2020 年第五届“数维杯”大学生
数学建模竞赛论文

题　　目　舆情监测情感倾向分析建模

摘　要

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。本文就如何对舆情的情感倾向进行预测，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进行探讨。

针对问题一，基于对海量数据的处理，更有效地筛选某一主题的网络舆情，应用了 KMP 算法，利用编程更快地定位到所需主题的位置，然后分别得到相关关键词的分布情况，最后利用 SPSS 绘制图形，可以得到关注人数对某一主题的关注情况。

针对问题二，提出了一个新的抓取数据的方法，挖掘有价值的数据进行深层次分析。首先建立了三维文档向量模型（3DVM），其次统计了附件中发表时间、评论人数及关注人数的数据信息，利用自适应 KNN 追踪器追踪到这些数据，将 3DVM 与 KNN 追踪器结合使用。利用 SPSS 绘制三种数据信息的三维图，可以得出舆情的相关数据情况。综上可以得出 3DVM+KNN 追踪器使系统的漏报率与误报率较低，为全新的数据抓取方法，属于人们可接受范围。

针对问题三，将采用情感倾向文字分析数学问题，是对于解决此类问题一般数学方法的分析，也称为倾向性分析。采用多方面多层次手段对基于网络、论坛等在线社交网络产生的主观评论文本内容进行分析、处理、归纳和推导，从中挖掘出用户网民针对主题、人物、事件等表达的评论、观点和意见的过程。再运用文本预处理、以及数据挖掘的方法，对文本库进行处理，进而获取用户所需的特定信息。最后进行文本预处理是文本挖掘的主要环节，主要包括对原始文本集合进行去噪处理、中文分词、去停用词，词性选择等一系列流程。

针对问题四，基于以上问题的综合考虑，采用网络舆情三维空间模型对此次疫情的传播时间、规模及网民情感倾向进行综合判断，提出了舆情监测评价指标体系并进行运用。为确定各个指标的权重，采用了 Delphi 法和层次分析法。通过 SPSS 软件绘制了此次疫情各个参数的三维空间图，最终划分出的监测等级分为轻度级、警示级、严重级、危险级 4 个等级。

关键词　KMP 算法；三维文档向量模型；自适应 KNN 追踪器；情感倾向分析模型

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

网络舆情信息繁杂多样，若在进行网络舆情分析时，能够重点关注价值较大的网络信息，则将为舆情的情感倾向预测带来事倍功半的效果。所以针对某一主题的舆情筛选方法尤为重要。

问题 1 属于基于关键词匹配技术的数学问题，对于解决此类问题一般使用检测重复或相似关键词的算法,例如利用空间向量模型法或基于词典的分词方法和KMP 算法。对于附件中所提供的海量信息，一次性地把文件中的所有内容读入到内存中，有时是不可能的，而且这么做可能没必要，所以我们需要对数据进行取样研究。

对于问题 1 所要求我们提供的针对某一主题的舆情筛选方法，对于所给的主题，我们可以很快得到此类舆情的全部文章或者评论。

由于以上原因，我们可以首先使用一个基于 KMP 算法关键词匹配方法，对关键词搜索结果的准确度，效率分别进行分析。

2.2 问题 2 的分析

问题 2 研究的意义在于探索一个全新的数据抓取方式，对具体的内容中有价值的数据进行深层次分析，显示其背后所展现的情感倾向，以便更好地商讨对策。

问题 2 属于研究全新抓取数据的方式的数学问题，对于解决此类问题我们可以采用多维向量+数据追踪的方式来解决。

由于以上原因，我们可以将首先建立一个三维文档向量模型的数学模型，然后对数据进行追踪，建立 KNN 追踪器，并对几种不同的模型方法进行比较，对结果分别进行预测，最终得到最优模型组合。

2.3 问题 3 的分析

问题 3 属于情感文字分析数学问题，是对于解决此类问题一般数学方法的分析。情感文字分析，也称为倾向性分析。是指采用多方面多层次手段对基于网络、论坛等在线社交网络产生的主观评论文本内容进行分析、处理、归纳和推导，从中挖掘出用户网民针对主题、人物、事件等表达的评论、观点和意见的过程。

文本收集处理是指对非结构化的自然语言文本进行处理并采用一定的技术从中发现和提取特定信息的过程。通过对网络评论文本资源的收集建立文本库，再运用文本预处理、以及数据挖掘的方法，对文本库进行处理，进而获取用户所需的特定信息。

最后进行文本预处理是文本挖掘的主要环节。主要包括对原始文本集合进行去噪处理、中文分词、去停用词，词性选择等一系列流程，本文通过文本挖掘软件对预处理后的网络舆论数据进行情感分析。

2.4 问题 4 的分析

问题 4 研究的意义在于提供一个充分考虑疫情的传播时间、规模及网民情感倾向

的舆情处理等级的划分方法。这样的划分对于政府或者企业可以更快的提出应对策略，解决问题。

问题 4 属于多种条件决定舆情等级的数学问题，对于解决此类问题我们可以运用舆情三维空间模型，提出了舆情监测评价指标体系，并以 Delphi 和层次分析法加以辅助来解决。

三、模型假设

1. 假设题目所给的以及所查找的数据真实可靠。
2. 假设只是对于某一关键词的匹配度搜索，对某一主题的预测。
3. 假设算法是有足够空间能够换取时间上的查找效率问题。

四、定义与符号说明

符号定义	符号说明
W_i	情感词
S_w	情感值
α_i	舆情主体指标的权重
β_i	舆情信息指标的权重
γ_i	舆情传播指标的权重
σ_i	舆情受众指标的权重

五、模型的建立与求解

数据的预处理：

1 .随机提取附件中的部分数据。

2 .通过 eclipse 对数据进行分组测试，每 5 分钟测试一组，共 40 组。

3 .对汽车，风险，无人驾驶，安全等关键词进行提取，采取 40 组采样，每组 1000 行。

4 .为保证词表的相对准确性，将抽取部分情感词表进行构建。

5.1 问题 1 的模型建立与求解

5.1.1KMP 算法的建立

模型建立的内容要点如下：KMP 算法[1]是一种线性时间复杂度的字符串匹配算法，它是对 BF（Brute-Force，最基本的字符串匹配算法）的改进。对于给定的原始串 S 和模式串 T，需要从字符串 S 中找到字符串 T 出现的位置的索引。KMP 算法可在一个主文本字符串 S 内查找一个词 W 的出现位置。此算法通过运用对这个词在不匹配时本身就包含足够的信息来确定下一个匹配将在哪里开始，从而避免重新检查先前已经匹配过的字符。

5.1.2KMP 模型的求解

在此过程，主要采用循环遍历的方法利用 KMP 算法对内容关键词和文件中的关键词进行匹配，并记录匹配成功的个数。具体算法实现为:

1.定义数组 Keywords 用来存储待检测关键词组,从数据库中查询关键词记录,并保存到 keydata 结果集中,同时初始化 Keydata[i].count =0(其中 i 表示记录的下标，count 为关键词匹配的个数),KeyCount 表示 Keydata 结果集中元素的数量，Keywords 数组的长度为 Keywords_length。 初始化 i=0(假设第一条记录从 0 开始)。

2.判断 i 是否小于 KeyCount,若小于则继续执行下面的步骤,若超出 KeyCount 范围，则输出 Keydata i]中的 count 结果。

3.选取 Keydata 结果集中的第 i 个关键词组串元素记录即 keys = Keydata [i],假设 keys 中含有的关键词分别为 key1 , key2, key3,.., keyn。

4.定义 j 为 Keywords 数组的下标，初始化 j=0。.

5.定义 newKey 为 Keywords 数组的第 j 个关键词记录。

6.将 newKey 依次与 key1 ,key2, key3,-,keyn 使用 KMP 算法进行匹配，若有匹配成功的记录,将 Keydata [i].count 加 1。

7.将 j 加 1,判断 j 是否超出 Keywords_length 范围,若在范围内，跳转到步骤 5,若超出范围,执行下面的步骤。

8.将 i 加 1,跳转到步骤 2 执行。

图 1 是将关键词内容输入，并显示结果。

```
5  public class TextFileSearchTest {
6
7      public static void main(String[] args) {
8
9          TextFileSearch search = new TextFileSearch();
10         search.SearchKeyword(new File("D:\\A题附件1数据.csv"), "汽车 ");
11     }
12
13 }
14
```

图 5-1 执行图

图 2 是选取关键词后，得到关键词匹配后的结果。

图 5-2 关键词匹配图

图 5-3 汽车舆情分析图

将部分预处理数据通过 MATLAB 软件得到关于汽车这一主题的舆情，通过分析发现安全这一主题词一直随着汽车这一主题词出现，反而风险，无人驾驶等关键词只是在关键地方会有出现。（编程代码-代码使用 java 编写与算法思想一致详见附件）。

5.1.3 结果

对于问题 1 的求解，利用了 KMP 算法，将数据进行处理，得到了关于汽车这一主题，人们对于它的评论和关注点，有助于汽车行业的发展，由结果得知这种对于网络舆情某一主题的筛选还是有效的。

5.2 问题 2 的模型建立与求解

5.2.1 三维文档向量模型的建立

我们需要解决的问题是提供一个全新数据的抓取方法，题目要求是其中尽量包含发表时间、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。剔除其他数据后选用三维文档向量的模型进行分析。

具体步骤 1：向量空间模型(简称 VSM)[2]是文档建模中被广泛采用的经典模型。在 VSM 中，一篇文档可以用一个向量表示：设有文档集合 D，则 D 中的每一个数据 d_i 均被表示为一个范化的矢量$\{w_{i1}, w_{i2}, \ldots, w_{in}\}$，其中 w_{ij} 是第 j 维对应的特征词 t_j 在 d_i 中的权重。该文档模型在应用于新闻报道时，并没有对不同类别的特征词进行区分，而在新闻报道中，不同类别的特征词其重要性是不同的，例如表示事件发生时间的特征词和表示事件内容的特征词其权重应该是不同的。为此，我们将舆情内容的特征词分为三类：第一类是发表时间的特征词；第二类是评论人数的特征词；第三类是关注人数的特征词。基于此分类，提出了一种突出特征词类别的信息报道向量空间文档模型，称之为三维文档向量模型(3-Dimension Document Vector Model，简称 3DVM)：三维文档向量模型[3],是一种针对信息报道特点的文档模型，它将每篇报道分解为信息发表时间、信息评论人数、以及信息关注人数三个维度，用三个向量分别表示三个维度上的文档特征。

具体步骤 2：实验证明在表示话题模型时，离散点模型要优于中心点模型。

具体步骤 3：3DVM 与自适应 KNN 追踪器相结合，漏报率 F 与误报率 M 是话题检测与追踪（TDT）领域除查准率、查全率、F-Measure 等传统指标之外常采用的度量指标，一个好 TDT 系统应该同时具有偏低的 F 和 M 值。

5.2.2 三维文档向量模型的求解

本组实验在简单距离向量追踪策略下比较了三种话题/文档模型，结果如图 5-1 所示。

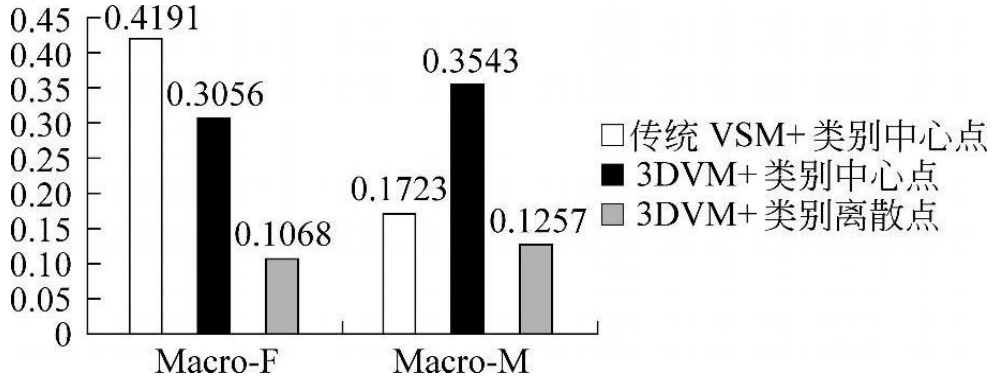

图 5-4 话题/文档模型比较实验结果

表 5-1 3DVM 与自适应 KNN 追踪器实验表

	Macro-F	Macro-M
传统 VSM+自适应 KNN	0.0297	0.1483
3DVM+自适应 KNN	0.0199	0.0756

本组实验在相同的自适应策略下对三种常用追踪器进行了比较，分别是简单向量距离追踪器、Rocchio 追踪器以及本文介绍的 KNN 追踪器。实验结果如图 5-2 所示。

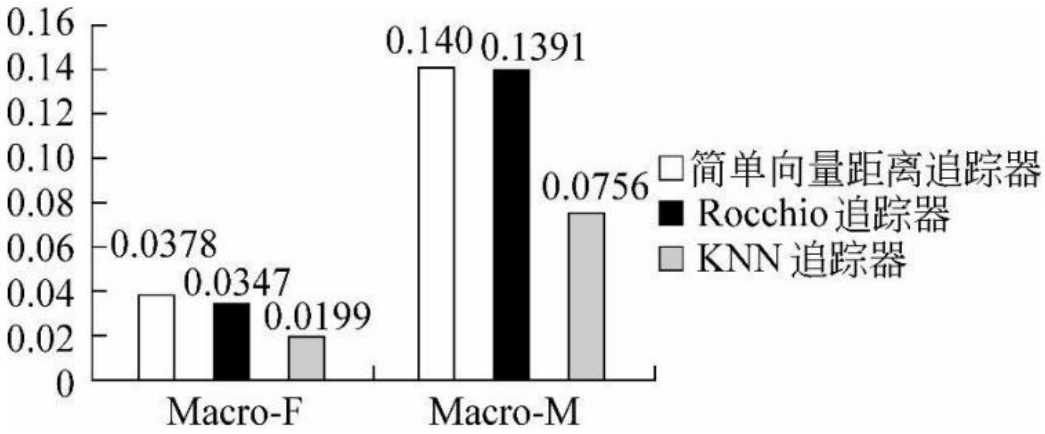

图 5-5 追踪器比较实验结果

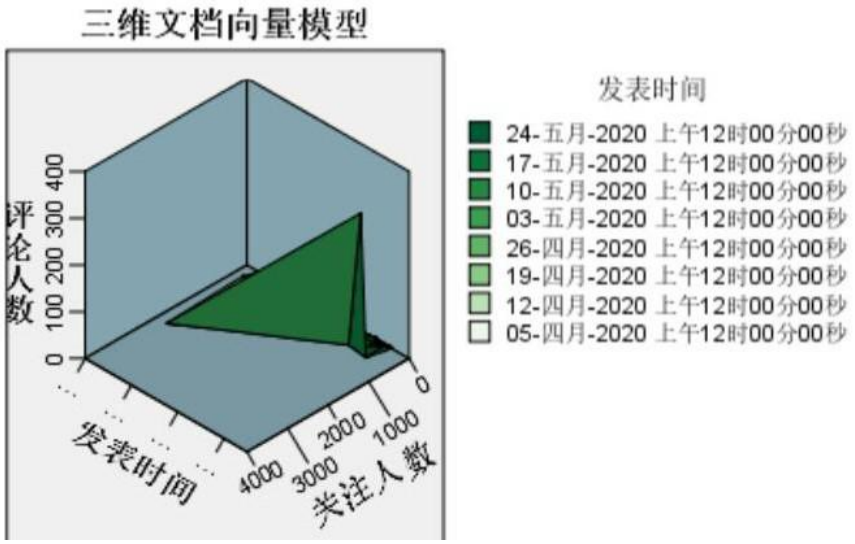

图 5-6 三维文档向量模型立体图

图 5-7 三维文档向量模型图

5.2.3 结果

可以看到，3DVM +类别离散点模型系统的宏平均漏报率与误报率都是最小的，因为 3DVM 区分发表时间、评论人数、关注人数，更加突出了文档的细节，计算文档相似度更加准确，使正确追踪到的文档数目增加；类别离散点模型将话题的中心平均分配到各个种子报道，能更好地排除噪声，保证了系统较低的误报率，而类别中心点模型中 3DVM 对细节的重视导致系统对噪音数据不敏感，使误报率增加。

实验结果表明 3DVM 提高了 KNN 追踪器的性能，同时也说明了自适应 KNN 追踪器的合理性，因为此时系统的漏报率与误报率较低，属于人们可接受范围。

5.3 问题 3 的模型建立与求解

5.3.1 情感倾向模型的建立

具体步骤 1：情感倾向分析模型构建

情感倾向判断的目的是对主观性文本内容进行情感类别的判定。因此构建情感分类对于情感倾向分析的意义不言而喻，首先通过情感词表找到与之相匹配的情感词，进而根据情感词的值计算出每个词汇的情感极性值。而整个文本的情感倾向就是所有句子情感值的整合。

具体步骤 2：文本收集处理

文本收集处理是指对非结构化的自然语言文本进行处理并采用一定的技术从中发现和提取特定信息的过程。如下图所示，首先通过对网络评论文本资源的收集建立文本库.再运用文本预处理、以及数据挖掘的方法，对文本库进行处理，进而获取用户所需的特定信息。

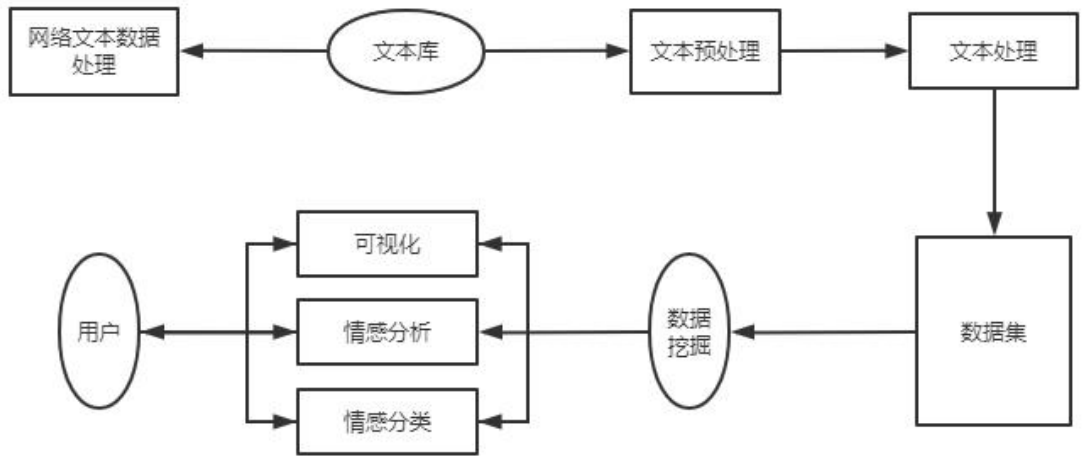

图 5-8 文本收集处理图

具体步骤 3：文本预处理

文本预处理是文本挖掘的主要环节之一。主要包括对原始文本集合进行去噪处理、中文分词、去停用词，词性选择等一系列流程，本文通过文本挖掘软件对預处理后的网络舆论数据进行情感分析。

具体步骤 4：文本情感计算规则

先将文本分解成句子集合 S，即 D=$\{s_1, s_2, \cdots s_n\}$。首先计算得出每一个句子 s_i 的情感值 $F(S_i)$，通过对所有句子情感值的整合得到文本 D 的情感值 F(S)。如式(4)、(5)所示：

$$F(S_i)=\sum S_{w_i} \quad (4)$$

$$F(S)=\sum F(S_i) \quad (5)$$

其中，S_{w_i} 为句中情感词 W_i 的情感值。

若 F(S)>0，则判定文本为正向情感；

若 F(S)<0，则判定文本为负向情感；

若 F(S)=0，则判定文本为中性情感。

其次，通过情感词表计算得出情感词 W_i 的情感值 S_{w_i}，

表示如下：

$$S_{w_i} = P_{w_i} - N_{w_i} \quad （6）$$

在公式（6）中，P_{w_i} 和 N_{w_i} 由公式（7）、（8）计算得出：

$$P_{w_i} = \frac{fp_{w_i}}{(fp_{w_i} + fn_{w_i})} \neq \frac{N_p}{(N_p + N_n)}$$

$$N_{w_i} = \frac{fn_{w_i}}{(fp_{w_i} + fn_{w_i})} \neq \frac{N_p}{(N_p + N_n)}$$

其中，N_p 表示情感词典中正向词汇的数量，N_n 为负向词汇的数量；fp_{w_i} 为情感词汇 W_i，与正向情感词汇数量的比例 fn_{w_i}。与之相反。通过公式(6)计算情感值 S_{w_i}，若 $S_{w_i}>0$，则 W_i 为正向情感词；若 $S_{w_i}<0$，则 w 为负向情感词；若 $S_{w_i}=0$，则 w 为中性情感词。再次，为保证情感值的准确性，加入程度副词为情感值，如式(9)所示：$S_{w_i} = (P_{w_i} - N_{w_i}) * (1 \pm \sigma) * N_e$。其中 N_e 为否定系数，若情感词 W_i 紧邻否定词(不"、“没”、“非”等)，情感将发生反转，故将否定系数 N_e 设置成-1，σ 为调节系数，若情感 W_i 紧邻“非常，极其”等程度副词时，则情感得分为 $S_{w_i} = (P_{w_i} - N_{w_i}) * (1+\sigma) * N_e$；若情感 W_i 紧邻“稍微，一般”等程度副词时，则情感得分为 $S_{w_i} = (P_{w_i} - N_{w_i}) * (1-\sigma) * N_e$。

表 5-2 部分情感词表[4]

Mav of Modem Information

抽取部分情感词表

总	分	代码	情感词	总	分	编码	情感词
乐	快乐	PA	1726	哀	思	PF	595
	安心	PE	956		失望	NJ	828
好	尊敬	PD	1169		愧疚	NH	465
	赞扬	PH	8985		伤感	NB	1894
	相信	PC	816	怒	愤怒	NA	736
	喜爱	PB	1361	恶	烦闷	NE	1367
	祝愿	PK	571		厌恶	ND	1870
惊	惊奇	PC	583		贬责	NN	7915
慎	慌	N1	732		嫉妒	NK	386
	恐惧	NC	1052		怀疑	NL	480
	羞	NG	439		全体总		34926

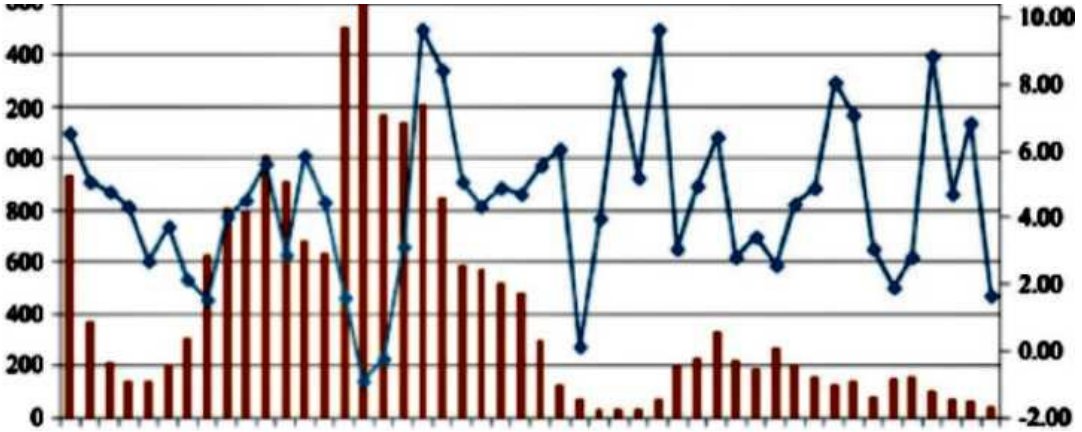

图 5-9 特定时期内某事件舆论网民平均情感倾向时序图

5.3.2 情感倾向模型的求解

从另一角度对网民情感倾向性做时序分析，如下图所示，此事件整体虽以正向情感为主，也有负向情感多次显现，有时甚至达到了高度。

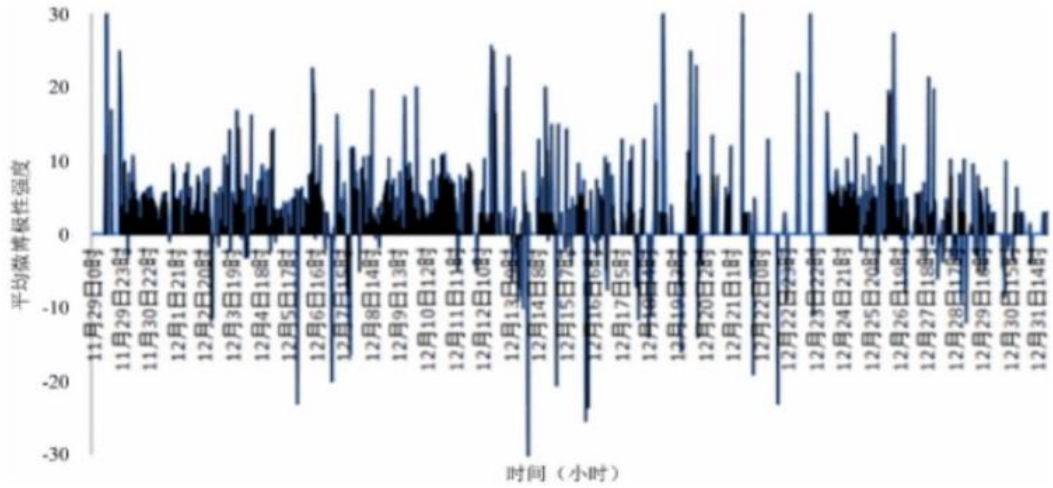

图 5-10 情感倾向时序变化图

通过对之前问题所做的数据论述分析，我们分析得到，网络舆情的倾向分析涉及到文本的情感倾向分析技术，通过对网络舆情监控与分析系统中的关键问题进行了详细的方法设计，由数据可以看出出于某一事件后在网络的传播中，网民的情感倾向被先有言论带动，之后逐渐趋于一致化，随时间推移网络舆论渐渐淡出，趋势下降。

5.3.3 结果

构建多元主体参与的协同治理机制，应加强网络舆论法治宣传教育，提升网民的责

任意识和自觉性，规范信息传播行为，净化网络信息传播环境。一方面，各地应加强网络安全普法宣传，使网民善于识别、自觉抵御网络不良信息，远离网络安全隐患，切断虚假、低俗、不利于社会安全等负面网络信息的传播渠道，营造风清气正的网络舆论环境。另一方面，要不断完善舆论监督机制。应注重对舆情反馈机制的构建，使大家能够更好地参与到舆论监督之中，同时提高舆论引导和舆情应对能力，要建立统一高效的网络安全风险报告机制和相关的网络舆论监控系统，加强对突发敏感舆情的预警和研判，在最短的时间内作出回应。不能使不良言论任由其发展、并加以引导网民们情感倾向逐步转向对政府或企业有利的指导处境。

5.4 问题 4 的模型建立与求解

5.4.1 三维空间模型的建立

我们需要解决的问题是提出一种等级划分方法，题目要求是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，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。选用三维空间模型进行分析，弥补了传统的二维及一维模型的不足之处。

具体步骤 1：用信息空间的方法，将网络舆情的三维空间[5]构造出来，便于了解网络舆情的内涵及其表达形式，从而为网络舆情监测指标体系设计提供理论依据。信息空间模型中，可编码、可抽象和可扩散构成了空间的 3 个维度。舆情要素、舆情受众、舆情传播分别对应着信息空间中的抽象维、编码维和扩散维，其中舆情要素还可细分为舆情信息和舆情主体。

具体步骤 2：根据舆情的三维空间关系，针对网络自身的特点，综合前面两种指标体系的优点，运用层次分析法，提出了舆情监测评价指标体系[5]。层次结构模型主要包括目标层、准则层和指标层。

具体步骤 3：监测指标体系的运用。进一步要确定的是各个指标的权重，可采用 Delphi 法和层次分析法来确定各项指标的权重。记准则层的特征向量为（ω_1，ω_2，ω_3，ω_4）^{T}，舆情主体 A 指标的权重分别是α_i，舆情信息 B 指标的权重分别是β_i，舆情传播 C 指标的权重分别是γ_i ，舆情受众 D 指标的权重分别是σ_i。最后进行综合分值的计算，其公式为:

$$Q=(A，B，C，D)\cdot(\omega_1，\omega_2，\omega_3，\omega_4)^T=$$

$$\omega_1\cdot A+\omega_2\cdot B+\omega_3\cdot C+\omega_4\cdot D=\omega_1\cdot\sum_{i=1}^{5}\alpha_i\cdot A_i+\omega_2\cdot\sum_{i=1}^{3}\beta_i\cdot B_i+\omega_3\cdot\sum_{i=1}^{4}\gamma_i\cdot C_i+\omega_4\cdot\sum_{i=1}^{5}\sigma_i\cdot D_i$$

5.4.2 三维空间模型的求解

层次分析法中的准则层对应着评价体系中的 4 个二级指标，包括有舆情主体、舆情信息、舆情传播和舆情受众，指标层再由准则层的隶属关系进行细分可得。

目标层	准则层	指标层
舆情综合指数	舆情主体 A	号召力 A1 主体转载率 A2 主体评论率 A3
	舆情信息 B	舆情敏感度 B1 舆情危害度 B2 舆情关注度 B3
	舆情传播 C	推荐度 C1 点击率 C2 点击频度 C3 排名指数 C4
	舆情受众 D	地理分布度 D1 区域稳定度 D2 受众共鸣度 D3 负面回应指数 D4 参与频度 D5

图 5-11 舆情监控指标评价体系的结构模型图

将预处理的数据（疫情传播时间、规模及网民情感倾向）带入上述模型，通过 SPSS 软件得到如下结果。

图 5-12 舆情三维空间分析图

5.4.3 结果

微博舆情的监测等级分为轻度级、警示级、严重级、危险级 4 个等级，并分别用蓝、黄、橙、红四个颜色加以区分。

表 5-3 微博舆情监测等级表

IV 级（蓝）	III 级（黄）	II 级（橙）	I 级（红）
轻度级	警示级	严重级	危险级

六、模型的评价及优化

6.1 误差分析

6.1.1 针对于问题 1 的误差分析

通过对附件数据的处理，对于一次性研究海量数据来说，会有匹配准确率问题出现，而且相对关键词的捕捉，以及对一整个某一主题的研究，可能会有一定查找困难。

6.1.2 针对于问题 2 的误差分析

在建立三维文档向量模型时，只考虑了信息的发表时间、评论人数以及关注人数，舍弃了对其他有价值的信息的分析，创建了基于上述三种数据类型的理想化模型。若要对多种数据进行分析，可采用多维度的向量模型进行构建，实际只是拓宽了维度，对于本身的分析没有较大的影响。

6.1.3 针对于问题 3 的误差分析

通过对之前问题所做的数据论述分析，我们分析得到，网络舆情的倾向分析涉及到文本的情感倾向分析技术，通过对网络舆情监控与分析系统中的关键问题进行了详细的方法设计，同时对现有文本倾向分析研究文本分类的关键，包括分词、停用词处理、文本特征选择和文本表示等，对文本分类和数据统计操作的整个流程有了清晰的认识。通过这种技术方式可以更加精确的捕捉到目前网络舆论的倾向程度而基于此作出积极回应问题。

6.1.4 针对于问题 4 的误差分析

针对问题 4 的误差来源主要有：问题 4 所需要的数据数量不够庞大，可能会造成结果分析不够准确，不能够较全面的表达当下疫情的全部特点，以及没有对疫情间受众的情感倾向做全面的分析，通过自适应 KNN 追踪器可有效地解决上述问题，达到所要求的结果。

6.2 模型的优点

1.三维文档向量模型：在一定程度上改进了传统向量空间模型的不足，突出了抓取重要的数据(发表时间、评论人数、关注人数或者其他具体的内容)，有利于区分相似主题报道。

2.情感文字分析模型：采用多方面多层次手段对基于网络、论坛等在线社交网络产生的主观评论文本内容进行分析、处理、归纳和推导，挖掘信息检索效率较高。

6.3 模型的缺点

1.三维文档向量模型：建立在人们可接受较低的漏报率与错误率的条件下。

2.KMP 算法依旧使用的是用空间换取时间，对于其检索还是效率不高的问题。主观性过强。

3.情感倾向性文字分析模型在文本的情感倾向性研究中缺乏对多种情感共现的转折句式的有效分析,为此提出一种专门对转折句式进行有效情感倾向性分析的方法。

6.4 模型的推广

1.三维文档向量模型及三维空间模型：都是基于三维空间设计出来的模型，在一定程度上改进了传统向量空间模型的不足。在此基础上，可以增加所需要的数据，构建三维模型，使人们对数据的结构，以及各方面的影响因素有了更直观、清晰且深刻的认识。

2.文本的情感倾向性研究中可以在一定基础上开发一种新的可以对复杂句式进行有效情感分析的情感分类模型。该模型充分分析了汉语中复杂句式的结构特点，通过已有资源构建中文情感词典，关联词表，否定词表，并提出了一种复杂句式模型来匹配各种复杂句式。最后将该复杂句模得到新的针对复杂句式的情感分类模型。该模型能更好的分析各种复杂句式的情感。

3.对于 KMP 算法的进一步研究，在此基础上利用 KMQT 算法可以进行优化。该算法的基本思想为：首先使用基于词典的分词方法中的逆向最大匹配法将题干内容分解出若干词组，并选择其中的部分词语作为题干内容的关键词，然后利用 KMP 算法对新增加的试题关键词和题库中的关键词进行字符串匹配，最后按照匹配成功的关键词的个数对结果集进行排序。

参考文献

[1] 程维刚,王宁,田勇.基于关键词匹配技术的相似试题检测方法研究[J].北华航天工业学院学报,2015,25(03):24-26.

[2] G.Salton, A.Wong , C .S .Yang .A vector space model for automatic indexing [C] // Communications of the ACM .1975, 18(11) :613-620 .

[3] 中文信息学报 基于三维文档向量的自适应话题追踪器模型 张 辉, 周敬民, 王 亮, 赵莉萍 北京航空航天大学 软件开发环境国家重点实验室, 北京 100191 2010 年 9 月 70-76 页

[4] 王子文，马静，网络舆情中的“情感倾向时序变化”问题研究[J].政治学研究，2011（3）:11-12页，11.；曾润喜，徐晓林，网络舆情突发事件预警系统、指标与机制情报杂志，2009, 28（3）:9-11页，10；表 5-2 情感词表摘录自知网 How Net,大学理工大学等情感词表相结合。

[5] 情报杂志 微博舆情监测指标体系研究 高承实 荣 星 陈 越 (解放军信息工程大学电子技术学院 郑州 450004) 2011.09 67-70 页

附录

1. KMP 算法源程序代码如下：

```
package cn.youzi.test;
import java.io.Closeable;
import java.io.File;
import java.io.FileReader;
import java.io.IOException;
import java.io.LineNumberReader;
import java.util.HashMap;
import java.util.Map;

/**
 * 对文本文件的关键词进行搜索
 * @author Abel
 *
 */
public class TextFileSearch {

public void SearchKeyword(File file,String keyword) {
//参数校验
verifyParam(file, keyword);
Map<Integer, Integer> map = new HashMap<>();
//行读取
LineNumberReader lineReader = null;
try {
lineReader = new LineNumberReader(new FileReader(file));
String readLine = null;
while((readLine =lineReader.readLine()) != null){
//判断每一行中,出现关键词的次数
int index = 0;
int next = 0;
int times = 0;//出现的次数
//判断次数
while((index = readLine.indexOf(keyword,next)) != -1) {
```

```
next = index + keyword.length();
times++;
}
if(times > 0) {
//map.put(lineReader.getLineNumber(), times);
System.out.println("第"+ lineReader.getLineNumber() +"行" + "出现 "+keyword+" 次数:
"+times);
}
}
System.out.println(map);
} catch (IOException e) {
e.printStackTrace();
} finally {
//关闭流
close(lineReader);
}
}

/**
 * 参数校验
 *
 * <br>
 * Date: 2014年11月5日
 */
private void verifyParam(File file, String keyword) {
//对参数进行校验证
if(file == null ){
throw new NullPointerException("the file is null");
}
if(keyword == null || keyword.trim().equals("")){
throw new NullPointerException("the keyword is null or \"\" ");
}

if(!file.exists()) {
```

```
throw new RuntimeException("the file is not exists");
}
//非目录
if(file.isDirectory()){
throw new RuntimeException("the file is a directory,not a file");
}

//可读取
if(!file.canRead()) {
throw new RuntimeException("the file can't read");
}
}

/**
 * 关闭流
 * <br>
 * Date: 2014年11月5日
 */
private void close(Closeable able){
if(able != null){
try {
able.close();
} catch (IOException e) {
e.printStackTrace();
able = null;
}
}
}

}
```

调用

```
package cn.youzi.test;

import java.io.File;
```

```
public class TextFileSearchTest {

public static void main(String[] args) {

TextFileSearch search = new TextFileSearch();
search.SearchKeyword(new File("D:\\A 题附件 1 数据.csv"), "安全 ");
}

}
```

2.三维文档向量模型的 SPSS 源程序：

```
GET DATA
  /TYPE=XLSX
  /FILE='C:\Users\20419\Desktop\工作簿 1.xlsx'
  /SHEET=name 'Sheet1'
  /CELLRANGE=FULL
  /READNAMES=ON
  /DATATYPEMIN PERCENTAGE=95.0
  /HIDDEN IGNORE=YES.
EXECUTE.
DATASET NAME 数据集 1 WINDOW=FRONT.
GGRAPH
  /GRAPHDATASET NAME="graphdataset"
    VARIABLES=关注人数[LEVEL=scale] 发表时间[LEVEL=scale] 评论人数[LEVEL=scale]
    MISSING=LISTWISE REPORTMISSING=NO
  /GRAPHSPEC SOURCE=VIZTEMPLATE(NAME="Surface"[LOCATION=LOCAL]
    MAPPING( "color"=" 发 表 时 间 "[DATASET="graphdataset"] "x"=" 发 表 时 间
"[DATASET="graphdataset"]
    "y"="评论人数"[DATASET="graphdataset"] "z"="关注人数"[DATASET="graphdataset"]
"Title"='三维文档向量模型'))
    VIZSTYLESHEET="Traditional"[LOCATION=LOCAL]
    LABEL='三维文档向量模型'
    DEFAULTTEMPLATE=NO.
```

3.三维空间向量模型的 SPSS 源程序：

```
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS CI(95) R ANOVA CHANGE
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT 网民情感倾向
  /METHOD=ENTER 日期 新确诊人数
  /PARTIALPLOT ALL
  /RESIDUALS HISTOGRAM(ZRESID) NORMPROB(ZRESID).
```