2020 年第五届“数维杯”大学生
数学建模竞赛论文

题　　目　自媒体下舆情监测情感倾向的分析与研究

摘　要

本文主要是针对网络舆论的管理与预测，通过对舆论的预测，可以有助于企业能够了解媒体或网民对相关事件或者品牌的舆情情感倾向分布和情感倾向趋势，同时能快速识别负面情感倾向的文章或评论，及时对口碑进行维护。

针对问题 1，本题使用 **TF-IDF 算法**，通过使用 python 筛选网络文章和网民评论的所有词汇，得出了在所有文章中使用频率最高的 50 个词汇，进一步去掉无意义词汇，进行归纳分析得出，在电池汽车领域，其主要舆情关注的为电动车科技方面的问题，其次是电动车的品牌。

针对问题 2，通过对话题热度方面对于舆情进行预测，根据话题的受关注程度，话题点击数评论量等，文章观点的争论，结合**高斯模型**进行话题热度趋势的拟合，通过结合 **Gamma 分布**，进一步对于在高斯分布态势拟合不足进行优化，很好的对于话题热度进行预测。

针对问题 3，要求提供一种能够合理引导网民情感倾向并逐步转向对政府或企业有利的干预方法，首先基于 **SEPPM 模型**对热点话题特征以及网民的行为特点构建舆情传播模型，然后利用 MATLAB 编程得到不同话题近似模型的参数估计值，发现舆情信息传播呈现出不同特征的形态，其形态特征与关键参数紧密相关，受到热度、传播者及其它相关性舆情信息的影响。最后，模拟政府有效注入话题，会使得热点事件数更加均衡，得到有利的干预。

针对问题 4，本题以舆情的传播时间、规模和网民的情感倾向为主体构建层次结构，然后采用**层次分析法**根据所收集的数据对方案列出判断矩阵，进行一致性检验得到 $CR=0.0516<0.1$ 满足一致性，再通过 MATLAB 计算权重。本题采用通过比较算术平均法、几何平均法和特征值法，选择特征值法计算权重，并通过排序和综合分析，得出网络舆情处理等级分为 4 个，严重程度由低到高分别为蓝色、黄色、橙色和红色。

关键词 ： TF-IDF 算法；高斯模型；Gamma 模型；SEPPM 模型；层次分析法

目　　录

一、问题重述

公共危机事件爆发时，如拍石击水，相关信息在短时间内迅速传播，引起群众的广泛关注。其中负面报道或者主观片面的一些失实评判常常在一定程度上激发人们普遍的危机感，甚至影响到政府及公共单位的公信力， 影响到企业的形象及口碑。如果不及时采取正确的措施分析和应对，将对相关部门或者企业造成难以估计的后果。所以关注相关舆情对政府或者企业来说非常重要。情感倾向分析是舆情分析技术中的重要内容。通过舆情的情感倾向预测，有助于企业能够了解媒体或网民对相关事件或者品牌的舆情情感倾向分布和情感倾向趋势，同时能快速识别负面情感倾向的文章或评论，及时对口碑进行维护。请您针对舆情的情感倾向分析问题展开如下的分析建模：

问题 1：附件 1 中我们通过技术手段抓取了部分媒体或网民评论的数据，您能否提供一个针对某一主题的舆情筛选方法；

问题 2：您能否提供一个全新数据的抓取方法，其中尽量包含诸如发表时间、 评论人数、关注人数及具体内容等具有深层次分析价值的数据；

问题 3：不同的舆情对不同的人群存在着不同的价值，期间不同的人员在舆情传播过程中起到了不同的作用。如果不能够合理的处理舆情，而是采用诸如删除评论等模式，则网民们可能还会以另外一种形式继续传播舆情。为此请大家提供一种能够合理引导网民们情感倾向逐步转向对政府或企业有利的干预方法；

问题 4：不同舆情的传播速度具有一定的差异，管理部门检测到的舆情时间点并不固定，对于政府或企业而言对处于不同阶段的舆情需要进行干预的等级不同，您能否提供一个充分考虑疫情传播时间、规模及网民情感倾向的舆情处理等级的划分方法。

二、问题分析

2.1 问题 1 的分析

因为要考虑到网民在网上发言方向较为复杂，因此基于所给的文件，我们要了解网民主要关注点在何处，因此我们使用 TF-IDF 算法，通过 python 找出网民在发言时所使用最多的词汇，通过去除无意义的词汇，从而对网民所关注的舆论方面进行研究总结，最终归纳出舆论方向。

2.2 问题 2 的分析

话题热度一般包括发表时间、点击数、评论数以及观点争议等，这些都会影响到话题热度预测，因此利用高斯模型进行拟和，但是由于话题在现实中并不是对称的，因此为了弥补高斯模型中的缺陷，我们通过引进 Gamma 模型，从而进一步对于话题预测模型进行完善，从而达到要求。

2.3 问题 3 的分析

针对问题三，首先基于 SEPPM 模型对热点话题特征以及网民的行为特点构建舆情传播模型，然后利用 MATLAB 编程得到不同话题近似模型的参数估计值，其次观察舆情信息传播呈现出的不同特征的形态，以及其关键参数。最后，模拟政府有效注入话题，获取得到此时舆情呈现的形态，利用 MATLAB 进行图形绘制。

2.4 问题 4 的分析

因为要考虑舆情的传播时间、规模和网民的情感倾向，并基于这三个方面对舆情进行处理等级的划分，对舆情进行分级则需要收集舆情的三个方面的影响程度，若要量化影响程度则可通过层次分析法建立对应的层次结构，根据细分的方案列出判断矩阵，再计算出相应的权重值，对权重值分析，即可得到舆情等级的划分。

三、模型假设

1. 假设题目所给的数据真实可靠；
2. 假设不存在网络水军控评，恶意诋毁；
3. 假设网民发言军符合国家规定，不存在因违法被删帖的情况。
4. 突发公共事件，会带来模型监控的不确定性。
5. 所有网民的发言都是真实有效。

四、定义与符号说明

符号定义	符号说明
n_{ij}	词条出现的次数
n_{kj}	文章中总共词数
$\|D\|$	语料库中的文件总数
$S(p_i)$	话题 p 的第 i 篇文章的热度得分
x_i	第 i 篇帖子的点击数
y_i	第 i 篇帖子的评价数
y	话题热度的影响值
x	话题中观点或不统一程度
N	观点总数

五、模型的建立与求解

数据的预处理：

由于所给附件较大，电脑资源不足以读取，因此将附件数据进行切分，通过使用 python 逐步读取切分的文件，将所得结果进行汇总，并且删除无意义文字以及乱码，最终所得数据进行问题分析。

5.1 问题 1 的模型建立与求解

随着互联网的发展，互联网已经走到了每个人的身边，每个人对于事件的发展都有着自己独特的看法，一千个人有一千个哈姆雷特，当这些看法汇集一起时所形成的舆论导向对于国家和企业有着重要的影响。因此通过筛选出对于企业最佳的舆论导向，并进行有效的引导，将会增加企业的获利并缓解企业在舆论中的危机。因此本问通过 TF-IDF 算法筛选所收集到的舆论，并且进行一定的模型分析，给出最佳的舆论筛选。

5.1.1 TF-IDF 算法的建立

TF-IDF(term frequency－inverse document frequency，词频-逆向文件频率)是一种信息检索与文本挖掘的常用技术。TF-IDF 是一种较常用的数学统计方法，通过评估一个词对于一个文件或者是一个资料库的重要程度。因为一个词出现频率越多，则代表其越重要，反之，当其出现频率较少时，其重要程度相对较低。

因此可以总结得出，如果某个单词在一篇文章中出现的频率 TF 高，并且在其他文章中很少出现，则认为此词或者短语具有很好的类别区分能力，适合用来分类。

(1) TF 词频（Term Frequency）

词频（TF）表示关键词在文本中的出现的频率

这个数字通常被归一化（一般为词频数除以文章总共词数），以防止其偏向成长。

其公式为：

$$TF_{ij} = \frac{n_{ij}}{\sum_k n_{kj}} \tag{1}$$

其中 n_{ij} 表示词条出现的次数，n_{kj} 表示文章中总共词数。

(2) IDF 逆向文件频率（Inverse Document Frequency）

逆向文件频率 IDF 指在某一特定词语的 IDF，可以由文件总数除以包含该词语的文件总数目，再将得到的词进行取对数得到。如果包含词条的文档越少，则 IDF 就会越大，这就说明了词条具有很好的类别区分能力。

$$IDF_i = \log \frac{|D|}{|\{j : t_i \in \mathrm{d}_j\}|} \tag{2}$$

其中 $|D|$ 是语料库中的文件总数，$|\{j : t_i \in \mathrm{d}_j\}|$ 表示包含词语 ti 的文件数目（即 ni,j ≠0 的文件数目）。如果该词语不在语料库中，就会导致分母为零，因此一般情况下用

$1+|\{j:t_i \in d_j\}|$

(3) TF-IDF 关系：TF * IDF

在某一份特定的文件中的高词频率中，以及该文件的的所有低频词语的集合，可以产生得到高权重的 TF-IDF，由此 TF-IDF 倾向于过滤掉常见的词语，保留重要的词语。

$$TF-IDF = TF * IDF \quad (3)$$

TF-IDF 算法是非常容易理解，并且很容易实现，但是其简单结构并没有考虑词语的语义信息，无法处理一词多义与一义多词的情况。

5.1.2 TF-IDF 算法的实现

通过分析题目所给的文件，通过进行 python 的编程，将文件导入得出其最初结果，并且通过人工筛选无意义的词汇，最终得出如下结果。

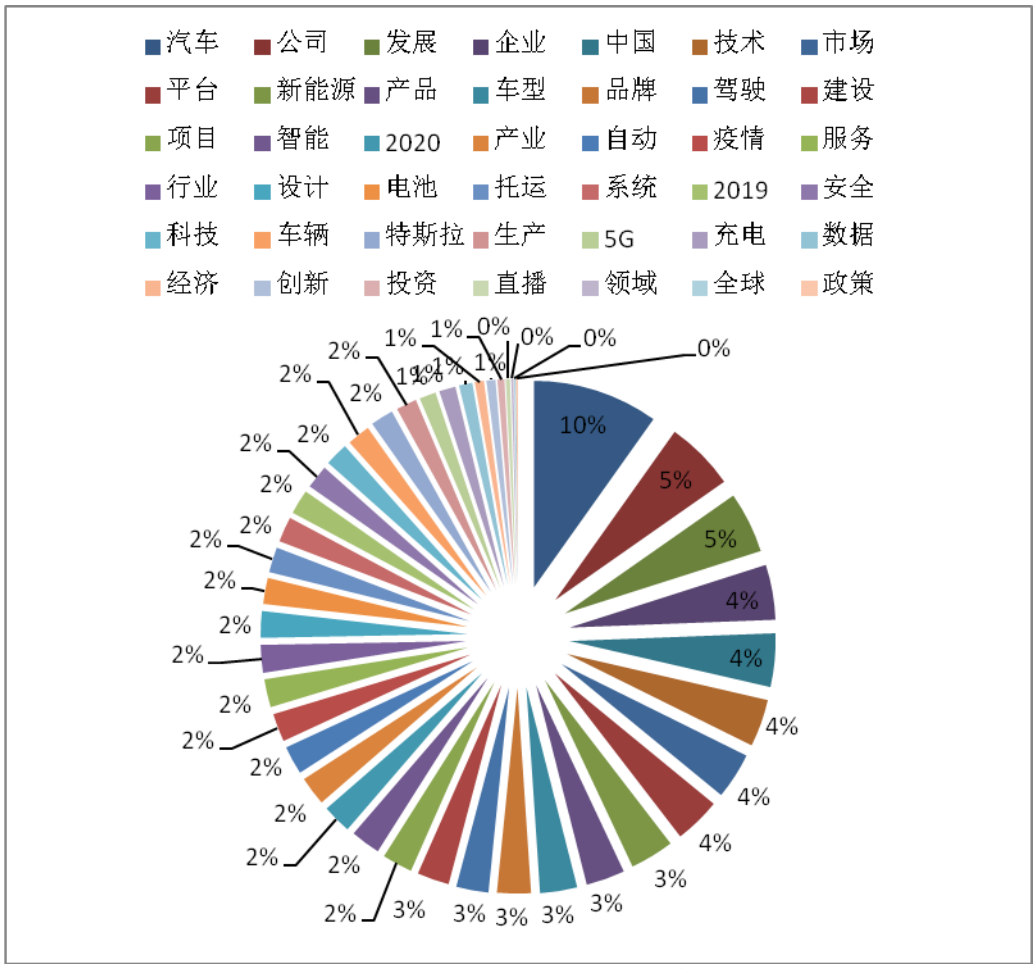

图 5-1　TF-IDF 运行结果图

从表一所得结果发现，在所给的附件中，网民对于电动汽车领域的关注较多，其其他观点均是以电动车所进行的发散讨论，在所讨论的 42 个领域，其数据较为发杂乱，对于进一步的分析有一定的阻碍。

因此，虽然图一所得的结果较多，但是其很多关键词会有一定重合，通过对于电动汽车领域进行分析，可以对于一些相似词汇进行进一步归纳，例如“5G”，“充电”均

属于汽车技术方面，从而进一步找到各个词的共性，方便进行下一步的分析。

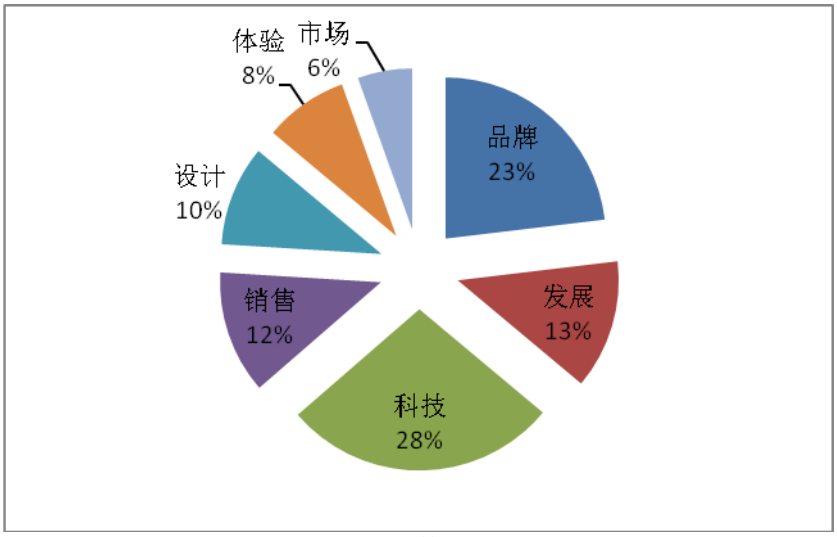

图 5-2 归纳数据图

通过分析图一、图二发现，对于我国的电动车市场，网民所关注的依旧为电动车的科技方向，电动车的科技方面，对于网民而言，电动汽车虽然已经出现了多年，但是相对于传统能源汽车，其还是一个较为新鲜的商品，担心其是否技术过硬，如电池续航是否较长，充电技术是否快等问题，因此商家可进行在电动汽车的科技舆论方面的引导，从而打开自己的市场；而网民对于电动车品牌同样是关注较多，目前各个汽车公司均推出了电动车，但是讨论最多的为特斯拉电动汽车，说明其他厂家在品牌营销方面仍不足，因此商家对于品牌的曝光率应进行提高，积极引导舆论。

5.2 问题 2 的模型建立与求解

网络话题热度的分析和预测是实现网络舆情监测的主要手段之一随着网民大幅增加，互联网每天都会产生成千上万的话题，不同的话题所涉及的内容和关注程度不同，因此不同话题产生的舆情对社会的影响也会不一样。

5.2.1 话题热度的模型建立

在互联网的舆论分析中，话题热度很大程度上对于话题的重要程度会产生影响，在本问中我们将从话题的关注程度、点赞数以及评论数作为话题热度的指标根据已有模型进行建模。点击数越多说明话题越吸引网民的参与与关注，根据话题的热度的点击数与评论数，我们得出以下公式。

$$S(p_i)=\omega_1\frac{x_i}{average(x_i)}+\omega_2\frac{y_i}{average(y_i)}+\omega_3\frac{\frac{y_i}{x_i}}{\max(\partial)} \quad (4)$$

其中，$S(p_i)$是指话题 p 的第 i 篇文章的热度得分；x_i，y_i表示第 i 篇帖子的点击数和评价数；$average(x_i)$，$average(y_i)$表示所有帖子的点击数和评论数的均值；ω_1，ω_2，ω_3为权重因子；$\max(\partial)$表示所有元组评价数和点击数的最大比值。因此热度值越大，

则越有可能成为热点话题。

话题热度的预测与文本处理中热点话题检测和 TDT 技术有关。话题热度预测主要是了解其未来趋势的变化。相比于其他话题分析研究，话题热度预测能够预测分析话题的发展趋势，并判断其是否有可能形成热点话题，这在网络舆情分析中有很大的作用。

话题热度的趋势预测方法中大都采用预测模型，比较常用的预测模型有高斯、时间序列模型等。在高斯模型的预测方法中，话题的热度随着实践变化趋近与正态分布。图 1 是经过高斯进行拟合后，话题热度的可能趋势。

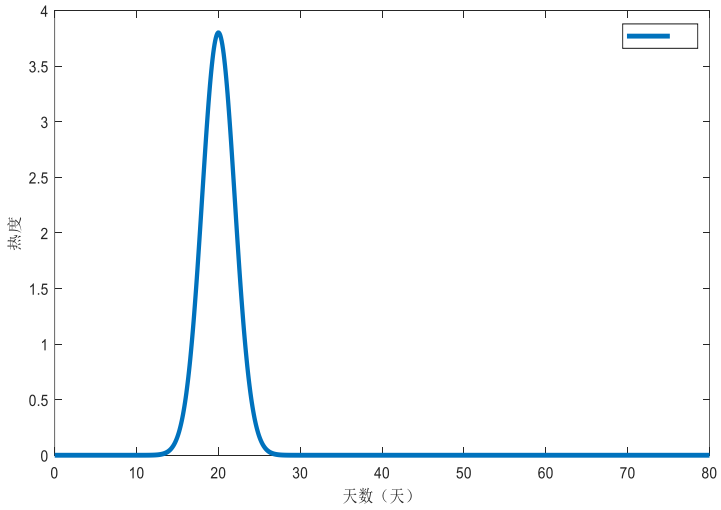

图 5-3 高斯模型拟合后的话题态势预测趋势

由于高斯函数的导函数存在，所以可以通过一阶求导的方式对高斯模型进行话题预测，根据极值点分析话题的热度是否处于上升或者下降状态。

在之前的算法中，并没有考虑到观点的不同对于热度的影响。实际上，当观点分歧越大时，其热度会越大，会进一步的增加点击率和评论量。因此我们必须对于观点分歧进行考虑。因此我们认为分歧越大对于热度影响越大，反正则没有影响。因此我们可通过幂函数进行描述，如下。

$$y=(x+\alpha)^{\beta} \tag{5}$$

其中，y 表示对话题热度的影响值；x 表示话题中观点或不统一程度，α,β 为调节参数，其中 $0<\alpha<1$，$\beta<0$。

话题中观点倾向课通过支持态度与反对态度之间的差值进行度量。差值越小，说明意见倾向越大，因此可将上式改为如下。

$$y=\frac{N}{\delta}(|m-n|+\alpha)^{\beta} \tag{6}$$

其中，m 和 n 分别代表支持数和反对数，考虑到观点总数对于结果的影响，增加调节参数 δ（$0<\delta$），N 表示观点总数。

但是进一步分析发现，当话题帖子数不止只有一个，因此在考虑话题点击数与回复式的数量很大，所以得到以下式子：

$$S(p_i)\begin{cases}\omega_1\dfrac{x_i}{average(x_i)}+\omega_2\dfrac{y_i}{average(y_i)}+\omega_3\dfrac{\frac{y_i}{x_i}}{\max(\partial)}+\omega_4(\dfrac{y_i}{\delta}(|m_i-n_i|+\alpha)\beta),i>1\\ \omega_1\log x_i+\omega_2\log y_i+\omega_3(\dfrac{y_i}{\delta}|m_i-n_i|+\alpha)\beta),i=1\end{cases} \tag{7}$$

5.2.2 话题热度趋势的拟合模型

话题的热度的发展趋势一般是波动的状态，现有的话题热度趋势主要预测模型为高斯模型，其公式如下：

$$y(x)=\alpha e^{-(\frac{x-\beta}{\gamma})^2} \tag{8}$$

其中，α,β,γ 是模型参数，可通过实际数据拟合。
然而高斯模型在话题态势拟合中存在着不足，经过研究发现，Gamma 模型可以解决。其分布函数如下：

$$\begin{cases}\Gamma(\beta)=\int_0^{\infty}t^{\beta-1}e^{-t}dt\\ f(x)=\dfrac{(\alpha x)^{\beta-1}}{\Gamma(\beta)}\alpha^{e-2x},0\le x\le\infty\end{cases} \tag{9}$$

其中 Gamma 模型可通过调节参数对各种曲线进行拟合，如下图所示：

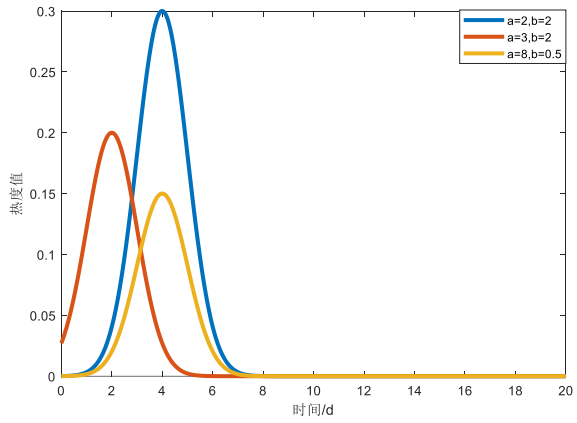

图 5-4 不同参数下 Gamma 分布曲线

对于融入观点因子的话题热度的建模方法，能够准确的去拟合话题热度的发展趋势，为预测提供更为合理的模型。但是先存着的话题热度分析并没有完全统一的指标，还可能存在其他影响话题热度发展变化的因素，这是在今后可以考虑的。

5.3 问题 3 基于 SEPPM 模型对舆情传播过程的研究

本问要求在舆情传播过程中，提供一种能够合理引导网民们情感倾向逐步转向对政

府或企业有利的干预方法。首先对舆情信息传播过程机理进行研究，然后建立了 SEPPM 模型进而根据热点话题特征及网民的行为特点构建舆情传播模型并与 SpikeM 模型结果进行比对试验。最后观察在外部因素影响下的变化趋势，得出需要如何进行引导使舆情传播对政府或企业有利的方法。

5.3.1 决策变量的确定

考虑到舆情本身的话题的吸引力、过去时刻对于当前舆情热度的影响、话题热度的衰减度对舆情传播的影响。定义如下的决策变量：

- ◆ p：背景发生率与自激效应相对大小
- ◆ μ：话题对参与讨论的用户吸引程度
- ◆ k_0：缩放因子，表示已受影响的节点对当前话题热度影响
- ◆ ω：背景发生率与自激效应相对大小

5.3.2 SEPPM 模型建立

SEPPM 模型：不同舆情信息热点的本身对于不同网民具有不同的吸引力，网民本身传播行为受到过去时期舆情信息热度和其他用户传播行为的影响，因而根据热点话题特征及用户行为特点，综合构建舆情信息传播模型，描述舆情信息传播随时间推移，在外部、本身等综合影响因素下的变化趋势。

首先假设 t_0 时刻消息产生，t_1 时刻消息开始传播，传递消息 $\{p_1, p_2, p_3\}$，以此类推 t_2 时刻传递消息 $\{p_4, p_5,\}$，t_3 时刻传递消息 $\{p_6, p_7, p_8, p_9,\}$，计算每个时刻传递舆情信息的总数量，得到舆情传递随时间的变化趋势，得到计数过程如图所示：

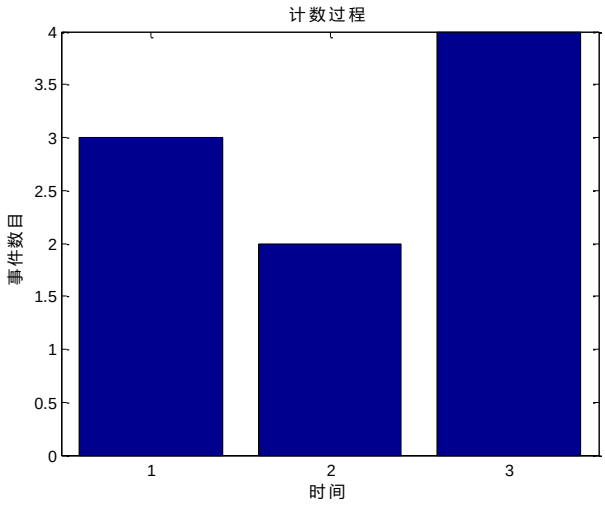

图 5-5 舆情信息传播的计数过程

由于舆情信息传播存在随时间的变化趋势，当前舆情信息依赖于已往舆情信息，舆情信息传播存在线性强度函数：

$$\lambda(t)=\mu+k_0\int_{-\infty}^{t}v(t-s)dN \tag{10}$$

且舆情传递随时间推移存在衰减函数：

$$v(t)=\sum_{j=1}^{Q}w_j e^{-\infty_j t}i_{R+} \tag{11}$$

将两公式进行代换得到一般形式：

$$\lambda(t)=\mu+k_0\sum_{t>t_i}we^{-w(t-t_i)} \tag{12}$$

计算函数对数似然估计函数，求解参数 μ 、k_0 、w 使似然估计函数最大化，求解得到完整模型：

$$\lambda(t)=p\mu+(1-p)k_0\sum_{t>t_i}we^{-w(t-t_i)} \tag{13}$$

基于 SEPPM 模型及其舆情信息于自媒体平台上传播的计数过程，以新浪微博为例，热点话题在时间 $(0,t]$ 产生事件个数，其中 $t\in[0,+\infty),S_1,S_{2,...,}S_{N(t)}$ 表示用户讨论产生的话题事件（即用户对热点话题进行评论、转发等讨论）。假设 t_0 时刻产生了一个热点话题，则产生的事件数目为 N_0 ，故我们可以得到话题事件满足随机计数过程 $\{N_t,t\geq 0\}$ ，该过程满足算式如下：

$$P\{N_0=0\}=1; \tag{14}$$

对于任意实数 $t\geq 0$ 和 $h\geq 0$ ，有

$$\begin{aligned}&P\{N_{t,t+h}=1|N_t,S_1,S_2,...,S_{N(t)}\}\\&=\lambda(t,N_t,S_1,S_2,...,S_{N(t)})h+o(h)\end{aligned} \tag{15}$$

$$P\{N_{t,t+h}\geq 2|N_t,S_1,S_2,...,S_{N(t)}\}=o(h) \tag{16}$$

$$\lambda(t,N_t,S_1,S_2,...,S_{N(t)})=\lim_{x\to\infty}h^{-1}P\{N_t,t+h\geq 1|N_t,S_1,S_2,...,S_{N(t)}\} \tag{17}$$

5.3.3 算法设计

*Step*1：定义 $\lambda_i\leftarrow\mu,t=0,J=1,I=0;$

*Step*1：首先，描述第一个事件 $U\to[0,1],-1/\lambda_{(i)}\log(u)\to X$

*Step*2：如果 $t+X>t$ ，则转到第 8 步继续问题

*Step*3：当 $t+X>t$ 时

*Step*4：问题存在 $U\to u_{[0,1]},-1/\lambda_{(i)}\log(u)\to X$

*Step*5：当 $U<=\lambda(t)/\lambda_{(j)}$ ，并且 $I=I+1,S(I)=t$ 时继续原有的步骤

*Step*6：当不满足第 6 步要求时，返回第 2 步

*Step*7：当 $J=K+1$ 条件成立时，停止算法计算

*Step*8：当计算过程出现 $X=(X-t_j)\lambda_j/\lambda_{(j+i)}$ ， $t=t_j$ ， $J=J+1$ 时

*Step*9：返回第 3 步骤继续，最后输出模拟过程

5.3.4 基于 SEPPM 模型的结果分析

以微博四个不同话题为例，对网络上热点话题传播进行研究，利用 Java 软件对微博进行模拟搜索以及抓取影响力大的“意见用户”信息传递的详细数据（编程代码详见附件 5），并通过聚类技术得出话题传播呈现不同特征，表现出不同时刻产生话题事件的不同形态。

得出当某一网络讨论热点总数为 $N=10000$ 时，得出话题对应的近似模型参数值如下表：

表 5-1 微博不同话题近似模型参数估计

话题	μ	ω	κ_0	p
1	0.01	0.2	0.1	0.5
2	0.001	0.2	0.1	0.5
3	0.01	0.7008	0.1083	0.0446
4	0.01	0.5931	0.0924	0.1

通过实例进行 SEPPM 模型模拟实验，发现舆情信息传播呈现出不同特征的形态，其形态特征与关键参数紧密相关，受到热度、传播者及其它相关性舆情信息的影响。

5.3.5 拟合实验的分析和检验

基于 SEPPM 模型，建立出对于舆情信息在传播平台进行传播过程中，受到相关因素影响问题的模型，定义话题吸引力、话题热度衰减度等四个关键参数，讨论系列相关影响因素在舆情信息传播过程中的作用效果，利用 SEPPM 模型算法对以微博微实例的对象进行分析，得到舆情信息传播依赖于自身及过去传播趋势的影响，受到与其高度关联的舆情信息的作用影响，在时间序列上呈现簇状分布的特点，在相关因素作用下呈现幂律上升和下降的特征。

对比检验：为客观衡量 SEPPM 对于实际传播模式分析的准确性，选择 SpikeM 模型（改进至 SIS 模型）和 SEPPM 模型进行比较性实验，为评价两种模型拟合真实话题传播过程，运用 MATLAB 软件进行了算法拟合实验（代码详见附件 5），得到如下图的结果：

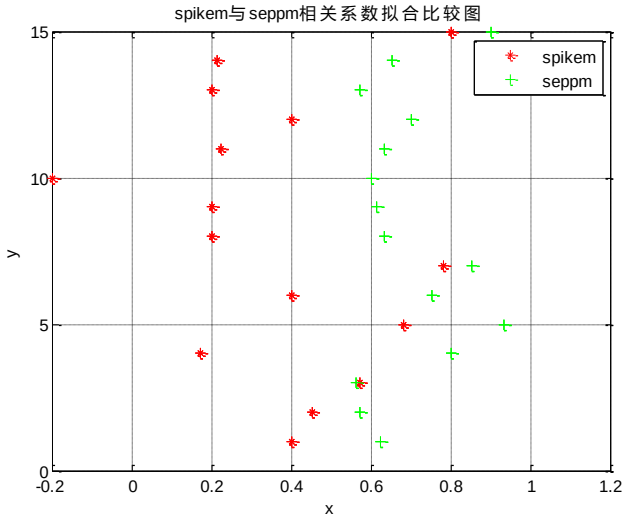

图 5-6 拟合实验比较图

经过比较发现 SEPPM 模型相关系数高于 SpikeM 的相关系数，存在 2 个较为接近的相关系数，说明 SEPPM 模型拟合效果优于 SpikeM 模型，且能够拟合出话题的不同传播模式，与仿真结果一致，具有较为准确的舆情信息传播分析性能。

加入事件干预前后的测试结果

为评估分析加入事件干预前后在同一自媒体平台下传播的差异，采用 SEPPM 模型，选取 2 个代表性话题舆情信息作为真实数据实验对象，发现二者呈现出不同的传播差异，存在多个峰值和单个峰值的不同形态，如图所示：

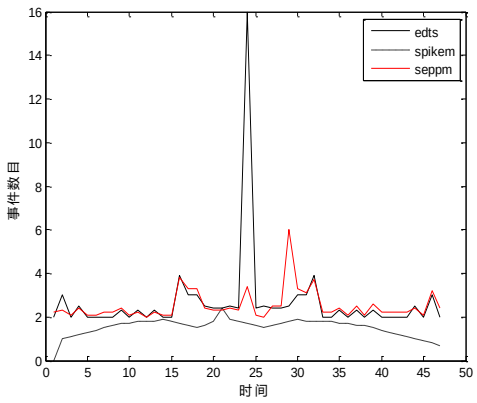

图 5-7 政府或者企业未注入话题

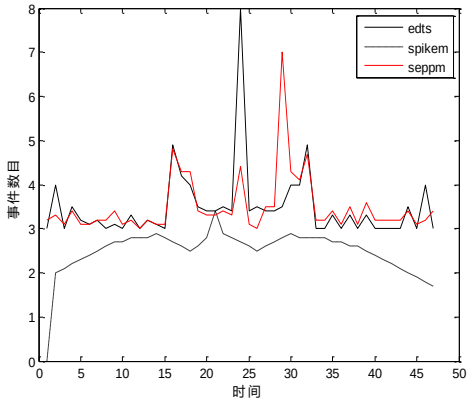

图 5-8 政府或者企业注入话题

由图 7 和图 8 可知，比较得到当政府注入话题后会使整个时间线的时间热点时间数目显示的更加均衡，这样可以有效分担舆情带来的影响。

5.4 问题 4 的模型建立与求解

5.4.1 基于层次分析法的模型建立

在网络舆情泛滥的今天，政府或企业需要根据不同阶段的的舆情进行不同程度的干涉。现在充分考虑舆情的传播时间、传播规模和网民的情感倾向这三种情况对舆情的处

理等级进行划分。

使用层次分析法（AHP）将网络舆情分成目标层、准则层、方案层的层次分析结构，通过对网络舆情的相关研究，构建出以三个主体，四个等级，总共三十四个指标的建立。其对应的层级结构如图表 2。

表 5-2 网络舆情评估层次结构

目标层	准则层	子准则层	方案层
网络舆情评估 A	传播时间 B1	传播速度 B11	极快 B111
			较快 B112
			略快 B113
			较慢 B114
			极慢 B115
		扩散阶段 B12	发酵期 B121
			爆发期 B122
			持续期 B123
			回落期 B124
		持续时间 B13	一天之内 B131
			一周之内 B132
			一个月之内 B133
			一个月之后 B134
	传播规模 B2	舆情平台 B21	微博 B211
			QQ 和微信 B212
			门户网站新闻评论 B213
			论坛和社区 B214
			聚合新闻 B215
		事件类型 B22	灾难 B221
			娱乐新闻 B222
			国家新闻 B223
			其他 B224
		行为倾向 B23	支持政府或企业 B231
			保持中立 B232
			煽动反政府或企业言论 B233
			鼓动暴力 B234
	网民情感倾向	点击量 B31	点击总数 B311
			实时点击增长量 B312
		回帖数 B32	回帖总数 B321
			回帖质量 B322
			评论风向 B323
		网民态度 B33	肯定态度 B331
			批评态度 B332
			中立态度 B333

通过数据收集和分析，在以此基础筛选出有效的评估指标并且确定权重。因为现在

网民主要以 90 后和 00 后为主要力量。本次收集总共收集了 100 份网络舆情相关数据，并以此进行筛选。

通过统计各个方案的分数值 a，问卷总数（100 份舆情数据）b。利用计算公式$S=\frac{a}{b}$，可以得出各个方案所占据的重要程度值 S。通过分析数据结果，将方案筛选中 S 值小于 0.2 的重要程度值进行剔除，用大于 0.2 的作为网络舆情评估评估方案。如表 3。

表 5-3 网络舆情方案筛选结果

目标层	准则层	子准则层	方案层	S 值	筛选
网络舆情评估 A	B1	B11	B111	0.88	✓
			B112	0.95	✓
			B113	0.35	✓
			B114	0.22	✓
			B115	0.47	✓
		B12	B121	0.13	×
			B122	0.26	✓
			B123	0.78	✓
			B124	0.86	✓
		B13	B131	0.73	✓
			B132	0.35	✓
			B133	0.9	✓
			B134	0.07	×
	B2	B21	B211	0.92	✓
			B212	0.92	✓
			B213	0.65	✓
			B214	0.55	✓
			B215	0.18	×
		B22	B221	0.42	✓
			B222	0.85	✓
			B223	0.43	✓
			B224	0.09	×
		B23	B231	0.48	✓
			B232	0.18	×
			B233	0.57	✓
			B234	0.62	✓
	B3	B31	B311	0.63	✓
			B312	0.59	✓
		B32	B321	0.43	✓
			B322	0.52	✓
			B323	0.88	✓
		B33	B331	0.37	✓
			B332	0.67	✓
			B333	0.19	×

在表 2 中，因为 B121、B134、B215、B224、B232 和 B333 的 S 值小于 0.2，所以表明这六项方案在所收集的数据中在网络舆情的影响占比不大，所以在之后的权重分析中将这六项方案剔除。

在使用层次分析法用舆情数据进行结果的筛选，对不同的方案之间进行比较判断他们的重要程度建立判断矩阵。为了度量两两元素的相对重要程度，建立了九分制度量的方法，如表 4。

表 5-4 网络舆情方案重要程度判断标志

重要性程度	赋值
i,j 两个元素同样重要	1
i 元素比 j 元素的 S 值高(0.1,0.2]（i 比 j 稍微重要）	3
i 元素比 j 元素的 S 值高(0.3,0.4]（i 比 j 明显重要）	5
i 元素比 j 元素的 S 值高(0.5,0.6]（i 比 j 强烈重要）	7
i 元素比 j 元素的 S 值高(0.7,0.8]（i 比 j 极其重要）	9
i 元素比 j 元素的 S 值增加量在区间(0.0,0.1]、(0.2,0.3]、(0.4,0.5]、(0.6,0.7]分别赋值为 2、4、6、8	

假定判断矩阵 B 中有 $b_{11}\cdots\cdots b_{1j}\cdots\cdots b_{i1}\cdots\cdots b_{ij}$ 个方案，那么它的判断矩阵如式（18）所示：

$$\begin{bmatrix} b_{11} & \cdots & b_{1j} \\ \cdots & \cdots & \cdots \\ b_{i1} & \cdots & b_{ij} \end{bmatrix} \tag{18}$$

式（18）中 b_{ij} 就是方案 i 和 j 两两比较相对重要性的结果。

判断矩阵的权重计算方法有算术平均法、几何平均法和特征值法。n 为方案个数。

算术平均法：

$$w_i = \frac{1}{n}\sum_{j=1}^{n}\frac{b_{ij}}{\sum_{k=1}^{n} b_{kj}} (i = 1,2,\cdots,\ n) \tag{19}$$

几何平均法：

$$w_i = \frac{\left(\prod_{j=1}^{n} b_{ij}\right)^{\frac{1}{n}}}{\sum_{k=1}^{n}\left(\prod_{j=1}^{n} a_{kj}\right)^{\frac{1}{n}}}, (i = 1,2,\cdots,n) \tag{20}$$

特征值法：

（1）构建判断矩阵 B 后，通过公式（21）计算判断矩阵的权重值和最大特征值列向量的归一化，得到新的矩阵 A，向量为 a_{ij} ：

$$a_{ij} = \frac{b_{ij}}{\sum_{k=1}^{n} b_{kj}} (i, j = 1,2,\cdots,n) \tag{21}$$

（2）将矩阵 A 按照公式（22）对第 i 行进行加总：

$$\overline{w_i}=\sum_{j=1}^{n}a_{ij}(i,j=1,2,\cdots,n) \tag{22}$$

（3）按照公式(23)将判断矩阵行和向量进行归一化后得到排序后的权重结果 w_i ：

$$w_i=\frac{\overline{w_i}}{\sum_{i=1}^{n}w_j}(i,j=1,2,\cdots,n) \tag{23}$$

（4）最大特征值 $\lambda_{\max}$ 按公式（24）计算：

$$\lambda_{\max}=\sum_{i=1}^{n}\frac{(Bw)_i}{nw_i}(i=1,2,\cdots,n) \tag{24}$$

（5）由于在进行两两比较的时候，不可能做到完全一致的评判标准，为了确认判断矩阵能否使用，需要对它进行一致性检验。一致性检验算法为式（25）。

$$CI=\frac{\lambda_{\max}-n}{n-1} \tag{25}$$

式子中 CI 为一致性指标，n 为矩阵的维数，$\lambda_{\max}$ 为矩阵的最大特征值数。

（6）当矩阵维度数比较大时，通过式（26）进行修正。

$$CR=\frac{CI}{RI} \tag{26}$$

其中 CR 为修正后的一致性指标，RI 为平均随机一致性指标。当 CR<0.1 时，认为该矩阵满足一致性要求。

5.4.2 层次分析法模型的求解

以本文网络舆情为例，写出判断矩阵的权重值如表 5。

表 5-5 网络舆情判断矩阵与权重

网络舆情	传播时间	传播规模	网民情感倾向	w_i
传播时间	1	$\frac{1}{2}$	$\frac{1}{2}$	0.1958
传播规模	2	1	2	0.4934
网民情感倾向	2	$\frac{1}{2}$	1	0.3108

可知特征值法误差最小，所以将表中数据带入上述模型用 MATALAB 计算可知表 5 中的一致性比例 $CR=0.0516$，$\lambda_{\max}=3.0536$。可以知道 $CR=0.0516<0.1$，表明本文的判断矩阵满足一致性，并且得到了传播时间、传播规模和网民情感倾向的权重值。同理可得其他判断矩阵的权重值为：

$$B1-B1i=\{0.2493,0.1571,0.5936\}^T$$

$$B2-B2i=\{0.625,0.2385,0.1365\}^T$$

$$B3-B3i=\{0.4286,0.4286,0.1429\}^T$$

$$B11-B11i=\{0.4038,0.4038,0.0578,0.0318,0.1027\}^T$$

$$B12-B12i=\{0,0.065,0.3641,0.5736\}^T$$

$$B13-B13i=\{0.0719,0.6491,0.279,0\}^T$$

$$B21-B21i=\{0.4061,0.4061,0.1151,0.0727,0\}^T$$

$$B22-B22i=\{0.0986,0.745,0.1564,0\}^T$$

$$B23-B23i=\{0.4286,0,0.4286,0.1429\}^T$$

$$B31-B31i=\{0.667,0.3333\}^T$$

$$B32-B32i=\{0.102,0.1721,0.7258\}^T$$

$$B33-B33i=\{0.2,0.8,0\}^T$$

在对每个方案层对应上一层元素的权重值求解，再利用 AHP 总排序法可以算出对应总目标的权重值，通过一次性检验可以知道上述的所有判断矩阵的结果符合一致性检验。本题通过 AHP 法计算出影响网络舆情方案的权重，再使用 AHP 总排序法得到 28 个方案的相对于总目标的权重值。如表 6。

表 5-6 方案层与子准则层相对于总目标的权重值

目标层	准则层	子准则层	权重	方案层	权重
A	B1	B11	0.0488	B111	0.0197
				B112	0.0197
				B113	0.0028
				B114	0.0016
				B115	0.005
		B12	0.0308	B122	0.002
				B123	0.0111
				B124	0.0176
		B13	0.1162	B131	0.0084
				B132	0.0754
				B133	0.0324
	B2	B21	0.3084	B211	0.1252
				B212	0.1252
				B213	0.0355
				B214	0.0224
		B22	0.1177	B221	0.0116
				B222	0.0877
				B223	0.0184
		B23	0.0673	B231	0.011
				B233	0.020
				B234	0.0363
	B3	B31	0.1332	B311	0.0888

				B312	0.0444
		B32	0.1332	B321	0.0136
				B322	0.0229
				B323	0.0967
		B33	0.0444	B331	0.0089
				B332	0.0355

5.4.3 结果

在网络舆情发生后，根据层次分析法计算出的各方案层的对应权重值，可以对权重值进行舆情处理等级划分。本题综合目前网络舆情处理研究现状以及上述网络舆情评估体系权重，可以完成舆情处理等级评分表，如表 7。

表 5-7 舆情处理等级平方表

处理等级	蓝色	黄色	橙色	红色
危险等级	较小危险区间	一般区间	危险区间	严重危险区间
权重区间	(0,0.25]	(0.25,0.5]	(0.5,0.75]	(0.75,1]

对应蓝色和黄色等级大众的关注度还不高，政府或企业可采取遏制舆情来源同时能够及时的在主流媒体渠道发布相应舆情事件消息等多种有效措施进行引导。在橙色和红色等级的舆情政府和企业需要多渠道的发表正式声明，将相关事件公开化和透明化处理。同时合理管控相应的媒体渠道，逐步降低舆情热度的影响，合理调控网络舆情的正常发展。

六、模型的评价

6.1 模型的优点

(1)基于优化后的 SEPPM 模型，所得到的数值更加准确；
(2)合理利用层次分析法模型进行分析，可以有效的对于不同层次舆论进行管理；
(3)使用 TF-IDF 算法检测，能有效找出数据之间的关系；
(4)结合影响舆论各个指标，能较好的研究舆论发展与管理之间的关系。

6.2 模型的缺点

(1)没有将各个平台数据进行分析对比；
(2)该模型的建立需要依靠大量的数据支持。

参考文献

[1]薛福亮,刘丽芳.基于 TF-IDF 和情感强度的细粒度情感分析——餐饮评论为例[J].信息系统工程,2020(03):83-84.

[2]卢珺珈,张宏莉,张玥.基于 BBS 的热点话题发现与态势预测技术的研究[J].智能计算机与应用,2012,2(02):1-5.

[3]李良强,李开明,白梨霏,曹云忠,吴亮.网购农产品评论中的消费者情感标签抽取方法研究[J].电子科技大学学报(社科版),2018,20(04):1-7.

[4]赵龙文,公荣涛,陈明艳,姚海波.基于意见领袖参与行为的微博话题热度预测研究[J].情报杂志,2013,32(12):42-46.

[5]田启燕.基于层次分析法的西安市网络舆论引导路径探究[J].电子测试,2014(24):161-162.

[6]洪宇,张宇,刘挺,李生.话题检测与跟踪的评测及研究综述[J].中文信息学报,2007(06):71-87.

[7]汪宏健.用 MATLAB 进行曲线拟合的方法[J].铜陵学院学报,2003(02):77-78.

附 录

附件一：问题一的代码（问题一）

功能：对附近进行数据分析

```
import  jieba

txt = open("D:\\A 题附件 1 数据.csv", "r", encoding='utf-8').read()
words = jieba.cut(txt)        # 使用精确模式对文本进行分词
counts = {}        # 通过键值对的形式存储词语及其出现的次数

for word in words:
    if  len(word) == 1:      # 单个词语不计算在内
        continue
    else:
        counts[word] = counts.get(word, 0) + 1      # 遍历所有词语，每出现一次其对应的值加
1

items = list(counts.items())#将键值对转换成列表
items.sort(key=lambda x: x[1], reverse=True)      # 根据词语出现的次数进行从大到小排序

for i in range(50):
    word, count = items[i]
    print("{0:<5}{1:>5}".format(word, count))
```

附件二：（问题三）

功能：seppm 与 spikem 拟合程序

```
x=1:15;
y=-0.2:0.2:1;
spikem=[0.4 0.45 0.57 0.17 0.68 0.4 0.78 0.2 0.2 -0.2 0.22 0.4 0.2 0.21 0.8];
plot(spikem,x,&apos;r*&apos;);
hold on
x=1:15;
y=-0.2:0.2:1;
seppm=[0.62 0.57 0.56 0.8 0.93 0.75 0.85 0.63 0.61 0.6 0.63 0.7 0.57 0.65 0.9];
plot(seppm,x,&apos;g+&apos;);
title(&apos;spikem 与 seppm 相关系数拟合比较图
&apos;);xlabel(&apos;x&apos;);ylabel(&apos;y&apos;);
legend(&apos;spikem&apos;,&apos;seppm&apos;)%曲线注释，依次对应曲线 1、曲线 2 等
grid on  %绘制网格
hold on  %保持原来绘制的图形，然后在绘制曲线不会覆盖原曲线
```

附件三：（问题四）

```
%层次分析代码
clc;
clear;
disp('请输入判断矩阵 A(n 阶)');
A=[1 1/2 1/2;2 1 2;2 1/2 1]
[n,n]=size(A);
x=ones(n,100);
y=ones(n,100);
m=zeros(1,100);
```

```
m(1)=max(x(:,1));
y(:,1)=x(:,1);
x(:,2)=A*y(:,1);
m(2)=max(x(:,2));
y(:,2)=x(:,2)/m(2);
p=0.0001;i=2;k=abs(m(2)-m(1));
while k>p
i=i+1;
x(:,i)=A*y(:,i-1);
m(i)=max(x(:,i));
y(:,i)=x(:,i)/m(i);
k=abs(m(i)-m(i-1));
end
a=sum(y(:,i));
w=y(:,i)/a;
t=m(i);
disp('权向量');disp(w);
disp('最大特征值');disp(t);
%以下是一致性检验
CI=(t-n)/(n-1);RI=[0 0 0.52 0.89 1.12 1.26 1.36 1.41 1.46 1.49 1.52 1.54 1.56 1.58 1.59];
CR=CI/RI(n);
if CR<0.10
disp('此矩阵的一致性可以接受!');
disp('CI=');disp(CI);
disp('CR=');disp(CR);
else
disp('此矩阵的一致性不可以接受!');
end
```