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](#_Toc98280967)

[ABSTRACT 4](#_Toc98280968)

[1 绪论 7](#_Toc98280969)

[1.1研究背景 7](#_Toc98280970)

[1.2研究的目的和意义 7](#_Toc98280971)

[1.3国内外研究现状 8](#_Toc98280972)

[1.4文本组织架构 9](#_Toc98280973)

[2 研究理论基础 10](#_Toc98280974)

[2.1 事件传播的生命周期 10](#_Toc98280975)

[2.2 事件主题提取 10](#_Toc98280976)

[2.3 基于余弦相似度的相关性计算 11](#_Toc98280977)

[3 技术支撑分析 12](#_Toc98280978)

[3.1 数据获取 12](#_Toc98280979)

[3.2 数据预处理 12](#_Toc98280980)

[3.3 数据分析工具 12](#_Toc98280981)

[4 实验流程 13](#_Toc98280982)

[5 实验结果与分析 14](#_Toc98280983)

[6 总结与展望 15](#_Toc98280984)

[7 参考文献 16](#_Toc98280985)

# 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. **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. **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

在互联网事件提取方面，如何快速提取出准确相关的事件并聚合是国内外学者关注的一个问题。1969年由Gerard和McGill两人首先提出著名的向量空间模型(VSM)，现在已经被广泛应用到文本分类、聚类等方面。使用SVM分类器来进行二元判断在各个领域都被国内外学者广泛使用。闵可锐等人采用SVM算法设计了一个话题识别与跟踪系统，该系统可以将海量的数据分类整理并聚合形成多个话题；Lapin M使用加权SVM的方案提高深度学习算法的预测性能和减少训练所需的数据量[14]；

对于使用社交媒体上的数据来研究，国内外学者研究的领域也是非常多。Signorini A等人，利用社交媒体Twitter获取的数据，结合传染性疾病领域的专业知识，研究疾病(H1N1)的当前疫情以及疾病的传播模式；Achrekar H等人更是通过Twitter数据对流行性感冒的传染发展趋势做出预测。Ferrari L等人，通过社交媒体数据的签到定位数据，结合城市规划相关知识，研究不同区域的居民行为活动模式；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[15]；

对于从海量数据中提取感兴趣的内容并加以研究分析上，研究人员在各个领域上进行了不同程度的尝试。例如，Anthony Stefanidis和Andrew Crooks在相关的研究中，通过对社交网络中数据的收集和分析，提出了一个获取地理空间信息的框架，以及由此产生的混合分析能力，以支持对人类活动相关的态势感知[16]。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[17]。David Crandall,Lars Backstrom等人利用网络上含有地理标记的照片，在全球范围内发现各种有趣的城市和地标建筑[18]。Yu Liu等人从遍布全国370个城市的50多万人的签到数据中提取人群在城市间的活动信息，分析旅行与空间相互作用的潜在模式[19]。Nagel等人通过研究网络空间消息与现实世界活动的相互关系，对Twitter中特定关键词汇的统计分析，发现了疾病的发生与微博数据中某些关键词汇的趋势分布存在着相关关系，验证了社交媒体在疾病监控方面的潜力[20]。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[21]。陈宁、彭霞、黄舟基于图片分享网站中来自用户分享的含有地理标签的图片信息，结合了空间聚类和文本语义挖掘等算法对热门景点进行提取，融合历史天气信息分析得到景点的热度分布规律，为游客提供相关景点的多角度游览规律[22]。在曹劲舟、武红宇的研究中，利用微博位置签到数据实现对城市中POI的快速更新。

## 1.4文本组织架构

# 研究理论基础

## 2.1 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。

## 事件主题提取

* + 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。

* + 1. **停用词过滤**

停用词是指在微博中使用次数很高，但没有实际的意义，只能表现发布者的一些心理活动的词语。这类词对于实验过程会有比较大的影响，过滤这些词语能够提高实验的效率，减小干扰噪音。

* + 1. **使用LDA模型进行主题提取**

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

## 2.3 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

1. 齐亚杰. 基于网络新闻流的热点事件检测与跟踪算法研究[D]. 杭州电子科技大学, 2018.
2. 刘林浩. 网络热点新闻事件挖掘和跟踪分析方法的研究与实现[D]. 中南大学, 2010.
3. 李想. 基于社交媒体的灾害事件提取与时空分析[D]. 兰州交通大学, 2018.
4. 周剑峰. 一种基于 FA-SVM 的热门微博特征选择及预测方法研究[J]. 计算机应用与软件, 2018, 12.
5. 李玮瑶, 赵凯. 基于特征提取的网络热点事件挖掘算法[J]. 计算机与现代化, 2015 (5): 17-20.
6. Lauw H W, Lim E P, Pang H H, et al. Social network discovery by mining spatio-temporal events[J]. Computational & Mathematical Organization Theory, 2005, 11(2): 97-118.
7. Lu X S, Zhou M C, Qi L, et al. Clustering-algorithm-based rare-event evolution analysis via social media data[J]. IEEE Transactions on Computational Social Systems, 2019, 6(2): 301-310.
8. 许旭阳, 李弼程, 张先飞等.基于事件实例驱动的新闻文本事件抽取[D]. , 2011.
9. 姚占雷, 许鑫. 互联网新闻报道中的突发事件识别研究[J]. 现代图书情报技术, 2011 (4): 52-57.
10. 高浩铭,洪磊,祁嘉诚,等. 新冠肺炎热点事件下的Twitter文本事件分析和用户画像[J]. 信息技术与信息化,2021(5). DOI:10.3969/j.issn.1672-9528.2021.05.006.
11. 刘晓东. 话题检测与跟踪系统的构建[D]. 北京邮电大学,2011.
12. 张美珍. 话题检测与跟踪算法的研究[D]. 2010. DOI:10.7666/d.y1781431.
13. Huang W, Liu H, Zhang Y, et al. Railway dangerous goods transportation system risk identification: Comparisons among SVM, PSO-SVM, GA-SVM and GS-SVM[J]. Applied Soft Computing, 2021: 107541.
14. 熊振华. 基于专业个人微博的事件提取研究[D]. 内蒙古科技大学,2015. DOI:10.7666/d.D665965.
15. Tversky A, Gati I. Studies of similarity [J]. Cognition and categorization, 1978, 1(19): 79-98.
16. Li Yong, Qian Mengjiong, Jin Depeng. Revealing the efficiency of information diffusion in online social networks of micro-blog [J]. Information Sciences, 2015: 293(1):383-389.
17. Wang Wenshuai, Du Ran, Cheng Yaodong. Topic mining method on massive microblog data [J]. Computer Engineering and application, 2014, 50(22): 32-37.
18. Lu Feifei, Xu Yabin, Li Zhuo, et al. Analysis of characteristics of social networks in terms of mocroblog impact [J]. Journal of Computer applications, 2013, 33(12):3359- 3362.
19. Guillermo, Jose Antonio, et al. Using latent semantic analysis and the predication algorithm to improve extraction of meanings from a diagnostic corpus [J]. The Spanish journal of psychology, 2009, 12(2):424-440.
20. Zhang Y, Tsai FS, Tang WY. Multilingual novelty detection [J]. Expert System With applications, 2011, 38(1):652-658.
21. Islam A, Inkpen D. Applications of corpus-based semantic similarity and word segmentation to database schema matching [J]. VLDB Journal, 2008, 17(5): 1293- 1320.