学号 2018302110232

密级 \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

武汉大学本科毕业论文

**时空关联的互联网事件提取技术**

院（系）名 称：武汉大学计算机学院

专 业 名 称 ：计算机科学与技术

学 生 姓 名 ：李伟

指 导 教 师 ：陈旭

二○二二年四月

**郑 重 声 明**

本人呈交的学位论文，是在导师的指导下，独立进行研究工作所取得的成果，所有数据、图片资料真实可靠。尽我所知，除文中已经注明引用的内容外，本学位论文的研究成果不包含他人享有著作权的内容。对本论文所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确的方式标明。本学位论文的知识产权归属于培养单位。

本人签名： 日期：

摘 要

这是摘要。

关键词：关键词1；关键词2；关键词3；

ABSTRACT

Content

**Key words：**Key word1； Key word2； Key word

目 录

[摘 要 3](#_Toc100343856)

[ABSTRACT 4](#_Toc100343857)

[1 绪论 8](#_Toc100343858)

[1.1 研究背景 8](#_Toc100343859)

[1.2 研究的目的和意义 9](#_Toc100343860)

[1.2.1 研究社交媒体中各种事件的有效提取和分类判别 9](#_Toc100343861)

[1.2.2 统计不同事件因时空因素影响发生的变化趋势 9](#_Toc100343862)

[1.3 国内外研究现状 9](#_Toc100343863)

[1.3.1 微博等社交媒体事件提取 9](#_Toc100343864)

[1.3.2 基于LDA的主题提取 10](#_Toc100343865)

[1.3.3 结合时空信息的事件提取 10](#_Toc100343866)

[1.4 文本组织结构 11](#_Toc100343867)

[2 研究理论基础 12](#_Toc100343868)

[2.1 互联网事件概念 12](#_Toc100343869)

[2.2 互联网事件传播特点 12](#_Toc100343870)

[2.2.1 传播主体多元、交互能力强 12](#_Toc100343871)

[2.2.2 信息碎片化、表现形式多样 12](#_Toc100343872)

[2.2.3 传播速度快、生命周期短 13](#_Toc100343873)

[2.3 事件传播的生命周期 13](#_Toc100343874)

[2.4 事件提取 14](#_Toc100343875)

[2.4.1 数据清洗 14](#_Toc100343876)

[2.4.2 停用词过滤 15](#_Toc100343877)

[2.4.3 特征选择 15](#_Toc100343878)

[2.4.4 特征加权选取 16](#_Toc100343879)

[2.4.5 主题提取 16](#_Toc100343880)

[2.4.6 相似度计算 18](#_Toc100343881)

[3 技术路线分析 20](#_Toc100343882)

[3.1 数据获取 20](#_Toc100343883)

[3.2 数据预处理 20](#_Toc100343884)

[3.2.1 分词和过滤停用词 20](#_Toc100343885)

[3.3 主题提取 20](#_Toc100343886)

[3.4 数据可视化 21](#_Toc100343887)

[4 实验与结果分析 22](#_Toc100343888)

[4.1 流程概述 22](#_Toc100343889)

[4.2 实验过程 23](#_Toc100343890)

[4.2.1 数据获取 23](#_Toc100343891)

[4.2.2 数据预处理 25](#_Toc100343892)

[4.2.3 词频统计 26](#_Toc100343893)

[4.2.4 主题提取 27](#_Toc100343894)

[4.2.5 相似度计算 33](#_Toc100343895)

[4.3 实验结果分析 35](#_Toc100343896)

[4.3.1 时间维度 36](#_Toc100343897)

[4.3.2 空间维度 37](#_Toc100343898)

[4.3.3 时空维度 38](#_Toc100343899)

[5 总结与展望 39](#_Toc100343900)

[5.1 完成的工作 39](#_Toc100343901)

[5.2 不足和改进之处 39](#_Toc100343902)

[6 参考文献 41](#_Toc100343903)

[7 致谢 42](#_Toc100343904)

# 1 绪论

## 1.1 研究背景

在现在的时代，互联网占据了人们主要的时间，在Data Reportal机构在2021年9月的一份全球数字报告中显示，目前全世界共有48亿互联网用户，相当于世界总人口的61%。全球移动用户已达到52.7亿，约占世界总人口的比例的67%，仅在去年一年全球就增加了1.17亿移动用户。全球范围内平均每人一天使用互联网高达7小时。这说明人们投入在网上的时间已经接近一天中除去睡眠时间8小时的剩余时间中近一半的比例。越来越多的互联网用户为学者们使用数据挖掘等技术研究特定社会现象提供了丰富的数据来源。

同时，社交媒体在人们使用互联网的比例也越来越大。社交媒体也成为许多人工作、生活中重要的一部分，用户活跃度高。在16至64岁的互联网用户群体中，99%用户使用社交媒体进行社交关系维护以及信息沟通，88%用户频繁在社交媒体上互动或发布信息，40%用户把社交媒体用作工作用途。使用社交媒体来分享自己的观点和看法，已经成为大多数人娱乐生活的一部分。发生一些事件，就很容易在社交媒体上引起网民对这些话题的关注和讨论，同时在不同地区、不同时间，社交媒体上话题的讨论度和关注力度都有所不同，研究这些因素能够帮助学者分析互联网事件发展的某些影响因素。例如在2021年，Omicron从被检测出，到被世界卫生组织命名，再到世界上许多国家开始出现Omicron变异株的感染者，再到国内开始出现本土的Omicron感染者，国内的社交媒体，如微博、知乎等，不同地区、不同文化、不同防疫政策下的人们对于Omicron话题的讨论有一定的变化曲线，这些变化能够帮助学者研究网民对疫情、防疫等的态度以及疫情的严重程度等等。

由于互联网上各类的讨论网站有很多，本文选择用户数量较多、使用便捷和进入门槛较低的微博作为数据来源的社交平台，使用数据挖掘、数据预处理、可视化等技术研究某个事件中时间和空间变化的趋势会随事件的发展有什么变化。

## 1.2 研究的目的和意义

### 1.2.1 研究社交媒体中各种事件的有效提取和分类判别

社交媒体中蕴含着大量的不同主题的相关信息，对这些信息加以提取利用可以有效了解诸多社会现象的变化趋势。然而由于这些社交媒体数据量庞大，噪声多，更新速度快等特点以及中文语言的随意性，多义性、非结构性的特征，如何准确有效的互联网事件就成为解决问题的关键。本文选取带有加权的LDA主题模型提取算法进行文本主题提取，并建立主题模型库，并通过与主题模型库进行文本相似度计算来过滤大量内容重复的事件。再采用文本分类和聚类思想，对同一主题的事件用更小的粒度进行分类。

### 1.2.2 统计不同事件因时空因素影响发生的变化趋势

在目前的社会中，想要从繁杂的信息中了解到一个完整的话题从产生，到成为热门事件，这中间有什么因素助推导致全网火爆呢？例如德尔塔病毒，当它首次在社交媒体中出现时，主流社交媒体上对它的讨论分布在什么地方，讨论的是什么事件，当病毒蔓延开来，人们对于其的讨论趋势又变成了什么样，是因为本地出现了相关的病例还是因为别的因素导致的。

本次论文是想从时间和空间特征上对提取的特定的互联网事件进行分析，拥有不同时空特征的事件之间会有什么相互的影响，彼此之间的影响是正面的还是负面的。从话题的产生到结束，从时间维度和空间维度观察其分布曲线，并找出导致变化较大的影响因素，得出相关结论。

## 1.3 国内外研究现状

### 1.3.1 微博等社交媒体事件提取

社交媒体能够快速在互联网上引起热点关注和讨论，有助于帮助掌握公众对于事件的反应。

韩连金在尝试检测微博谣言的课题中提出了用预训练模型BERT的方法，有效处理中文中的一词多义的方法【】；李纲、徐伟、王馨平等人提出了一种结合LDA主题模型和最大熵互信息模型提取微博热点事件并依据时间轴进行简要的信息罗列的思路【】；陈超将微博特征细化，使用改进的TF-IDF算法提取关键词，综合使用相似度计算、改进的聚类算法、以及微博的时间特征来提取个人微博的Timeline事件【】；Ferrari等人从基于位置的社交网络出发，利用概率主题模型，通过分析纽约的1300万条Tweets数据，在社交网络的碎片化信息中提取城市格局和社会人群的活动规律【】；王波、甄峰、张浩借助微博的签到数据，分别从时间、空间以及活动3个方面分析城市活动空间的动态变化，进而实现对城市活动区域的划分【】；

### 1.3.2 基于LDA的主题提取

基于LDA模型的主题提取方法在事件演化分析中得到了较为广泛的使用。安璐等根据相关性公式改进的LDA模型，在MERS爆发期间微博和微信上信息的传播分为5个阶段，分阶段地分析了利益相关这在这不同的阶段可能会关注的主题分布【】；熊振华等人从专业个人微博上使用LDA算法提取微博上的事件信息，然后用K-Means聚类算法进行聚类分析【】；杨嘉韵、张慧明对疫情相关的微博文本，使用了LDA模型和语音规则，将主题和情感结合进行分析，分析什么因素对网络用户的情感的影响更大；石晶,范猛,李万龙等人在文本分割的基础上，利用LDA为语料库及文奉建模，实用Clarity度量相似度大小，根据词语的香农信息提取该片段的主题词【】；张晨逸,孙建伶,丁轶群等人在对微博这种带有一些结构化的社会网络关系信息的文本进行研究时，提出了基于LDA的微博生成模型MB-LDA，综合考虑微博的联系人关系和文本关联关系，辅助进行主题挖掘【】；

### 1.3.3 结合时空信息的事件提取

微博的时间特征能够充分体现微博的实时性的特点，而空间特征则能够体现微博传播的范围广的特点，通常被用来研究突发性的公共事件舆情变化趋势。韩珂珂、刑子瑶、刘哲等人在研究重大公共卫生事件的舆情时，实现了基于空间特点多尺度舆情分析算法，可以全方位展示不同区域的舆情分布及特点【】；李想以“九寨沟地震”为例使用LDA主题提取算法建立灾害事件主题模型库，利用开源网络地图API的地理编码功能，统计分析灾害事件的时空属性，以便为灾害应急响应提供决策支持；

## 1.4 文本组织结构

# 2 研究理论基础

## 2.1 互联网事件概念

《现代汉语词典》中将事件定义为“历史上或社会上发生的不平常的大事情”。

而互联网事件，顾名思义，就是通过互联网传播信息的一系列公共事件。汤普森在哈贝马斯的交往行为理论基础上，加入了对当代社会日常生活的理解，提出了公共性的一种新的概念[100]。在他的理论中，大众传媒包括报纸、电台等等传媒工具创造的可见性或者公共性，重新定义了公共与私人生活之间的界限，并且超过了共同在场的公共性。在这个互联网与生活紧密结合的时期，由于各种社交平台上充斥着个人生活的分享与讨论，所以在本文中研究的事件，通常都指：有一定社会影响力，引起社会普遍关注，进而能引起社会议论、社会波动的事实或事情。

互联网事件与寻常公共事件相比有几个特性：1. 传播范围广。在互联网上大肆传播的事件，经过信息平台的个性化推荐、热点推送、实时热榜查询等功能的渲染之后，影响的范围已经不局限在事件利益相关的人群中了，关注网络动态的人能够快速获取到相关事件的信息，并快速向其他人群扩散。2. 具有群体性。互联网事件在时间和空间上的分布特点能表明，在传播过程中影响的是一个群体，不单指具体的某些人。

## 2.2 互联网事件传播特点

### 2.2.1 传播主体多元、交互能力强

互联网事件的信息传播主体往往包含多种角色，如普通网民、各种民间数字媒体、官方媒体、政府部门、企业媒体等等，会引起各种形式的交互交流。由于网络受众有极大的自由以某种立场加入到事件传播系统中，同时还能在事件的出传播中充当信息再加工者的角色，因此事件的交互性极强。

### 2.2.2 信息碎片化、表现形式多样

在互联网上，每个人都是一个发声个体，在互联网上编辑、转发、点赞的任何一个帖子都会最终形成一个信息碎片，由于含有的主观因素非常多，不同个体使用相同的词语或者内容表达出来的意思及造成的影响都是有所不同的。而这些千人千面的信息碎片在短时间内借助互联网技术传播出去，形成一个完整事件的时间轴、信息链。在新冠病毒变异株在世界某地首次被检测出，然后被世卫组织命名后，这类相关的新闻很快就被各种关注国外新闻的大v媒体转发，然后政府部门也开始纷纷转发，相关话题迅速登上微博热搜榜，进而持 续发酵，在短短几天内获得上亿的点击量，形成热议讨论。普通网民点击浏览后，又迅速传播到身边的亲朋好友，又在互联网上引起持续的讨论。

### 2.2.3 传播速度快、生命周期短

由于微博帖子信息篇幅小，阅读时间需求少，因此传播代价低，再加上各种信息传播媒介中，互联网的传播速度极快，所以互联网事件一旦在网络世界产生，有一定因素的助力下，能够在极短时间内让现实世界中相隔非常远的人在同一时间了解到该事件。但由于互联网上信息种类繁杂，一个事件如果没有后续的相应的事件推动，热度很容易被其他事件取代，热度降低的过程和热度升高的过程都是一个正反馈的过程，当热度出现下降的趋势了，无其它因素干扰的话，凭借事件本身的生命周期来看，在短短几小时中就可能没有多少人继续讨论该事件。由此可见，互联网事件虽然传播范围广、速度快，但其本身的生命周期较短，并不能长久形成广泛地的关注和讨论。

## 2.3 事件传播的生命周期

上世纪八十年代中期,国外学者芬克（Fink）第一次引入了危机管理的生命周期理论,从生命周期角度出发,对危机管理进行了描述。

在Fink的危机四阶段模型理论基础上，将事件信息传播的生命周期分为征兆期、爆发期、持续期和恢复期。

从互联网事件的发展趋势来看，互联网事件传播的生命周期也能大致分为这四个时期。林弘以“西安奔驰事件”举例对社会热点事件舆论发展趋势及传播特点进行了研究【】。

在事件的征兆期，该事件最先在百度贴吧中发布，被网友转发到新浪微博后，经过官微的编辑发帖后逐渐进入人们的视野。该帖子由于逻辑清晰、叙述观点有理有据，引起了小部分的讨论，同时引起了有关部门的关注。

在事件的爆发期，当事女主和奔驰官方及4S店完全不同的表态引起了网友大量的讨论、转发。同时，各大官方媒体也在新浪微博上进行发文评论，事件的热度井喷式上升。

在事件的持续期，在这一阶段互联网事件已经从网络媒体传播到了包括电视、广播电台等传统媒体，传统媒体以不同于网络媒体的传播方式进行了详细的报道，让更多尚未了解的人们开始关注到该事件的发展。随着双方的协商的具体细节开始被曝光，越来越多的人开始关注该事件。但由于互联网事件的快速更新，该事件被其它新热搜事件吸引了部分流量，导致热度不会高于爆发期的峰值，但仍然会处于一个较高的水平上。

在事件的恢复期，这一阶段，由于“西安奔驰事件”该事件最终以双方达成协议终止，此时事件的曝光度已经基本被其它事件取代，网民对其的关注和讨论程度都已经大幅下降，最后的热度也就逐渐下降。

图表, 折线图

描述已自动生成

图2.1 互联网事件传播生命周期

## 2.4 事件提取

### 2.4.1 数据清洗

在得到的微博帖子数据中，有很多是对本研究无用的信息，如各类表情符、连接符、图片链接、视频链接等等，这些信息在实验中需要去除，并尽可能保留原有信息的完整。

本次实验会采用正则表达式匹配的形式，去除掉一些特殊的符号和无用信息，避免对实验产生比较大的噪声干扰。

### 2.4.2 停用词过滤

在自然语言处理中，我们通常把停用词、出现频率很低的词汇过滤掉。这个过程其实类似于特征筛选的过程。当然停用词过滤，是文本分析中一个预处理方法。它的功能是过滤分词结果中的噪声。比如：的、是、啊等。在英文里，我们经常会遇到比如“the”，“an”，“their”等这些都可以作为停用词来处理，但是也考虑自己的应用场景。当然如果出现一种出现频率特别低的词汇对分析作用不大，所以一般般也会去掉。把停用词、出现频率低的词过滤之后，即可以得到一个我们的词典库。

本次使用的停用词表是在哈尔滨工业大学停用词表、百度停用词表、四川大学机器智能实验室通用词表、中文停用词表结合起来扩展的停用词库，同时结合搜狗词库中的诸多专业词汇表作为用户词库，在进行数据预处理时，结合分词工具对所有博文分词后再过滤停用词、过滤使用频率非常低的词语（例如各种方言形成词语）。

### 2.4.3 特征选择

在预处理过后，微博数据仍然有比较大的规模，直接使用向量表示文本数据，维度会非常多，影响后续的实验效率。所以我们需要对文本进行特征选择，降低特征的表示维度数量。在进行调研之后，发现可以使用CHI方法来进行特征提取，这可以把数据维度降到千级别，能够较大地提升过滤效率。

CHI方法的核心思想是：词和类别之间符合分布。统计量体现了词和类别之间的相关性，值越高，词和类别之间的独立性越小，相关性越强，那么就表示这个词对该类别的贡献占比越大，同时表示包含该词的文本属于该类别的概率越大，这会给之后的主体提取中提供很大的便利。而如果的值为0则表示词和类别不相关。

### 2.4.4 特征加权选取

在特征选择之后，需要给特征进行加权处理，以相应选择出实验所需的特征倾向。本实验采用TF-IDF算法，该算法由Salton首次论证提出，主要核心思想是：一个特征词在文本中出现的频率越高，说明这个特征词对文本的影响越大，区分文本内容的能力越强；一个特征词在文本中出现的类别越多、范围越广，说明它表示文本不同的能力越弱。TF-IDF算法在自然语言处理领域应用范围很广，其主要计算公式为：

其中，表示特征项权重；表示特征项在文档中出现的频率；表示特征项的文档倒数。

在本文中，会对微博的标签的加权做特殊处理，因为在使用过程中，发现标签携带的信息最终能够概括成博文的主题的概率比较大，所以结合具体情况来看，在特征加权这一步会给予以“#”构成的标签更高的权重。

### 2.4.5 主题提取

在数据挖掘相关研究中，主题概率模型有很多种。BLEI等在2003年提出了LDA（Latent Dirichlet Allocation）模型，是目前使用最为广泛的概率主题模型。随后诞生的CTM（Correlated Topic Model）模型，与LDA类似，但CTM只能用来研究成对主题间的相似性。基于此局限性，Li等在2006年提出了PAM模型，不仅可以描述词之间的相似性还可以描述主题之间的相关性，在挖掘特殊文本结构的热点时，相比LDA模型，PAM概率主题模型具有更强的表达能力。然后王亚民、苏凯等提出了BTM模型，克服了短文本的稀疏性问题但由于BTM的建模速度较慢等因素，相比其他概率模型，较容易产生冗余主题。

在本文，因为课题研究的特征主要是时间和空间特征，对主题的划分精度不是特别严格，所以本文选择了LDA主题模型来提取博文的主题。

#### 2.4.5.1 LDA模型

LDA主题模型算法的理论基础是贝叶斯理论。LDA根据词的共现信息的分析，拟合出词语—文本—主题的分布，进而将词、文本都映射到一个语义空间中。主题模型认为每个文本拥有一个或多个主题，而每个主题都有对应的词分布，通过概率选取出主题，可以得到每个文本的分布。主题模型将没有直接联系的词语与文档使用主题将两者串联起来。

LDA是一个多层的产生式概率模型，包含词、主题和文档三层结构。通过浅层的主题将词和文档关联起来。文档可以由潜在主题的多项式分布来表示，主题可以由词语的集合的多项式分布来表示。文档中的每一个主题Topic的分布都是基于Multinomial分布，先验是基于Dirichlet分布（Multinomial分布的共轭先验）；同样，每个主题Topic下单词都是基于Multinomial分布，先验是基于共轭先验的Dirichlet分布[23]。

图示

描述已自动生成

图2.2 LDA模型图

在本文中，由于数据获取的过程是使用关键字匹配进行爬取的，所以很大程度，博文的语义和关键字并不匹配，需要使用主题提取算法，将博文的语义浓缩成多个主题词，这样能快速得到相关的博文信息，便于后续研究进行。

由于微博这种属于短文本、关键词表现特征稀疏的预料，对以处理长文本的传统LDA模型来说不太合适，所以我们在此基础上使用了基于相关性改进的LDA模型来进行主题提取。

#### 2.4.5.2 基于相关性改进的LDA模型【】

改进后相关的计算公式为：

其中，表示主题，表示语料库中的词语，表示需要调整的参数。即表示在参数为的情况下，主题与语料库中词语的相关性。表示词语被划分到主题的概率。表示词语在主题分布矩阵中的边际概率。

当被调整到接近1时，该主题下词频多少更能体现主题特征，此时就接近于传统的LDA模型；当被调整到接近0时，更倾向选择具有排他性和特殊性的即词频较低，具有其他特殊特征的词语作为主题的表征词语。

引入改进后的LDA模型的优势在于**：**

（1）新增了参数，是一个可变参数，在[0,1]之间调整。通过调整参数，选择恰当的词语作为表征话题，从而得到每条博文的主题分布矩阵。

（2）改进后的模型能够避免短文本过低的词频带来的影响，提高主题提取的准确度。

### 2.4.6 相似度计算

相似度是衡量两条博文之间关系的重要指标。相似度计算一般会考虑词语之间、句子之间、段落之间的相似程度，对于微博这种中文短文本、发表格式不统一的文本来说，我们需要着重考虑语义的相似程度。正常情况下，文本表达出来的主题一样，这样的文本基本会被认为是相似度很高的，那我们就能将其划分成一个主题。

#### 2.4.6.1 基于Cosine 相似度的相似度计算

基于Cosine 相似度的相似度计算的原理是：两个向量空间中两个向量夹角间的余弦值作为衡量两个个体之间差异的大小，余弦值接近1，夹角趋于0，表明两个向量越相似，余弦值接近于0，夹角趋于90度，表明两个向量越不相似。

余弦相似度用向量空间中两个向量夹角的余弦值作为衡量两个个体间差异的大小。相比距离度量，余弦相似度更加注重两个向量在方向上的差异，而非距离或长度上。

余弦相似度的特点：

（1）对用户的绝对的数值不敏感。

（2）计算时不考虑数据之间的共同评分项数量，即使仅仅有极少相同评分项，也有可能获得很大的相似度结果。

（3）只要各个评分项之间越趋向于对应成比例，而不论数值差异如何，则相似度越趋近于1.000。

余弦相似度更多的是从方向上区分差异，而对绝对的数值不敏感，更多的用于使用用户对内容评分来区分用兴趣的相似度和差异，同时修正了用户间可能存在的度量标准不统一的问题。

# 3 技术路线分析

## 3.1 数据获取

数据获取阶段主要是用python的requests、beautifulsoup、lxml等工具包编写爬虫程序。目前爬虫主要流程都已比较成熟：分析目标网站的Url构成特点；分析网页html Dom树的结构，找到所需信息的路径；构建请求header，使用requests包构建request请求网站，使用beautifulsoup包解析得到的网页文件，按照xpath找到对应的信息；将结果保存起来，并构建下一个请求。

## 3.2 数据预处理

### 3.2.1 分词和过滤停用词

分词工具：python的jieba扩展库。jieba分词算法使用了基于前缀词典实现高效的词图扫描，生成句子中汉字所有可能生成词情况所构成的有向无环图(DAG)，再采用了动态规划查找最大概率路径，找出基于词频的最大切分组合，对于未登录词，采用了基于汉字成词能力的HMM模型，使用了Viterbi算法。jieba分词算法有三种模式：精确模式、全模式、搜索引擎模式。从功能上烤炉，我们使用精确模式就能满足实验的要求。

停用词表：使用哈尔滨工业大学停用词表、百度停用词表、四川大学机器智能实验室停用词表、中文停用词表结合起来扩展的停用词库，同时结合搜狗词库中的诸多专业词汇表作为用户词库。对于停用词过滤的要求，要求能够将微博中人们大量使用的语气词去除的同时，还要将不在用户词库中的出现频率极低的词语去除掉。另外由于本实验不对博文的情感进行分析，所以我们对表达语气的词语也会过滤，例如“好”、“非常”等程度副词。

## 3.3 主题提取

TF-IDF和LDA模型建立，本实验都依赖于sklearn第三方库。sklearn是python的机器学习库，具有各种分类、回归、聚类算法，包括支持向量机（SVM）、k-means等等算法，在各种科研领域被广泛使用。它包含了从数据预处理到训练模型的各个方面，拥有可以用于监督和无监督学习的方法，其中大部分可以分成估计器和转化器两类。其中转化器可以用于对数据的处理，例如向量化、数据降维、特征选择、特征加权等等。

## 3.4 数据可视化

本文选择pyecharts作为可视化工具。pyecharts是一个用于生成Echarts图表的类库。Echarts是百度开源的一个数据可视化JS库，主要用于数据可视化，凭借着良好的交互性，精巧的图表设计，得到了众多开发者的认可。pyecharts实际上就是Echarts与Python的结合，可以实现用后端代码实现在前端展示图表效果的功能。Pyecharts提供了30多中常见图表，以及丰富的地图文件，可以为地理数据可视化上提供强有力的支持。

# 实验与结果分析

## 4.1 流程概述

图示

中度可信度描述已自动生成

图4.1 流程概述图

如图4.1，在实验中我们首先根据关键词在微博上检索所有包含该关键词的博文并爬取对应的数据信息。由于仅包含关键词的博文并不一定描述相同的互联网事件，所以我们需要提取出相应的主题词语并进行相似度计算才能真正将一个互联网事件提取出来，之后再根据时间和空间特征进行数据分布的研究。

## 4.2 实验过程

### 4.2.1 数据获取

图示

描述已自动生成

图4.2 数据获取流程图

如图4.2，在实验中，为了避免在爬虫过程中被反爬机制打断，我们设置了一个cookie池，用来不断更新请求的cookie信息，减少一个cookie爬取的频率。同时我们设置了每隔一定时间随机停顿几秒钟的设定，以便让爬虫能够顺利爬取更多的信息。

由于微博的特性，空间特征即发布位置，不是每条博文都会展示，所以我们选择通过该用户最近10条的博文来粗略估计他（她）发帖时所在的位置，如果仍未获取到位置信息，将在数据预处理阶段过滤这种博文信息。

本次实验中，选取了从2021年12月1日到2022年3月31日的包含关键词“疫情”的相关博文，收到微博本身API访问限制、爬取时间过长连接不稳定等等因素，一共爬取到599639条，其中包含完整时间和地理位置信息的共有62373条，表明大部分用户并不会在博文中带上自己的地理位置信息。相关实验图片如下：

文本

描述已自动生成

图4.3 数据爬取总数

文本

描述已自动生成

图4.4 数据爬取符合要求数

### 4.2.2 数据预处理

图形用户界面, 图示

描述已自动生成

图4.5 数据预处理流程图

预处理流程中，提取出了博文的话题标签，由于该信息在微博中带有非常重的主观情感，具有提炼关键信息的作用，我们会在TF-IDF计算中为其特别增加权重，所以在预处理阶段就将话题标签提取出来。

在结合四种停用词表后，经过去重，停用词库已经扩展到了2312个，经过测试，已经能够将寻常出现的语气词、程度副词、中英文的上下文过渡词等等过滤。



图4.6 停用词表

在经过过滤及分词之后，我们得到了噪声非常少的数据集合，如图4.7。

报纸上的文字

描述已自动生成

图4.7 预处理后的数据

### 4.2.3 词频统计

该过程是在获取后所有的文本中统计词语出现的频次，以便在这文本中粗略统计各种词语分布情况。从62373条博文中，我们总共获取了1952574个无重复的词语。在进行词频统计之后，我们得到了表4.1。

|  |  |  |  |
| --- | --- | --- | --- |
| 序号 | 词语 | 出现次数 | 出现次数/总词语数 |
| 1 | 疫情 | 84061 | 4.31% |
| 2 | 加油 | 26749 | 1.37% |
| 3 | 上海 | 15391 | 0.79% |
| 4 | 西安 | 15090 | 0.78% |
| 5 | 防控 | 13003 | 0.69% |
| 6 | 核酸 | 11770 | 0.67% |
| 7 | 希望 | 9409 | 0.48% |
| 8 | 结束 | 7767 | 0.40% |
| 9 | 检测 | 7763 | 0.40% |
| 10 | 深圳 | 7133 | 0.37% |
| 11 | 工作 | 7099 | 0.37% |
| 12 | 病例 | 6740 | 0.35% |
| 13 | 2022 | 6713 | 0.34% |
| 14 | 隔离 | 6652 | 0.34% |
| 15 | 确诊 | 6627 | 0.34% |
| 16 | 人员 | 5965 | 0.31% |
| 17 | 谭方宜 | 5813 | 0.30% |
| 18 | 新冠 | 5706 | 0.30% |
| 19 | 生活 | 5490 | 0.28% |
| 20 | 抗疫 | 4855 | 0.25% |

表4.1 出现频次前20的词语

在得到词频统计后，我们绘制出了相关的词云图，如图4.8所示，在众多和“疫情”相关的博文中，出现最多的词语有疫情、核酸、防控、上海等词语。

图片包含 播放器, 前, 球, 游戏

描述已自动生成

图4.8 词云图

### 4.2.4 主题提取

在多次通过困惑度曲线初步确定主题数范围为30个，=0.3时取每个主题下的前30个表征次进行分布展示。聚类结果，以Topic1为例，如图4.9所示，LDA模型提取出的50个主题及其对应的部分表征词如表4.2所示。

图表

中度可信度描述已自动生成

图4.9 LDA聚类可视化结果

|  |  |  |  |
| --- | --- | --- | --- |
| 主题 | 表征词 | 主题 | 表征词 |
| Topic1 | 防控、措施、做好、口罩、管控 | Topic16 | 春天、解封、家里、三月、上班、几天、杭州、演唱会 |
| Topic2 | 感觉、喜欢、事情、炬火、朋友 | Topic17 | 学生、老师、面对、老百姓、开放、反复、警察、受伤害 |
| Topic3 | 阳性、检测、奥密克戎、疾控中心、转运 | Topic18 | 回家、快点、在家、早上、好吃、假期、感恩、咖啡 |
| Topic4 | 防疫、志愿者、物资、苏州、蔬菜、现场、 | Topic19 | 医院、患者、生命、妈妈、医生、父亲、透析、孕妇、救治 |
| Topic5 | 求助、市场、企业、公司、发展、能力、十九、未来 | Topic20 | 西安、抗疫、武汉、学院、春暖花开、青年、退散、同心、咸阳、共抗 |
| Topic6 | 无症状、感染者、本土、境外、输入、治愈 | Topic21 | 小区、感谢、广州、辛苦、宁波、谢谢、文学、封闭、致敬、领导 |
| Topic7 | 郑州、血透、小时、控制、高速、行程 | Topic22 | 结束、早点、长沙、无碍、廊坊、宅家、抑郁症 |
| Topic8 | 视频、微博、早日、平安、希望、宝贝、幸福、平平安安 | Topic23 | 加油、吉林、快递、赶紧、隔离、战胜、泉州、爸爸、居家、可怕 |
| Topic9 | 郴州、谭方宜、桂阳、众志成城、资兴、联通、国旗 | Topic24 | 期间、转发、南京、日常、重庆、济南、超话、打卡、拜托 |
| Topic10 | 上海、浦东、外溢、辟谣、浦西 | Topic25 | 百合、回来、昨天、情绪、祖国、度过、回复、打电话、关心、飞行、想念 |
| Topic11 | 香港、国家、病毒、疫苗、英国、影响、接种、导致 | Topic26 | 天津、国际、记录、开学、心情、顺利、江苏、校园、和平 |
| Topic12 | 中国、旅游、美国、旅行、出门、张真源、文化、江湖、消费 | Topic27 | 哈尔滨、人间、永远、难过、emo、珍惜、明天、听课、飞机、演出 |
| Topic13 | 深圳、学校、孩子、手机、沈阳、燎原、何去何从、考试 | Topic28 | 长春、期待、买菜、病人、外卖、印象、过年、日本、新年、愿望、春节、航班、紧张 |
| Topic14 | 新闻、发布会、家属、战争、昆明、灾难、介绍、突发 | Topic29 | 人民、全国、内衣、封城、过后、创文、家乡、资兴市、网课、联盟 |
| Topic15 | 北京、大学、电影、分享、评论、大学生、一周、美食、外面、封校 | Topic30 | 确诊、今日、新增、病例、韩国、海外、昨日、澳大利亚、人数、累计、增加 |

表4.2 LDA提取微博主题及其表征词

文本

描述已自动生成

图4.10 LDA参数设置

实验参数见图4.10，模型评估部分本文使用困惑度来评估。perplexity是一种信息理论的测量方法，是用来度量一个概率分布或概率模型预测样本的好坏程度，通常被用来比较概率模型。在经过多次比较后，我们绘制出了困惑度曲线，如图4.11，发现在20到30区间困惑度较低，模型拟合效果比较好，综合实际聚类效果考虑下我们选择了主题数为30作为最终主题参数。

图表, 折线图

描述已自动生成

图4.11 困惑度评估

在LDA过程中，模型有文档-主题的概率分布输出，本课题选择了三个文档将其选出，将其列出，人工确认了下每个文档和选出的topic的相关关系，发现有些关联程度及其紧密，能够包含许多主题的表征词；而有些博文则与确认的主题相关程度比较低，观察其概率分布，发现这类的博文的主题概率分布比较均匀，在30个主题概率分布中没有明显的分布峰值，而这类博文的原内容特点是带了相关“疫情”的标签，但实际上是转发的内容，所以这类的博文主题提取的效果比较一般。

表格

描述已自动生成

图4.12 文档-主题概率分布图

文本

描述已自动生成

图4.13 文档-主题对应关系图

### 4.2.5 相似度计算

可以观察到，从LDA获得的结果来看，有很多主题的表征词很多是重复的或者是同义词，而由于数据来源于包含关键词“疫情”的博文，所以很多博文提取出的主题是相似度比较高的，这种情况我们应该通过相似度计算，再次对主题进行聚类，同时更新上节中的文档-主题对应关系，真正提取出从“疫情”这个关键词下搜索出的互联网事件。

在LDA提取阶段会计算主题-词语概率分布矩阵，表明对于每个主题，表征词语在文本中的概率分布情况。在用余弦相似度计算时，我们选在将该矩阵的每一个n维向量都与其他向量做相似度计算，当结果大于阈值0.5时，可认为两个主题相似，对其进行合并。流程图如图4.14：

图示

描述已自动生成

图4.14 相似度计算流程图

实验结果如下：总共有30个主题，最后因相似而合并的有十几个主题。最终剩下的主题如表4.2.3所示。

|  |  |  |
| --- | --- | --- |
| 主题 | 与主题相应的描述 | 合并的主题 |
| Topic 1 | 没有疫情该有多好 | 无 |
| Topic 2 | 疫情上网课、线上考试 | 无 |
| Topic 3 | 疫情爆发、新增病例 | 无 |
| Topic 4 | 上海疫情情况 | 无 |
| Topic 5 | 即将清明，各地的疫情防控措施 | 无 |
| Topic 6 | 一名抗疫医护人员被一核电站员工因醉驾、超速等撞倒，不治身亡 | 无 |
| Topic 7 | 网友呼吁疫情赶紧结束 | 无 |
| Topic 11 | 奥密克戎的相关公开信息，呼吁人们加强疫苗接种，建设疫苗屏障 | 无 |
| Topic 14 | 想念疫情之前的日子 | Topic 12、Topic 13、Topic 15 |
| Topic 30 | 希望在疫情结束后能够恢复之前的生活，在节假日出去旅行、去演唱会 | 无 |
| Topic 23 | 在疫情期间，网友们居家办公或隔离的生活写真 | Topic 20、Topic 21 |
| Topic 22 | 各地网民支持抗疫，鼓励人们坚持戴口罩，做好防疫措施 | Topic 19、Topic 17 |
| Topic 18 | 位于中高风险地区的人们去做核酸 | Topic 24、Topic 29 |
| Topic 25 | 疫情已经开始三年了，什么时候才能结束 | Topic 9、Topic 8 |
| Topic 19 | 国内疫情、空难，国外战争，平安健康才是最宝贵的 | Topic 17 |
| Topic 16 | 祝福身处疫情爆发区的家人、朋友们身体健康、平安快乐 | Topic 26、Topic 28 |
| Topic 27 | 描述网友们疫情之后的旅游清单及其想做的事情 | 无 |

表4.3 合并后的主题及其描述

对表4.3中合并后的主题进行分类，分类结果如下：其中，Topic 1、Topic 7、Topic 14、Topic 30、Topic 27都是描述希望疫情尽早结束，不要再影响自己的生活；Topic 2、Topic 23是在描述在疫情期间人们的生活现状如何；Topic 3、Topic 4、Topic 5是在描述全国各地发生疫情及疫情的具体情况；Topic 6描述的是一位抗疫人员因为他人违反交通法规而失去生命，引起人们的关注、讨论；Topic 22、Topic 18可以归类为政府在疫情期间的防疫措施和网友们对其的态度；Topic 19、Topic 16描述的是网友们对别人和自己的祝福，希望在疫情以及其他灾难频繁出现的2022年中能够平安健康；Topic 25描述的是疫情持续的时间已经很长和它带来的影响是巨大的。

|  |  |  |  |
| --- | --- | --- | --- |
| 归类 | 描述 | 包含的主题 | 聚类在分类的微博数量 |
| 归类1 | 希望疫情结束 | Topic 1、Topic 7、Topic 14、Topic 30、Topic 27 | 17126 |
| 归类2 | 疫情期间生活现状 | Topic 2、Topic 23 | 12465 |
| 归类3 | 全国疫情严重情况 | Topic 3、Topic 4、Topic 5 | 4321 |
| 归类4 | 抗疫人员的不幸去世 | Topic 6 | 13054 |
| 归类5 | 抗疫措施和配合态度 | Topic 22、Topic 18 | 7532 |
| 归类6 | 祝福平安健康 | Topic 19、Topic 16 | 14642 |
| 归类7 | 持续时间长 | Topic 25 | 6032 |

表4.3 相似主题合并归类表

## 4.3 实验结果分析

从上节中的表4.2.4来看，在发表的微博数量上，主题为“希望疫情尽早结束”的相关话题在7个分类中占比最高，而主题为“各地疫情的严重程度”相关的话题占比最低。经过分析，原因是本文中选择研究的数据是带有地址的博文数据，而发布这些信息的一般为官方的微博账号，这种账号一般是不会带上私人的地理信息的，而普通民众在发微博时不会将各地的疫情写进自己的博文内容，所以在这里的数据分类显示占比比较小。在搜索了2021年12月至2022年3月间发生的疫情相关的事件后，发现在这期间，国内疫情一直在反复出现，深圳、北京、上海、哈尔滨、长春等城市因为疫情的出现，当地城市的人们出行就会收到限制，而这段时间正好处于旧的一年的末尾和新的一年的开端，再加上元旦、春节等假期的因素，人们渴望在这个时期能够像往常一样随意出行游玩，而不是受限在某一个地方。这就是“希望疫情尽早结束”的相关话题占比最高的原因。

### 4.3.1 时间维度

图表, 折线图

描述已自动生成

图4.15 全部主题事件的时间分布图

在Fink的危机四阶段模型理论基础上，将事件信息传播的生命周期分为征兆期、爆发期、持续期和恢复期。如图4.15来看，“疫情”这个大主题下的互联网事件总体上来一直是增加的，特别是在临近节假日的时候，例如元旦、春节、清明等，关于疫情的讨论话题会出现一个爆发的趋势。然后在节后关于疫情的讨论又会出现一个小幅度的下降趋势。这是因为在重大事件的发展周期中，关键事件的增长会引起用户行为的爆发式增长，同时结合用户的心理和情感，用户发表相关的博文的数量就会快速增长。

以元旦举例，在元旦前夕，关于希望出去旅行（Topic27、Topic1）的博文数量大幅增加，这是受了元旦假期的影响，生命周期就逐渐进入了爆发期。到了元旦期间，关于这类的博文数量一直保持较高的水平。在假期结束后，该话题热度稍有下降，开始进入了恢复期。但在1月5日深圳等地出现了本土的确诊病例，Topic3这样的话题又引起强烈讨论，又将Topic27、Topic1的热度拉高，生命周期又从恢复期之间变成了爆发期，热度也一度达到了之前爆发期时的峰值。这表明之前的互联网事件进入了另一个互联网事件的影响范围，并且互相了影响彼此的生命周期。所以从时间分布情况来看，不同主题的事件会在各个纬度影响与之相关的事件，所以在看待上文的事件生命周期特点，不能忽略别的事件的影响因素。

从总体的生命周期来看，因为疫情不断反复，同时由于2022年3月份上海等地的疫情比较严重，很多人都被要求隔离、居家办公、持续做核酸，会有比较强烈的诉求去微博等社交平台上发博，从以“疫情”为主题的事件的生命周期来看，目前一直处于持续期，处于增长趋势。

### 4.3.2 空间维度

图形用户界面, 应用程序

描述已自动生成

图4.16 讨论度前二十城市空间分布图

从图4.16来看，上海、西安、深圳、北京等城市在对于疫情的讨论中热度非常高，其中上海的讨论度在四个城市中遥遥领先。根据资料显示，在2022年3月22日左右，由于疫情感染人数突增，如图4.17所示，在当日，相关话题的博文数量及点赞、转发、评论等数量急剧上升。由此可见，由于疫情爆发的影响，在互联网上疫情相关事件的空间分布，有着聚集在事件爆发地周围的趋势。

而图4.16中所示其他城市，其中西安经查阅资料，虽然在数据采集期间没有出现比较大规模的确诊，但西安在2021年出现过因疫情导致封城的事件，所以位于西安的人们对于该类事件关注度更高，发微博的数量占比更大一些。由此可见，该空间分布与人们是否有类似的经历相关联。

而我们发现，出现在图4.16中的城市中，大多城市是省会或者直辖市、经济特区，属于经济较为发达的地区，从数据上显示，生活在这部分地区的人们在微博等社交平台上发博、发声的比例更大。由此可见，该空间分布与人们生活水平及表达意愿息息相关。

图形用户界面, 应用程序

描述已自动生成

图4.17 3月22日相关讨论空间分布

### 4.3.3 时空维度

结合时间维度和空间维度的分析来看，在微博上有关“疫情”的话题及相关事件在网络上的传播，受着诸多因素的影响。相关话题下的不同事件可以互相促进各自的生命周期发展，例如在元旦期间关于出行的事件和之后某地出现疫情的事件之间，后者促进了前者的发展，导致原本已处于恢复期的事件再次进入了爆发期，引起了广泛的讨论。其他因素，诸如不同地区的不同生活方式、该地区有没有经历过类似的事件等等，这些因素都会直接或间接地影响互联网事件的传播生命周期以及发展趋势。

# 总结与展望

## 5.1 完成的工作

本文完成了在中文社交媒体新浪微博上提取互联网事件的工作。中文互联网上包含了许多对提取事件有干扰的因素，同时也存在着许多内容、表达主旨相似的博文，本文排除了这些干扰因素的印象，并对提取出来的事件从时间和空间维度进行数据可视化分析，研究事件产生的原因、发展的趋势及其热度、讨论度改变的原因。

本文在提取微博上的互联网事件中主要进行了以下几方面的工作：

（1）预处理。预处理阶段排除了互联网社区常使用的emoji表情和各种视频、图片的链接，以及@别的用户相关的干扰信息。然后对过滤后的内容实用分词工具进行中文分词，然后从分词结果中过滤停用词。

（2）计算词频和选择特征关键词。在对预处理后的结果使用TF-IDF算法来选择加权后的关键词，同时进行词频统计分析。

（3）LDA建模分析。利用LDA概率模型数据中提取主题和主题表征词，然后利用提取出的30个主题来分析研究的数据集中博文的内容聚集在哪些事件上，并对其进行归类统计，利用LDA的文档-主题概率分布，将文档的主题确认为概率最大的主题。

（4）计算相似度。在对使用LDA模型提取出来的主题后，我们使用了基于余弦相似度的方法计算主题间的相似度，然后合并相似度大于0.5的主题，并将对应的事件合并。

在提取完互联网事件后，我们使用各种可视化工具研究相应事件的时间和空间分布，研究影响事件发展的因素有哪些和对时间、空间分布的影响。从实验结果来看，本文的研究是具有价值的。

## 5.2 不足和改进之处

本文在几个方面上还存在着不足的地方：

（1）预处理的程度还不够。微博博文中可能存在着一些隐藏的广告信息和外部链接，需要对这一步进行更全面的处理，才能得到更优质的数据。

（2）LDA建模过程中调整参数和迭代次数目前都不太完美。由于数据量规模很大的原因，在实验中迭代次数只设为了50，理想情况下，应该设为几百的迭代次数，能够训练出更加准确的模型。

（3）数据选择局限在了新浪微博。由于国内众多社交媒体中新浪微博的用户和内容数量较多，所以本课题选择的数据均来自于新浪微博，缺少了其他社交媒体的数据来源，所以在之后的研究中可以进一步探讨国内其余平台及国内外不同社交媒体的互联网事件发展变化，进行多维的对比分析。

总体上看，在这次课题中还有很多地方需要之后去努力完善，为研究互联网事件变化趋势提供有价值的参考。

# 参考文献

# 致谢