

本科毕业论文



论文题目： 基于网络爬虫和数据挖掘技术的

新冠疫情可视化分析

姓 名： 曲俊杰 学 号： 16307110125

院 系： 软件学院

专 业： 软件工程

指导教师： 徐迎晓 职 称： 高级工程师

单 位： 复旦大学

完成日期： 2020 年 5 月 25 日

# 目录

[目录 1](#_Toc41426265)

[摘要 3](#_Toc41426266)

[ABSTRACT 4](#_Toc41426267)

[第一章 引言 5](#_Toc41426268)

[1.1 新冠疫情可视化分析的研究背景 5](#_Toc41426269)

[1.2 项目结构概述 5](#_Toc41426270)

[1.3 章节结构介绍 6](#_Toc41426271)

[第二章 项目相关技术介绍 8](#_Toc41426272)

[2.1 网络爬虫spider技术简述 8](#_Toc41426273)

[2.2 可视化技术简述 9](#_Toc41426274)

[2.3 中文分词技术简述 10](#_Toc41426275)

[2.4 机器学习相关技术简述 11](#_Toc41426276)

[2.5 本章小结 11](#_Toc41426277)

[第三章 数据收集模块的设计和实现 12](#_Toc41426278)

[3.1 实际项目中的爬虫策略 12](#_Toc41426279)

[3.1.1 基于每日上报页面的爬虫策略 12](#_Toc41426280)

[3.1.2 基于搜索页面的爬虫策略 12](#_Toc41426281)

[3.2 针对搜狗搜索引擎的反反爬虫策略 13](#_Toc41426282)

[3.2.1 隐藏分页的反制策略 13](#_Toc41426283)

[3.2.2 包装真实URL的反制策略 14](#_Toc41426284)

[3.2.3 验证用户Cookie的反制策略 14](#_Toc41426285)

[3.2.4 封禁可疑IP地址的反制策略 14](#_Toc41426286)

[3.2.5 其他反反爬虫策略 14](#_Toc41426287)

[3.3 数据的具体内容和编排方法 14](#_Toc41426288)

[3.4 本章小结 14](#_Toc41426289)

[第四章 疫情数据的可视化分析 15](#_Toc41426290)

[第五章 疫情预测的相关尝试 16](#_Toc41426291)

[第六章 总结和展望 17](#_Toc41426292)

[参考文献 18](#_Toc41426293)

[致谢 19](#_Toc41426294)

# 摘要

2020开年，新冠疫情肆虐全球，给人们带来了空前的恐慌，短短三个月内，数以百万计的居民生命安全遭受威胁，亿万人的生活受到不同程度的影响，中国作为14亿人口的泱泱大国，充分使用了应对2003年非典疫情时的经验，采用地区强制隔离等政策抗击疫情，全国上下齐心协力，共同战疫。中国人民抗击疫情的坚定决心和显著成效给世界各国极大的鼓舞和信心。

项目将从当前世界关于新冠疫情的研究入手，简述该项目的知识背景和实现意义，通过网络爬虫的技术手段收集整理疫情数据，在保证数据准确性的基础上，使用可视化技术对数据进行多维度的直观呈现，并使用echart制图技术和javascript技术结合搭建每日疫情上报页面，使用机器学习相关技术做疫情发展趋势的简单预测。

此外，项目也对疫情中媒体的作用予以关注，使用相关度判别方法筛选整理出近千篇微信公众号文章，使用中文分词程序结合疫情中出现的具体问题作词频分析，利用wordcloud技术生成新冠疫情词图，揭示新冠疫情中媒体视角下的关注热点。

论文将对项目开发具体过程中遇到的问题和解决方案做一个总结，并对项目可供延伸改进的地方做一些猜想和尝试。

关键词 新冠疫情，网络爬虫，数据可视化，词频分析，机器学习

# ABSTRACT

# 第一章 引言

## 1.1 新冠疫情可视化分析的研究背景

2019年12月，新型冠状病毒肺炎（简称新冠肺炎，COVID-19）在中国湖北省武汉市爆发，中国政府迅速组织人力物力救援武汉，防控疫情。2020年1月，世界卫生组织（World Health Organization WHO）将新型冠状病毒疫情认定为“国际关注的突发公共卫生紧急事件”。

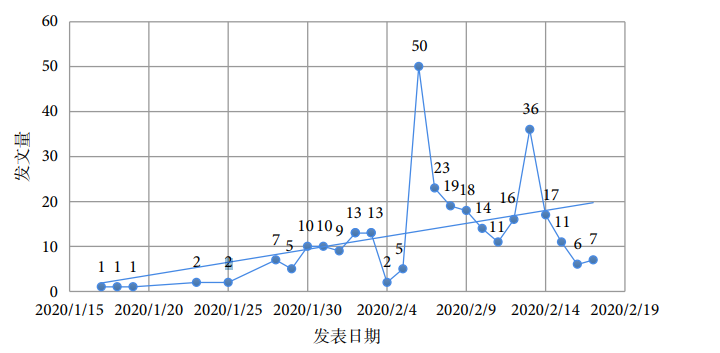
 新冠病毒爆发以来，各领域工作者积极投身科研工作，发表了大量相关研究论文，辅助实际防疫工作中的决策，为快速有效遏制疫情做出了巨大贡献，也为本项目提供了详实的理论基础。

图1-1 新型冠状病毒相关研究发表数量时间趋势

疫情期间，世界各国疫情相关数据都公开透明，提供了多种可供查看和收集的方法，本次项目也尝试整理收集了不同国家和地区的疫情相关数据供研究参考。使用适当的可视化方法处理这些数据，将使得疫情数据更加清晰直观，更具价值。

媒体是本次疫情中传导信息、增强信心的重要力量，当前关于新冠疫情的相关研究中数据和医疗研究占大多数，本项目对疫情中的媒体文章予以适当关注，分析出民众关注热点，为更好应对类似突发状况累积宝贵经验。

## 1.2 项目结构概述

本项目主要分为如下几个部分：

1. 数据收集部分，项目研究涉及的疫情数据分为两个主要模块，一是世界各国疫情相关人数（包括新增/累计确诊人数、新增/累计治愈人数、新增/累计死亡人数等）的汇总，通过官方提供的数据接口可以收集取得；另一模块是疫情期间的公众号文章，通过网络爬虫技术进行爬取。
2. 数据清洗部分，疫情相关人数按国家/地区/日期等维度进行分类，形成划分清晰的原始数据；而微信公众号文章则按相关度筛选出近千篇具有代表性的文章作为数据集。
3. 数据可视化部分，数据可视化是本次项目的核心部分，如何清晰直观地描述数据，使用了Matplotlib和echarts技术描述人数相关数据，展现其变化趋势并作对比总结；使用中文分词技术和词云技术处理微信公众号文章，作词频分析和热点分析。
4. 疫情预测部分，尝试用机器学习手段预测疫情发展的趋势走向，对本次数据集的构建和使用进行反思和思考，提出一些可行的改进方式。

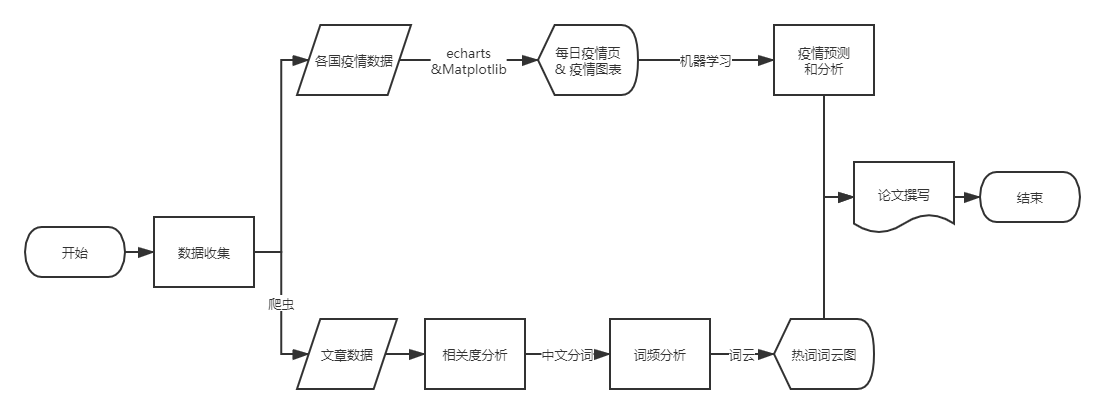


图1-2 “基于网络爬虫和数据挖掘技术的新冠疫情可视化分析”项目流程图

## 1.3 章节结构介绍

本文将以代码具体实现为基础，按照代码模块划分相关章节，具体说明项目实践过程中遇到的问题和解决方案，并融入对技术的应用和思考，从新冠疫情的第一手数据入手进行可视化探索。

**第一章 引言**

介绍当今新冠疫情的相关研究进展，以及数据的主要来源和论文的框架。

**第二章 项目相关技术介绍**

介绍论文的代码实现中使用到的具体知识。

**第三章 数据收集模块的设计和实现**

介绍利用网络爬虫收集数据中遇到的问题和解决方案。

**第四章 疫情数据的可视化分析**

介绍本项目中疫情数据的可视化手段及其优劣，以及不同可视化手段的技术应用分析、效果呈现。

**第五章 疫情预测的相关尝试**

介绍如何利用收集到的数据进行预测疫情发展趋势的尝试以及一些延伸内容。

**第六章 总结和展望**

本次研究项目的总结。

# 第二章 项目相关技术介绍

## 2.1 网络爬虫spider技术简述

自互联网问世以来，网络信息资源一直保持着迅猛增长的态势，为了便捷人们方便快速的获取有效信息，网络爬虫应运而生。简单来说，网络爬虫就是一个自动下载网页的自动化脚本，帮助我们在短时间内获取大量网页信息的工具。

根据系统结构和实现技术，网络爬虫可大致分为以下几种类型：通用网络爬虫、主题网络爬虫、增量式网络爬虫和深层网络爬虫。因为我们的目标信息相对明确，可以基于搜索引擎给出的结果进行爬虫，所以采用主题网络爬虫和深层网络爬虫结合的形式进行数据收集。

主题爬虫是按照预先定义的爬行主题，在给定初始URL种子集后，根据一定的分析算法，对爬行网页进行主题相关分析，过滤不相关网页，在不断抓取相关网页的过程中，将主题相关链接放入等待队列，直到满足设定条件为止。

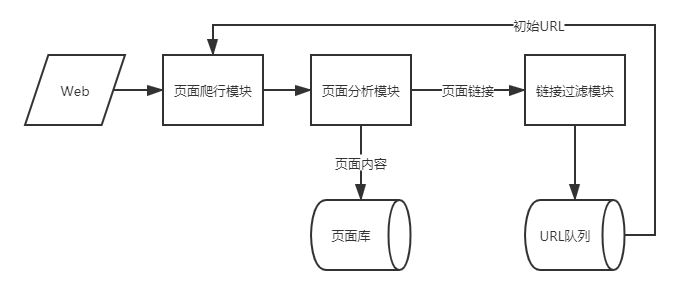


图2-1 主题网络爬虫体系结构图

关于主题爬虫的爬行策略——针对等待序列的访问顺序的多种处理方式以及主题相关性的判断是主题爬虫中的重点，常分为传统启发式方法、语义分析方法、经验爬行（基于前次爬行结果迭代）方法。如果我们基于搜索引擎的结果进行爬取，则可以将传统启发式方法和经验爬行方法相结合得到理想的结果集，直接利用搜索引擎标签的优先策略（热度、时间顺序等）免去部分设计优先方法的繁杂。

深层网络爬虫作为辅助手段，主要针对那些大部分内容不能通过静态链接获取、隐藏在搜索表单后的页面，这些页面需要用户提交特定关键词才能取得，例如用户注册后可见的页面，或是根据用户喜好有所区别的推荐页面等。

深层网络爬虫的重点是表单/COOKIE的构建填写，针对不同的网站需要不同的应对策略，将在第三章中作具体阐述。

## 2.2 可视化技术简述

我们使用爬虫技术收集来的数据大多数数字构建成的表格，缺乏直观性且数据特征不明显，需要使用可视化手段突显出其数据特征。图数据由于自身的结构特征，可以很好地表示事物之间的关系，便于多方向对比，使用Matplotlib库中的画图方法初步处理疫情数据可以形成多种图表，带来直观感受，也为后续的曲线拟合带来便利。

在大数据处理中，好的可视化分析还应该具备一定的交互性，以本次疫情地图为例，疫情地图极大了减少了数据呈现所需要的空间，增强了视觉效果，使用echarts技术和一些JS技巧可以帮助我们制作自己的疫情地图。

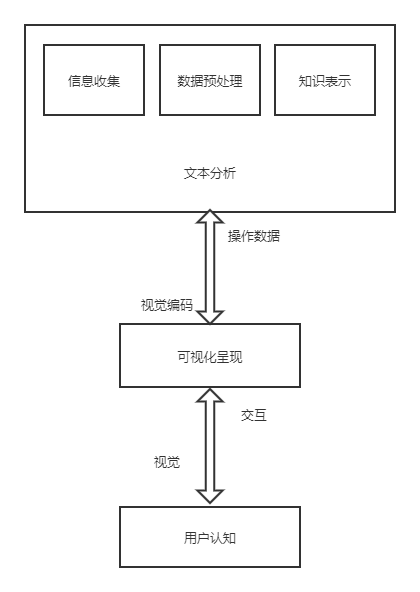
 本项目中还用到了文本可视化的相关知识，文本可视化服务于人们对文本进行分析和理解的基本过程，相当程度上依赖于自然语言处理技术，本项目中主要使用基于词频的文本可视化来探寻疫情发生以来的媒体热点，使用wordcloud技术构建词云图。

图2-2 文本可视化的基本框架

关于可视化技术部分的应用和思考将在第四章中做详细阐述。

## 2.3 中文分词技术简述

自然语言处理是人工智能的一个重要分支，中文分词是中文自然语言处理中的一项基础性工作，主要难点在于英文中的一个单词就是一个词，而汉语是以单个字为单位进行书写的，词语之间没有类似空格的区分标记，需要通过其他方法进行划分。

目前常见的中文分词方法可以分为三大类：基于字符串匹配的分词方法，基于理解的分词方法，以及基于统计的分词方法。随着大规模语料库的建立，统计机器学习方法不断研究和发展，基于统计的分词方法逐渐成为主流。

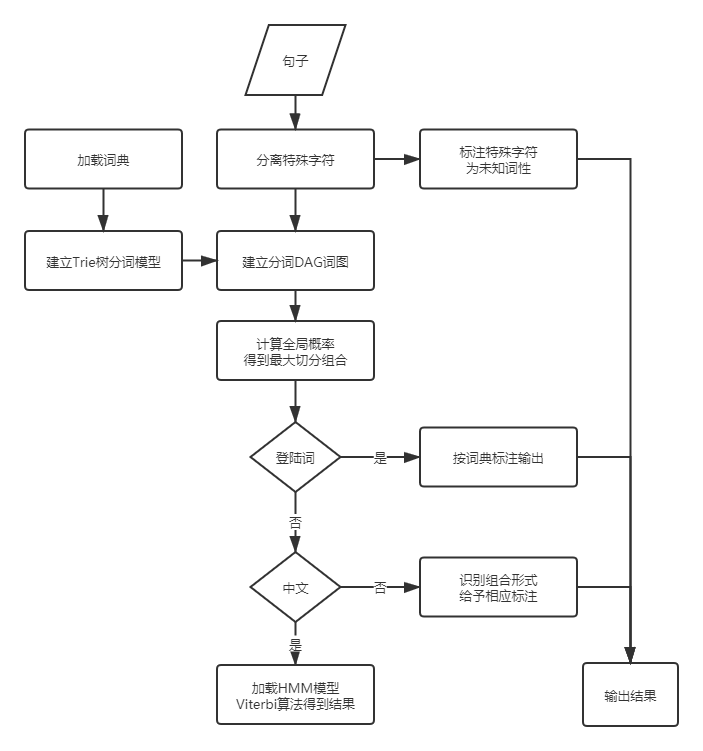
 项目中采用结巴中文分词的精确模式进行分词，结巴中文分词的优势在于在一般情形下精确度高，支持用户自定义词典（绑定某些较长的词），对于未登录词使用Viterbi算法独立成词，由于疫情中出现了很多新词汇，这种对于未登录词的敏感性正是我们所需要的。

图2-3 结巴分词流程示意图

## 2.4 机器学习相关技术简述

在项目的最后部分，尝试使用线性回归、贝叶斯曲线拟合、支持向量机（SVM, Support Vector Machines）等机器学习方法进行疫情曲线的拟合用于做疫情趋势的预测。

## 2.5 本章小结

本章介绍了本项目实施中所需要的主要技术及其运作基本原理，包括主题网络爬虫、结巴分词技术、词云可视化技术、echarts绘图可视化技术和SVM机器学习技术等。

# 第三章 数据收集模块的设计和实现

## 3.1 实际项目中的爬虫策略

### 3.1.1 基于每日上报页面的爬虫策略

 疫情人数相关数据，可以在百度、腾讯等疫情每日上报页面中爬取，由于数据都显示在主页面上，编写代码模拟访问该网页遍历相关变量名就可以汇总成数据表格，不需要二级跳转。

图3-1 以中国为例的疫情数据获取爬虫代码

这里使用到python的requests库模拟浏览器向服务器发送请求获取数据，对HTML网页文本直接解析就可以得到目标数据。

### 3.1.2 基于搜索页面的爬虫策略

疫情期间的公众号文章来源不一，需要更为复杂的爬虫策略，我们采用主题网络爬虫和深层网络爬虫相结合的方式进行爬取。

首先，从源URL库进行爬虫数据量太大，数据相关性不强，且不同网站的反爬虫策略不同，不利于我们编写通用爬虫代码，为了获取足量的有效数据，本模块将采用基于微信公众平台搜索功能的网络爬虫：若已知主题关键词，可以试用搜狗的微信搜索功能接口；若要收集特定公众号的文章，可以使用公众号内部的搜索链接进行搜索。

 通过浏览器的开发者工具，可以观察结果检索页面的网页结构，我们的一级目标是列表中文章URL，再使用xpath获取文章URL的详细值（超链接值），在搜索结果的对应项中，我们不难找到相应的链接信息。

图3-3 用于访问文章页面的真实URL

图3-2 搜狗搜索页面的文章URL信息

通过对比可以发现两个URL不是同一URL，抓包可以发现中间发生了一些跳转，这里搜狗搜索引擎做了一些反爬虫限制，如果我们补上https://mp.weixin.qq.com的域名访问图3-2中的URL，会跳转到访问出错页面，指出这是一个异常访问请求。我们需要用到一些反反爬虫手段解决这个问题。

## 3.2 针对搜狗搜索引擎的反反爬虫策略

搜狗搜索引擎使用的反爬虫策略主要如下几种：隐藏分页、包装页面URL、验证用户cookie、封禁可疑IP地址，本节简要介绍不同策略的反爬虫原理和应对手段。

### 3.2.1 隐藏分页的反制策略

搜狗搜索引擎对未登录的用户限制访问内容量，只允许未登录用户访问检索结果的前十页内容。

由于requests库是模拟浏览器访问搜索引擎页面，他继承了你使用浏览器的用户状态（不包括Cookie），因此我们只需要用个人账号在浏览器上保持登录状态，进行任意搜索即可。

### 3.2.2 包装真实URL的反制策略

包装真实页面URL是搜索引擎最常见的反爬虫策略，而且以40-50天为周期更换不同的包装方法，这种反扒策略成本较低且便于更换，可以有效拦截部分通配的自动爬虫脚本。

 我们可以在正常搜索页面中访问目标结果页，使用Fiddler抓取数据包观察这种数据结构。请求参数中除了URL等常规参数，还包含k和h参数，这两个参数的构造方式作为JavaScript脚本就放置在html文件的末端。

图3-4 搜索结果页中URL的包装方法

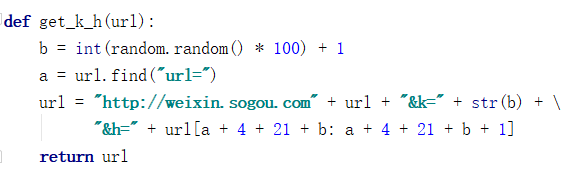
 因此我们在爬虫时模拟这种构造方式生成对应的k、h值补充在URL末尾，通过黏贴到浏览器测试，我们正常访问到了文章页面，则说明构造出的URL被正确回应。

图3-5 构造真实URL的代码

### 3.2.3 验证用户Cookie的反制策略

### 3.2.4 封禁可疑IP地址的反制策略

### 3.2.5 其他反反爬虫策略

## 3.3 数据的具体内容和编排方法

## 3.4 本章小结

# 第四章 疫情数据的可视化分析

# 第五章 疫情预测的相关尝试

# 第六章 总结和展望

# 参考文献

[1]张龙浩,李柏宏,贾鹏,蒲剑,白蓓,李音,朱培嘉,李雷,曾国军,赵欣,董珊珊,刘梦菡,张楠.新型冠状病毒（SARS-CoV-2）全球研究现状分析[J].生物医学工程学杂志,2020,37(02):236-249.

[2]孙立伟,何国辉,吴礼发.网络爬虫技术的研究[J].电脑知识与技术,2010,6(15):4112-4115.

[3]于娟,刘强.主题网络爬虫研究综述[J].计算机工程与科学,2015,37(02):231-237.

[4]唐家渝,刘知远,孙茂松.文本可视化研究综述[J].计算机辅助设计与图形学学报,2013,25(03):273-285.

[5]任磊,杜一,马帅,张小龙,戴国忠.大数据可视分析综述[J].软件学报,2014,25(09):1909-1936.

[6]龙树全,赵正文,唐华.中文分词算法概述[J].电脑知识与技术,2009,5(10):2605-2607.

[7]丁世飞,齐丙娟,谭红艳.支持向量机理论与算法研究综述[J].电子科技大学学报,2011,40(01):2-10.

[8]余豪士,匡芳君.基于Python的反反爬虫技术分析与应用[J].智能计算机与应用,2018,8(04):112-115.

# 致谢