文章编号：2095-2163(2021)04-0035-04 中图分类号：G206；C912 文献标志码：A

基于爬虫技术与智能算法的网络舆情监测

雍龙泉[1,2]，贾 伟[1]，张建科[3]
（1 陕西理工大学 数学与计算机科学学院，陕西 汉中，723001；2 陕西省工业自动化重点实验室，陕西 汉中 723001；
3 西安邮电大学 理学院，西安 710121）

摘 要：采用网络爬虫技术从百度指数获取某一“热门事件”的数据，并对这些数据进行预处理，进而建立网络舆情的 Logistic 微分方程模型。结合已有数据，采用智能算法确定微分方程解中的 3 个关键参数；最后应用于网络舆情预测。

关键词：网络舆情；爬虫技术；百度指数；Logistic 微分方程模型；智能算法

Network public opinion monitoring based on crawler technology and intelligent algorithm

YONG Longquan[1,2], JIA Wei[1], ZHANG Jianke[3]
(1 School of Mathematics and Computer Science, Shaanxi University of Technology, Hanzhong Shaanxi 723001, China;
2 Shaanxi Key Laboratory of Industrial Automation, Hanzhong Shaanxi 723001, China;
3 School of Science, Xi'an University of Posts and Telecommunications, Xi'an 710121, China)

【Abstract】 The data of a "hot event" is obtained by web crawler technology from Baidu Index, and the data are further preprocessed, then the Logistic differential equation model of network public opinion is established. Three key parameters in the solution of differential equation are determined by intelligent algorithm based on the given data. Finally, this method is applied to network public opinion prediction.

【Key words】 network public opinion; crawler technology; Baidu Index; Logistic differential equation model; intelligent algorithm

0 引 言

互联网技术的发展到今天，社交网络大肆兴起，人们越来越习惯于使用社交网络媒体，也越来越倾向于借助这一平台来实时分享自己的信息，发表言论、抒发情感。但是，当某一事件发生时，必将在社交网络中广泛传播，由于社交网络用户的爆炸式增长，就很可能会产生舆情，舆情将会对民众产生巨大的影响，甚至对社会安全产生一定的威胁[1-4]。因此，研究突发事件网络舆情的传播特性及演化过程，建立数学模型来探讨突发事件网络舆情演化规律及动力学分析，具有重要的现实意义[5-8]。

本文采用较为流行的网络爬虫技术从百度指数爬取某一“热门事件”的数据，并对这些数据进行预处理；进而建立 Logistic 微分方程数学模型，利用已有数据，采用智能算法确定微分方程解中的 3 个参数，最后应用于网络舆情预测。

本文的思路如图 1 所示。

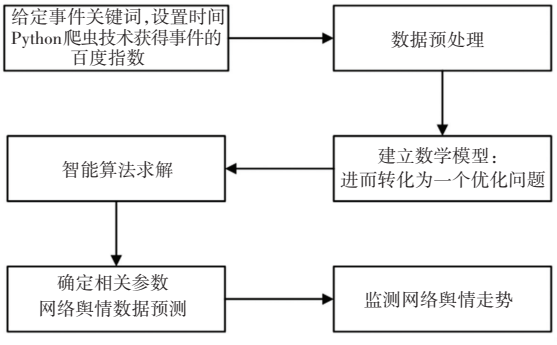

图 1 网络舆情监测流程

Fig. 1 Network public opinion monitoring process

下文以近期“苟晶”事件为例，说明该方法的应用。

1 实现过程

1.1 获取百度搜索指数

打开网址 http://index.baidu.com/，输入关键词“苟晶”；设置时间范围为近 30 天，即 2020.6.24-2020.7.23（时间范围可以手动设置）；得到百度搜索指数如图 2 所示。

基金项目：国家自然科学基金（11401357）；陕西省教育厅重点科学研究计划项目（20JS021）；陕西理工大学科研项目（SLGYQZX2002）；陕西理工大学教学改革研究项目（SLGYJG2015）；陕西省重点研发计划项目（2021SF-480）。

作者简介：雍龙泉（1980-），男，博士，教授，主要研究方向：最优化理论与算法、智能优化算法；贾 伟（1977-），男，硕士，讲师，主要研究方向：智能优化算法；张建科（1978-），男，博士，副教授，主要研究方向：舆情分析。

收稿日期：2020-12-03

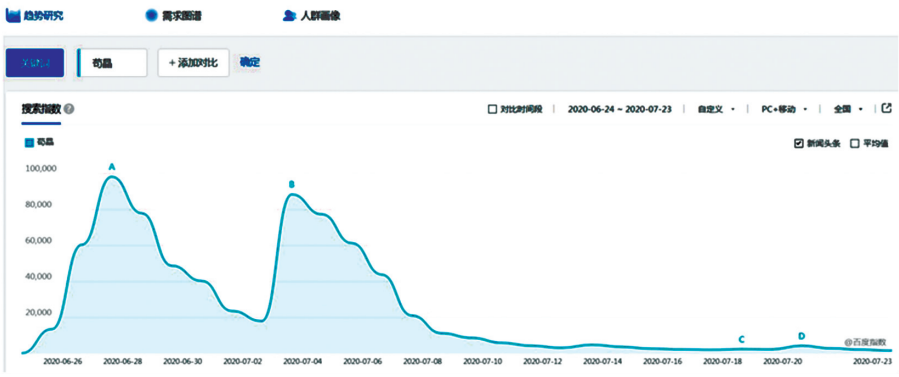

图 2 百度搜索指数

Fig. 2 Baidu Search Index

百度指数,主要包括搜索指数、资讯指数、媒体指数。百度搜索指数是以网民每天在百度的搜索量为数据基础,以关键词为统计对象,科学分析并计算出各个关键词在百度网页搜索中搜索频次的加权和;它能形象地反映该关键词每天的变化趋势,根据使用百度搜索来源的不同,搜索指数分为 PC 端和移动端。

我们采用 Python 爬虫技术[9-10],获得该时间段内的数据见表 1 前 3 列。为了建立数学模型,下面对这些数据进行初步处理。

1.2 数据处理

把 2020-06-24 当做第一天,即 $t_1 = 1$, 其余依次类推,2020-07-23 便是第 30 天;百度搜索指数,也即每天的关注量,从数学上而言即单位时间信息量 dx/dt; 每天关注量的累计, 即网络舆情信息量的和 $x(t_i)$; 简单处理后得到表 1 的第 4 与第 5 列数据。

表 1 数据列表

Tab. 1 Data list

搜索关键词	时间	百度搜索指数:即单位时间信息量	时间序列 t_i	每天关注量累计:即网络舆情信息量的和 $x(t_i)$	预测结果
苟晶	2020-06-24	0	1	0	56 101.46
苟晶	2020-06-25	13 359	2	13 359	78 068.84
苟晶	2020-06-26	60 274	3	73 633	107 255.7
苟晶	2020-06-27	98 110	4	171 743	144 904.9
苟晶	2020-06-28	77 799	5	249 542	191 660
苟晶	2020-06-29	48 581	6	298 123	247 060.3
苟晶	2020-06-30	40 154	7	338 277	309 166.6
苟晶	2020-07-01	23 496	8	361 773	374 611.7
苟晶	2020-07-02	17 834	9	379 607	439 227.1
苟晶	2020-07-03	88 207	10	467 814	499 049.1
苟晶	2020-07-04	77 227	11	545 041	551 224.1
苟晶	2020-07-05	61 178	12	606 219	594 412.8
苟晶	2020-07-06	43 701	13	649 920	628 642.9
苟晶	2020-07-07	20 934	14	670 854	654 850.4
苟晶	2020-07-08	11 044	15	681 898	674 389.1
苟晶	2020-07-09	8 520	16	690 418	688 669
苟晶	2020-07-10	5 757	17	696 175	698 954.5
苟晶	2020-07-11	4 140	18	700 315	706 285.5
苟晶	2020-07-12	2 995	19	703 310	711 471.5
苟晶	2020-07-13	4 728	20	708 038	715 120.6
苟晶	2020-07-14	3 542	21	711 580	717 678.8
苟晶	2020-07-15	2 521	22	714 101	719 467.4
苟晶	2020-07-16	2 164	23	716 265	720 715.7
苟晶	2020-07-17	1 908	24	718 173	721 585.8
苟晶	2020-07-18	2 313	25	720 486	722 191.7
苟晶	2020-07-19	2 149	26	722 635	722 613.4
苟晶	2020-07-20	4 124	27	726 759	722 906.7
苟晶	2020-07-21	2 699	28	729 458	723 110.7
苟晶	2020-07-22	1 980	29	731 438	723 252.6
苟晶	2020-07-23	1 478	30	732 916	723 351.2

1.3　建立数学模型

大疫当前，数学能做什么[11]？国内已有一些学者对突发事件网络舆情进行了研究，张一文等利用系统动力学建模探究事物自身演化机理，为控制非常规突发事件网络舆情扩散，引导非常规突发事件舆情传播提供有力依据[12]；宋海龙等根据突发事件网络舆情具有自由性、互动性、即时性、隐匿性、群体极化性等特点，探讨了形成、高涨、波动和最终淡化 4 个阶段网络舆情的引导和控制问题[13]。学习过数学建模或者生物数学的人都知道，(整数阶或分数阶)传染病模型，包括 SIR、SEIR 模型[14-15]等，从数学上而言，就是微分方程(组)。换句话说，只要是与时间变化相关，则建立的模型要么是差分方程模型(离散问题)，要么是微分方程模型(连续问题)，此外还有时滞微分方程模型等。

网络舆情的演进规律，遵循如下微分方程模型：

$$\begin{cases}\dfrac{dx}{dt}=rx(t)\left(1-\dfrac{x(t)}{K}\right)\\ x(0)=x_0\end{cases}. \tag{1}$$

这里 $x(t)$ 表示网民对某一热门事件进行交流而形成的网络舆情信息量的和，单位时间信息量和值的相对变化率为 r，$x(t)$ 的上限为 K。这个模型也称为 Logistic 模型，广泛应用于生态系统、经济系统、传染病模型等。计算可以得到该微分方程的解为：

$$x(t)=\frac{K}{1+\left(\dfrac{K}{x_0}-1\right)e^{-rt}}. \tag{2}$$

取初始值 $x_0=50$，$K=11\ 000$，$r=1$，函数 $x(t)$ 的图像如图 3 所示。

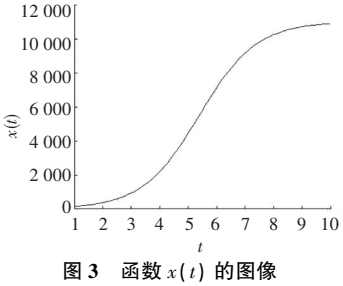

图 3　函数 $x(t)$ 的图像

Fig. 3　Image of function $x(t)$

在舆情建模与仿真相关文献里面，有的学者把舆情的传播划分为 3 个阶段，分别称之为舆情的产生阶段、发展阶段、衰退阶段；有的学者把舆情的传播划分为 6 个阶段，分别称之为舆情的潜伏期、成长期、蔓延期、爆发期、衰退期、消亡期；也有些学者把舆情的传播划分为 5 个阶段，分别称之为舆情的潜伏期、萌动期、加速期、成熟期、衰退期。划分为 5 个阶段的较为常见，这方面的研究见文献[5-6]，在此不再详述。

网络舆情信息量的和 $x(t)$，其一般形式为：

$$x(t)=\frac{c_1}{1+c_2e^{-c_3t}}. \tag{3}$$

这里 $c_1=K$；$c_2=\left(\dfrac{K}{x_0}-1\right)$；$c_3=r$。关键问题是算出参数 c_1、c_2、c_3 的值，这是一个非线性拟合问题。下面把非线性拟合问题转化为一个最优化问题，然后采用智能优化算法进行求解。

分别建立如下最小误差优化模型：

$$\begin{cases}\min\sum_{i=1}^{n}f_i^2(c_1,c_2,c_3)\\ \text{s.t. } f_i(c_1,c_2,c_3)=x(t_i)-\left[\dfrac{c_1}{1+c_2e^{-c_3t_i}}\right],i=1,2,\cdots,n.\end{cases} \tag{4}$$

$$\begin{cases}\min\sum_{i=1}^{n}|f_i(c_1,c_2,c_3)|\\ \text{s.t. } f_i(c_1,c_2,c_3)=x(t_i)-\left[\dfrac{c_1}{1+c_2e^{-c_3t_i}}\right],i=1,2,\cdots,n.\end{cases} \tag{5}$$

模型(4)采用非线性最小二乘，模型(5)采用非线性最小一乘。

1.4　智能算法求解

令 $X=(c_1,c_2,c_3)$，模型(4)与(5)便是一个无约束优化问题 $\min f(X)$，下面采用正弦余弦算法(Sine Cosine Algorithm，简称 SCA)来确定参数 X。SCA 算法步骤如下：

步骤 1　初始化

设置种群规模 N，空间维数 D，控制参数 a，最大迭代次数 $T_{\max}$；在可行域空间中随机初始化 N 个个体组成初始种群；$t=1$；计算当前每个个体的适应值，并记录最优个体位置 $P(t)$；

步骤 2　种群更新

while($t<T_{\max}$)

for $i=1$ to N do　//对每一个个体进行更新

for $j=1$ to D do　//对每一维上进行更新

根据式 $r_1=a-at/T_{\max}$ 计算 r_1 的值；

随机产生 $r_2\in U[0,2\pi]$，$r_3\in U[0,2]$，$r_4\in U[0,1]$；

if $r_4<0.5$

$$X_i^j(t+1)=X_i^j(t)+r_1\sin(r_2)\left|r_3P^j(t)-X_i^j(t)\right|; \tag{6}$$

else

$$X_i^j(t+1)=X_i^j(t)+r_1\cos(r_2)\left|r_3P^j(t)-X_i^j(t)\right|; \tag{7}$$

end if

end for

end for

越界处理;

计算每个个体的适应值并更新种群的最优个体位置 $P(t)$;$t = t + 1$;

end while

步骤 3 输出解

SCA 算法最显著的特点是基于正弦函数(6)和余弦函数(7)值的变化来达到寻优目的,其结构简单,容易实现,在 SCA 算法中,主要参数有 4 个:r_1、r_2、r_3、r_4。其中,最关键的是 r_1,控制算法从全局搜索到局部开发的转换。有关正弦余弦算法详细分析,见文献[16]。

对模型(4),SCA 优化结果为:

X = (723 575.959 657 036,17.119 765 978 283 3,

0.363 894 261 998 827)。

1.5 网络舆情数据预测

从以上便得到:

c_1 = 723 575.959 657 036, c_2 = 17.119 765 978 283 3,

c_3 = 0.363 894 261 998 827;

于是 $x(t)$ 的表达式为:

$$x(t) = \frac{723\ 575.959\ 657\ 036}{1 + 17.119\ 765\ 978\ 283\ 3e^{-0.363\ 894\ 261\ 998\ 827t}}. \quad (8)$$

代入 t_i,$i = 1,2,\cdots,30$,预测得到的数据见表 1 的最后一列。

为了较为直观的反映网络舆情的发展趋势,图 4 给出了百度搜索指数,图 5 给了网络舆情信息量的拟合曲线与误差图。

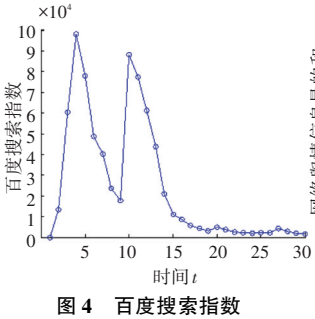

图 4 百度搜索指数

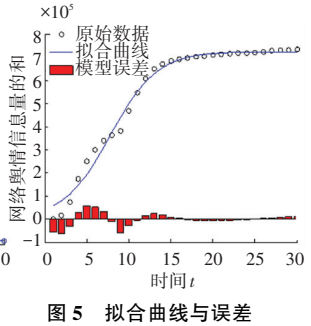

图 5 拟合曲线与误差

Fig. 4 Baidu Search Index Fig. 5 The fitting curve and error

从图 5 可以看出,前期存在误差,后期误差很小。主要原因在于,对于突发事件,网络舆情前期不稳定、波动大,所以误差较大;越到后期,网络舆情较为一致,所以后期误差较小。

2 结束语

正弦余弦算法对目标函数的可导性无限制,因此不论是采用可导的非线性最小二乘模型(4),还是采用不可导的非线性最小一乘模型(5),SCA 算法都能够获得近似一致的结果;这为舆情传播建模与仿真开辟了新的方向。

参考文献

[1] 覃伊蕾,王清泉. 新媒体时代的突发公共卫生事件舆情演化特点及治理措施——以新冠肺炎疫情为例[J]. 新闻研究导刊,2020,11(7):13-16.

[2] 邹静. 在重大突发公共事件中传统媒体如何应对网络舆情——以湖北广电集团抗击新冠肺炎疫情宣传为例[J]. 当代电视,2020(4):57-60.

[3] 陈兴蜀,常天祐,王海舟,等. 基于微博数据的"新冠肺炎疫情"舆情演化时空分析[J]. 四川大学学报(自然科学版),2020,57(2):409-416.

[4] 赵耀,王建新. 基于多元主体共在与信息即时公开的新冠肺炎疫情网络舆情的思考[J]. 中国矿业大学学报(社会科学版),2020,22(2):88-100.

[5] 兰月新,邓新元. 突发事件网络舆情演进规律模型研究[J]. 情报杂志,2011,30(8):47-50.

[6] 兰月新,夏一雪,刘冰月,等. 网络舆情传播阶段精细化建模与仿真研究[J]. 现代情报,2018,38(1):76-86.

[7] 张鹏,兰月新,李昊青,等. 突发事件网络谣言危机预警及模拟仿真研究[J]. 现代情报,2019,39(12): 101-108,137.

[8] 赵剑华, 万克文. 基于信息传播模型-SIR 传染病模型的社交网络舆情传播动力学模型研究[J]. 情报科学, 2017, 35(12): 34-38.

[9] 夏火松,李保国. 基于 Python 的动态网页评价爬虫算法[J]. 软件工程,2016,19(2):43-46.

[10] 熊畅. 基于 Python 爬虫技术的网页数据抓取与分析研究[J]. 数字技术与应用,2017(9):35-36.

[11] 梁进. 大疫当前,数学能做什么?[J]. 科学,2020,72(2):57-60.

[12] 张一文,齐佳音,马君,等. 网络舆情与非常规突发事件作用机制——基于系统动力学建模分析[J]. 情报杂志,2010,29(9):1-6.

[13] 宋海龙,巨乃岐,张备,等. 突发事件网络舆情的形成、演化与控制[J]. 河南工程学院学报(社会科学版),2010,25(4):12-16.

[14] 马知恩, 周义仓, 王稳地,等. 传染病动力学的数学建模与研究[M]. 北京: 科学出版社,2006.

[15] 杨光. 传染病动力学模型及其控制[M]. 沈阳:辽宁大学出版社,2009.

[16] 雍龙泉,黎延海,贾伟. 正弦余弦算法的研究及应用综述[J]. 计算机工程与应用,2020,56(14):26-34.