson correlation between the epidemic distribution of the two countries and the search popularity of the epidemic topic has a similar pattern, that is, the search popularity of the average epidemic topic has significantly negative correlation with the number of confirmed cases of COVID-19 in each region. In addition, the average topic search interest in China has a significant negative correlation with the GDP of each province, but the average topic search popularity in the United States has a significant positive correlation with the per capita GDP of each state.

Keywords

Google Trends, Search Popularity, COVID-19, Geographical Distribution, Pearson Correlation

1. 引言

通过网络搜索引擎数据进行严肃的学术课题研究不乏先例，尤其在定量分析社会事件的传播和影响方面。例如，Zhao 等[1]结合了基于百度搜索查询对手足口病 HFMD 病例进行实时估计。Lei Qin 等[2]收集社交媒体搜索指数(SMSI)预测新型冠状病毒(COVID-19)的新可疑或确诊病例数，SMSI 可能是有效的早期预测指标，它将使政府的卫生部门能够确定潜在的高风险暴发地区。Yongqing Zhao 等[3]探索了百度指数 BDI 与土地覆被变化的关系，表明 BDI 中的关键搜索词是与“耕地占用税”和“建设用地规划许可证”密切相关的。Wang 等[4]探索百度指数监测网络肾结石相关信息的搜索行为以及了解中国人口特征和优先事项方面的价值。Schootman 等[5]检查了 2004 年 1 月至 2014 年 4 月的关于癌症筛查和准备工作的信息搜索的谷歌趋势数据，并作谷歌趋势和 BRFSS 数据之间的相关性分析。Abel 等[6]使用谷歌趋势数据来测试在欧美实施的因 COVID-19 大流行采取的隔离是否导致福利相关主题搜索字词发生变化。使用差异和回归不连续性设计来评估隔离的因果关系，他们发现欧洲和美国的无聊搜索强度大大增加。他们还发现孤独感、忧虑和悲伤的搜索量显着增加，而压力、自杀和离婚的搜索反而下降了。

搜索引擎和内容自动推荐是数字社会居民获取知识和信息的重要手段，与此衍生的相关网络指数也是计算社会学的重要研究内容[7] [8] [9] [10]。尤其是搜索引擎因用户主动获取信息的特性使得搜索引擎指数成为预测和分析社会事件传播和影响力的重要参考。

在谷歌趋势中提供了单个关键词(组)的搜索频度的地理分布，例如以关键词“离婚”为例，图 1 展示近五年中国各省区该词的搜索热度分布：

在研究离婚率这样的主题时或可以使用“离婚”作唯一的关键词(组)分析各地的搜索热度，但一般情况下人们关心的主题会涉及到多个关键词(组)，如何得到这些关键词(组)搜索分布的合成效果是本文致力于研究的课题。

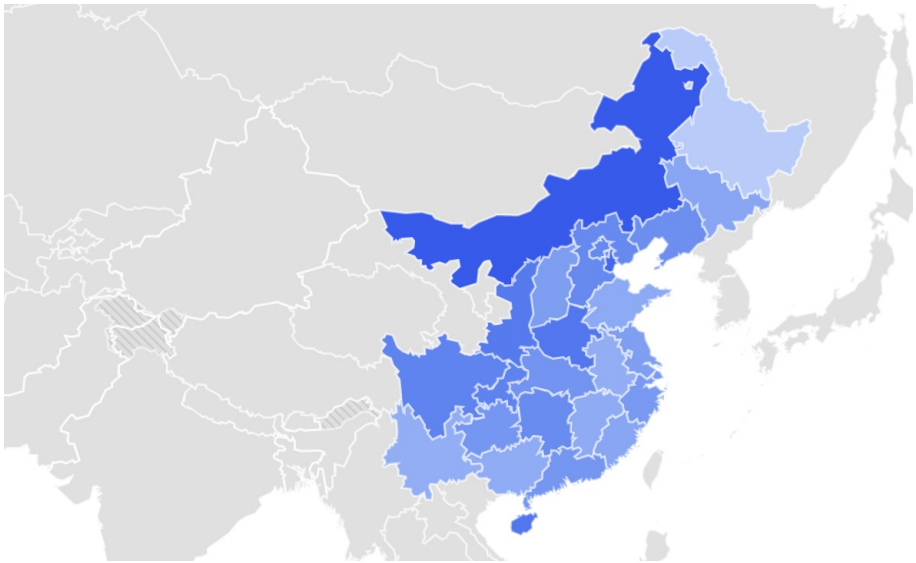

Figure 1. Distribution of search popularity for “divorce” in various provinces and regions in China in the past 5 years
图 1. 近 5 年中国各省区“离婚”搜索热度分布

本文方法的阐述是通过新冠疫情主题的实例进行展开的，并研究各地网民疫情搜索行为和当地确诊病例数以及各地 GDP 产值的关系，揭示居民借助网络搜索引擎积极学习应对新冠病毒的知识对于抑制疫情蔓延的正向作用，启发公共卫生管理部门应更加重视新冠疫情的网络内容的发布，以及 GDP 大省更应着力提升互联网运用在居民防疫抗疫方面的助力。

后文第 2 节介绍主题搜索热度的地理分布的合成算法，第 3 节进行算法应用的讨论，第 4 节进行全文总结。

2. 词族热度的地理分布

2.1. 算法背景

使用谷歌趋势进行某主题调研时，常常希望得到该主题的搜索热度依行政地域的分布，即在关注主题的时间趋势时如何获取主题的频度热度的空间分布信息。目前谷歌趋势只提供了某关键词(组)的相关主题的排名，没有直接的关于主题的时间趋势和空间分布的输出。主题是由多个相关的网络用户使用频度较高的关键词(组)构成，对于单个关键词(组)谷歌趋势提供时间趋势和地域分布，而对于多个关键词(组)谷歌提供趋势比较曲线，以及多个关键词(组)的各词热度比例的地理分布，并非是合成的热度绝对量的分布。

那么如何获得一个主题对应的关键词(组)词族？首先为该主题配置一个根词属性的关键词，然后输入到谷歌趋势进行检索，根据输出的相关查询，生成该主题的词族，这个根关键词称为入口检索词。图 2 展示了“离婚”入口检索及其相关查询输出的示意。

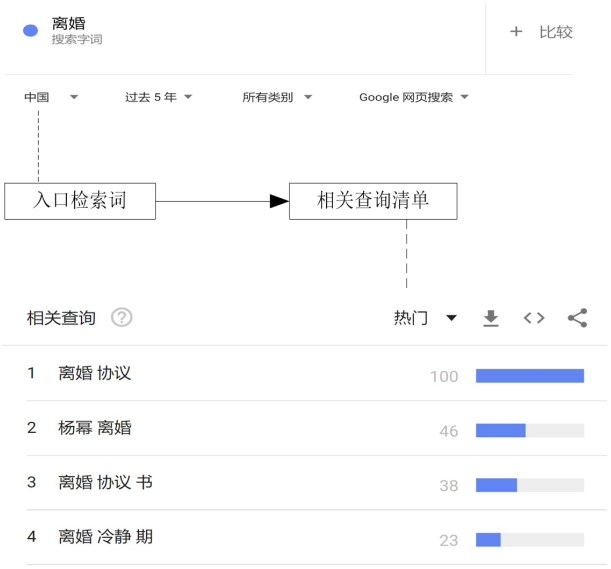

Figure 2. Entry search terms and related query output interface
图 2. 入口检索词及其相关查询输出界面

以新冠疫情主题为例，设计入口检索词为“病毒”，并以谷歌趋势对关键词“病毒”所作的相关查询排名清单给出新冠疫情的主题词族：{病毒，冠状 病毒，新型 冠状 病毒，新冠 病毒，冠状 病毒 肺炎}。利用多关键词(组)的趋势比较功能(谷歌趋势每次只能作至多 5 个关键词的比较)，得到比较趋势曲线和各词热度比例的地理分布，见图 3、图 4 所示。

在图 3 中给出该词族的 5 个关键词(组)各自在 2019 年 6 月底到 2020 年 6 月底一年期间的搜索热度，每星期一个采样点，一年共 52 个数据点。每个时间点对应的搜索热度值已经由谷歌系统正则化，以“病毒”的某峰值点数据为 100，其余都取与该峰值的相对比值。从图 3 看出词族中各词热度强度也是按照词族中排列的顺序降低的：“病毒”>“冠状 病毒”>“新型 冠状 病毒”>“新冠 病毒”>“冠状 病毒 肺炎”。图 4 按中国各省区地域给出了各词热度的比例关系，从中看出除西藏自治区外每个省的总值都是 100，并给出了各词所占的比例。顺便指出西藏在该词族的各词(组)搜索热度皆为零，当然是系统正则化处理时比值小于 1 而归零的。否则，假定西藏在“病毒”的某时间点搜索热度为 1，其它时刻、其余词(组)皆为 0，则在图 4 的比例关系的总值也为 100。可见图 4 虽然给出了多关键词(组)的热度比例的地理分布，却不是通常主题调研者所关心的多关键词(组)合成的热度绝对量的分布。

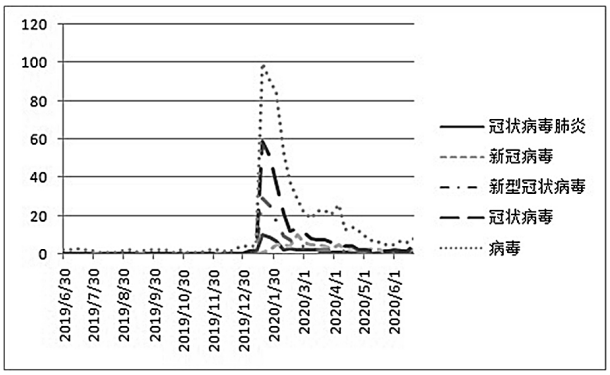

Figure 3. Comparison of the trend of multiple keywords (group) of the words family
图 3. 词族多关键词(组)的趋势比较

图 5~9 是该词族各关键词(组)搜索热度在 2019 年 6 月底到 2020 年 6 月底的地理分布情况。从中看出各词(组)分布中皆以 100 作峰值，且显然各词的 100 代表的实际搜索次数值肯定是不同的，它是以各词(组)的地理分布中的峰值次数作正则化处理得来的。因此直接将各词(组)的分布依各地域项相加是没有意义的，不是合成的搜索热度地理分布。

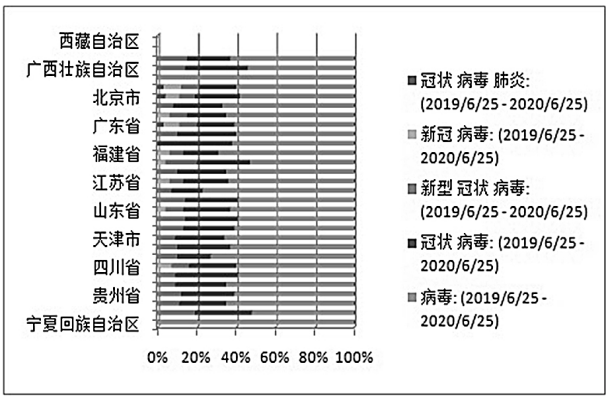

Figure 4. Geographical distribution of the popularity of each word
图 4. 各词热度比例的地理分布

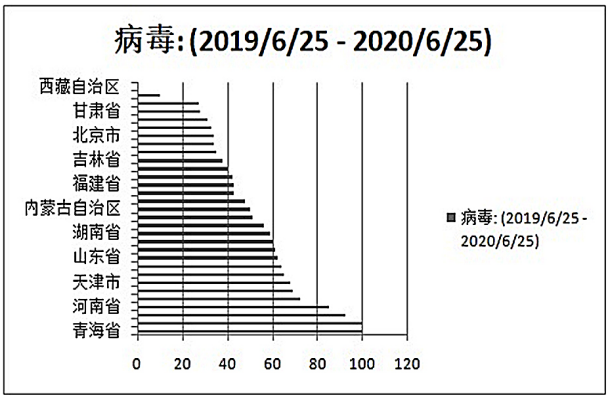

Figure 5. "Virus" geographical distribution
图 5. "病毒" 地理分布

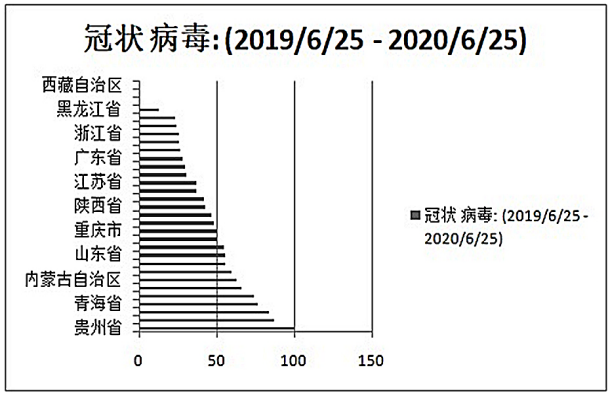

Figure 6. Geographical distribution of "coronavirus"
图 6. "冠状 病毒" 的地理分布

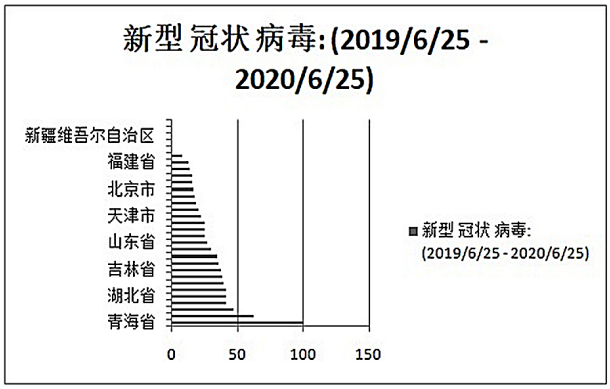

Figure 7. Geographical distribution map of "new coronavirus"
图 7. "新型 冠状 病毒"的地理分布

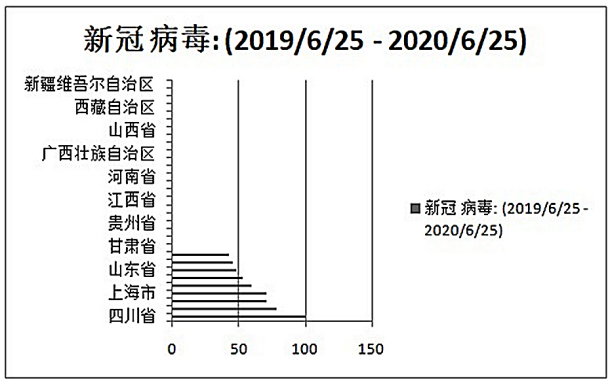

Figure 8. Geographical distribution of the "new crown virus"
图 8. "新冠 病毒"的地理分布

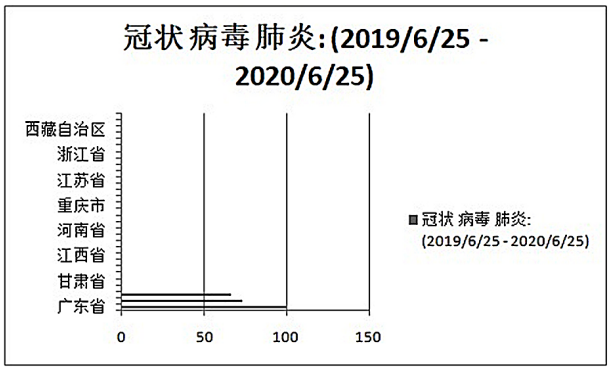

Figure 9. Geographical distribution of "coronavirus pneumonia"
图 9. "冠状 病毒 肺炎"的地理分布

2.2. 算法推导

为推导各词(组)合成的热度绝对量的分布，记 w_i^j 为 j 词(组) i 地域的分布值(如图 5~9)，$w_i^j(t)$ 为 t 时刻 j 词(组)i 地域的分布值，它们是 0 到 100 的自然数，而且满足

$$w_i^j = \sum_t w_i^j(t) \tag{1}$$

记词族的地域分布为 W_i，

$$W_i = \sum_j w_i^j \kappa(j) \tag{2}$$

其中$\kappa(j)$是引入的未知参数，符合限定条件：$100 \cdot \kappa(j) = j$词(组)的峰值地域的搜索次数。为导出$\kappa(j)$的表达式考虑各词(组)趋势比较，记I_t^j表示t时刻j词(组)的趋势值(如图3)，易知

$$I_t^j = \sum_i w_i^j(t)\kappa(j) \tag{3}$$

且$\sum_j I_t^j$即为词族的时间趋势。

进一步得$\sum_t I_t^j = \sum_t \sum_i w_i^j(t)\kappa(j) = \sum_i \sum_t w_i^j(t)\kappa(j) = \sum_i w_i^j \kappa(j)$，

即

$$\kappa(j) = \frac{\sum_t I_t^j}{\sum_i w_i^j} \tag{4}$$

于是

$$W_i = \sum_j w_i^j \kappa(j) = \sum_j w_i^j \frac{\sum_t I_t^j}{\sum_i w_i^j} = \sum_j \frac{w_i^j}{\sum_i w_i^j} \sum_t I_t^j \tag{5}$$

可见各词(组)合成的地域分布W_i是由各词(组)趋势比较曲线关于时间t的积分$\sum_t I_t^j$的加权和，权系数$\frac{w_i^j}{\sum_i w_i^j} = \frac{w_i^j \kappa(j)}{\sum_i w_i^j \kappa(j)}$是每个地域的$j$词(组)实际搜索次数占所有地域该词(组)的搜索次数的比例。但W_i并不等同于i地域的词族的实际搜索次数，它依旧如趋势值I_t^j那样是被正则化影响的刻画搜索频度的数值。

由式(5)可知作出词族的地域分布需要首先作出多词(组)的趋势比较并计算每个词(组)的趋势积分，然后逐一获取每个词(组)的地理分布生成各词(组)各地域的权重系数，最终加权求和而成，图10给出了这一算法过程框图；图11是基于图3和图5~9经由图10所示算法作出的该词族的各地域搜索热度分布图。

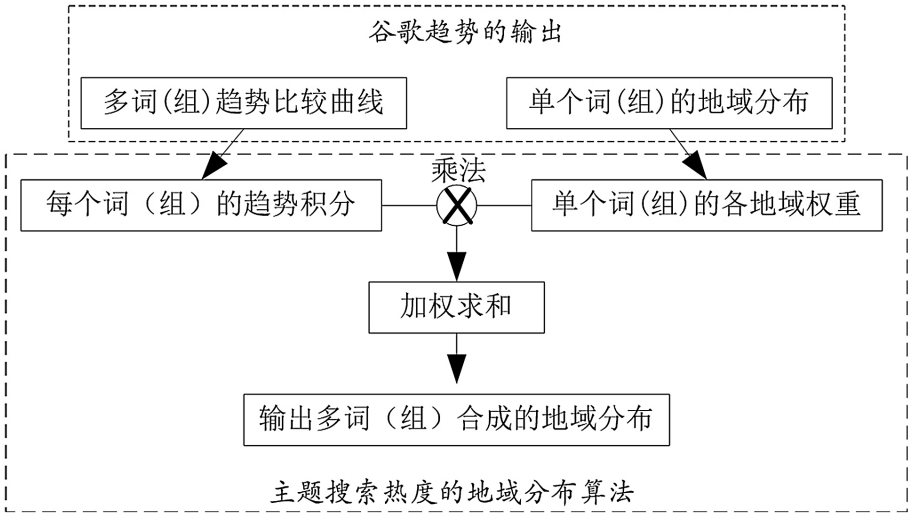

Figure 10. The geographical distribution algorithm of topic search popularity
图10. 主题搜索热度的地理分布算法

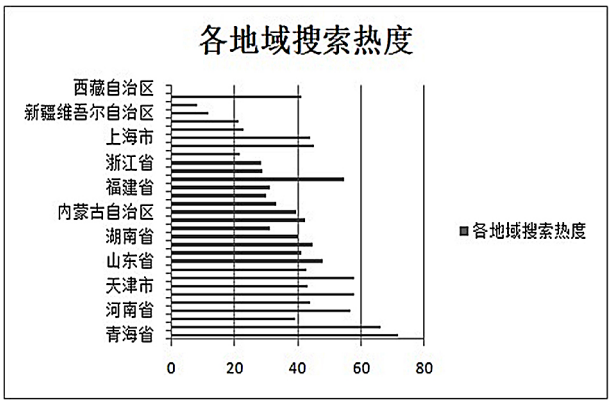

Figure 11. Search popularity distribution of the word family by region
图 11. 该词族的各地域搜索热度分布

3. 算法应用

上节图 11 给出的地域搜索热度是该地区网民总体行为的结果，人口大省拥有的网民数也多，自然就有更多的搜索量。为此，考虑地域的平均搜索热度，

$$\overline{W_i} = W_i / P_i \tag{6}$$

其中 P_i 指 i 地域的网民数，在得到网民数困难情况下可以使用当地人口数替代。地域的平均搜索热度从一定程度上反映了地区居民对疫情发展和防治的关注程度，通过网络搜索引擎学习新冠病毒的预防知识对于抑制疫情蔓延无疑是有用的。我们采用中国各省区、美国各州截至同期的确诊数与当时一年内的地域平均搜索热度作相关性分析，证实了网络搜索引擎对新冠病毒蔓延的抑制作用，在中美两国地域平均搜索热度和各地新冠确诊数有着显著的负相关性，美国的显著性比中国更强。

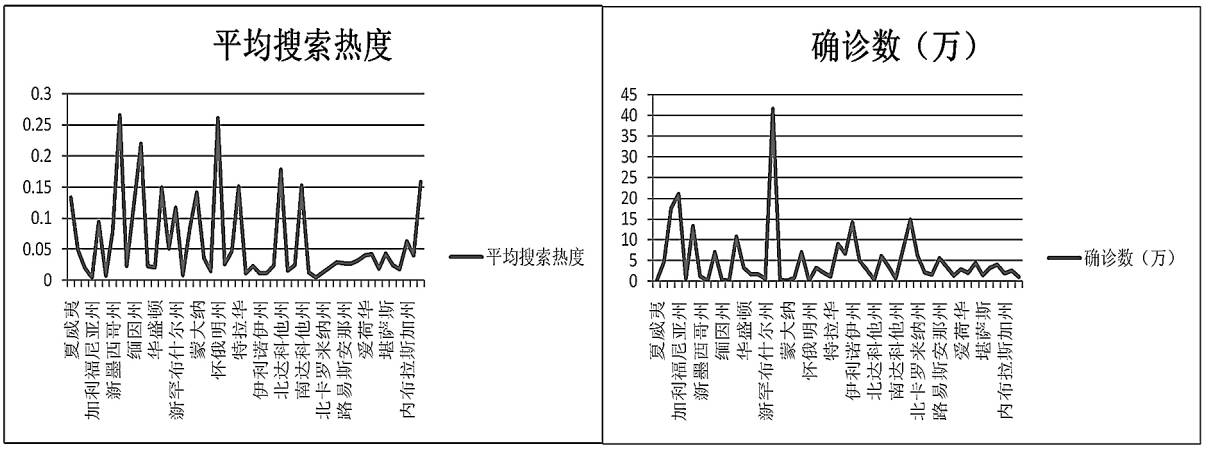

Figure 12. Average search popularity and number of new crown diagnoses in each state in the U.S.
图 12. 美国各州平均搜索热度和各州新冠确诊数

通过谷歌趋势找出入口检索词“virus”的相关查询排名清单构成词族{virus, corona virus cases, corona virus symptoms, COVID virus, corona virus update}作为英语疫情主题的五个关键词(组)，应用热度地理分布算法得到美国各州的平均搜索热度，如图 12 (左端)。在利用式(6)作平均时是使用美国各州人口数，而非各州网民数。图 12 右端各州确诊数是截至 2020 年 6 月底的数据。利用 SPSS 软件对两组数据进行皮尔逊线性相关性分析：

Table 1. Correlation between the average search popularity and the number of confirmed diagnoses in each state in the United States
表 1. 美国各州平均搜索热度和各州确诊数的相关性

		平均搜索热度	各州确诊数(万)
平均搜索热度	皮尔逊相关性	1	−0.453**
	显著性(双尾)		0.001
	个案数	51	51
各州确诊数(万)	皮尔逊相关性	−0.453**	1
	显著性(双尾)	0.001	
	个案数	51	51

**.在 0.01 级别(双尾)，相关性显著。

从表 1 可以看出平均疫情主题搜索热度和确诊数是负相关的，显著性水平非常突出，相关性的线性率也达到中等以上程度。

使用中国各省网民数作平均得到的平均搜索热度与各省区确诊数的相关性分析如表 2。

Table 2. The correlation between the average search popularity and the number of confirmed cases in 29 provinces and regions (excluding Hubei, Tibet, Hong Kong, Macao and Taiwan) in mainland China
表 2. 中国大陆 29 省区(不含湖北、西藏和港澳台)平均搜索热度和各省区确诊数的相关性

		平均搜索热度	各省区确诊数
平均搜索热度	皮尔逊相关性	1	−0.385**
	显著性(双尾)		0.039
	个案数	29	29
各省区确诊数	皮尔逊相关性	−0.385**	1
	显著性(双尾)	0.039	
	个案数	29	29

**.在 0.05 级别(双尾)，相关性显著。

为进一步证实平均搜索热度与确诊数的负相关性是本质性的，采用各省区人口数作成平均搜索热度，这个版本的平均搜索热度可以看成原始版本的一个噪声扰动，那么在数据带有扰动时可以预计线性率和显著性水平均会降低。比较表 2 和表 3 可见，负相关性不变，线性率由 0.385 降至 0.380，而显著性水平由 0.039 降至 0.042 (显著性水平数值越小，显著性强度越高)。

Table 3. Correlation between the average epidemic concern index (averaged by population) and the number of confirmed cases in 29 provinces and regions in Mainland China (excluding Hubei, Tibet, Hong Kong, Macao and Taiwan)
表 3 中国大陆 29 省区(不含湖北、西藏和港澳台)平均疫情搜索热度(按人口数平均)和各省区确诊数的相关性

		平均搜索热度(人口数平均)	各省区确诊数
平均搜索热度	皮尔逊相关性	1	−0.380**
	显著性(双尾)		0.042
	个案数	29	29
各省区确诊数	皮尔逊相关性	−0.380**	1
	显著性(双尾)	0.042	
	个案数	29	29

**.在 0.05 级别(双尾)，相关性显著。

由表 1 和表 2 比较显示中美两国的地域平均搜索热度与当地确诊数均呈现显著负相关性，但美国的显著性水平更高。

自新冠疫情爆发以来，国内外相关新闻报导指出了疫情传播和城市居民出行以及地区 GDP 产值的关系。这里显示各地疫情主题的搜索热度与地区 GDP 产值也具有一定的相关性。

中国大陆 31 省区平均疫情主题搜索热度与各省 GDP 值有显著的负相关，相关性分析见表 4。由于

西藏的平均疫情关注指数为 0 是一个(谷歌趋势数据正则化导致的)显著的噪声，考虑去除西藏后的 30 省区平均疫情主题搜索热度与该 30 省区 GDP 值的相关性，见表 5。从表 5 到表 4 是数据增加一个噪声(对相关性分析不利)，所以显著性水平 p 值是由 0.021 弱化到 0.035，相关性系数绝对值由 0.420 降到 0.380，由此可见平均疫情关注指数的地域分布和各地 GDP 值的负相关性是客观事实。

Table 4. Correlation between the average search popularity of the epidemic and the GDP value of each province in mainland China
表 4. 中国大陆 31 省区平均疫情搜索热度和各省区 GDP 值的相关性

		平均疫情搜索热度	各省 GDP 值
平均疫情搜索热度	皮尔逊相关性	1	−0.380**
	显著性(双尾)		0.035
	个案数	31	31
各省 GDP 值	皮尔逊相关性	−0.380**	1
	显著性(双尾)	0.035	
	个案数	31	31

**.在 0.05 级别(双尾)，相关性显著。

Table 5. Correlation between average search popularity of epidemics in 30 provinces and regions (excluding Tibet) in Mainland China and GDP value of the 30 provinces and regions
表 5. 中国大陆 30 省区(不含西藏)平均疫情搜索热度和该 30 省区 GDP 值的相关性

		平均疫情搜索热度	各省 GDP 值
均疫情搜索热度	皮尔逊相关性	1	−0.420**
	显著性(双尾)		0.021
	个案数	30	30
各省 GDP 值	皮尔逊相关性	−0.420**	1
	显著性(双尾)	0.021	
	个案数	30	30

**.在 0.05 级别(双尾)，相关性显著。

至于美国各州 GDP 值和各州平均疫情搜索热度却没有显著相关性，但考虑各州人均 GDP 却与平均疫情搜索热度的分布呈现显著的正相关，分析结果见表 6。考察中美两国平均疫情搜索热度和各地(人均)GDP 值的相关性分析结果，两者的显著性水平差异不大，但相关性指向相反。这也确实反映了中美两国 GDP 结构的差异，因为美国的三产服务业 GDP 比重较大，而互联网网络搜索服务正是属于服务业范畴。

Table 6. Correlation between the average search popularity of each state in the United States and the per capita GDP of each state
表 6. 美国各州平均疫情搜索热度和各州人均 GDP 的相关性

		平均疫情搜索热度	各州人均 GDP
平均疫情搜索热度	皮尔逊相关性	1	0.329**
	显著性(双尾)		0.018
	个案数	51	51
各州人均 GDP	皮尔逊相关性	0.329**	1
	显著性(双尾)	0.018	
	个案数	51	51

**.在 0.05 级别(双尾)，相关性显著。

4. 总结

以谷歌趋势提取某个主题的搜索热度的地理信息是多关键词(组)时间趋势外的空间分布情况，是弥补时间趋势不足的更全面的信息展示，具有重要的战略决策价值。新冠疫情发展的趋势监测和地理分布对

于疫情蔓延的控制治理具有重要的情报意义。网络搜索引擎提供实时的“动态”数据，对这些数据的处理和分析给公共卫生管理部门提供了更有价值的疫情趋势规律，为打赢防疫抗疫阻击战提供助力。对于GDP 大省其信息化、网络化运用水平未必与 GDP 数量排位一致，需要加强运用互联网信息化手段去提升疫情防控水准。

基金项目

宁夏回族自治区自然科学基金(2020AAC03278)资助。

参考文献

[1] Zhao, Y., Xu, Q., Chen, Y., *et al*. (2018) Using Baidu Index to Nowcast Hand-Foot-Mouth Disease in China: A Meta Learning Approach. *BMC Infectious Diseases*, **18**, Article No. 398. https://doi.org/10.1186/s12879-018-3285-4

[2] Qin, L., Sun, Q., Wang, Y.D., Wu, K.-F., Chen, M.C., Shi, B.-C. and Wu, S.-Y. (2020) Prediction of Number of Cases of 2019 Novel Coronavirus (COVID-19) Using Social Media Search Index. *International Journal Environmental Research and Public Health*, **17**, 2365. https://doi.org/10.3390/ijerph17072365

[3] Zhao, Y.Q., Li, R.D. and Wu, M.Q. (2020) Correlation Studies between Land Cover Change and Baidu Index: A Case Study of Hubei Province. *International Journal of Geo-Information*, **9**, 232. https://doi.org/10.3390/ijgi9040232

[4] Wang, T., Xia, Q., Chen, X. and Jin, X. (2020) Use of Baidu Index to Track Chinese Online Behavior and Interest in Kidney Stones. *Risk Management and Healthcare Policy*, **13**, 705-712. https://doi.org/10.2147/RMHP.S245822

[5] Schootman, M., Toor, A., Cavazos-Rehg, P., *et al*. (2015) The Utility of Google Trends Data to Examine Interest in Cancer Screening. *BMJ Open*, **5**, e006678. https://doi.org/10.1136/bmjopen-2014-006678

[6] Brodeur, A., Clark, A.E., Flèche, S. and Powdthavee, N. (2020) COVID-19, Lockdowns and Well-Being: Evidence from Google Trends. IZA Institute of Labor Economics, IZA DP No. 13204. https://doi.org/10.1016/j.jpubeco.2020.104346

[7] 徐映梅, 高一铭. 基于互联网大数据的 CPI 舆情指数构建与应用[J]. 数量经济技术经济研究, 2017(1): 94-112.

[8] 杨艳红, 曾庆, 等. 基于谷歌趋势的乙型肝炎预测模型[J]. 上海交通大学学报医学版, 2013, 33(2): 204-209.

[9] 陈叶旺, 王华珍, 等. 基于百度百科与文本分类的网络文本语义主题抽取方法[J]. 小型微型计算机系统, 2012, 33(12): 2605-2610.

[10] 刘海鸥, 黄文娜, 等. 移动社交网络情境化推荐关键问题研究综述[J]. 小型微型计算机系统, 2020, 41(9): 1812-1819.