Advances in Applied Mathematics 应用数学进展, 2022, 11(5), 2764-2774
Published Online May 2022 in Hans. http://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2022.115292

Hans汉斯

基于ARIMA模型研究舆情传播的特点和规律

——以微博平台早期数据为例

张　敏

上海工程技术大学，上海

收稿日期：2022年4月23日；录用日期：2022年5月17日；发布日期：2022年5月24日

摘　要

社交网络舆情通过网络平台传播，是网民对社会和生活中的一些热门话题，所持有的具有较高影响力、倾向性的看法与观点的集合。其中微博作为当今最热门的社交媒体之一，实时更新的热搜榜单成为人们的日常谈资。文章利用Python软件对爬取的部分微博文本数据进行数据预处理，并有针对性地筛选数据，提取特征字段信息，从中挖掘高价值的舆情主旋律，然后建立时序分析模型，对数据特征进行归纳总结，所得结果能够清晰地表明舆情传播有三阶段：产生－扩散－消减，从而挖掘舆情传播的特点与规律，合理抓住这些导向性内容的演变时段对于信息检索、舆情控制、影视宣传等都具有重要的意义和实用价值。

关键词

舆情传播，ARIMA模型，时序分析

Research on the Characteristics and Laws of Public Opinion Communication Based on ARIMA Model

—Taking the Early Data of Microblog Platform as an Example

Min Zhang

Shanghai University of Engineering Science, Shanghai

Received: Apr. 23rd, 2022; accepted: May 17th, 2022; published: May 24th, 2022

Abstract

Social network public opinion is spread through network platforms, and it is a collection of highly

文章引用: 张敏. 基于 ARIMA 模型研究舆情传播的特点和规律[J]. 应用数学进展, 2022, 11(5): 2764-2774.
DOI: 10.12677/aam.2022.115292

influential and tendentious views and opinions held by netizens on some hot topics in society and life. Microblog is one of the most popular social media nowadays, and the hot search list updated in real time has become people's daily conversation. In this paper, Python software is used to preprocess some crawled microblog text data, filter the data pertinently, extract feature field information, mine high-value public opinion themes from them, and then establish a time series analysis model to summarize the data features. The results can clearly show that there are three stages of public opinion communication: generation-diffusion-reduction, so as to mine the characteristics and laws of public opinion transmission. Therefore, it is of great significance and practical value to excavate the characteristics and rules of public opinion communication and reasonably grasp the evolution period of these guiding contents for information retrieval, public opinion control and film and television propaganda.

Keywords

Public Opinion Communication, ARIMA Model, Time Series Analysis

1. 引言

网络舆情是在传统媒体舆论传播的基础上，以网络为载体，通过互联网的方式将舆论放大，来表达网民的观点、态度和情感[1]。网络舆情是对社会舆情的反映，也是公众舆论在互联网上的映射[2]。随着计算机和网络技术的快速发展，互联网日渐成为各种信息传递的载体。人们在上面主动的获取、发布、共享、传播各种观点性信息(包括新闻评论、产品评论、情感微博、网络社区等)。据 2021 年微博第二季度财报显示，截至 2021 年 6 月，微博月活跃用户规模已升至 5.66 亿，日活跃用户达到 2.46 亿。微博逐渐变成网络舆情的重要诞生地和发酵场。网络舆情信息挖掘是指借助相关信息处理技术，识别、提取舆情文本中的热点词汇、倾向性词汇，为舆情引导工作提供依据的信息挖掘活动[3]。

文章选取微博 2009 年正式上线至 2012 年这四年的数据研究舆情传播的时序分析问题，对此问题构建模型开展研究，分析受众面窄时网络舆情传播的特点，为之后新平台的开发提供一定的理论基础。通常解决时序分析问题使用的时间序列模型有：自回归模型 AR、移动平均模型 MA、自回归移动平均模型 ARMA、自回归差分移动平均模型 ARIMA [4]，这些模型均使用了 Box-Jenkins 算法，而其中的 ARIMA 模型只依赖数据本身，但需要时序数据具有稳定性，所以对于稳定的时序数据我们选择 ARIMA 模型来分析问题。

2. 数据预处理

现有脱敏处理的微博记录 22+万条，分别放在 22 个 TXT 文件中，我们将其整合进一个 CSV 文件，并按不同问题所列要求进行数据处理，构建模型并分析所得结果(见图 1)，从而更好的挖掘舆情传播的特点和规律，为研究舆情热点传播提供一定的思路。

2.1. 数据提取

首先对提供的微博记录共 22 个数据文件进行数据提取，每条数据有 person_id (所属人物的 id)、id (文章编号)、article (正文)、discuss (评论数目)、insertTime (插入时间)、origin (来源)、time (正文发布时间)、

transmit (转发)共 8 个属性(见表 1)，得到数据总量为 227,566 条。

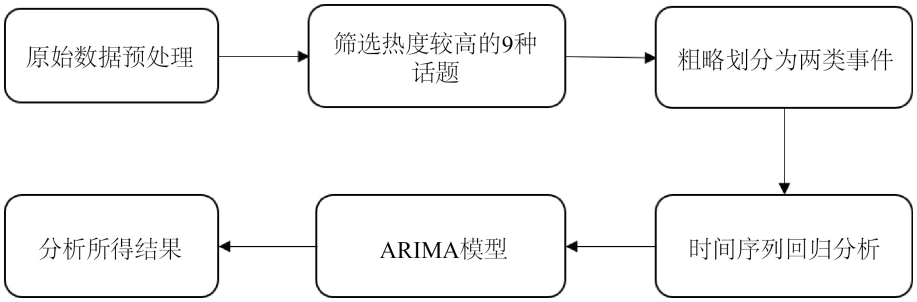

Figure 1. Problem solving flow chart
图 1. 解题流程图

Table 1. Preliminary extracted data (part)
表 1. 初步提取出的数据(部分)

person_id	id	article	discuss	insertTime	origin	time	transmit
1646051850	35761	昨天上午……	0	2011/11/29 14:43:04	新浪微博	2011/9/17 9:20:24	0
1646051850	35760	京城第一贪……	1	2011/11/29 14:43:04	新浪微博	2011/9/17 9:22:50	0
1646051850	35759	北京时间15 日……	0	2011/11/29 14:43:04	新浪微博	2011/9/17 9:25:16	0
……	……	……	……	……	……	……	……

2.2. 数据统计

统计每个 person_id 的数据量，按数据量由大到小进行排序(表 2)。其中 person_id 为 1646051850 的用户发布量最高有 12,029 条发布量。总共有 44,177 条数据代表 44,177 个 person_id。

Table 2. person_id and its release statistics (part)
表 2. person_id 及其发布量统计(部分)

person_id	发布量
1646051850	12,029
1288915263	9977
0	8078
1644114654	6255
……	……

2.3. 数据去重

通过观察发现，同一个 person_id 存在多条属性相同的数据，将 article、discuss、origin、time、transmit 相同，insertTime、id 不同的数据认为是重复数据(表 3)，该重复的数据量较大，故需对数据进行了去重处理。去重后共有 124,137 条数据(表 4)。

Table 3. Duplicate data (part)
表 3. 重复数据(部分)

person_id	id	article	discuss	insertTime	origin	time	transmit
1646051850	236814	毋庸讳言，作为……	2	2012/1/2 14:27:39	新浪微博	2011/9/27 12:50:24	0
1646051850	150380	毋庸讳言，作为……	2	2012/1/1 18:39:49	新浪微博	2011/9/27 12:50:24	0
1646051850	35501	毋庸讳言，作为……	2	2011/11/29 14:42:38	新浪微博	2011/9/27 12:50:24	0
……	……	……	……	……	……	……	……

再次按每个人发布的条数重新进行统计排序，其中 person_id 为 0 的用户发布量最高，发布了 6571 条数据，person_id 为 1641561812、1644114654、1646051850 的用户依次有 2970、2527、2224 条发布量。统计共有 44,177 个 person_id。

Table 4. Person_id and its publication statistics after the recallation (part)
表 4. 去重后的 person_id 及其发布量统计(部分)

person_id	发布量
0	6571
1641561812	2970
1644114654	2527
1288915263	2075
……	……

2.4. 数据预处理小结

1) 处理前共有数据 227,566 条，去重后数据共有 124,137 条。

2) 数据中共有 44,177 名用户(person_id 为 0 的视为一名用户)。

3) 其中共有 171 名用户发布量大于 10 条，共有 118 名用户发布量大于 100 条，共有 17 名用户发布量大于 1000 条。

4) 使用时间进行排序可知，最早数据发布时间为 2009/08/26，最晚数据发布时间为 2012/02/09。其中 2011/03/11 之前数据量仅为 5478 条。

5) 其中，在 person_id 为 0 对应的数据中，discuss、transmit 属性没有内容，猜测其为匿名用户发布，但仍将其视为一个用户。

以下所有数据基于本章预处理后的数据。

2.5. 数据处理

为了表示热点随时间的演化过程，我们将词频较高的关键词及其类别筛选出来并抽取其中 9 种，并按照热度－时间的关系作图，见图 2。可知在 2009 年 08 月 26 日至 2011 年 02 月 20 日期间，微博纪录的信息近似于零，那时微博这个社交平台才刚推出，大家对它的认知度和熟悉度不高，因而历史纪录不多。2011 年 02 月 20 日之后，随着一些事件的发生，人们发现在互联网上发声更有利于事件的传播和解决，微博纪录的数据逐渐增多。

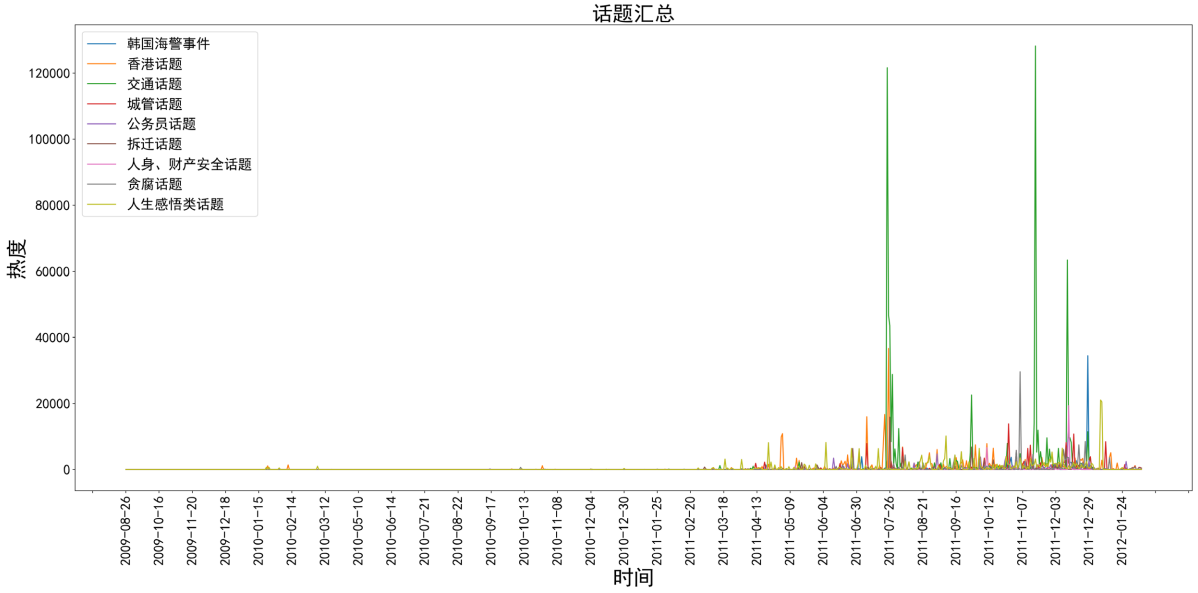

Figure 2. Summary of 9 topics
图 2. 9 种话题汇总

通过查阅文献我们学习到，热点事件在社会化媒体中的话题演变有 3 级：话题传播初期酝酿阶段；社会网络的关键传播阶段；网络媒体传播的协同阶段[5]。在图 2 所示的 9 种热点话题中，有持续性事件和突发性事件，这边我们取香港话题与交通话题进行话题传播分析。

首先，2011 年 03 月 01 日至 2012 年 02 月 05 日期间，香港问题热度一直在线(见图 3，图 4)，我们认为这个话题是一个持续性事件。在图 4 中，我们发现加了“内地”“大陆”这两个关键词后，整体数据走向只有小小的波动，此类政治性话题的热度不会随着时间而冷却。

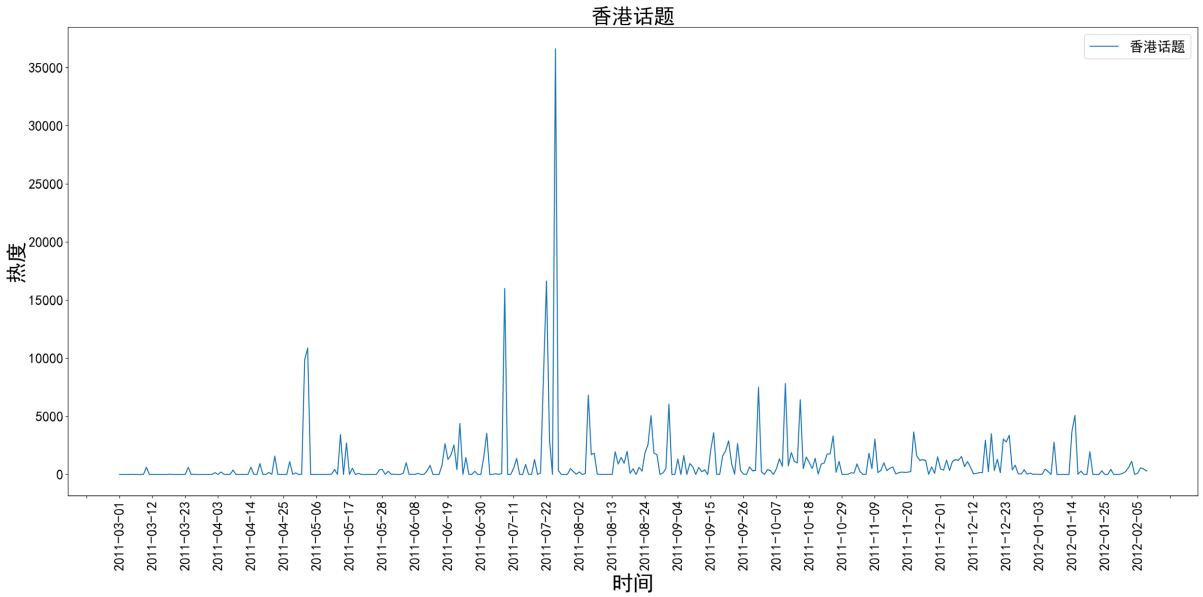

Figure 3. Hong Kong topic
图 3. 香港话题

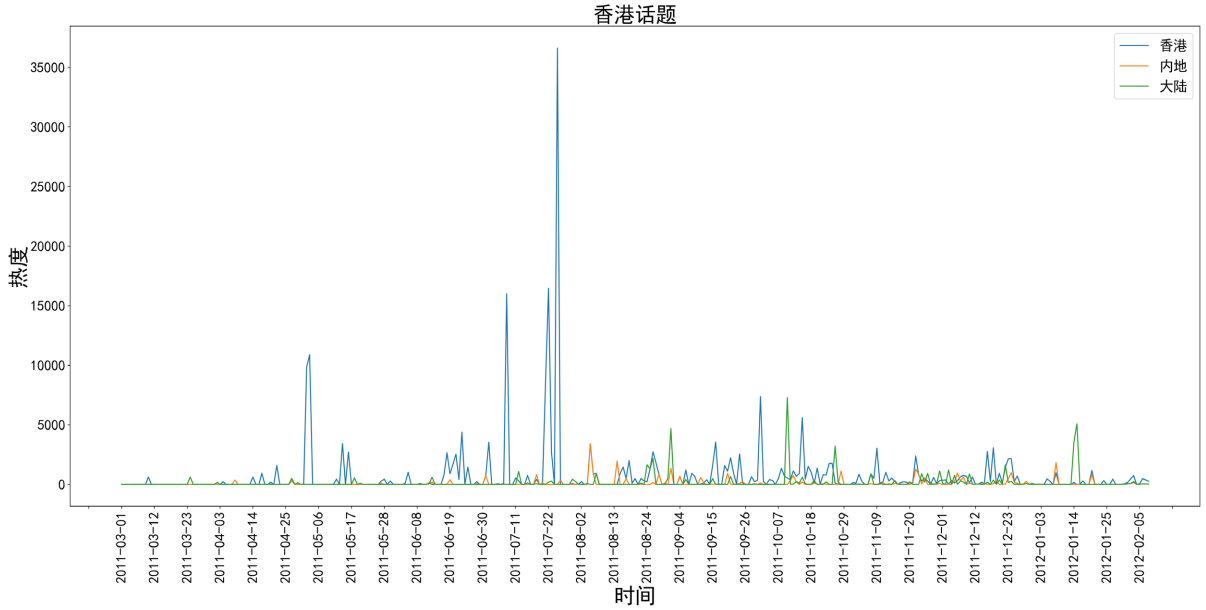

Figure 4. Hong Kong topics + keywords
图 4. 香港话题 + 关键词

其次，2011 年 7 月 25 日，交通话题(见图 5，图 6)突然被网民热议，虽然之后也有过几次数据大爆发，不过随着时间的推移它的热度最后会趋于 0，我们便将这个话题归为突发性事件。联系实际，交通安全类话题往往是在事件出现后，鉴于不同的成长背景及生活环境，来自四面八方的人们会激烈的表达自己的观点，在互联网的推动下，事件不断发酵至爆发，等热度一过，大家又回归平和的生活，直到同类事件的发生。

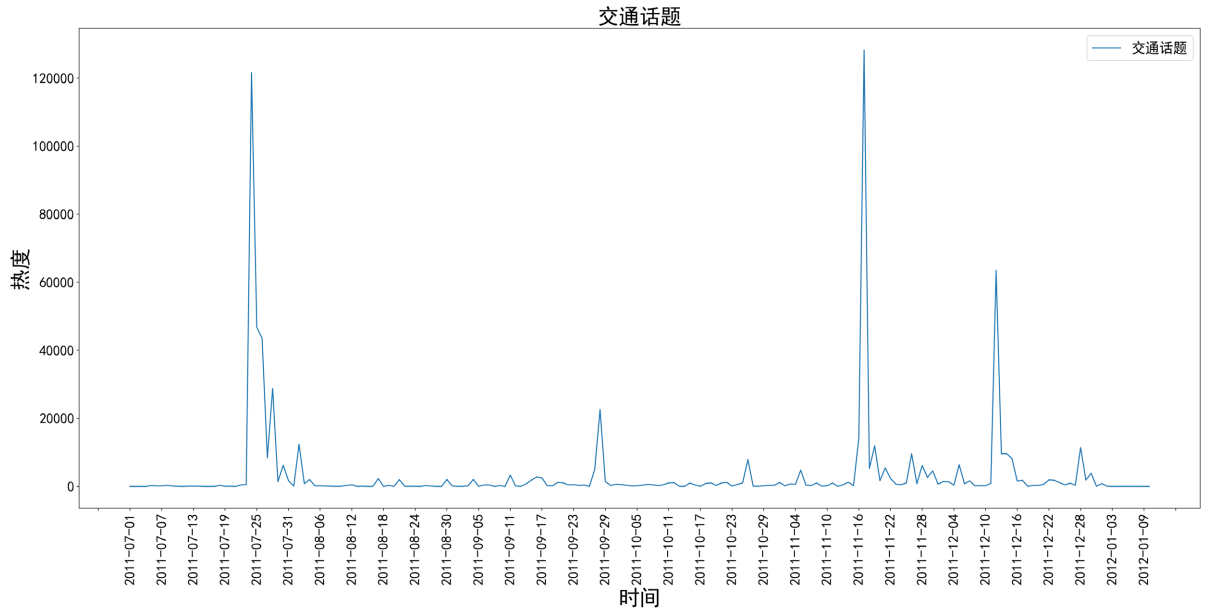

Figure 5. Traffic topic
图 5. 交通话题

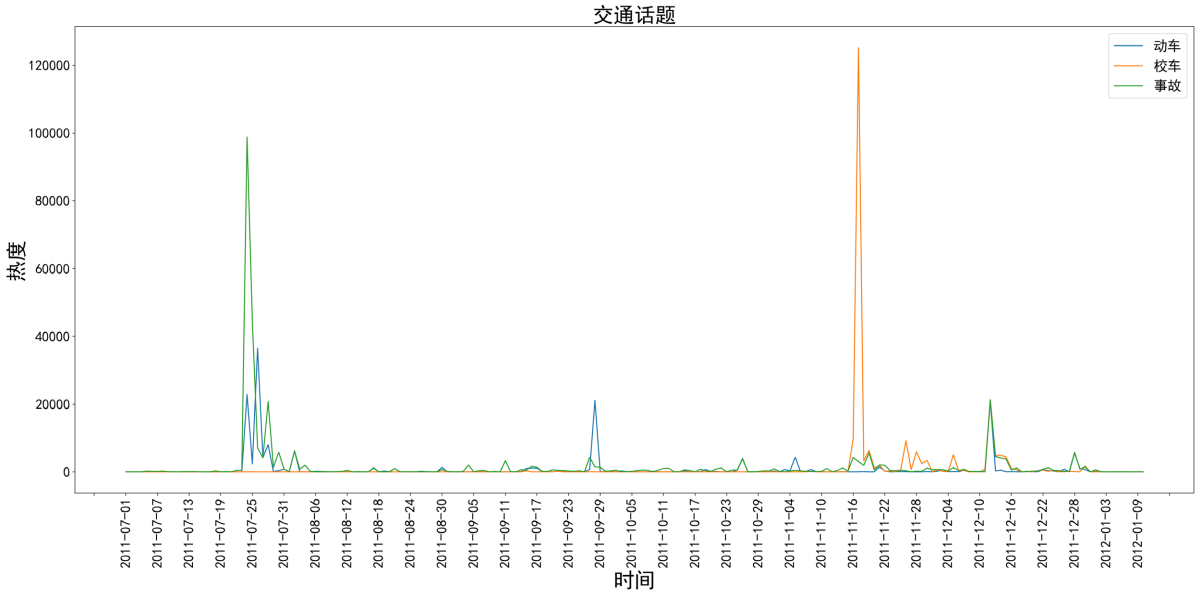

Figure 6. Traffic topics + keywords
图 6. 交通话题 + 关键词

因此，通过简单的数据预处理，我们了解到，对于不同性质的事件，舆情传播的时效性也不同。持续性事件因为事件发生的频繁，舆情传播的战线拉长，传播面广，不太容易热度骤冷；而爆发性事件由于事件本身性质的恶劣和突然，舆情传播的速度特别快，从而战线较短，“速战速决”。

从所给数据来看，舆情传播的力量是无穷的，如果想要深入剖析舆情传播的规律还需要预测未来的数据。因而我们要借助数学模型来具体分析，同时微博数据存在缺失，缺失的数据不会对时序分析造成影响，因而假定以某一热点话题为例推出舆情传播的特点。

3. ARIMA 模型

时间序列预测是通过观察分析历史数据来预测未来的值。ARIMA 模型是时间序列预测分析方法之一，可以较好地预测以时间为基准的数据。它的基本思想是将预测对象随时间推移而形成的数据序列视为一个随机序列，用一定的数学模型来近似描述这个序列[6]。

3.1. 公式

查阅资料，我们了解到 ARIMA 模型包含 3 个部分，即自回归(AR)、差分(I)和移动平均(MA)。ARIMA 模型记作 ARIMA(p, d, q)，其中 AR 是“自回归”，p 为自回归项数；MA 为“滑动平均”，q 为滑动平均项数，d 为使之成为平稳序列所做的差分次数(阶数)，L 是滞后算子(Lag operator)。“差分”一词虽未出现在 ARIMA 的英文名称中，却是关键步骤。

ARIMA(p, d, q)模型是 ARMA(p, q)模型的扩展[7]。ARIMA(p, d, q)模型可以表示为：

$$\left(1-\sum_{i=1}^{p}\phi_i L^i\right)(1-L)^d X_t=\left(1+\sum_{i=1}^{p}\theta_i L^i\theta_i L^i\right)\varepsilon_t \tag{1}$$

$$d\in Z,\ d>0$$

3.2. 参数的确定

1) 参数 p、q 的确认

在时间序列分析中，通常采用自相关函数(ACF)、偏自相关函数(PACF)来判定 ARMA(p,q)模型的系数和阶数。

自相关函数(ACF)描述时间序列观测值与其过去的观测值之间的线性相关性。偏自相关函数(PACF)描述在给定中间观测值的条件下时间序列观测值与其过去的观测值之间的线性相关性，简介见表 5。

Table 5. Characteristic coefficient of stationary random time series model
表 5. 平稳随机时间序列模型特征系数

平稳随机时间序列模型	自相关系数	偏自相关系数
AR(p)	拖尾	P 阶截尾
MA(q)	Q 阶截尾	拖尾
ARMA(p, q)	拖尾	拖尾

p 由显著不为 0 的偏自相关系数的数目决定，此时序列的偏自相关函数表现为拖尾性，即当 $k > p$ 时，偏自相关系数的值都在置信区间以内；

q 由显著不为 0 的自相关系数的数目决定，此时序列的自相关函数表现为截尾性，即当 $k > p$ 时，自相关系数的值都在置信区间内。

这里的拖尾是指以指数率单调或振荡衰减，截尾是指从某个开始非常小(不显著非零)。

2) 参数 d 的确认

差分是求时间序列{rt}在 t 时刻和 $t-1$ 时刻的差值，把 rt 与 $t-1$ 时刻的值 $r(t-1)$的差值记做 dt，则得到了一个新序列{dt}，为一阶差分；对新序列{dt}再做同样的操作，则为二阶差分。

3.3. 模型求解

ARIMA 模型解题具体分析步骤如下：

Step 1：求出数据序列的样本自相关系数(ACF)和样本偏自相关系数(PACF)的值。

Step 2：根据样本自相关系数和偏自相关系数的性质，选择适当的 ARMA(p,q)模型进行拟合。

Step 3：估计模型中位置参数的值。

Step 4：检验模型的有效性。如果模型不通过检验，转向 Step 2，重新选择模型再拟合。

Step 5：模型优化。如果拟合模型通过检验，仍然转向 Step 2，充分考虑各种情况，建立多个拟合模型，从所有通过检验的拟合模型中选择最优模型。

以香港问题为例，选取“香港”、“内地”、“大陆”这三个关键词抽取数据作图(见图 7~12)。

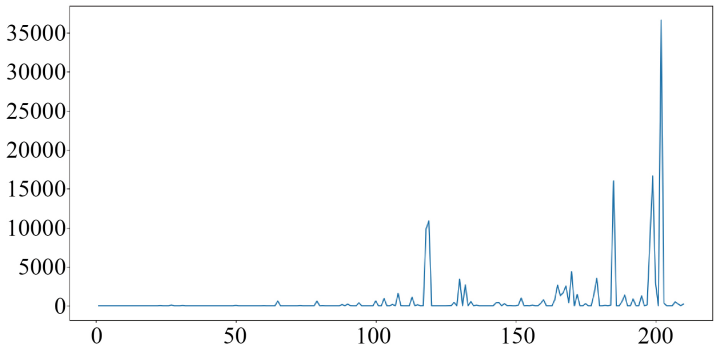

Figure 7. Visualization diagram
图 7. 可视化图

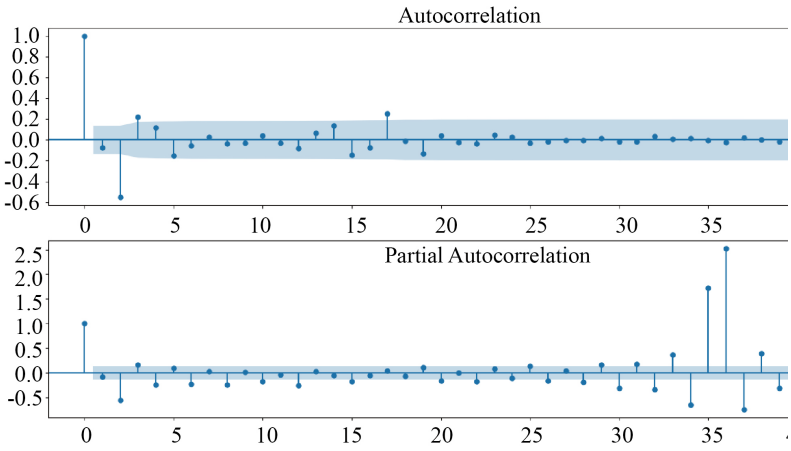

Figure 8. Autocorrelation diagram (ACF diagram) and partial autocorrelation diagram (PACF diagram)
图 8. 自相关图(ACF 图)和偏自相关图(PACF 图)

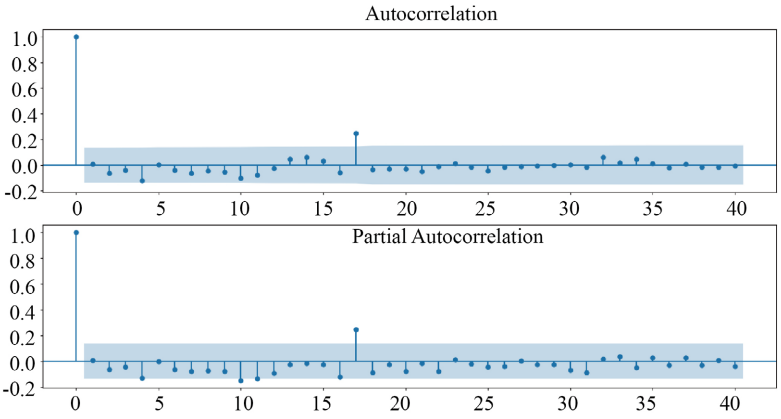

Figure 9. Autocorrelation and partial autocorrelation diagrams of generated residuals
图 9. 产生的残差的自相关和偏自相关图

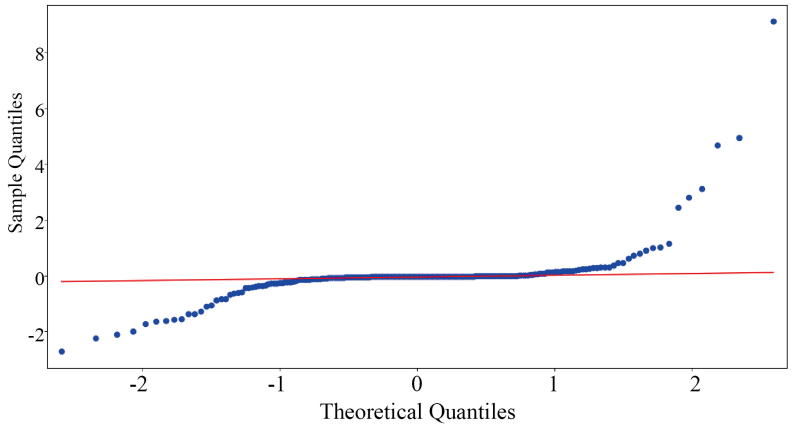

Figure 10. Scatter diagram of normal distribution
图 10. 正态分布散点图

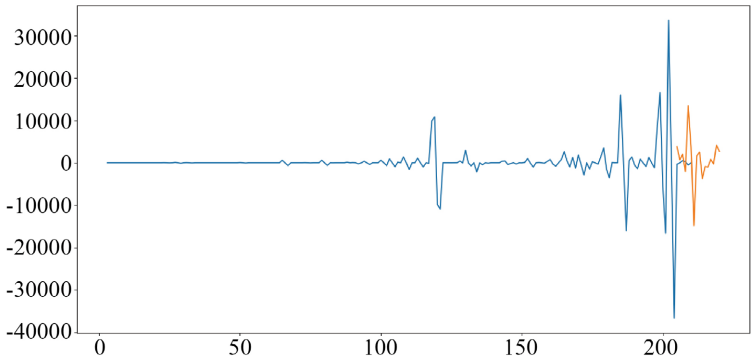

Figure 11. Prediction sequence diagram of generated stationary series
图 11. 生成的平稳序列的预测时序图

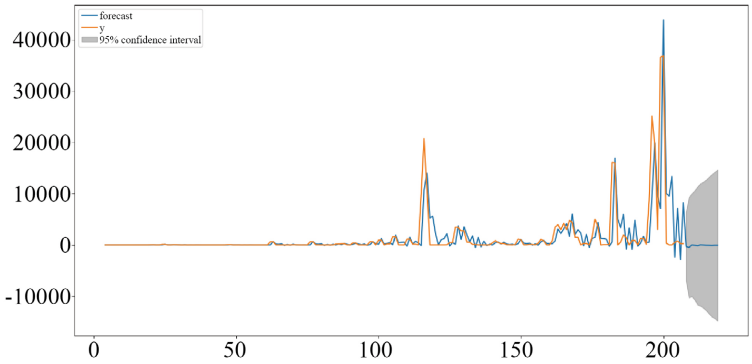

Figure 12. Timing prediction diagram of original data
图 12. 原始数据的时序预测图

从图 11 和图 12 可知，预测数据跟原始序列趋势基本一致，即在之后香港话题仍会是热点事件话题。

3.4. 小结

从所得结果来看，舆情的生命周期为产生－扩散－消减，因事件性质的不同，舆情生命周期的长短也不同。持续性事件舆情存活时间长，爆发性事件舆情存活时间短，但是它们都会在爆发期将舆情推到顶峰。信息总是在不断积累的过程中，从源头上讲，传统的新闻媒体在网络新闻平台和新媒体平台上采用不同的信息发布方式，个人用户的信息发布更加灵活。

分析不同层次的网络平台，关于各类事件的发酵与传播有着不同的特点，比如，作为一种“短频快”的媒体，微博日活用户量日益增多，它可轻易将两个甚至多个平台无缝连接，由此让事件迅速传播。而对于一些受众面较窄的网络平台，事件的传播时间较长，传播面也较窄，但是针对性可能较强。因而，通过研究舆情传播的规律和特点，新闻类机构可合理借助这些平台，更好地控制和指导非常规突发事件网络舆论的传播，正确引导社会的舆论风向。

4. 模型评价与改进

4.1. 模型优缺点

4.1.1. 优点

文章选用的 ARIMA 模型，它是时间序列预测分析方法之一，模型构建简单，只需要内生变量而不需要借助其他外生变量。

4.1.2. 缺点

文章选用 ARIMA 模型，它要求时序数据是稳定的，或者通过差分化之后是稳定的；本质上只能捕捉线性关系，不能捕捉非线性关系。因此我们实验结果的准确度有待提高。

4.2. 模型的改进与推广

可增加灰色关联度分析，得出各个属性与时间变化的关系，增加组合优化模型，防止模型过拟合。后续研究可以选择其它受众面广的平台，选取一定量的有效数据开展比较分析，为舆情传播的研究提供强有力的依据。

参考文献

[1] 朱毅华, 张超群. 基于影响模型的网络舆情演化与传播仿真研究[J]. 情报杂志, 2015, 34(2): 28-36.

[2] 田占伟, 隋玚. 基于复杂网络理论的微博信息传播实证分析[J]. 图书情报工作, 2012, 56(8): 42-46.

[3] 刘娟, 郝云强, 尹雪雪. 网络舆情信息挖掘关键技术分析[J]. 信息记录材料, 2021, 22(3): 94-95.

[4] 赵嘉宝, 陈杰, 安霞, 孙占海, 张学东. 基于 ARIMA 模型的吐鲁番市葡萄产量预测分析[J]. 江苏科技信息, 2019, 36(31): 34-39.

[5] 王奕文, 刘昕, 曹帅, 王丰. 基于关联规则的热点事件时序分析方法[J]. 计算机与现代化, 2018(8): 108-113.

[6] 李潇瀛, 方鸽, 李昌均. 基于 FCM-ARIMA 的多阶段退化设备寿命预测研究[J]. 计算机仿真, 2021, 38(8): 33-36+74.

[7] 郑永坤, 刘春. 基于 ARIMA 模型的二手房价格预测[J]. 计算机与现代化, 2018(4): 122-126.