西安電子科技大學

学　　院：计算机科学与技术学院

专　　业：计算机技术

选课班级：05 班

上课班级：05 班

学　　号：19031211696

姓　　名：吕亚龙

发明名称

网络爬虫在舆情监测中的应用研究

摘要

随着科技的发展，互联网已经成为人们生活中不可或缺的交流工具，针对现实生活中的某些热点、焦点话题，人们往往以信息化的方式发表自己的观点，给网络舆情监督工作带来一定的困难和挑战。

本文主要以网络爬虫在舆情监测中的应用为研究内容，以微博这一信息交互平台为研究对象，爬取微博用户对微博热搜榜前十的话题的评论，对评论进行情感分析，并进行数据展示，同时也设计了微博热搜榜榜首话题持续时间的趋势图。

近年来，随着移动终端的快速发展，人们对社交平台软件的需求越来越强烈，网络舆论已经成为社会舆论最重要的一部分，为了能够及时的发现舆情，并做出相应的应对措施，从而创造一个良好的网络环境。

摘要附图

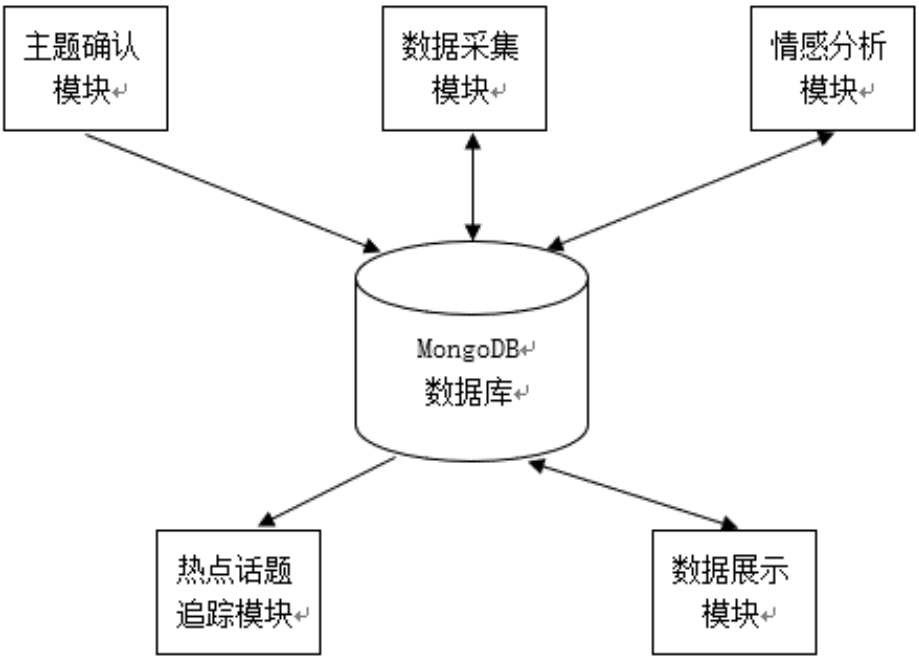
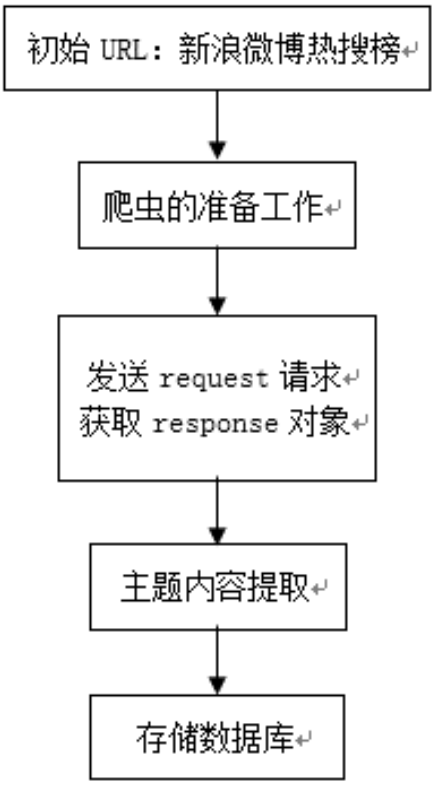


图 1 系统体系架构图



0 图 2 主题确认模版

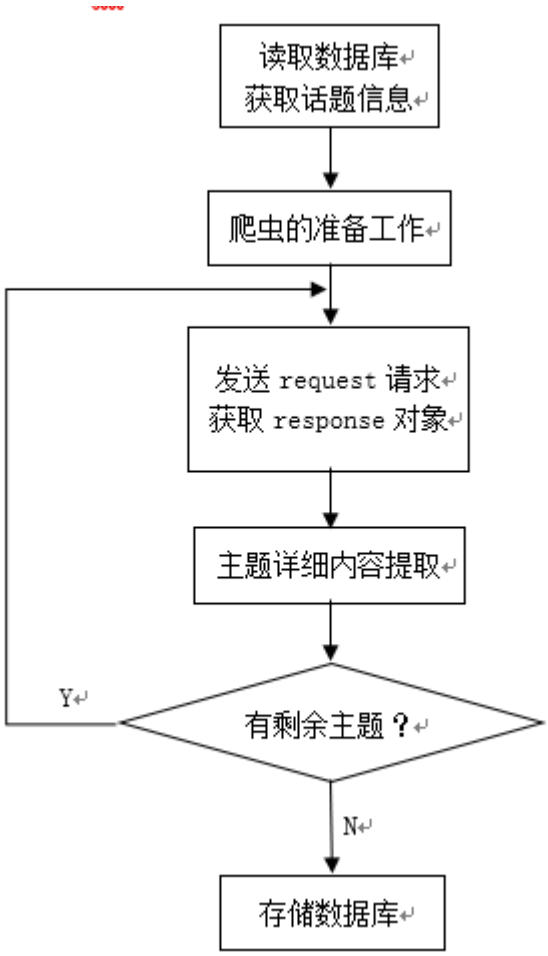


图 3 数据采集模块流程图

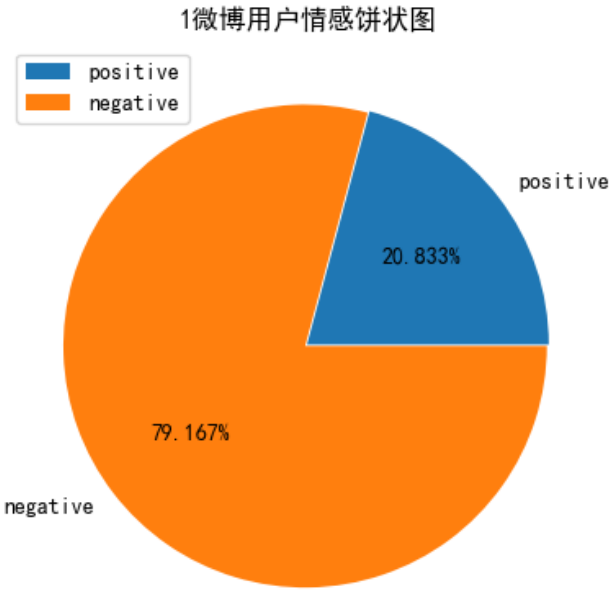


图 4 用户情感饼状图

权利要求书

1.基于社交网络的事件传播分析，其特征在于，具体步骤如下：

步骤一、通过编写分布式爬虫，从微博热搜榜上搜取微博热搜榜上的数据，并深入每一条热搜微博，获取与热搜内容相关所有微博数据，并存入数据库中；

步骤二、针对每一条微博都要采集相应的特征数据：热搜关键词、微博名称、微博图像、微博内容、微博的点赞量、转发量、评论量等；

步骤三、用抓取到的数据，结合理论研究，对于微博数据的特征：事件趋势、微博内容、意见领袖、核心传播人、传播途径、数据类型等数据进行整理；

步骤四、根据相关特征数据，绘制微博的事件趋势图、词云图、传播图、事件传播途径、意见领袖等图表；

步骤五、基于已绘制的图标和对于事件的统计分析结果、制作针对热搜关键词的微博事件分析报告。

2.如权利要求 1 中所述的基于社交网络的事件传播分析，其特征在于步骤一中分布式爬虫的编写。需要模拟登陆微博账户，解决微博账户的账户加密、验证码识别、整理微博页面上的数据并进行清理、获取有关数据、去除无关数据，于网络延迟，很有可能导致网页请求失败，进而中断请求进程，这样数据很容易失去，这里采用重试的方法再次请求，这样可以提高获取的数据量。批量微博账号导入，登录多个账号获取 Cookies，并维护一个 Cookies 池，每次请求随机从中得到 Cookies，这样缓解因集中访问导致数据获取失败。另外，还增加 IP 代理和随机用户代理功能，以次提高数据获取量。将微博账号存放在 Redis 中，登录后得到 Cookies 也存放在 Redis 中，并维护这个 Cookies 池，动态更新，保持请求的稳定。同时，将清洗好的数据存储在本地 Mysql 中。

3.如权利要求 1 所述的基于社交网络的事件传播分析，其特征在于步骤二中所述的事件微博特征包括：热搜关键词、微博名称、微博图像、微博内容、尾部的点赞量、转发量、评论量。需要对微博数据进行实时的跟踪分析。

4.如权利要求 1 所述的基于社交网络的事件传播分析，其特征在于步骤四中说的事件图表包括微博的事件趋势图、词云图、传播图、事件传播途径、意见领袖的绘制。

5.如权利要求 1 所述的基于社交网络的事件传播分析，其特征在于步骤五中的事件报告生成。采用百度开源可视化组件 ECharts 绘制柱状图、散点图等，结合 Django Web 框架以及第三方开发库 django-echarts 进行整合。

说明书

网络爬虫在社交在舆情监测中的应用研究

技术领域

[0001]　本发明属于数据分析领域，涉及一种社交网络的信息抓取和微博事件的传播分析方法。

背景技术

[0002]　近年来，互联网行业迅猛发展，以大数据、人工智能为首的数据驱动产品更是打破原有的现状，计算机行业正在迈向更高的阶段。科技在发展的同时，也给人们带来了新的交流方式和途径，逐渐的改变人们的生活方式，展现了不可抗拒的力量。截止到 2017 年 12 月，我国网民已有 7.72 亿，手机用户 7.53 亿，比例高达 97.5%。新浪微博用户发展报告[1]指出，新浪微博月活跃用户 3.76 亿，移动端占 92%。每天有数以亿计的网民活跃期中，通过微博发布、评论、转发各种新鲜事。

[0003]　如此多的用户，利用社交网络迅速分发的优点，在极短的时间内集聚网民的意见，影响着社会。随着大数据的越来越热门，这些社交网络上的数据也成为研究的重点。在国内，微博是社交网络的代表，微博上的事件也是实时的更新，体现着网络舆论的动向。微博作为一把双刃剑，让大家之间的相互距离越来越近，方便了与世界的交流，更及时的获取信息，也给社会安宁，言论安全带来新的挑战。同时，在自媒体时代，人人都是信息的创造者和传播者，各种信息层出不穷，给管理者带来极大地困难。

[0004]　关于微博的信息传播，在国内外已有不少的专家学者进行了研究。其中大多数的研究人员都是对信息的传播理论进行研究。传播树的生成算法，其中著名的有两个：独立级联模型和线性阈值模型，它们分别从概率和阈值的角度描述了影响传播的过程。也有研究人员应用信号分析框架，对分析模型进行分解，通

过信号进行话题识别。还有研究人员，通过挖掘微博文本信息，通过情感分析，识别微博的情感极性。

发明内容

[0005]　本发明爬取微博的热点事件，即被人们热搜的事件，抓取相关博文分析事件的传播路径，如何传播，以及可能对网上舆情带来的影响，重在解决整个事件的传播路径进行分析。最后，进行数据可视化，展示结果。通过 Web 页面的形式将抓取的数据，经过清洗，分析，形成图表等信息，生成微博事件传播报告。

[0006]　具体步骤如下：

[0007]　步骤一、通过编写爬虫，从微博热搜榜上搜取微博热搜榜上的数据，并深入每一条热搜微博，获取与热搜内容相关所有微博数据，并存入数据库中；

[0008]　步骤二、用抓取到的数据，结合理论研究，对于微博数据的特征：事件趋势、微博内容、意见领袖、核心传播人、传播途径、数据类型等数据进行整理；

[0009]　根据相关特征数据，绘制微博的事件趋势图、词云图、传播图、事件传播途径、意见领袖等图表。

附图说明

[0010]　图 1 为本发明基于社交网络的事件传播分析的微博数据采集方法的流程图。

具体实施方式

[0011]　下面结合附图对本发明的具体实施方法进行详细说明。

[0012]　本发明针对的社交网络是新浪微博，在目前的已有的理论基础上，试图通过现有阶段的微博内容，分析微博事件的时间起源和事件传播过程。通过各种数据特征，如点赞量、评论量、转发量、微博内容的分词处理等，进行综合统计分析，预测下一阶段微博事件的可能传播方向和传播趋势。

[0013]　具体步骤如下：

[0014] 步骤一、对微博页面进行分析，通过模拟登陆微博账户，首先解决微博账户的账户加密，验证码识别等问题，登陆成功后获取微博账户的 Cookies，利用 Cookies 不断请求页面微博页面的数据。

[0015] 步骤二、由于采集下来的数据很多是不完整的，多半需要自己再进行清理，比如微博正文部分，其数据中有很多是链接、表情、与话题无关数据等等，都需要自己的再次处理才能进行使用。

[0016] 步骤三、由于有登录的要求，微博有验证码的识别，针对验证码采用云打码平台，将采集到的验证码提交给第三方打码平台，这样可以提高验证码识别效率。

[0017] 步骤四、由于网络延迟，很有可能导致网页请求失败，进而中断请求进程，这样数据很容易失去，这里采用重试的方法再次请求，这样可以提高获取的数据量。批量微博账号导入，登录多个账号获取 Cookies，并维护一个 Cookies 池，每次请求随机从中得到 Cookies，这样缓解因集中访问导致数据获取失败。另外，还增加 IP 代理和随机用户代理功能，以次提高数据获取量。

[0018] 步骤五、将微博账号存放在 Redis 中，登录后得到 Cookies 也存放在 Redis 中，并维护这个 Cookies 池，动态更新，保持请求的稳定。同时，将清洗好的数据存储在本地 Mysql 中。

[0019] 步骤六、数据可视化主要是从 Mysql 数据库中取出微博热搜榜数据和对应的微博详情数据，利用 Echarts 绘制对应的图形，包括微博时段热搜，微博事件详情过程，散点图，事件传播树的生成。

[0020] 步骤七、事件报告将 Echarts 和 Django 结合，将可视化的图表和微博事件相结合，以微博事件报告的形式进行展示，可以查看热搜榜上对应的热搜微博对应事件在该时段的传播过程。

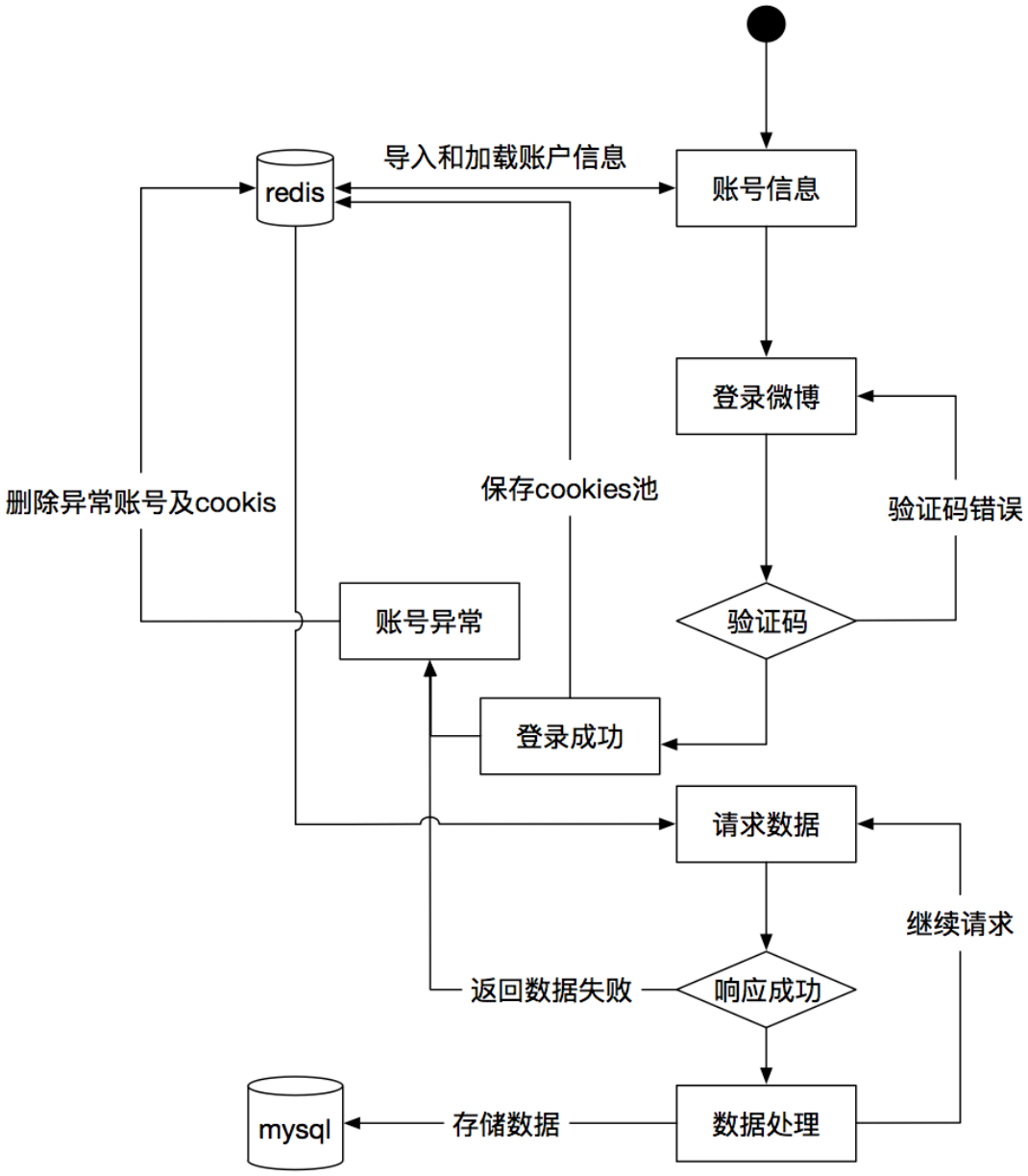


图 1 基于社交网络的事件传播分析的微博数据采集方法的流程图

专利来源

专利名称：网络爬虫在舆情监测中的应用研究

专利来源：本科毕业设计

目录