题 目  **一种面向微博主题的自动画像系统**

# 摘 要

摘 要：智能手机等终端设备带给人们便捷性的同时，也无时无刻将人们的想法遍及整个互联网，会造成信息的自我公开，随着互联网用户普及率即将到达100%，人们的想法也随之拥有潜在的公开透明，当你的想法表达到互联网上，可能你喜欢的产品不一定出现在你的礼物中，但会出现在你的个性化广告中，你会拥有一台最懂你的智能手机。这些都是画像存在的结果。

微博主题画像是有关部门在监测网络舆情过程中根据已获得微博信息数据对某些特别关注的主题进行描述和刻画，类比于公安机关的侦察画像可了解到，侦察画像是公安机关寻找犯罪嫌疑人的有力法宝，以能够在现实空间中轻而易举地抓获准确的犯罪嫌疑人；而在互联网空间中，网络不是法外之地，一切违反法律的言论甚至谣言都应受到应有的处罚，于是我们很有必要采取措施去监控网络动态，及时捕捉微博敏感数据，及时处理数据并描绘出可疑分子，并及时的分析研判，最后达到发现可疑事件的微博主题发出预警。在中国知网，以“微博画像”为主题进行搜索可获得相关文献255篇，发表时间为2011年到2022年，梳理这些文献可以发现微博画像技术的研究主题内容及发展阶段。通过检索可以发现，大部分研究方向都偏向于微博用户画像的研究，而对于微博主题画像的研究则少之又少，所以特别需要对微博主题进行研究，通过整体全面地从大局上描绘出整个主题的画像，作为一种面向主题的模型，给主题加上标签，通过可视化形成画像，从中发现可疑点。“神笔警探”林宇辉，专注模拟画像，侦破无数大案，甚至在没有任何目击者的情况下，根据已有信息描绘极为相似的嫌疑人画像。互联网上可利用的数据更多，需要我们去挖掘利用。本课题以此为背景，本人结合专业所学知识，对微博主题画像进行研究，通过软件系统的形式实现研究目的，对我国当前互联网舆情治理具有重要的理论意义和现实意义。

微博作为中国最具代表性的社交媒体平台之一，其用户数量和影响力不断增长。微博画像作为微博用户的头像，成为了用户在微博上展现自己的重要方式之一。本文以微博画像为研究对象，从微博画像的发展历程、微博画像的功能与意义、微博画像的设计与表现等方面进行了探讨。通过对微博画像的研究，本文认为微博画像不仅仅是用户在微博上的头像，而是一种表达个性、传递信息的载体，具有重要的社交功能和文化意义。

微博主题画像系统是一种能够根据微博内容和用户信息，对用户进行自动分类和画像的系统。本文提出了一种基于机器学习和自然语言处理技术的微博主题画像系统设计，能够对微博用户进行分析，识别其主题兴趣和社交特征，并将其归类为不同的用户画像。本系统包括数据采集、数据预处理、特征提取、分类建模、用户画像生成和可视化展示等多个模块。实验结果表明，本系统能够有效地对微博用户进行分类和画像，为用户提供更好的社交体验和推荐服务。

关键词：信息传播、节点行为、舆情博弈、演化模型

The research of virtual currency

**Abstract:** With the development of network technology, microblog, WeChat, Zhihu and other social platforms have become a place for netizens to express their opinions and feelings on the hot issues in real life. Prediction of public opinion on the Internet plays an important role in supervision. Understanding and analyzing the evolution of public opinion events on the Internet can find potential risks in the development of public opinion and avoid misstatements to guide public opinion. In this paper, from the perspective of public opinion game, evolutionary game dynamics is applied to the specific behavior network analysis. Through the analysis of the existing network members' behavior, the evolution model of public opinion game is constructed, so as to accurately predict the evolution direction of network public opinion and realize the systematic analysis of the propagation trend.

In the aspect of research content, from the perspective of public opinion game, this paper applies evolutionary game dynamics to the specific behavior network analysis, and constructs the evolution model of public opinion game through the behavior analysis of existing network members, so as to accurately predict the evolution direction of network public opinion and realize the systematic analysis of communication trend. The main work of this paper is divided into the following parts: introduction of the current situation of public opinion network development, comparison and construction of public opinion game evolution model, verification and analysis of public opinion game evolution model. In the comparative construction of the evolution model of public opinion game, the two modeling ideas of replication mutation dynamics and selection drift dynamics are analyzed in detail, and how to collect and clean the information dissemination data and several typical evolution modes (behavior avalanche, behavior aggregation, behavior shaking) are introduced. In the process of verification and analysis, the data are carried out from two aspects of the whole individual Visualization makes the expression more diverse and the result clearer.

This paper obtains a large amount of social network information data, and evaluates and analyzes the results of different public opinion game evolution models under the current data set, and finally establishes a public opinion game evolution model suitable for information dissemination networks, and finally realizes public opinion prediction in social networks. Detecting and regulating the evolution of trends and trends in social groups provides new perspectives.

**Key words:** Information dissemination, public opinion game, evolution model

# 目 录

1. [引言 1](#_bookmark0)
2. [博弈演化的发展2](#_bookmark1)
   1. [元胞自动机模型2](#_bookmark2)
   2. [多数决定模型2](#_bookmark3)
   3. [ISING 模型 3](#_bookmark4)
   4. [有界信任模型3](#_bookmark5)
   5. [SZNAJD 模型 4](#_bookmark6)
   6. [总结 4](#_bookmark7)
3. [博弈演化模型构建6](#_bookmark8)
   1. [行为网络博弈演化的概念6](#_bookmark9)
   2. [模型相关概念的介绍7](#_bookmark10)
   3. [复制-突变动力学模型8](#_bookmark11)
      1. [复制动力学 8](#_bookmark12)
      2. [突变动力学 9](#_bookmark13)
      3. [复制突变动力学 10](#_bookmark14)
   4. [选择-飘移动力学模型11](#_bookmark15)
   5. [模型比较与选取11](#_bookmark16)
   6. [选择漂移模型中的典型突变网络14](#_bookmark17)
      1. [复制突变动力学 15](#_bookmark18)
      2. [复制突变动力学 15](#_bookmark19)
      3. [复制突变动力学 16](#_bookmark20)
4. [博弈演化模型验证分析17](#_bookmark21)
   1. [数据准备17](#_bookmark22)
      1. [数据的选择 17](#_bookmark23)
      2. [数据获取及数据爬虫实现 17](#_bookmark24)
      3. [数据清洗 18](#_bookmark25)
   2. [模型验证 21](#_bookmark26)
   3. [个体行为展示22](#_bookmark27)
5. [结论27](#_bookmark28)

致 谢 **错误！未定义书签。**

参考文献： **错误！未定义书签。**

## 引言

网络舆情是网民将某些社会现象或者一些生活中存在的问题的一些看法在社交网络上公开表达，当这些观点达到一定数目形成一个共同的观点，并且产生了较为深远的影响，如果不加以控制，会对我们的正常生活以及社会秩序带严重后果，在大数据环境下舆情信息的监管及对舆论的正确引导都是非常重要的，保护公民的网络安全，避免谣言及虚假信息的传播，是警察的使命与义务。网络舆情可以认为是一个集合，这个集合具体包括在互联网上进行传播的人们对于某一事件的所有认知、态度、情感和行为倾向。我们生活在一个突发事件频发，消息共享速度极快的大数据网络时代下，突发事件是指突然发生，造成或可能造成严重的社会危害，需要采取应急处置措施予以应对的公共事件。突发事件的网络舆情演化是指“没有对事件源和次发生事件进行人工应急干扰的原扩散路径”。突发事件爆发后，一定会引起网络舆情发生，在各种网络平台上信息快速传播和蔓延下，政府部门要发挥常规的传播模式的效力，还要积极有效的给予矫正型的传播影响作用，才能更好有效地控制网络舆情发酵。为此，必须要做到日常网络舆情的预警、完善舆情监管监督机制、加强舆情的导控。

当前网络舆情的研究主要就是用经典理论为基础然后用仿真软件进行仿真，然后用现 有公开数据集验证模型的有效性。尽管现在大型网站都有数据库但是数据集的获得缺很难，所以未来的研究应该借助信息挖掘的先进技术获得大量实例然后再和理论模型来相结合。 其次现有的理论模型还是相对单一，都是单一的从社会学、心理学以及管理科学来研究， 舆情演化注定是个跨学科的东西，需要这些学科已有机的结合在一起才能更好的研究舆情 演化。最后舆情研究的目的是为了控制舆情，我们不仅仅应该将研究停留在传播模型的构 建，对于舆情的预测以及引导有一个更深入的研究。

本文从一个全新的角度，以博弈论相关知识为基础进行相关研究，构建出合适于舆情演化的博弈模型。

## 博弈演化的发展

在过去的数年中，国内外的学者将社会学、社会心理学、计算机科学以及统计物理和管理学等学科知识用于舆论博弈演化研究，对各节点之间的交互以及各种观点形成的研究都取得了显著成效，提出了很多经典的网络舆情演化的理论模型，为之后的研究奠定了理论基础。目前现有的复杂网络上的舆情演化模型很大一部分都是在这些理论基础之上形成的，在此对当前主流的几大舆情研究模型进行简略介绍。

### 元胞自动机模型

基元胞自动机（ CA）， 这是一个运用离散数学相关知识建立的数学模型，最先运用于生物繁殖的相关研究中，其根据既定的繁殖规则然后对每个元胞个体进行赋值，从而决定其状态变化，实现对生物繁殖的研究。科学家 Sakoda 首次使用元胞自动机模型对二战之后的美国的一个搬迁中心进行研究，并分析了其疏散过程，此后此模型就广泛的被运用于社会科学领域。后来元胞自动机又被应用于投票模拟中，将选民的投票的结果与新闻媒体的报道和选举相关知识的更新结合起来，从而预测选民的最终意向。此外其在社会科学方面还有其他具体应用，不在进行赘述。戴建华等人[9]将个人观点倾向力和顽固力这两个模糊规则和元胞自动机进行了结合，从而建立起了模糊规则下网络舆情的传播模型。

元胞自动机所建立的模型， 其最大的特点就是个体的态度在结果中很明确，并且其人物之间的交互规则相对简单，所以易于上手使用，从而运用较多。但是也正是因为其规则的简单，使得它在当今的复杂的社会网络中难以适用，在深入

探讨复杂系统个体之间交互过程中捉襟见肘，难以胜任。

### 多数决定模型

多数决定模型（ Majority Rule Model）， 顾名思义， 该模型的核心理论就是人们常说的少数服从多数。在每次时间节点推进之前，当前个体会将邻居节点中占比最大的意见作为自己的观点发出，这是对网民浏览事件发表言论时的从众心理进行的模仿，即某个观点的可能性与其邻居中保持该观点的人数是成正比的。此模型经常会被用于研究投票选择或者观点辩论。在所有的多数决定模型之中，最出名的模型就是 Galam 模型。Galam[ 10 ]使用 Ising 模型对选举的过程进行了

模拟，在这个模拟过程中暗含一个前提，就是网络的初始的状态对最终的各观点的演化方向具有很大的影响。之后 Stauffer 把多数决定模型的传播规则改为点阵扩散，从而建立了可用于预测网络舆论的观点演化模型。陈福集等人从观点的动态变化出发将该模型从而用于舆情演化机制的研究。

Galam 模型可以很直观地看出来复杂网络的具体演化过程，并且可以通过对评论的控制从而引导网络中个体的偏好选择从而控制舆情走向，使其往好的我们期望的方向发展变化。但是由于现实生活中人们的知识水平性格习惯存在很大的差异，层次不齐，所以此模型仅仅是提供了相应的理论支撑，但如果将其运用于

实际的现实网络舆情演化预测的过程中，是很难实现的。

### Ising 模型

Ising 模型，又被称为伊辛模型，其最早被运用于物理学领域，以此模型中的+1 或-1 来实现对磁铁两极的表示，从而研究铁磁中物质相变的过程。因为在舆情传播中，我们的个体观点一般情况或者相对而言也是二值观点，可以用+1来表示支持，-1 来表示反对，并且观点的取值会受到其相邻节点影响，所以 Ising 模型也逐步被运用于观点传播之中。此模型曾被用于]研究小世界网络中的温度相变，帮助人们发现了 BA 无标度网络上的观点传播的时间序列与高斯分布非常吻合。Grabowski 通过把网民的空间位置以及其社交关系加入的模型的定义之中，再结合 BA 无标度网络结构，从而构建了出了可用于网络舆情模拟的意见形成模型。该模型将个体的权威性与此个体的链接数，即影响力联系在了一起，是对邻节点的选择更加合理。

虽然 Ising 模型因为影响力等因素的量化，所以应用起来有很大的局限性，但是它是真正意义上舆情演化模型的开端，对之后的舆情演化领域的研究模拟有着很大的促进推动作用。之后的 Aznajd 以及 Galam 模型都是在该模型的理论基础上修改提升而来的。

### 有界信任模型

有界信任模型（Bounded Confidence Model），此模型是将观点值刻画成了一个连续的区间。因为现实生活中我们的观点不可能仅仅是赞成或者反对这样的极端情况，反而可能会有很多其他可能的选择。从社会网络中网民作为节点交互的实际情况来看，只有属性类似的网民才会产生更多的交互，因此引入了一个连续态度模型。这类模型最为 HK 模型和 D 模型，此二者都将节点的观点抽象成

了区间上连续的实数，然后设定了观点阈值，通过邻居观点的值来和阈值进行比较，如果大于阈值则按照规则交互，但若是低于阈值则不进行交互。

但是在现实情况中，网络信息是非常庞大的，如此庞大的互联网社区显然和难实现对每一个节点观点的逐一遍历，这也就导致很难与其他用户之间的观点进行交互，或者模型运行速度极慢，甚至低于舆情发展本身的速度。

### Sznajd 模型

Sznajd 模型，此模型是一个观点动力学模型，因其最初提升就是以观点作为研究对象，所以其对社会网络领域做出相当大的贡献。Sznajd 模型认为个体观点取值只需要定义两个即可并且且每个节点在实际中只会和邻居进行交互，在交互的过程中个体可能会对邻居的邻居的策略复制。所以此模型的基本更新规则就是个体节点的观点只会受到网络中邻节点的影响而且如果邻节点中观点相同的人更多，则该节点更容易选择其此观点。Bonnekoh 使用将 Sznajd 模型用于了舆 情演化实践中，并在 BA 无标度网络的基础上建立了网络舆情演化模型，并认为在舆情机制开始之后，再新加入的到网络中的节点对网络舆情的走势不会产生较大的影响。

虽然 Sznajd 模型在舆情传播领域已经应用颇多，但是此模型考虑的社会网

络特性并不多。所以以其为基础建立的很多模型也都仅仅是建立在其理论基础上，以仿真数据进行验证，并没有再实际中检验过。

### 总结

通过对网络文献的梳理，我们不难发现复杂网络上的舆情演化模型可以分为两大类，一是离散的观点模型，也就是把所有的观点都限制在二值观点内，要么赞成要么反对。这种模型可以很好地解释二进制观点可以描述的问题，而且个体也没必要了解以前的观点。但是社会网络中观点的表达并非如此简单。因为人们的观点可能在两者之间，有很多种程度，所以出现了第二类观点演化模型，也就是连续观点模型。二者都为观点传播学提供了很好的理论依据，都产生了很多经典的模型。

当前网络舆情的研究主要就是用经典理论为基础然后用仿真软件进行仿真，很少用现实的舆情事件的数据集来验证模型的有效性，当前大型社交网络非常多但是数据集的获得确实相当的困难，所以以后的研究应该更多的建立在真实的数

据集上借助数据挖掘技术获得大量实例然后再和理论相结合。其次网络舆情的研究注定不是单一的学科研究， 我们必须将社会学、心理学、物理科学、管理科学等学科有机的结合起来才能很好地研究社会网络这个复杂系统。最好在有了很多的传播规律的基础上我们更应该对于舆情的预测以及引导有一个更深的研究。

## 博弈演化模型构建

### 行为网络博弈演化的概念

在分析具体模型之前，先对行为网络的博弈演化进行简单介绍，行为网络从博弈论 的角度来看，可以当作一个网络博弈，其中每个行为对应一个策略，而效用矩阵对应于收 益矩阵．行为网络的博弈描述为刻画和分析群体交互行为选择的复杂模式提供了一个有力 的工具，实际上，行为网络此前就已经被广泛用于描述个体在各类情境下的行为交互，包 括金融环境下的商品选择和社会文化演化中的语法、习惯选择等，此外，博弈论中解的概 念以及分析工具也为解释和预测个体的行为决策过程提供了一个有效的方法。本文挑选了 三种典型的行为网络进行介绍，以便大家更好的理解行为网络博弈演化的概念，介绍如下：

1. 完全同质对称型行为网络

这类行为网络的效用矩阵U 为

|  |
| --- |
| 𝑢𝑖𝑖 = 1,𝑢𝑖t = 𝑎,𝑎 ∈ 0,1 ,对于t ≠ 𝑖 |

从图论的角度看，上述效用矩阵对应于一个每个节点带有自环、节点之间所有连边权重相等的加权完全图．显然，在上述行为网络模型中，从自交互和两两交互的意义上，所有行为是相互等价的．

1. 完全异质对称型行为网络。这类行为网络的效用矩阵U 为

|  |
| --- |
| 𝑢𝑖𝑖 ∈ 1,+ ∞ ,𝑢𝑖t = 𝑎,𝑎 ∈ 0,1 ,对于t ≠ 𝑖 |

在这类行为网络模型中，不同行为之间的两两交互仍然是完全对称的，然而与同质型行为网络不同，在异质型行为网络中，每个行为自交互带来的效用可能相互不同，这种行为网络模型描述了行为具有相同两两交互效用却不同自交互效用的情形．

1. 非对称型行为网络 。 这类行为网络的效用矩阵U 为

|  |
| --- |
| 𝑢𝑖𝑖 = 1,𝑢𝑖t ∈ 0,1 ,对于t ≠ 𝑖 |

这里，对于i 不等于j,如果行为bi 与bj,之间不相连，那么uij=0；否则uij 是从区间[0， 1） 中均匀选取的值，这种行为网络模型描述了行为具有相同的自交互效用却不同的两两 交互效用的情形。

在上述三种不同类型的行为网络中，每个行为对应于一个严格纳什均衡点，也就是说，每个行为是其本身的最优响应，而且如果所有个体选择了某一同样的行为，那么任何个体不能通过改变其行为选择来获取更高的收益。

### 模型相关概念的介绍

行为占比x：以列表的形式定义，存储当前节各行为的占比数据。每一事件节点都要重新计算。

适应度f：以列表形式存储当前节点各行为的适应度情况。适应度是收益预期的量化结 果，以百分比的形式存在于模型公式中，从而影响下一时刻的行为占比。其中F 定义为浮点型，代表平均适应度，存储适应度f 的平均值。

收益矩阵u：收益矩阵u 以矩阵的形式存放，在适应度数据的量化上，我们用原始时间节点数据的收益情况来对收益矩阵进行赋值计算，在事件进程中，若无其他因素影响，则收益矩阵不会发生变化，为定值。而因行为占比每一时间节点都发生变化，所以当前时间节点行为的适应度需要重新计算。

突变矩阵Q：突变矩阵Q 是突变网络的数据体现，目前已有多种突变网络被提出，一般有两种形式，随机生成的静态随机矩阵，和与当前群体状态或适应度相关的动态矩阵。在信息传播的突变矩阵的量化赋值中，同样需要进行多方考虑，除本身数据外，还要考虑到人对自身行为的固执等，所以对角线数据值应当最大，本文提出了三种信息网络突变的具体量化方式：静态值的方式，即将在整个演化过程中突变矩阵数据不发生改变，只有保持自身的概率远远大于选择其他行为的概率，在多次行为选择之后，群体行为占比将趋于稳定，与最初的突变矩阵无较强关联，在复制动力学的影响下，但因模拟的是人的随机选择行为，所以模型中突变率也不宜太低，需要多次实验选择最佳值。与行为占比动态关联的方式，在人的选择过程中，更趋向去占比人数多的个体，所以下一节点数据完成后，将当前该行为占比与之前进行对比，若增加，则增加突变为该行为的概率，若减少则较少突变为该行为的概率。但需要设定最高值与最低值，若达到最值，则不予变化，以防止选择某行为的概率为零，此方式中仍需要让保持自身概率最大。第三中方式定义了v 作为突变率，并令存在出度的行为平分突变率v。此方式中，

所有行为保持自身的概率相等，某一行为改编为其他行为（如果存在出度）的概率相等。并且引入了突变率v，可以更好的调控突变动力学对整个模型的影响。在本文中，按需求选用三种情况。

在定义过相关概念之后，接下来进行具体的模型说明。

### 复制-突变动力学模型

复制-突变动力学是描述种群演化过程中的复制和突变两个基本机制的微分方程组。这个动力学方程通常用于对生物种群的演化过程进行建模，但也广泛应用于刻画语法演化、文化变迁、创新传播以及博弈中的群体的决策过程等，所以本文最先选择复制突变动力学作为建模算法。下面用复制突变动力学进行舆论博弈演化建模，并对其可行性进行分析，复制突变动力学包括复制和突变两个方面。下面将分别对复制和突变两个过程进行介绍：

* + 1. 复制动力学

复制动力学刻画了个体行为选择机制作用下群体状态的演化过程，模拟的是个体对高收益行为的模拟，具体实现思路如下：

首先计算得出各行为在当前集群中的收益预期（具体定义方式在后面进行说明），若某一行为，其收益预期大于平均收益预期，则该行为为高收益行为，若此行为收益预取小于平均收益预期，则该行为为低收益行为。通过设计算法实现个体对高收益行为的模仿，即在下一时间点，选择该行为的个体所占比例将增加，同理，选择低收益行为的个体数量将减少，公式为：。复制动力学计算过程代码如下：

|  |
| --- |
| 选择突变模型的实现： |
| 输入内容：行为总数T,初始行为占比列表x，收益矩阵u |
| 输出内容：下一时间节点的i 行为的占比 |
| 将数据带入如下公式中  𝑥 ← 𝑥𝑖(𝑓𝑖 − 𝜙)  𝑥𝑖 = 𝑥𝑖 + 𝑥𝑖 |

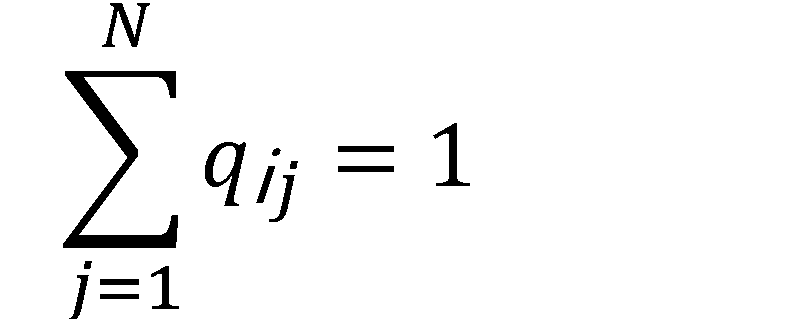
上式中的f 为行为适应度，是为了更好的量化收益预期而定义的，设行为所占比例x，行为总数为T，行为适应度f，所有行为的适应度存储于列表f，其中行为适应度𝑓𝑖即i 行为对当前集群的适应情况，并定义平均收益预期，以博弈论

方式建立收益矩阵 ，则 就是i 行为当面对j 行为时的收益情况。j 行为在集群中所占比例越大，行为i 对行为j 的收益越高，在此定义下，行为i 对行为j 的收益预期为行为j 的占比乘上，行为i 的总收益预期为行为i 对所有行为（包括行为i 本身）的预期之和。i 行为的适应度定义为行为i 的收益预期比上最高收益行为的收益预期。从而实现收益预期的量化。适应度定义过程代码如下：

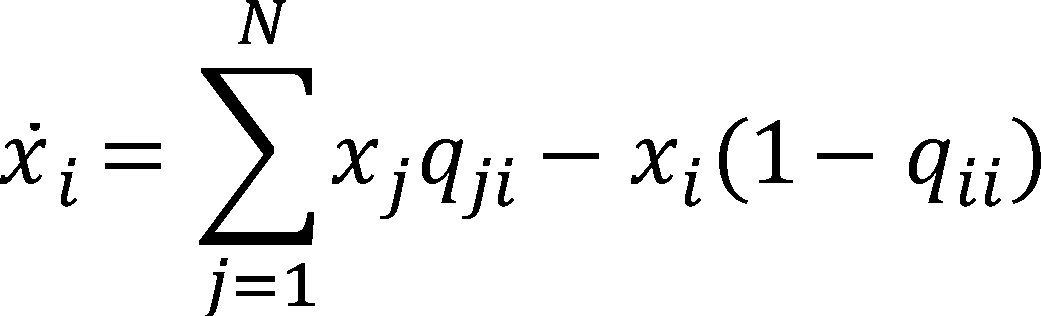
|  |
| --- |
| 适应度计算 |
| 输入：收益矩阵u，行为x，选择强度w |
| 输出：适应度列表f,平均适应度 |
| While i in range(len(x))    f.append() return f  𝑁  𝜙 = 𝑥𝑓𝑇 = ∑ 𝑓𝑖𝑥𝑖  𝑖=1  Return |

在此复制动力学模型中，行为i 占比的变化量为当前行为占比乘上行为i 的适应度与平均适应度的差值。显然，若行为i 适应度高于平均适应度，则行为 i 占比将会增加，反之，若小于平均适应度，则行为i 占比将会减少。从行为科学的角度看，上述机制，可以看作是个体对成功行为或高收益行为的模仿导致的。

* + 1. 突变动力学

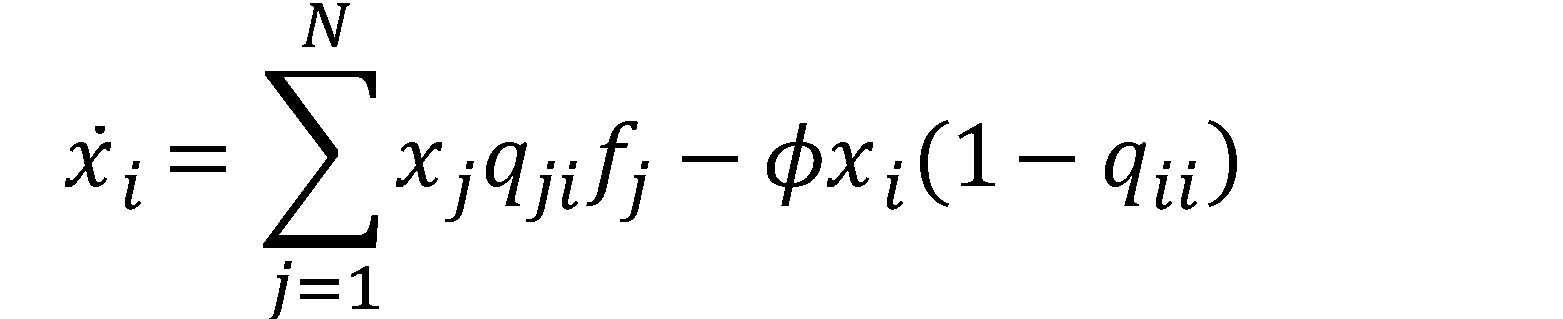
突变动力学描绘的是集群中个体行为的随机探索过程，即当前选择行为i 的个体，在下一个时间节点，会随机选择一个行为，此过程为突变过程。根据以上定义，在行为网络的舆情博弈过程中，下一时间节点，存在一个突变矩阵 ，显然其中，

由此可知，选择行为i 的个体占比为当前时间节点各行为占比与其突变为行为i 的概率乘积之和，减去当前节点行为i 占比与其突变成其他行为的概率乘积之和。下一时刻的行为占比为

 此过程在行为网络上可以理解为个体的探索试错过程导致的群体状态的演化过程。突变过程代码如下：

|  |
| --- |
| 突变动力学 |
| 输入内容：突变矩阵Q，个体行为占比x |
| 输出内容：下一时刻行为占比x |
| While i in range(len(x))    x.append(+) #计算下一时间节点的行为占比 |

* + 1. 复制突变动力学

在完成对复制动力学和突变动力学的单独介绍之后，现在将二者组成可以进行舆论博弈演化的复制-突变动力学模型，此过程就是将行为选择和行为探索两个过程合并为一个系统，可以理解为在行为选择的过程中，部分个体进行了行为探索，从而共同影响群体行为的变化，组

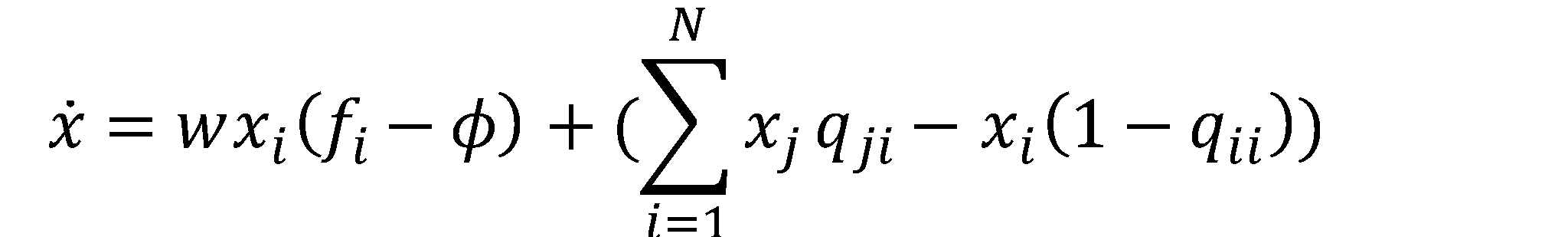
合微分方程为：

|  |
| --- |
| 复制-突变动力学模型 |
| 输入内容：当前节点行为占比x，突变矩阵Q,适应度数列f |
| 输出内容：下一时间节点的行为占比 |
| for i in range(0,T)    x.append(+) |

在此系统中，如果不考虑个体的突变过程，即突变矩阵中主对角线元素全为1，其他均为零的情况，则复制突变动力学退化为复制动力学，如果不考虑个体的选择机制，即所有个体的适应度f 均为1，那么在这种情况下复制突变动力学就退变为突变动力学。此过程通过调整f与X 决定。

### 选择-飘移动力学模型

复制动力学与突变动力学除了复制突变动力学中的组合方式外，还存在一种线性组合

的方式：。

在上式的组合方式中，复制动力学和突变动力学两者使相对对立的，即选择优势行为和随机选择一种行为使两个独立的过程。

选择漂移模型的具体实现：

|  |
| --- |
| 选择-漂移动力学模型 |
| 输入内容：当前节点行为占比x，突变矩阵Q,适应度数列f |
| 输出内容：下一时间节点的行为占比 |
| for i in range(0,T)    x.append(+) |

### 模型比较与选取

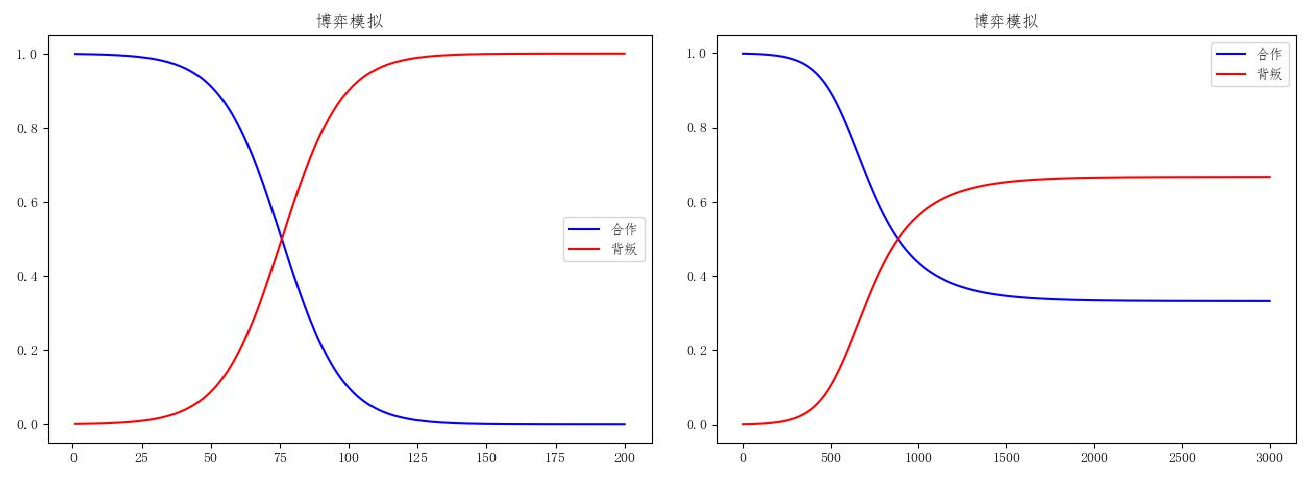
在此我们选用两个等价的囚徒困境博弈收益矩阵进行舆论博弈研究，分析的具体情况如下：

存在的两个等价的囚徒困境收益矩阵，。在博弈论中，两个收益矩阵所产生的收益预期是等价的，并且按照博弈论的相关分析，最终结果应该为

背叛行为多于合作行为，但二次呈现并存趋势。接下来我们首先将该数据带入到复制突变动力学模型中。

|  |
| --- |
| 复制突变动力学 |

|  |
| --- |
| 输入内容：收益矩阵u1，u2，突变概率矩阵Q，行为占比x |
| 输出内容：u1,u2 下时间节点的演化预测图 |
| 1. 导入相关第三方库 2. 导入自定义函数 3. F1Fadd(u1)#fadd 为适应度计算函数。 F2Fadd(u2)#fadd 为适应度计算函数。 4. 𝜙1 ← sum(f1)/len(f1)   𝜙2 ← sum(f2)/len(f2)  5  #分别对u1，u2 求出下一节点事件  占比   1. 调用matplotlib 库实现可视化   定义合作行为为红色，背叛行为为蓝色，及其他可视化参数   1. 输出两张图片 |



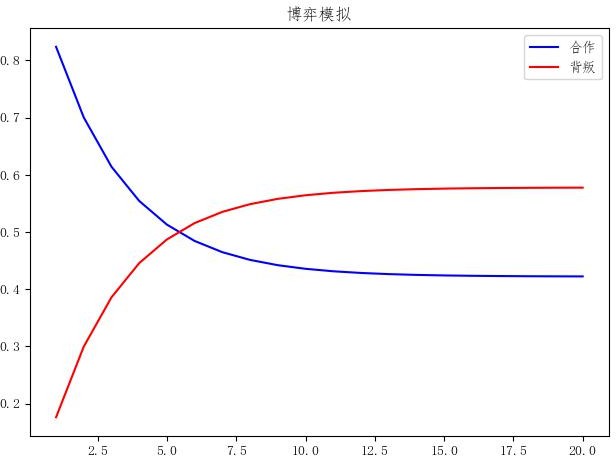
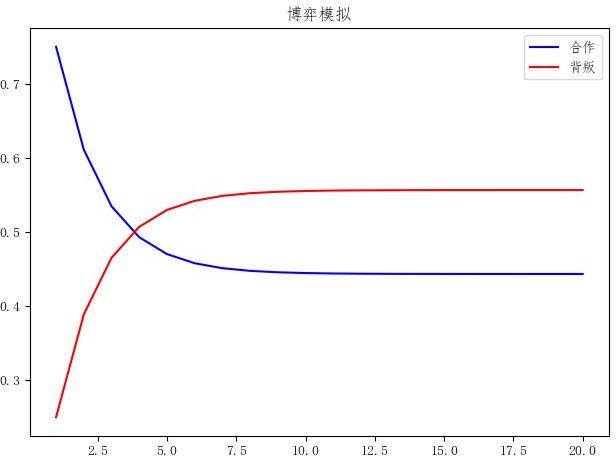
图一（a） 图一(b)

图一（a）为收益矩阵为u1 时的演化预期，图一(b)为收益矩阵为u2 时的演化预期。通过图（1）我们看到将两收益矩阵带入复制突变动力学模型的同一事件中，展现出了完全不同的结果，于是我们对模型算法的分析，发现产生上述原因的根本在于复制突变动力学中复制过程和突变过程存在强非线性耦合。以上述所选数据为例，收益矩阵某一列加上

某一常数后，如不考虑突变动力学，即突变矩阵对角线元素为 1 时，收益矩阵变化前后不 会收益预期不会变化，但若考虑突变动力学，则会产生一个以c 为系数的突变动力学方程，从而造成了收益预期的不合理性。可见在复制突变动力学模型中行为网络的效用矩阵在决 定个体行为选择的过程同时，也会对突变过程产生影响，具体地，突变动力学机制会在群 体行为演化中的作用能够被行为的效用放大或缩小，但显然突变动力学不应该受到收益矩 阵影响。但这很显然是不合理的，所以复制-突变动力学模型不适合于进行网络舆情的演 化研究。

下面，我们将数据带入到选择-漂移动力学模型中：

|  |
| --- |
| 选择-漂移动力学 |
| 输入内容：收益矩阵u1，u2，突变概率矩阵Q，行为占比x |
| 输出内容：u1,u2 下时间节点的演化预测图 |
| 1. 导入相关第三方库 2. 导入自定义函数 3. F1Fadd(u1)#fadd 为适应度计算函数。   F2Fadd(u2)#fadd 为适应度计算函数。   1. 𝜙1 ← sum(f1)/len(f1)   𝜙2 ← sum(f2)/len(f2)  5  #分别对u1，u2 求出下  一节点事件占比   1. 调用matplotlib 库实现可视化   定义合作行为为红色，背叛行为为蓝色，及其他可视化参数   1. 输出两张图片 |

图二（a） 图二（b）

图二所展示出的结果，二者仅在稳定时间点上存在不同，但两者最终的平衡状态是完 全相同的。这与我们的预期是相同的，在选择漂移模型的组合方式中，复制动力学和突变 动力学两者进行线性组合，从而避免强非线性耦合对模型的影响。同时在模型中加入选择 强度w 用来调整行为选择机制对整个演化动力学的影响，与选择强度相对应的突变强度由 突变矩阵中的突变概率调控。可见选择漂移动力学模型更符合与舆论博弈演化的需求情况。

进一步对二者进行比较分析，在选择漂移动力学模型的线性组合方式中暗含了一个基本假设，即个体的行为选择与行为探索是两个独立的过程。个体在面对一个新的事件时存在两种可能性，选择具有优势，即收益预期高的行为，或者选择随便选择一个行为。而在复制突变动力学中，收益对突变的影响，可以理解为个体在进行行为选择时产生差错从而造成突变，即先选择，后突变，这种模型更适合于细胞分裂，异常等生物种群行为，但对舆情的演化并不适用。而选择漂移模型更符合舆情演化时的具体情况。

因此在对两大模型的可行性合理性进行比较分析之后，最终选用选择漂移模型作为信息传播网络舆情博弈演化的研究模型。

### 选择漂移模型中的典型突变网络

我们对选择漂移模型的进一步的研究中，发现在对选择漂移模型的多次测试中存在几 中较为典型的突变网络：行为聚集，行为雪崩，行为震荡。在此对三者进行一下简单介绍。

* + 1. 复制突变动力学

行为雪崩：是指多种不同的种群在群体中共存，当所有行为占比相同时，称为完全雪崩，在选择漂移模型中行为雪崩只存在于完全同质对称性行为网络上，行为雪崩更多出现在选择强度较小时，若选择强度为 0，则行为占比趋近于平均，即完全行为雪崩.

* + 1. 复制突变动力学

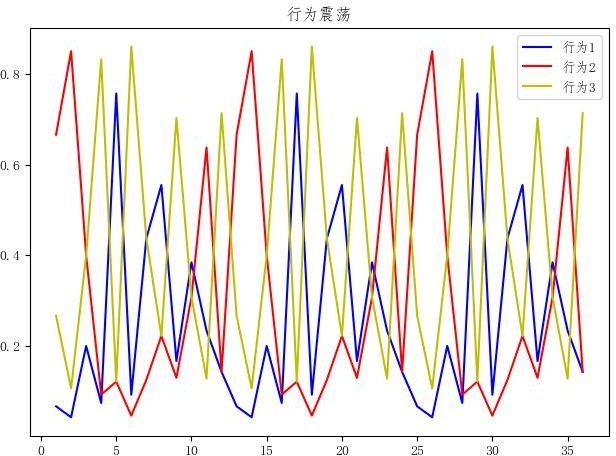
行为聚集：行为聚集是指莫一行为在群体中起到主导作用，若存在一行为占比为一，则被成为完全聚集，在选择漂移模型中行为雪崩只存在于完全同质对称性行为网络上，行为聚集更多出现在突变强度较小时，若突变强度为 0，则某一行为占比趋近于 1，即完全行为聚集。

将完全同质性对称的突变率带入模型中，不难发行，在先生成的公式中，行为的变化趋势完全由u，w，a 三者决定，在具体的案例中，a 一定为定值，所以说，我们可以通过调整u 与w 两个系数的值，从而改变最终的预期结果，随着w 增加，u 减少，群体行为从最初的单一主导行为慢慢向所以行为共存过度，即从行为聚集到行为雪崩，而当w，u 都是一个合适的适中值时，最初的行为占比将决定最初的聚集或雪崩现象，若在同质性对称网络中发生了行为聚集，那么此单一性行为一定是最初时刻行为占比最大的行为。下图为调整u，w 值得出的几种行为现象。

对同质性行为网络，u，w，a 三种满足一个特定比例时，将是聚集与雪崩的分界线，因不是本文研究的重点，故不在赘述。

* + 1. 复制突变动力学

行为震荡：行为震荡是指行为的周期性变化，主要出现在非对称性行为网络上，呈现一定的周期性。导入一个简单的数据生成如下所属的行为占比图：



图三

各行为的占比呈现周期性变化，这是因为非对称性行为网络上的行为突变概率的存在差异造成的。震荡程度由选择强度和突变强度共同决定。

## 博弈演化模型验证分析

### 数据准备

在对模型的验证中，我们首先需要收集原始数据并对其进行处理，所以这里本文先对数据的选择和数据的预处理进行介绍。

* + 1. 数据的选择

全世界有着数量众多的社交网络平台，需要选择一个用户数量最大，覆盖率最广的平 台进行研究，这样的平台才具有代表性，微博和知乎都是国内较大的社交网络平台。因此，选择二者作为社交网络研究对象，最具有代表性，研究效果最佳，本文选取了近期几起在 微博、知乎上引起热议的事件进行相关研究，使研究具有代表性。

* + 1. 数据获取及数据爬虫实现

前期使用python 编写相关爬虫程序，但因微博反爬虫及封号的相关影响改用八爪鱼平台进行相关数据的收集，在数据收集的过程中，对近期网络上引起热议事件进行分析比较，选出最具代表性的事件，收集其言论获赞数及相关评论，从而获得的相关数据，并导出到excel 表格中，获得原始数据集。

|  |
| --- |
| 数据收集 |
| 输入内容：网址，爬取的关键词 |
| 输出内容：存储相关数据的excal 表格 |
| 1. 模拟登录网址 2. 关键词命中 3. 数据整理并存储到数据库中 4. 从数据库导入到表格 |

因本文着力于舆论演化研究，因此在数据收集过程中，应突出时间变化，方便后期时间节点的区分，以所选肖战粉丝举报ao3 平台事件为例，在此事件相关数据收集的过程

中，以天为单位，分五次对 2 月 27 日至 3 月 3 日与此事件有关的言论数据进行收集，并单独存储。

* + 1. 数据清洗

通过爬虫代码收集到的原始数据以文字为主，具有规律性，难以实现相关操作，所以在进行正式的数据分析，模型带入之前，需要先行对数据进行预处理，首先，利用正则表达式将文本中出现的符号表情，颜文字，等无关信息进行删除，为进一步处理数据减轻难度。

|  |
| --- |
| 实现数据清洗 date\_clear： |
| 输入内容：导入的excal 文件 |
| 输出内容：清理完成后的字符变量txt |
| 1. 存储文本的变量txt← 具体需要导入的文本文件excal 2. 用正则清理txt 中的空行 3. 用正则清理日期时间等无实际作用的信息 4. 用正则过滤特殊字符等其他无关信息 5. 将过滤后的txt 返回 |

其次，因研究所选平台为微博，知乎，无法像英文内容一般直接区分词语，所以需要进一步对汉语文本进行处理，从而将微博正文的段落句子拆分成数个词组，并统计其出 现频率。实现理论如下：将文字数据的进行切片，并于现有字典进行比对，从而实现简单 分词，对于字典中不存在，但微博正文中文中出现的词语，可以通过前后文对比的方式确 定其是否为词语，首次出现时将此“词语”写入临时列表，如在后文中再次出现，则在临 时列表中删除，写入分词列表并统计频率，若此“词语”的某一切片命中了下文“词语”，则将此部分写入分词列表，从而实现对未知词语的分辨。本文在对汉语文本的预处理中选 用了第三方库jieba，jieba 分词采用更高级的方式对汉语文本进行分词并统计其频率， 因此在研究过程中均采用jieba 库对汉语文本进行预处理。

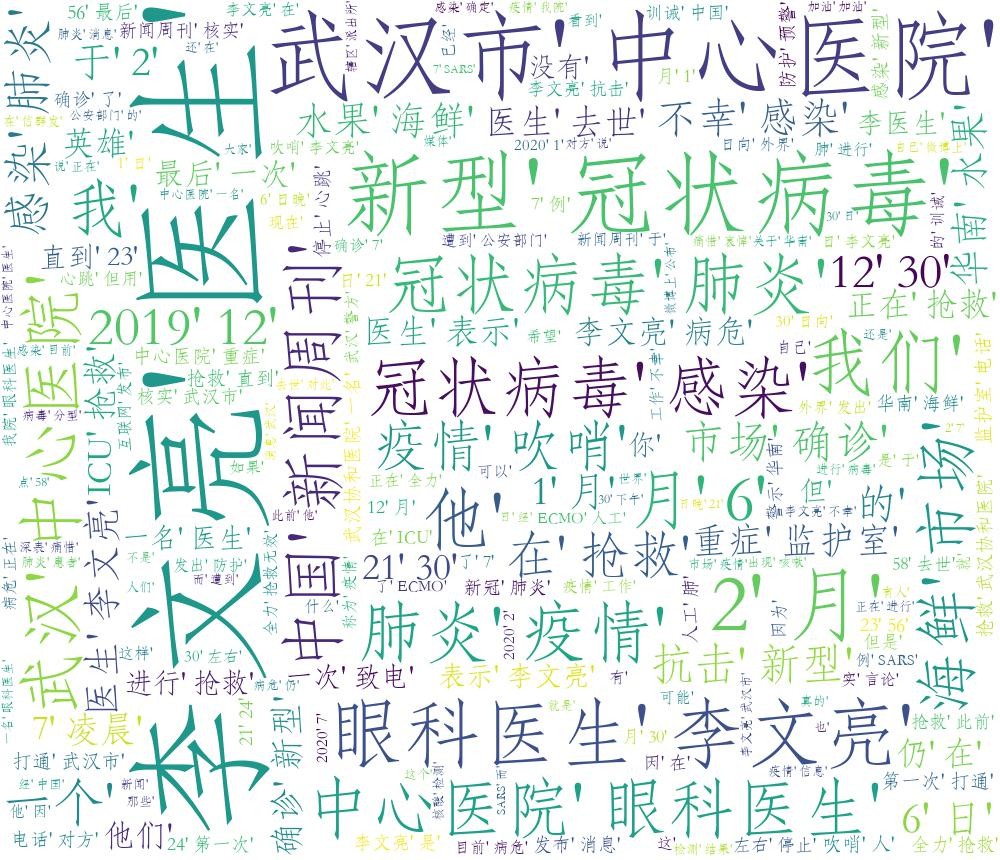
|  |
| --- |
| 分词的具体处理 |
| 输入内容：需要统计词频的具体内容txt |
| 输出内容：统计的词与其词频生成的字典count |

|  |
| --- |
| 1. 导入第三方库jieba 以便进行分词操作 2. 导入第三方库pandas 进行excal 表格处理 3. Dataframe 表格图df ← excal 表格处理 4. While df 文件内容列不为空时   存储所以内容的变量s ← df 文件的内容列  调用运用之前定义的清洗函数 date\_clea 进行数据清洗   1. ffci ←Jieba(s)分词词频导入列表 |

在完成分词及词频统计之后，还需设计算法对发文博主的立场进行判断，从而确定该节点的属性，在利用词语区分立场属性之前，需进一步对所获分词列表进行处理，删除 其中与立场属性无关的词语（如：非常、三月等），减轻之后步骤的工作量。实现方式为：对分词列表逐个进行遍历，并与事先定义的无关词列表进行比对，若存在无关词列表中则 将其删除。并完成词频统计，在分词最后输出词频最高的前 60 个词语，验证算法正确性。实现代码如下：

|  |
| --- |
| 中立属性的删除 |
| 输入内容：分词后存储词频的字典：ffci；中性词列表：vsxk |
| 输出内容：删除中性词后的词频字典：ffci |
| While ffci [i] in vsxk删除ffci[i]  Return ffci |

为更清晰的表明词频情况，以李文亮医生一事为例，调用word cloud 库将分词后词频生成的词云。



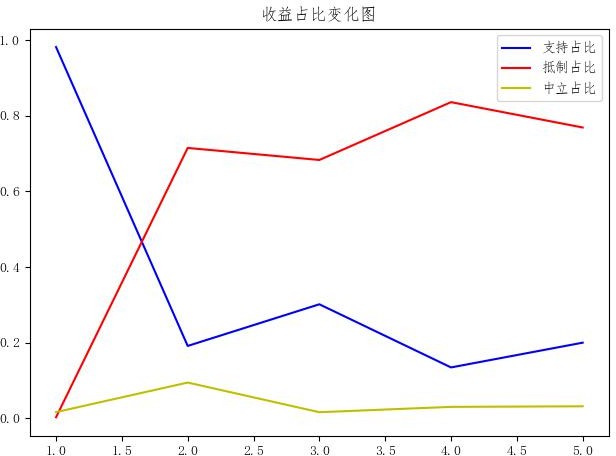
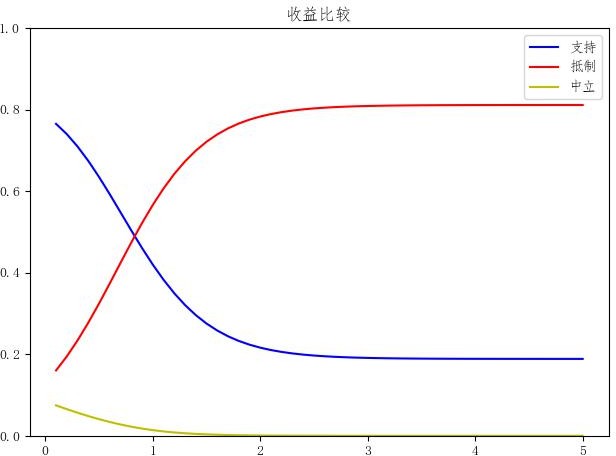
图四图四为“李文亮医生”舆情事件词云分析图

完成无关词排除后，采用同样的方式，遍历删除无关词之后的分词列表，与之前定义的属性词列表进行对比，从而确定属性。为降低反语，讽刺等文法对最终结果的影响，减小误差，可对命中数量进行设定，当分词词组与属性词组匹配词达到一特定值时，才修改当前发文博主属性。算法如下：

|  |
| --- |
| 属性的具体判定 |
| 输入：删除无关词后的词频字典ffci，各属性列表： |
| 输出：词属性的布尔值 |
| 1. For j in 遍历每一个属性列表   If ffci [i] in :逐一比较命中次数加 1  判断命中次数是否大约某一设定的值   1. 是的话返回true，不是返回fouse |

通过对数据的预处理，从而提高了文本数据的可操作性，降低工作难度，提高了在该数据集下验证模型的可行度。

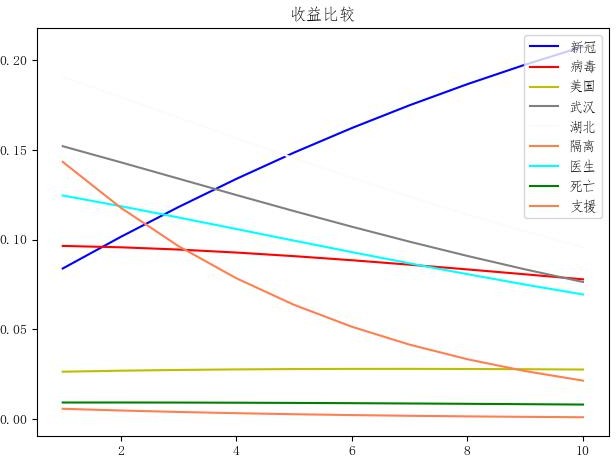
* 1. 模型验证

实现方式，以 2 月 27 日作为初始时间节点，根据演化模型推导出之后几天的演化趋 势，并与实际数据的变化趋势进行比较，判断其可行性。实际情形与模型预测情况如下图：

图五（a） 图五（b）

图五（a） 为肖战 227 时间预测趋势，图五（b）为实际趋势。从演化结果我们可以看出，预测结果与实际解决具有较强的相似性。此数据集在最初模型演化过程中，支持行为一直趋近于 0，通过对肖战粉丝群体的行为分析发现，肖战 227 案例具有特殊性，其部分粉丝不存在突变行为，故设定选择支持行为的最小值后，预测明显趋于合理。

除肖战案例外，我们同样对新型冠状病毒相关舆情进行分析，但因新冠病毒相关数据不存在肖战案例这般清晰的属性特征。故以关键词作为属性进行相关预测，此种方式也能更好的模拟多行为的博弈演化过程。行为收益由由初始节点的词频及相关发言获赞比决定。时间节点的选取为国家刚提出‘新冠’这一名词时，预测图如下：



图六

此图对新冠这一名词的出现频率的增长进行了合理预测，但其他行为的情况较为模糊，这是在定义收益矩阵时行为收益的定义方式造成的，除新冠名词外，其他词语收益差距不 大，所以造成的误差，不影响新冠增长结果的合理性。

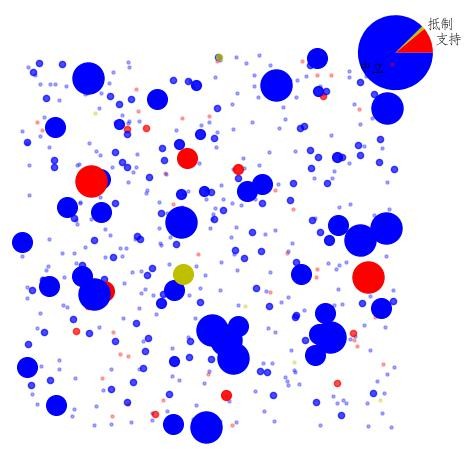
* 1. 个体行为展示

前面的折线图可以很好的看出在演化过程中，具体各行为的占比情况变化，但对变化过程中单个个体的情况无法进行清楚的描绘，所以为更好的看清整个变化中个体行为的变化过程，在此采用第三方库networkx 对行为进行动态模拟。从而更清晰的观察个体行为变化。

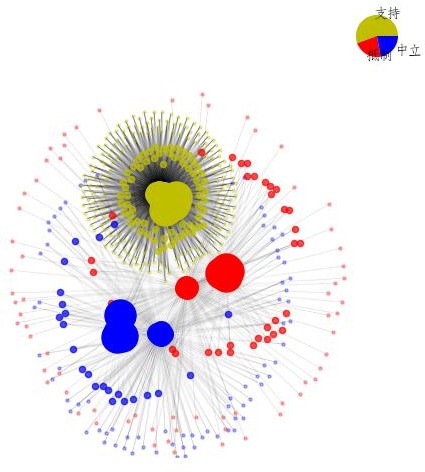
在运用networkx 展示具体的个体行为时，需要重新进行数据处理，本次选用的案例为肖战粉丝氪金光点，同样进行简单处理，将个体行为汇总为黑粉路三种情况，以收益强度确定节点大小，以个体行为划分节点颜色。使结果更为直观，同时用正文内容与属性列表匹配的词语数设定属性强度，属性强度完全由数字构成，方便体现，并在完成处理之后，运用选择突变模型对每个节点单独处理，具体实现如下：

|  |  |
| --- | --- |
| 个体行为数据可视化 | |
| 输入内容：文件fd，节点个数N | |
| 输出内容：原始节点数据可视化图，及趋于稳定前五张过程可视化图。 | |
| 1 | 导入相关库 |
| 2 | df 'xznet.xlsx'#读取文件df |
| 3 | for i in range(0,df.shape[0]):对每个节点属性进行设定 |
|  | if int(df['获赞'][int(i)]) > 7000 |
|  | df['强度'][int(i)] = '5'，设置强度 |
|  | …… #逐一添加 |
|  | if int(df['获赞'][int(i)]) < 100 |
|  | df['强度'][int(i)] = '1’ |
|  | if 立场判断函数(博文内容): #调用自定义函数添加立场 |
|  | df['立场'][int(i)] = '具体立场'设定立场 |
| 4 | N 具有节点个数 |
| 5 | 用matplotlib 输出初始节点行为占比 |
| 6 | for i in range(5):#循环五次，输出图形 |
|  | 调用plt.pie 绘制行为占比饼状图放在输出图的右上角 |
|  | 调用选择漂移模型计算各行为下一时间节点可能的比例矩阵 |
|  | X |
|  | 判断 与生成的随机数random.random()大小，从而决定该节点下一 |
|  | 时刻的行为 |
|  | 用matplotlib 输出图形 |
| 7 |  |

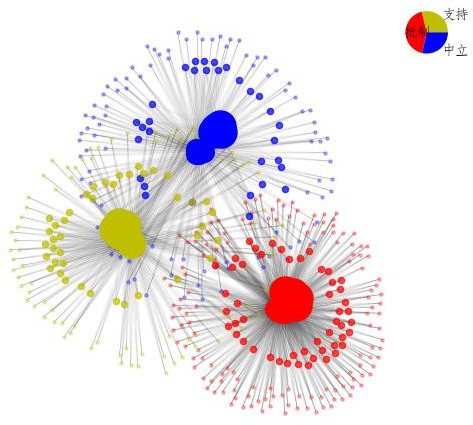
代码运行可视化后的结果如图：



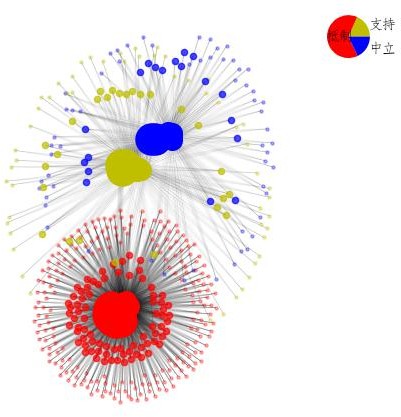
图六（a）



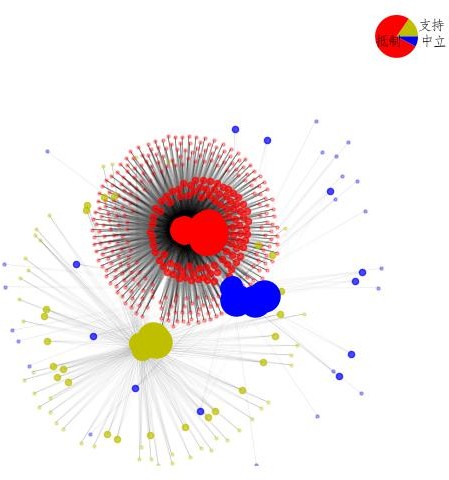
图六（b）



图六（c）



图六（d）



图六（e）

图六（a）为最初未建连接时的状态，其余四图为四个时间节点依次推进的结果。在代码中，节点被分为三大类，支持抵制和中立，然后每个点被定义了属性强度，由三大类决定颜色，属性强度决定节点大小，同类节点进行连接建立，不同节点的属性强度最大的节点进行连接建立，从而生成上图。根据当前时间节点和下一时间节点的行为占比，推算出个体变化的概率，并通过随机数进行调控。

从图六中可以清晰的看出在舆情博弈演化的过程中，个体行为和个体总数的变化情况，以一种更直观的方式对选择漂移动力学模型就行表述。

## 结论

综上所述，我认为选择漂移模型是进行信息传播网络中的舆情博弈演化的不二选择，它很好的模拟了网络用户发表言论的过程，运用此模型预测网络舆情的走势，可以更好的进行舆论监管和引导。

在模型的制作过程中，因此前未学习过相关的python 知识，所以需要先学习python 的基础语法，并选择合适的第三方库事项相关功能。同时，在研究之初，我查阅了大量资 料，了解了当前舆情演化模型研究的现状。并且为了能更灵活的运用博弈相关知识进行建 模，专门学习了大量博弈论和动力学相关知识，从演化博弈动力学，网络随机漂移过程， 网络常数的设定到复杂网络的合作涌现机制，复合网络演化博弈等等，最终发现行为网络 的博弈演化动力学与网络舆情博弈演化的过程十分相似，所以以其相关知识进行博弈演化 模型的制作。在模型的实现过程中，行为收益预期的定义及个体节点的属性界定难以量化，在对行为收益预期的定义中选择过数种方式对其进行测试，从常数值到与行为占比关联， 但都不如人意，最终在负责网络上的博弈及演化一书中找到了灵感，引入了新的量，收益 矩阵，并通过收益矩阵计算适应度，以某行为对当前舆情的适应情况来表示行为的收益预 期。行为属性的属性词列表的制作更是艰难，因汉语的语言魅力，褒词贬用，贬词褒用， 反语等等现象的存在，属性词的区分难以通过计算机直接实现，所以增加了很多工作量。 此外，鉴于网络大小不同，则人们模仿成功行为的可能性便不同，所以在选择过程中加入 了选择强度w 进行控制。突变矩阵的定义方式同样在多次实验后才确定下来。除模型本身 外，数据的收集也造成过一定的困扰，验证模型需要上万条的的数据才更为合理，但微博 知乎对收集数据都有一定的反爬虫措施，所以在收集过程中必须坚持少量多次的原则。并 且在数据选择上也需要进行斟酌，最终选取了最近极大热点话题最为研究对象。但方法总 比困难多，最终在不懈努力下，重要制作出了符合实际的博弈演化模型。

希望通过此模型能更好的实现舆论的监控引导，营造一个和谐积极向上的网络环境。