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](#_Toc99473262)

[ABSTRACT 4](#_Toc99473263)

[1 绪论 7](#_Toc99473264)

[1.1 研究背景 7](#_Toc99473265)

[1.2 研究的目的和意义 7](#_Toc99473266)

[1.3 国内外研究现状 8](#_Toc99473267)

[1.4 文本组织架构 9](#_Toc99473268)

[2 研究理论基础 10](#_Toc99473269)

[2.1 互联网事件概念 10](#_Toc99473270)

[2.2 互联网事件传播特点 10](#_Toc99473271)

[2.3 事件传播的生命周期 11](#_Toc99473272)

[2.4 事件主题提取 12](#_Toc99473273)

[2.5 基于余弦相似度的相关性计算 14](#_Toc99473274)

[3 技术支撑分析 15](#_Toc99473275)

[3.1 数据获取 15](#_Toc99473276)

[3.2 数据预处理 15](#_Toc99473277)

[3.3 数据分析工具 15](#_Toc99473278)

[4 实验流程 16](#_Toc99473279)

[5 实验结果与分析 17](#_Toc99473280)

[6 总结与展望 18](#_Toc99473281)

[7 参考文献 19](#_Toc99473282)

# 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。运用SVM支持向量机对实时获取的文本进行判别，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，是用改进的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【】；

### 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；

### 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；

## 文本组织架构

# 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[22]。

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 事件主题提取

### 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进行扩展的词表，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。

### 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。

### 2.4.5 使用LDA模型进行主题提取

// 用自己的语言形容，不然重复率肯定很高

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2.1 LDA模型图

## 2.5 基于余弦相似度的相关性计算

# 技术支撑分析

## 3.1 数据获取

## 3.2 数据预处理

## 3.3 数据分析工具

# 实验流程

# 实验结果与分析

# 总结与展望

# 参考文献