封面



毕业设计（论文）

基于数据挖掘的舆情警报系统

学生姓名： 唐可寅

学 号： 201710253110

所在系部： 电气信息系

专业班级： 17gb计算3班

指导老师： 王晓华 教授

日 期： 二O二一年六月

**Public opinion alarm system based on Data Mining**

By

Keyin Tang

June 2021

毕业设计（论文）任务书

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 系 部 | 电气信息系 | 指导教师 | 王晓华 | 职 称 | 项目老师 |
| 学生姓名 | 唐可寅 | 专业班级 | 17gb计职1 | 学 号 | 201710253110 |
| 论文题目 | 基于数据挖掘的舆情警报系统设计与实现 | | | | |
| 论  文  内  容  目  标  及  进  度  要  求 | 论文内容：  随着大量的舆论充斥网络，在互联网高速传播的帮助下，舆论己经成为一股非常强大的力量。由于网络具有虚拟的特点，一些非理性评论、负面报道或者谣言在网络上的传播极易激发网民的负面情绪，甚至会直接影响到社会大众对我国党以及政府的信任程度。因而，进一步深入研究基于数据挖掘技术的舆情监控系统的设计与实现具有十分重要的现实意义谐。  论文目标：  一、用户权限管理：用户登录注册，以及完善个人信息，用户发布信息新闻等等。  二、新闻管理：针对于用户发布的新闻进行检索，用过技术手段检测出危险字符，并加以拦截，同时向用户发出警告。  三、管理员模块：管理新闻关键字信息，管理用户相关信息，舆情信息等等。  进度要求：  2020年12月18号：学生完成选题；  2020年12月19号-2020年12月31号：由指导老师给学生下达任务书；  2021年1月4号-2021年1月17号：收集资料，完成开题报告撰写、外文文献翻译工作；  2021年1月18号-2021年1月31号：针对项目完成项目立项说明书、需求分析说明书、系统设计说明书（包含总体设计和详细设计）；  2021年2月1号-2021年3月28号（第5周前）：完成项目代码编写以及完成对应的编码进度表；  2021年3月29号-2021年4月25号：完善项目、进行深入研究，撰写论文初稿  2021年4月25号-2021年5月14号：指导教师对初稿进行批改，检查参考文献阅读情况，学生根据指导意见对初稿进行修改，并对项目进行进一步的完善。  2021年5月15号-2021年5月21号：论文定稿，项目完成，答辩资格审查，答辩资料准备，完成毕业设计（论文）重复率自查工作，并提供查重报告。  指导教师签名：  年 月 日 | | | | |
| 系 部  审 核 |  | | | | |

毕业设计（论文）学生开题报告

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 课题名称 | 基于数据挖掘的舆情警报系统设计与实现 | | | | |
| 课题类型 | B | | | 指导教师 | 王晓华 |
| 学生姓名 | 唐可寅 | 学 号 | 201710253110 | 专业班级 | 17gb计算3班 |
| 本课题的研究现状、研究目的及意义   1. 本课题的研究目的与意义   随着当代互联网和信息技术的飞速发展，网络技术的日新月异，网络已经逐渐融入到社会政治、经济生产和人民生活的各个方面，成为社会活动中不可或缺的一部分。信息技术的发展使得网络社交日益流行，并受到国内外广大群众的广泛关注和参与。每年通过在线社交网络注册的用户数都在不断增长。中国互联网络信息中心（CNNIC）于2020年9月29日发布第46次《中国互联网络发展状况统计报告》。报告显示，截至2020年6月底，我国总体网民规模达到9.60亿，互联网普及率为67.0%。网络化生存已成为这个时代的基本能力和基本事实。  随着第四代、第五代的移动信息系统（4G、5G）的快速发展和应用，网络已经成为继报纸、广播、电视后的“第四媒体”，发挥着强大、高效的舆论表达和传递能力。社交网络为每一位网络参与者都提供了信息交互的平台，每一位网络参与者都有能力在网络上成为一名自媒体人，网络媒体可以成为一个自媒体平台，但他们不需要通过政府主管部门的审批，没有记者证、出版物刊号、运营许可证。当代中国正处于中华民族伟大复兴的关键时期，各种社会矛盾仍然比较突出、严重。网络作为一个传递信息的高效率平台，会使舆论信息快速传递并扩散，并形成舆情，对社会稳定带来重大影响。网络参与者逐渐深度介入各类社会事件的发展和演变过程，网络舆情已成为当代社会稳定和谐的一种标志，是各国政府和社会关注的焦点。近年来，社会事件频繁通过网络舆情进一步扩散发酵，成为社会矛盾爆发的导火索，对社会稳定带来巨大影响，使得网络热点事件分析与研究对网络舆情监控管理以及预警具有十分重要的现实意义。  本次毕业设计（论文）希望设计一种基于数据挖掘的舆情警报系统，使用计算机科学行业各类前沿技术，并以大数据观念变革传统网络舆情管理思维，准确把握网络舆情的内在特征及其在演化过程中的潜在规律，对于新形势下做好网络舆情管理工作具有重要的理论意义和实践价值。   1. 本课题的研究现状：   舆论、舆情和网络與情是一个具有中国特色的概念，英语语言体系中对舆论、民意、民情与舆情等都不做严格区分，一般都使用“Public Opinion”表述。西方国家民众观点的表达是政治生活中自然构成的一部分，渠道畅通，西方舆情研究起源于社会学领域，是传播学的研究内容之一，主要集中在民众观点的传播与调查的机理方面，如总统选举、国会选举等选情预测，顾客对企业产品质量口碑和态度调查，较少出现中国热点事件网络舆情汹涌现象。  随着互联网和自媒体的普及，越来越多的政府，企业开始意识到舆情监测的重要性，中国的舆情行业开始蓬勃发展。网络时代的突发重大事件中，“权威发布”难以抢到第一时间，表面上看是事发单位部门和相关媒体的问题，实际上是传统的新闻传播规律已经不适应自媒体时代的信息传播速度。  政府舆情相对发展较快，企业舆情多与公关或客服绑定在一起，因为一个政府要处理的舆情事件比一个企业到多得多，对舆情各方面的需求度都比较高。但不论政府还是企业舆情，基本都是由于内部前期事件没有处理好，才逐渐走向网络（排除突发公共事件，如天津大爆炸、6·1东方之星旅游客船倾覆事件）。企业爆发的舆情，比如携程幼儿园事件、滴滴顺风车事件，都是内部发现问题处理，相关双方未达成和解，最终受害方寻求网络期望得到解决。也正是因此，企业舆情多侧重前期预防，而不是后期应对。而政府面对的突发舆情和不确定性，比企业相对要多。很明显会引发舆情的社会公共领域目前基本都已经显现，各个政府部门对此都有基本的认知，政府舆情目前更多是在后期应对，而不是前期预防。但是应对舆情的态度和处理水平有很大差别，从上到下，水平依次下降。  在线社会网络舆情是通过在线社会网络对各种社会现象、问题所表达的信念、态度、意见和情绪，往往具有极强的思想性和政治性，网络舆情一旦成为社会事件的放大器，凝聚共识，激发行动，必将影响社会。因此，无论是中央还是地方政府都非常重视网络舆情，科研院所和高校的学者从新闻学与传播学、图书馆情报与文献学、社会学、政治学、计算机以及这些学科的交叉等不同学科对网络舆情进行了相关研究。  关于网络舆情的分析与预警的研究，国外研究学者的研究方向主要集中在以下四方面：   1. 针对互联网对舆情或网络舆情的影响展开； 2. 侧重于危机事件应对和决策的研究，大多是以群体决策理论框架和批判理论角度出发； 3. 针对群体传播理论的网络传播行为研究； 4. 针对网络传播群体中人员社会背景、性别、角色等由此带来等影响的研究，关注重点大多是不平等和压迫现象； 5. 针对公众舆情对政府及政府对重大事件政治决策的评论及其影响。   国内研究学者的研究方向主要集中在网络舆情监控预警理论体系建立和完善、网络舆情预警等级指标体系建立和完善、网络舆情预警的政府应对策略研究这3个方面。 | | | | | |

课题类型： A-理论探究型；B-实践应用型。

|  |
| --- |
| 本课题的研究内容  随着大量的舆论充斥网络，在互联网高速传播的帮助下，舆论己经成为一股非常强大的力量。由于网络具有虚拟的特点，一些非理性评论、负面报道或者谣言在网络上的传播极易激发网民的负面情绪，甚至会直接影响到社会大众对我国党以及政府的信任程度。因而，进一步深入研究基于数据挖掘技术的舆情监控系统的设计与实现具有十分重要的现实意义。  本次毕业设计（论文）将首先对与本课题相关的关键技术，以及国内外相关研究作出介绍，然后阐述在实际工作中遇到的问题，并结合用户需求，提出有效的解决思路和方案，包括将可能使用的工具、方法等。利用诸如数据挖掘、文本情感分析、大数据技术、机器学习等行业前沿信息化技术手段实现一套集“新闻、论坛、博客、微博客等互联网信息实时采集与发现；信息传播渠道、跟踪与溯源；舆情监测分析与预警、发展预测与防范、引导与干预”于一体的高效、智能化业务工作系统。 |
| 本课题研究的实施方案、进度安排   1. 实施方案：   本次毕业设计（论文）为设计基于数据挖掘的舆情警报系统，将使用Java(JDK 1.8)语言进行开发，使用Spring Boot或Spring Cloud作为J2EE框架进行系统开发，使用MyBatis和MyBatis-Plus作为数据库ORM框架，使用Apache Shiro和JWT作为权限管理和安全认证框架，使用WebCollector、WebMagic、Htmlunit、jsoup等网络爬虫框架技术模拟浏览器采集国内外各大社交媒体平台用户、推文、评论等信息。系统使用上述各类技术并结合行业当前前沿算法、技术监测社交平台上的各类推文信息，检测舆论信息，并给出信息反馈，从而实现本次毕业设计（论文）要求等系统功能和能力。  本次系统设计预期划分6个功能模块，以帮助实现系统整体功能，系统各功能模块设计目标如下：   1. 数据采集模块：利用网络爬虫技术，爬取国内外各大社交媒体平台用户、推文、评论等信息，对数据编码、字段识别、字段抽取、数据去重等处理行动，最终数据入库； 2. 数据预处理模块：依据采集到的数据信息进行数据清洗、提取、降噪、分词等数据预处理行动； 3. 数据分析模块：对数据预处理模块得到的预处理数据，利用全文检索、文本情感识别、聚类等技术采取关键词提取、摘要提取、舆论信息辨别、推文和舆论信息特征分析、用户和推文影响力评级等数据分析行动； 4. 舆情预警和人工干预模块：利用行业前沿技术分别对推文和舆论信息特征结合数据分析行动产生的结果进行学习和分析，并最终产生舆论报警预警信息，同时提供人工对机器识别分析结果进行干预的能力； 5. 信息展示、报送和对外接口模块：提供新闻关键字、舆情信息、舆情报警预警等信息向平台用户展示的能力，提供相关信息主动向主管、监管部门主动报送的能力，提供对外公开接口允许第三方获取本系统服务信息的能力； 6. 系统基础运行和管理模块：提供后台管理能力，如系统参数管理、账户管理、用户登陆注册管理、权限管理、新闻关键字管理、舆论管理等。 7. 进度安排：   2020年12月18号：学生完成选题；  2020年12月19号 - 2020年12月31号：由指导老师给学生下达任务书；  2021年1月4号 - 2021年1月17号：收集资料，完成开题报告撰写、外文文献翻译工作；  2021年1月18号 - 2021年1月31号：针对项目完成项目立项说明书、需求分析说明书、系统设计说明书（包含总体设计和详细设计）；  2021年2月1号 - 2021年3月28号（第5周前）：完成项目代码编写及完成对应的编码进度表；  2021年3月29号 - 2021年4月25号：完善项目、进行深入研究，撰写论文初稿  2021年4月25号 - 2021年5月14号：指导教师对初稿进行批改，检查参考文献阅读情况，学生根据指导意见对初稿进行修改，并对项目进行进一步的完善。  2021年5月15号 - 2021年5月21号：论文定稿，项目完成，答辩资格审查，答辩资料准备，完成毕业设计（论文）重复率自查工作，并提供查重报告。 |
| 已查阅的主要参考文献   1. 会议类 2. Ge Hongbing. Research on the Construction of Early Warning Index System of Network Public Opinion Emergency Based on Computer Simulation[A]. 安徽工业大学.Proceedings of 2019 2nd International Conference on Manufacturing Technology, Materials and Chemical Engineering (MTMCE 2019)[C].安徽工业大学:香港环球科研协会,2019:6. 3. Yongchang Ren,Zhe Li. Internet Public Opinion Management in Big Data Era[A]. Institute of Management Science and Industrial Engineering.Proceedings of 2018 5th International Conference on Education,Management,Arts,Economics and Social Science(ICEMAESS 2018)[C].Institute of Management Science and Industrial Engineering:计算机科学与电子技术国际学会(Computer Science and Electronic Technology International Society),2018:5. 4. Yong-qing YUAN,Qin-ying WANG,Feng-ming LIU. Analysis of Internet Public Opinion Situation[A]. Advanced Science and Industry Research Center.Proceedings of 2018 International Conference on Modeling, Simulation and Analysis(ICMSA 2018)[C].Advanced Science and Industry Research Center:Science and Engineering Research Center,2018:6. 5. Fulian Yin,Beibei Zhang,Ge Su,Ruizhe Zhang. Research on the Public Opinion Pre-warning Based on the Logistic Model[A]. The Institute of Electrical and Electronics Engineers、IEEE Beijing Section.Proceedings of 2017 IEEE 8th International Conference on Software Engineering and Service Science[C].The Institute of Electrical and Electronics Engineers、IEEE Beijing Section:IEEE BEIJING SECTION(跨国电气电子工程师学会北京分会),2017:5. 6. MA Jing. Analysis on Governance of Government Online Public Opinion Big Data Perspective[A]. University of Ghana (UG)、Ghana Institute of Management and Public Administration(GIMPA)、University of Cape Coast (UCC)、American Society for Public Administration (ASPA)、University of Electronic Science and Technology of China (UESTC).Proceedings of 2017 International Conference on Public Administration(12th) &amp; International Symposium on West African Studies (1st)(Volume II)[C].University of Ghana (UG)、Ghana Institute of Management and Public Administration(GIMPA)、University of Cape Coast (UCC)、American Society for Public Administration (ASPA)、University of Electronic Science and Technology of China (UESTC):公共管理国际会议组委会,2017:6. 7. Xiaohong Jin,Hui Li,Hong Fu,Jing Zhang,Boda Ding. A Study on Early-Warning Response of Emergency Events Network Public Opinion Based on Analysis of Microbloggers' Behaviors[A]. IEEE Beijing Section，China、上海财经大学、同济大学.Proceedings of the 2016 IEEE International Conference on Progress in Informatics and Computing(PIC)[C].IEEE Beijing Section，China、上海财经大学、同济大学:IEEE BEIJING SECTION(跨国电气电子工程师学会北京分会),2016:4. 8. Zhao Liyong,Zhang Xu,He Min,Zhang Dan,Liu Wei,Liu Chunyang. Research on Public Opinion Index System of Chinese Microblog[A]. IEEE Beijing Section.Proceedings of 2014 IEEE 5th International Conference on Software Engineering and Service Science[C].IEEE Beijing Section:IEEE BEIJING SECTION(跨国电气电子工程师学会北京分会),2014:4. 9. 期刊论文 10. 杨菁,赵岩.网络舆情及其预警研究综述[J].江苏科技信息,2013(06):12-13. 11. 杨娜,褚茂胜.基于深度学习的网络舆情情绪监测系统设计[J].中国科技信息,2020(24):76-78. 12. 郑昌兴.关于突发事件网络舆情监测预警的探讨[J].电脑知识与技术,2020,16(33):66-68. 13. 梁泽鹏,李直达.基于机器学习的舆情预警模型的设计与实现[J].广东公安科技,2020,28(03):1-5. 14. 杨森,王黎,李超.基于主动监测引擎和大数据的网络舆情烈度演化分析系统[J].价值工程,2020,39(17):201-203. 15. 曾宇.基于大数据的网络舆情实时监测系统的构建[J].漳州职业技术学院学报,2020,22(02):92-99. 16. 段峰峰,陈淼.多元与智能:跨媒体网络舆情数据的分析与处理[J].传媒观察,2020(01):33-38. 17. 万倩,朱里越,欧阳峰.基于人工智能的广电舆情分析系统[J].广播与电视技术,2019,46(12):46-52. 18. 刘小满,王小辉.基于“互联网+”的网络舆情监控系统的设计与实现[J].电脑知识与技术,2019,15(31):37-40. 19. 白茹.基于云计算和Hadoop的网络舆情监控系统设计[J].电子设计工程,2019,27(16):141-144+150. 20. 洪小娟,宗江燕,于建坤,黄卫东.网络舆情监测系统的分析与设计[J].软件工程,2019,22(08):37-39+13. 21. 祁凯,彭程.基于OCS-EGM模型的网络集群行为监测及预警体系研究[J].情报杂志,2019,38(09):134-141+149. 22. 江瑾.网络舆情监控系统的设计和实现[J].信息与电脑(理论版),2019(13):63-65. 23. 王晨妮,王宇晨,张超,刘蓝静,孙钜晖,宋威.基于非结构化数据处理的网络舆情监测系统[J].科技创新与应用,2019(13):38-40. 24. 李猛哲,杨跃东.教育考试舆情监测指标体系设计[J].中国考试,2019(04):35-40. 25. Jian Li. Analyse the data tendency in the public opinion monitoring system[J]. Journal of Physics: Conference Series,2019,1187(5). 26. 潘晓英,陈柳,余慧敏,赵逸喆,肖康泞.主题爬虫技术研究综述[J].计算机应用研究,2020,37(04):961-965+972. 27. Chuiju You,Dongjie Zhu,2,Yundong Sun,Anshan Ye,Gangshan Wu,Ning Cao,Jinming Qiu,Helen Min Zhou. SNES: Social-Network-Oriented Public Opinion Monitoring Platform Based on ElasticSearch[J]. CMC: Computers, Materials &amp; Continua,2019,61(3). 28. 陈刚,李弼程,郭志刚,林琛.网络舆情监测预警系统模型与关键技术[J].信息工程大学学报,2019,20(01):116-121. 29. 董积有,李岸,王海艺.基于深度学习的高校舆情监测的探索与研究[J].计算机产品与流通,2017(08):195. 30. 张荣显,曹文鸳.网络舆情研究新路径:大数据技术辅助网络内容挖掘与分析[J].汕头大学学报(人文社会科学版),2016,32(08):111-121. 31. 刘丽员,杨昔阳.基于文本相关性的高校网络舆情监控系统的设计与实现[J].泉州师范学院学报,2016,34(02):50-54. 32. 柳向东,曹雨婷,李利梅.网络影响力预知模型:一种大数据下高校舆情监测与预警机制[J].深圳大学学报(人文社会科学版),2015,32(04):156-160. 33. 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. 34. 方星星,鲁磊纪,徐洋.网络舆情监控系统中主题网络爬虫的研究与实现[J].舰船电子工程,2014,34(09):104-107. 35. 刘红梅.垂直搜索引擎主题爬虫搜索策略研究[J].科技信息,2013(24):252-253. 36. 石鲁生,陈林,李凯.一种网络舆情的动态预警方法[J].天津师范大学学报(自然科学版),2012,32(02):59-65. 37. 张瑞.SMS网络舆情信息监控系统的设计与实现[J].现代情报,2012,32(03):68-71. 38. 王长宁,陈维勤,许浩.对微博舆情热度监测及预警的指标体系的研究[J].计算机与现代化,2013(01):126-129. 39. 宋海洋,刘晓然,钱海俊.一种新的主题网络爬虫爬行策略[J].计算机应用与软件,2011,28(11):264-267+293. 40. 叶昭晖,曾琼,李强.基于搜索引擎的网络舆情监控系统设计与实现[J].广西大学学报(自然科学版),2011,36(S1):302-307+334. 41. 高承实,荣星,陈越.微博舆情监测指标体系研究[J].情报杂志,2011,30(09):66-70. 42. 魏晶晶,杨定达,廖祥文.基于网页内容相似度改进算法的主题网络爬虫[J].计算机与现代化,2011(09):1-4. 43. 张立涛,綦振法.“五位一体”的网络舆论监控体系研究[J].中国软科学,2010(07):8-15. 44. 唐铁兵,陈林,祝伟华.基于Lucene的全文检索构件的研究与实现[J].计算机应用与软件,2010,27(02):197-199+230. 45. 施聪莺,徐朝军,杨晓江.TFIDF算法研究综述[J].计算机应用,2009,29(S1):167-170+180. 46. 陈丛丛. 主题爬虫搜索策略研究[D].山东大学,2009. 47. 黄旭,朱艳琴,罗喜召.基于内容评价的爬虫搜索策略研究[J].微电子学与计算机,2008(11):25-28. 48. 倪贤贵,蔡明.基于链接结构和内容相似度的聚焦爬虫系统[J].计算机工程与设计,2008(07):1709-1710+1763. 49. 蒋宗礼,徐学可,李帅.一种基于超链接引导的主题搜索的主题敏感爬行方法[J].计算机应用,2008(04):942-944+950. 50. 王娟.网络舆情监控分析系统构建[J].长春理工大学学报(高教版),2007(04):201-203+214. 51. 毛欣娟.网络舆论控制机制的构建及相关问题思考[J].江西公安专科学校学报,2007(06):49-52. 52. 刘金红,陆余良.主题网络爬虫研究综述[J].计算机应用研究,2007(10):26-29+47. 53. 梅中玲.基于Web信息挖掘的网络舆情分析技术[J].中国人民公安大学学报(自然科学版),2007(04):85-88. 54. 初建崇,刘培玉,王卫玲.Web文档中词语权重计算方法的改进[J].计算机工程与应用,2007(19):192-194+198. 55. 曾凡斌.重大突发事件中的BBS舆论特点与管理初探——对人民网“强国论坛”的个案观察[J].出版发行研究,2006(04):61-67. 56. 王甘武.用辩证的方法正确把握网上舆论导向[J].信息网络安全,2006(04):28-30. 57. 李正华.网络舆论监督的临界问题及相关法律责任[J].信息网络安全,2006(04):31-33. 58. 叶惠敏,戴冠中.基于综合集成方法的网上舆论倾向分析与评估系统方案[J].计算机工程与应用,2005(16):216-217+220. 59. 叶惠敏,唐三平.用于网上舆论观点抽取的几种方法[J].计算机应用研究,2005(05):256-257+260. 60. 张校乾,金玉玲,侯丽波.一种基于Lucene检索引擎的全文数据库的研究与实现[J].现代图书情报技术,2005(02):40-43+48. 61. 李晓明,朱家稷,闫宏飞.互联网上主题信息的一种收集与处理模型及其应用[J].计算机研究与发展,2003(12):1667-1671. 62. 鲁松,李晓黎,白硕,王实.文档中词语权重计算方法的改进[J].中文信息学报,2000(06):8-13+20. 63. 硕博论文 64. 王鑫. 基于BS模式的网络舆情监控系统设计与实现[D].电子科技大学,2020. 65. 段吉东. 基于热量模型的微博舆情实时监控系统研究[D].济南大学,2020. 66. 胡中皓. 基于规则和情感的法治舆情监测系统研究与实现[D].江西财经大学,2020. 67. 钱晨. 基于信息测度的舆情情感态势模型研究[D].南京邮电大学,2019. 68. 王晶. 关于舆情监测系统实时交互式可视化的研究[D].西安理工大学,2019. 69. 赵康. 面向主题的网络爬虫系统的设计与实现[D].北京邮电大学,2019. 70. 李寒冰. 基于深度学习的舆情监测系统[D].北京交通大学,2019. 71. 王天齐. 基于微博的涉警网络舆情风险评估模型研究[D].中国人民公安大学,2019. 72. 李迁. 公安部门互联网舆情监管系统设计与实现[D].大连海事大学,2017. 73. 柳淑婷. 基于社交网络的舆情关键技术研究[D].吉林大学,2017. 74. 赵丽梅. 网络舆情分析与监测系统的设计与实现[D].兰州大学,2017. 75. 黄宝成. 基于社交网络的舆情信息挖掘方法研究[D].哈尔滨工业大学,2017. 76. 李洪胜. 内容与链接结构相融合的主题爬虫研究[D].华侨大学,2016. 77. 林红静. 基于K-means的微博短文本聚类算法研究[D].海南大学,2016. 78. 罗路天. 垂直搜索引擎中主题网络爬虫算法的设计与研究[D].广东工业大学,2016. 79. 邹妍. 网络舆情监控与分析系统的设计与实现[D].吉林大学,2015. 80. 冯佳明. 深度词汇网络学习舆情监测关键技术的研究[D].北京化工大学,2015. 81. 邹建明. 面向广东基层公安的网络舆情监控系统设计与实现[D].电子科技大学,2014. 82. 张蕾. 基于机器学习的网络舆情采集技术研究与设计[D].电子科技大学,2014. 83. 董坚峰. 面向公共危机预警的网络舆情分析研究[D].武汉大学,2013. 84. 夏景隆. 基于版权服务的网络数据采集算法研究[D].北方工业大学,2013. 85. 周小丽. 基于网络爬虫和Lucene索引的互联网舆情监测系统设计与实现[D].吉林大学,2013. 86. 侯万友. 群体性突发事件微博舆情演化分析[D].哈尔滨工业大学,2013. 87. 吴健. 基于Hadoop的上市公司舆情挖掘系统的研究与实现[D].电子科技大学,2013. 88. 万源. 基于语义统计分析的网络舆情挖掘技术研究[D].武汉理工大学,2012. 89. 李斌. 基于Lucene的分布式全文检索在构件库系统中的实现[D].西安电子科技大学,2012. 90. 巨慧慧. 针对网络舆情的情感倾向性研究[D].哈尔滨工业大学,2010. 91. 蒋昌金. 基于关键词提取的中文网页自动文摘方法研究[D].华南理工大学,2010. 92. 杜阿宁. 互联网舆情信息挖掘方法研究[D].哈尔滨工业大学,2007. 93. 郑健珍. 定题爬虫搜索策略研究[D].厦门大学,2007. 94. 刘云峰. 基于潜在语义分析的中文概念检索研究[D].华中科技大学,2005. |
| 指导教师意见  指导教师签名：  年 月 日 |

学位论文原创性声明

本人郑重声明：所呈交的学位论文是本人在导师的指导下独立进行研究所取得的研究成果。除了文中特别加以标注引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写的成果作品。本人完全意识到本声明的法律后果由本人承担。

作者签名： 年 月 日

学位论文版权使用授权书

本学位论文作者完全了解学院有关保管、使用学位论文的规定，同意学院保留并向有关学位论文管理部门或机构送交论文的复印件和电子版，允许论文被查阅和借阅。本人授权省级优秀学士学位论文评选机构将本学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存和汇编本学位论文。

本学位论文属于

1、保密 □，在 年解密后适用本授权书。

2、不保密 □。

（请在以上相应方框内打“√”）

作者签名： 年 月 日

导师签名： 年 月 日

摘 要

正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文

关键词：数据挖掘 舆情 警报

Abstract

Times New Roman Times New Roman Times New Roman Times New Roman Times New Roman Times New Roman Times New Roman Times New Roman Times New Roman Times New Roman Times New Roman Times New Roman Times New Roman Times New Roman Times New Roman Times New Roman Times New Roman Times New Roman Times New Roman Times New Roman Times New Roman Times New Roman Times New Roman Times New Roman Times New Roman Times New Roman

**Keywords:** Data Option

目 录

目录

[摘 要 i](#_Toc917870524)

[Abstract ii](#_Toc1274860467)

[目 录 iii](#_Toc1135522750)

[1 引言 1](#_Toc43688361)

[1.1 二级标题 1](#_Toc1978359700)

[1.1.1 三级标题 1](#_Toc802171399)

[2 绪论 2](#_Toc192367127)

[2.1 课题研究背景与研究意义 2](#_Toc1151414754)

[2.2 国内外研究现状 2](#_Toc852627361)

[2.3 本课题研究目标 2](#_Toc2097163543)

[2.4 本课题研究内容 2](#_Toc378568990)

[2.5 论文组织结构 2](#_Toc1762452516)

[3 可行性研究 3](#_Toc1297493341)

[3.1 技术可行性 3](#_Toc1421630549)

[3.2 经济可行性 3](#_Toc441580521)

[3.3 社会可行性 3](#_Toc2087816062)

[3.4 本章小结 3](#_Toc41762054)

[4 系统需求分析 4](#_Toc1815172656)

[4.1 分析方法与工具 4](#_Toc454140110)

[4.2 面向对象的需求分析 4](#_Toc575947332)

[4.3 系统总体功能架构分析 4](#_Toc1238011895)

[4.4 系统功能性需求分析 4](#_Toc296863482)

[4.4.1 用户功能需求 4](#_Toc780029993)

[4.4.2 管理员功能需求 4](#_Toc1723911063)

[4.5 系统性能需求分析 5](#_Toc2071354164)

[4.6 本章小结 5](#_Toc392032831)

[5 可行性研究 6](#_Toc415961621)

[5.1 系统设计的原则和目标 6](#_Toc1007693162)

[5.2 总体设计 6](#_Toc1242933492)

[5.2.1 设计流程 6](#_Toc1409765675)

[5.2.2 总体功能设计 6](#_Toc744622374)

[5.2.3 平台框架 6](#_Toc1481028749)

[5.3 基于的框架与协议 6](#_Toc167232066)

[5.3.1 SpringMVC,Spring,Mybatis(SSM) 6](#_Toc1760722986)

[5.3.2 HTTP协议 7](#_Toc146570042)

[5.3.3 WebService 7](#_Toc238952785)

[5.4 服务器端设计 7](#_Toc285037605)

[5.4.1 系统层次结构 7](#_Toc1738494425)

[5.4.2 访问控制流程 7](#_Toc213299893)

[5.4.3 基类设计 7](#_Toc781094808)

[5.4.4 主要模块设计 7](#_Toc292903945)

[5.5 客户器端设计 8](#_Toc804084691)

[5.5.1 UI 设计 8](#_Toc136811066)

[5.5.2 主要模块设计 8](#_Toc1576083972)

[5.6 本章小结 9](#_Toc32531659)

[6 系统实现 10](#_Toc1298746475)

[6.1 开发环境 10](#_Toc1008217217)

[6.1.1 Android 开发环境搭建 10](#_Toc1460791289)

[6.1.2 服务器环境搭建 10](#_Toc1486141719)

[6.1.3 开发工具 10](#_Toc201572976)

[6.2 基类与工具类的设计与实现 10](#_Toc1255296313)

[6.3 主要功能的实现 10](#_Toc885784463)

[6.4 本章小结 11](#_Toc1022828637)

[7 系统测试 12](#_Toc74307824)

[7.1 系统测试的目的和要求 12](#_Toc1203599061)

[7.2 系统功能测试 12](#_Toc1740947134)

[7.3 系统性能测试 12](#_Toc633790763)

[7.4 系统测试结论 12](#_Toc602464621)

[7.5 本章小结 12](#_Toc237489542)

[8 总结与展望 13](#_Toc1462116268)

[8.1 总结 13](#_Toc132743655)

[8.2 展望 13](#_Toc1934583999)

[结 论 14](#_Toc1650855613)

[致 谢 15](#_Toc441568451)

[参考文献 16](#_Toc1884955572)

[目 录 18](#_Toc769538060)

[1 英文文献翻译 1](#_Toc1479652186)

[8.3 SNES: Social-Network-Oriented Public Opinion Monitoring Platform Based on ElasticSearch 1](#_Toc653657842)

[8.4 SNES：基于ElasticSearch的面向社交网络的舆情监测平台 16](#_Toc1648496089)

[9 专业阅读书目 30](#_Toc1587237876)

[9.1 书目一 30](#_Toc665118898)

[9.2 书目一 30](#_Toc1000936051)

1. 引言

正文[[1]](#footnote-0)正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文

表 1实验1A

|  |  |  |  |
| --- | --- | --- | --- |
| 试验号 | 温度 | 摩尔比 | 催化剂 |
| 1 |  |  |  |

* 1. 二级标题

正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文

* + 1. 三级标题

1. 绪论

正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文

* 1. 课题研究背景与研究意义

正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文

* 1. 国内外研究现状

正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文

* 1. 本课题研究目标

正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文

* 1. 本课题研究内容

正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文

* 1. 论文组织结构

正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文

1. 可行性研究

正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文

* 1. 技术可行性

正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文

* 1. 经济可行性

正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文

* 1. 社会可行性

正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文

* 1. 本章小结

正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文

1. 系统需求分析

正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文

* 1. 分析方法与工具

正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文

* 1. 面向对象的需求分析

正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文

* 1. 系统总体功能架构分析

正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文

* 1. 系统功能性需求分析

正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文

* + 1. 用户功能需求

正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文

* + 1. 管理员功能需求

正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文

* 1. 系统性能需求分析

正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文

* 1. 本章小结

正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文

1. 可行性研究

正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文

* 1. 系统设计的原则和目标

正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文

* 1. 总体设计

正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文

* + 1. 设计流程

正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文

* + 1. 总体功能设计

正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文

* + 1. 平台框架

正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文

* 1. 基于的框架与协议

正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文

* + 1. SpringMVC,Spring,Mybatis(SSM)

正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文

* + 1. HTTP协议

正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文

* + 1. WebService

正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文

* 1. 服务器端设计

正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文

* + 1. 系统层次结构

正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文

* + 1. 访问控制流程

正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文

* + 1. 基类设计

正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文

* + 1. 主要模块设计

正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文

* + - 1. 资源管理模块

正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文

* + - 1. 用户管理模块

正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文

* + - 1. 推送消息模块

正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文

* + - 1. 数据统计分析模块

正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文

* 1. 客户器端设计

正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文

* + 1. UI 设计

正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文

* + 1. 主要模块设计

正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文

* + - 1. 登陆注册模块

正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文

* + - 1. 图片上传回显

正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文

* + - 1. 资源发布模块

正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文

* + - 1. 资源显示模块

正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文

* + - 1. 报名支付模块

正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文

* 1. 本章小结

正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文

1. 系统实现

正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文

* 1. 开发环境

正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文

* + 1. Android 开发环境搭建

正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文

* + 1. 服务器环境搭建

正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文

* + 1. 开发工具

正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文

* 1. 基类与工具类的设计与实现

正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文

* 1. 主要功能的实现

正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文

* 1. 本章小结

正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文

1. 系统测试

正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文

* 1. 系统测试的目的和要求

正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文

* 1. 系统功能测试

正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文

* 1. 系统性能测试

正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文

* 1. 系统测试结论

正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文

* 1. 本章小结

正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文

1. 总结与展望

正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文

* 1. 总结

正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文

* 1. 展望

正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文

结 论

正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文

致 谢

正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文

签 名：

年 月 日

参考文献

[1] [俄]巴赫金.巴赫金全集（第三卷）[M].白春仁,晓河译.石家庄:河北教育出版社,1998.



毕业设计（论文）附件材料

基于数据挖掘的舆情警报系统

学生姓名： 唐可寅

学 号： 201710253110

所在系部： 电气信息系

专业班级： 17gb计算3班

指导老师： 王晓华 教授

日 期： 二O二一年六月

目 录

目录

[摘 要 i](#_Toc1953504803)

[Abstract ii](#_Toc1825228685)

[目 录 iii](#_Toc1962095047)

[1 引言 1](#_Toc172571597)

[1.1 英文文献原文题目](#_Toc1307907329) **[错误！未定义书签。](#_Toc1307907329)**

[1.1.1 三级标题 1](#_Toc355867811)

[2 绪论 2](#_Toc328342582)

[结 论 2](#_Toc1568286531)

[致 谢 3](#_Toc2124926886)

[参考文献 4](#_Toc993123392)

[目 录 6](#_Toc1181944860)

[1 英文文献翻译 1](#_Toc723527270)

[1.1 SNES: Social-Network-Oriented Public Opinion Monitoring Platform Based on ElasticSearch 1](#_Toc1270417576)

[1.2 SNES：基于ElasticSearch的面向社交网络的舆情监测平台 16](#_Toc1625781358)

[2 专业阅读书目 30](#_Toc2072843125)

[2.1 书目一 30](#_Toc1794680241)

[2.2 书目一 30](#_Toc1782988372)

1. 英文文献翻译
   1. SNES: Social-Network-Oriented Public Opinion Monitoring Platform Based on ElasticSearch

Chuiju You, Dongjie Zhu, Yundong Sun , Anshan Ye , Gangshan Wu , Ning Cao, Jinming Qiu and Helen Min Zhou

**Abstract:** With the rapid development of social network, public opinion monitoring based on social networks is becoming more and more important. Many platforms have achieved some success in public opinion monitoring. However, these platforms cannot perform well in scalability, fault tolerance, and real-time performance. In this paper, we propose a novel social-network-oriented public opinion monitoring platform based on ElasticSearch (SNES). Firstly, SNES integrates the module of distributed crawler cluster, which provides real-time social media data access. Secondly, SNES integrates ElasticSearch which can store and retrieve massive unstructured data in near real time. Finally, we design subscription module based on Apache Kafka to connect the modules of the platform together in the form of message push and consumption, improving message throughput and the ability of dynamic horizontal scaling. A great number of empirical experiments prove that the platform can adapt well to the social network with highly real-time data and has good performance in public opinion monitoring.

**Keywords:** Social network, public opinion monitoring, ElasticSearch, scrapy-redis.

1. **Introduction**

With the rapid development of Internet technology, the number of netizens has increased rapidly. According to statistics, the number of Weibo users in China reached 337 million in the first half of 2018. The number of instant messaging users represented by WeChat reached 756 million, and the usage rate has reached 94.3% [Li, Wei and Xi (2018)]. Social networks have entered an era that emphasizes user engagement and experience, users can create contents, share contents and interact with each other on the web anytime, anywhere. So, there are massive amounts of user data on social networks. Containing great value and reflecting the public opinion orientation of the whole society, these massive social media data attracts more and more researchers to conduct research and exploration on online public opinion and hot topics. Hindelang [Hindelang (1974)] found that the link between criminal justice and the public opinion is close. Burstein [Burstein (2003)] found that the impact of public opinion remains strong even the activities of political organizations and elites are considered. Firstly, data on social network is characterized by immediacy and rapid explosiveness, crawling data from social platforms in real-time is necessary for discovering public opinions and extracting hot topics timely and accurately.

Secondly, UGC (User generated content) is massive and unstructured, storing and retrieving unstructured data quickly is critical to a social network public opinion monitoring platform. Lastly, a platform is not static, but can be expanded and adjusted continually with ever-expanding of data scale and performance requirements, therefore, the scalability of the platform and the low coupling between modules are indispensable. Researchers have made some progress in the design of online public opinion monitoring platform. The framework of the public opinion monitoring system of Weibo based on Spark proposed by Shen [Shen (2018)] achieved good results in predicting the detonation of the Weibo content source, however, the system does not adapt well to the real-time nature of social media data because of lacking data acquisition module and cannot retrieve target information fastly because of the absence of search engine. Tang et al. [Tang (2013); Jie and Jungang (2009)] added a web crawler module in their platform, which can crawl the latest data on social net-works in real time. However, the modules in their platforms performing functions directly by calling each other are highly coupled and do not have good scalability.

In this paper, we propose a novel platform called the “SNES”, which is based on Elasticsearch (an open source distributed full-text retrieval system), and we design subscription module based on Apache Kafka (a high throughput distributed publish and subscribe message system) and Spark Streaming (a real-time streaming computing framework). A great number of empirical experiments prove that the platform can adapt well to social media data and obtains good performance in public opinion monitoring.

The SNES has the following advantages:

1. Crawling ever-changing social media data in real time.
2. Storing and retrieving massive data quickly and efficiently.
3. Excellent performance, scalability and fault tolerance.
4. **Related work**

Firstly, mining public opinion and extracting hot topics on social networks require massive and multi-source social media data, and the explosion and immediacy of social media data determine that data cannot be acquired and processed in the traditional way. Scrapy is an application framework for crawling web sites which can be used for a wide range of useful applications, like data mining, information processing or historical archival. Wang et al. [Wang and Guo (2012)] crawled the data of Taobao share-platform using Scrapy crawl architecture and analyzed the format of web pages. Scrapy-redis is a Scrapy-based distributed crawler component. It uses Redis to store and schedule requests, and store the items generated by the crawl for subsequent processing. Fan [Fan (2018)] designed and implemented a distributed crawling system based on scrapy-redis. Xie et al. [Xie and Xia (2014)] designed a topic-focused web crawler, which can crawl and gather the subject-related web pages as soon as possible. After crawling the data, we need to process the data, such as filtering out illegal data and converting data. Spark Streaming can process real-time streaming data from multiple data sources including Kafka, Flume, Twitter, ZeroMQ, Kinesis and TCP sockets with high-throughput, fault-tolerant by using advanced algorithms such as map, reduce, join etc. And finally, the processing results can be stored in the file system, database, etc. Tan et al. [Tan and Zhou (2018)] designed a real-time traffic processing platform based on Spark Streaming and verified that the system can be applied to large-scale high concurrent data streams in real time. Zhang et al. [Zhang, Li, Liu et al. (2018)] proposed a more effective strategy to schedule tasks and added it to the source code of Spark. Yan et al. [Yan and Wang (2018)] designed a realtime movie recommendation system based on Spark streaming computing, which can better meet the real-time needs of users. Therefore, the SNES architecture proposed in this paper incorporates a distributed crawler cluster module, which enables the platform to crawl ever-changing social media data in real time.

Secondly, mining social network public opinion and extracting hot topics require efficient storage and retrieval of massive UGC which are unstructured data. Elasticsearch is an open source, distributed, full-text search engine that supports RESTful interfaces. It can store, search and analyze large amounts of data in a very short time and it usually be used as a core engine with a complex search scheme. We can set up the same cluster name to form a distributed cluster in the same network segment and balance the load by adding more nodes to the cluster for horizontal expansion. ES supports full-text search, each field can be indexed, and the data of each field can be searched. ES adopts a noncentralized cluster design, after the central node in the cluster fails, ES will automatically select a new node as the central node and automatically migrate the data fragments to ensure the security and access of user data. Its many advantages make it a good performance in solving big data related problems and a preferred tool in enterprise big data solutions. Zhou et al. [Zhou and Han (2015)] designed and implemented elastic search cluster based on ES for an e-commerce system, which significantly reduced CPU usage, memory usage, maintenance costs of database server and greatly improved search efficiency and system stability. Li et al. [Li, Li, Zhang et al. (2018)] designed a highperformance chemical structure & data search engine called DCAIKU, built on CouchDB and ElasticSearch engines, which can handle both keyword search and structural search for millions of records with both high accuracy and low latency. Therefore, the SNES architecture proposed in this paper incorporates the open source high-distribution distributed full-text search engine, which can store and retrieve massive unstructured data in near real time.

Finally, a platform should have good scalability, fault tolerance and decoupling between modules. Apache Kafka is a distributed flow processing platform, which can publish, subscribe and persist data with high timeliness and fault tolerance, so it can be used as message middleware. Kafka has the following advantages over other mainstream messaging middleware including ActiveMQ and RabbitMQ:

1. Simple and intuitive operation.
2. Excellent message throughput.
3. Supporting load balancing and dynamic horizontal scaling with a fully distributed architecture.

These advantages of Kafka are necessary to build a highly available and efficient messaging subscription service. Lu [Lu (2018)] designed and implemented a Kafka-based messaging subscription service called MPS, extending the way of communication between micro-services and providing asynchronous decoupling between services by solving many problems, such as highly coupled HTTP and RPC calls between services within the platform, easily blocked request, unbuffered request. Yan et al. [Yan and Yu (2018)] used Kafka as a message platform to complete the transmission of massive data and concurrent real-time processing. Therefore, we design a subscription module based on Apache Kafka which connects the modules of the platform together in the form of pushing and consuming message to increase message throughput and the ability of dynamic horizontal scaling.

1. **The design of SNES**
   1. The overall design

In terms of multi-source data acquisition, SNES improves the speed of data crawling and the schedulability of crawling tasks by combining distributed crawling frameworks with subscription module based on Kafka message subscription service; in terms of massive data storage and retrieval, we design and add ES-centric full-text storage and search engines to optimize the storage and retrieval efficiency of the entire system platform; in the way of combining different modules, the Kafka-based message subscription service is designed, tasks and instructions are transmitted between modules in the form of messages, which greatly reduces the coupling between modules and improves the scalability of the platform. The overall topology of the architecture is shown in Fig. 1,

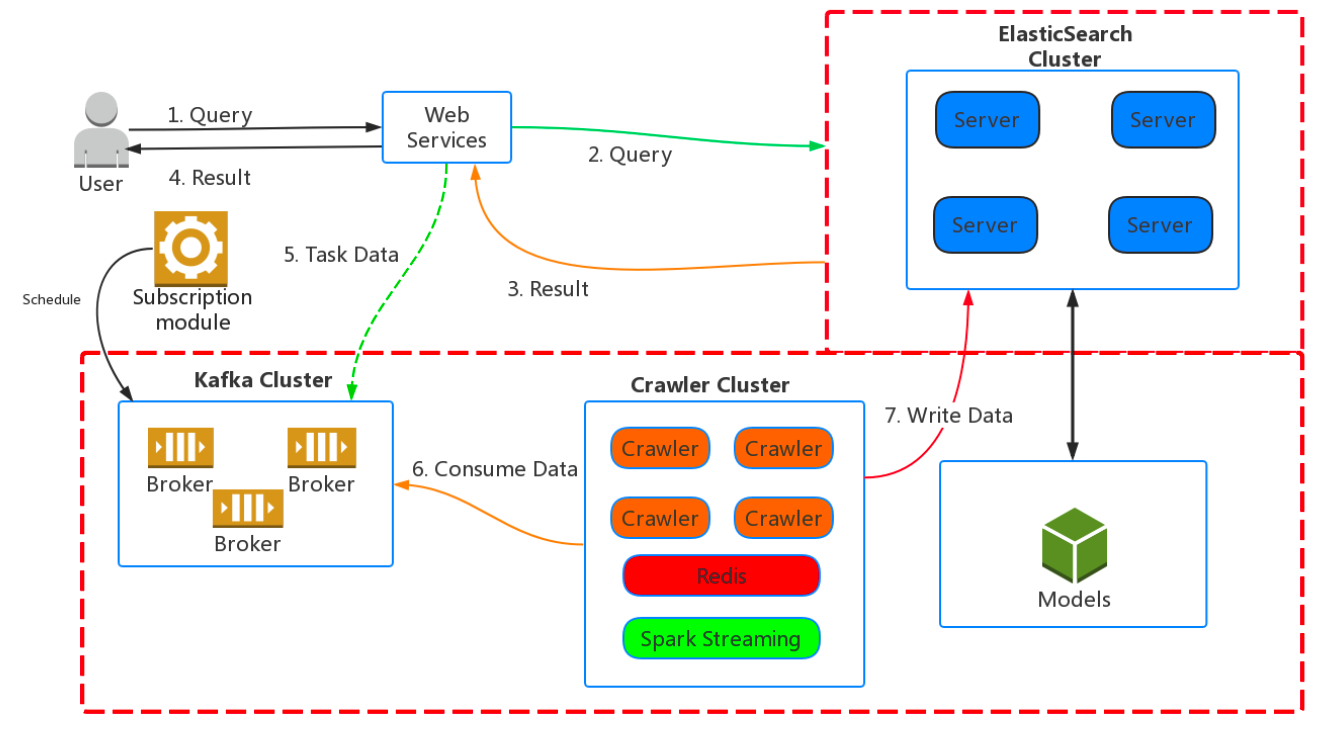


Figure 1: The overall topology of SNES

Considering the requirements of data size, performance and function for public opinion and hot topics mining on social network, we add the data storage and retrieval module, the message publishing subscription service module, the webservice logic processing module and the multi-source distributed information and stream processing module into SNES architecture. The work steps can be summarized as follows:

Step 1: Users search for specified content through web services;

Step 2: If there are hit information in ES, go to Step 7, else go to Step 3;

Step 3: The service issues a message crawling command through the message subscription service module;

Step 4: The data crawling module receives the relevant command and crawl the specified data;

Step 5: The stream processing module processes the crawled data;

Step 6: ES stores the processed data;

Step 7: Return the hit information to web services.

The network topology of the entire platform is shown in Fig. 2. There are four clusters, including multi-source data crawling cluster, storage and search engine, Kafka-based messaging service and Spark-based data processing cluster. Each module is connected through an intranet switch, which reduces the coupling and increases the rate of communication. The above parts are in the internal LAN environment, and in order to improve security, the data communication with the external network must pass through the firewall. In order to avoid the limit of IP access frequency from the target website, we use proxy IP services to climb domestic social networking sites such as Sina Weibo and WeChat, and we adopt virtual cloud hosting services to climb foreign social networking sites such as Twitter and Facebook. Fig. 3 depicts the internal structure of the system. It consists of four parts: the message publishing subscription service, the social network information crawling service, data storage and retrieval clusters and public opinion information processing models.

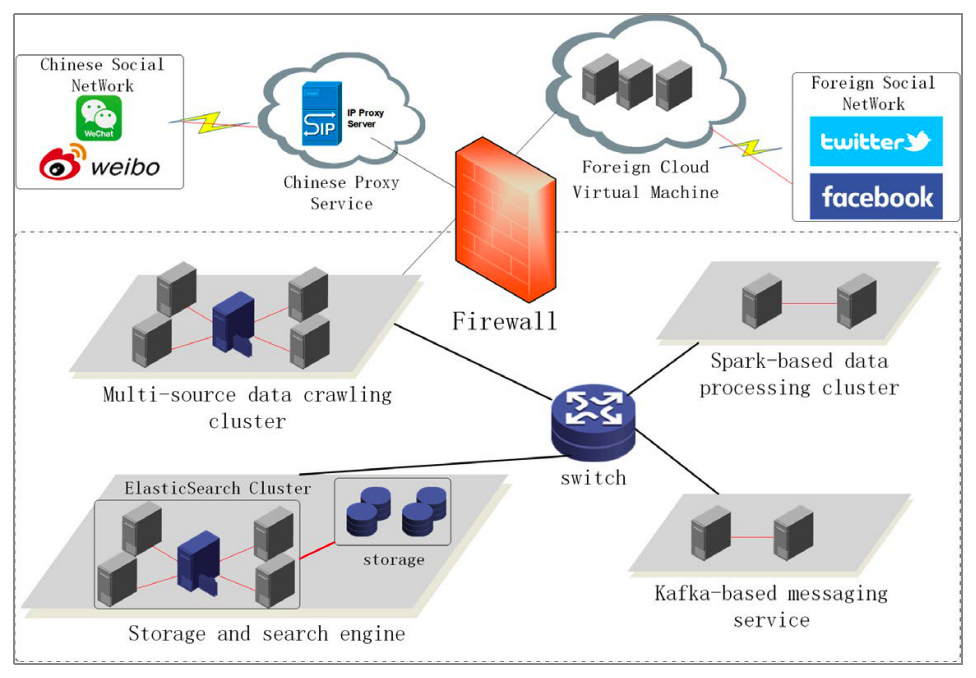


Figure 2: The network topology of SNES

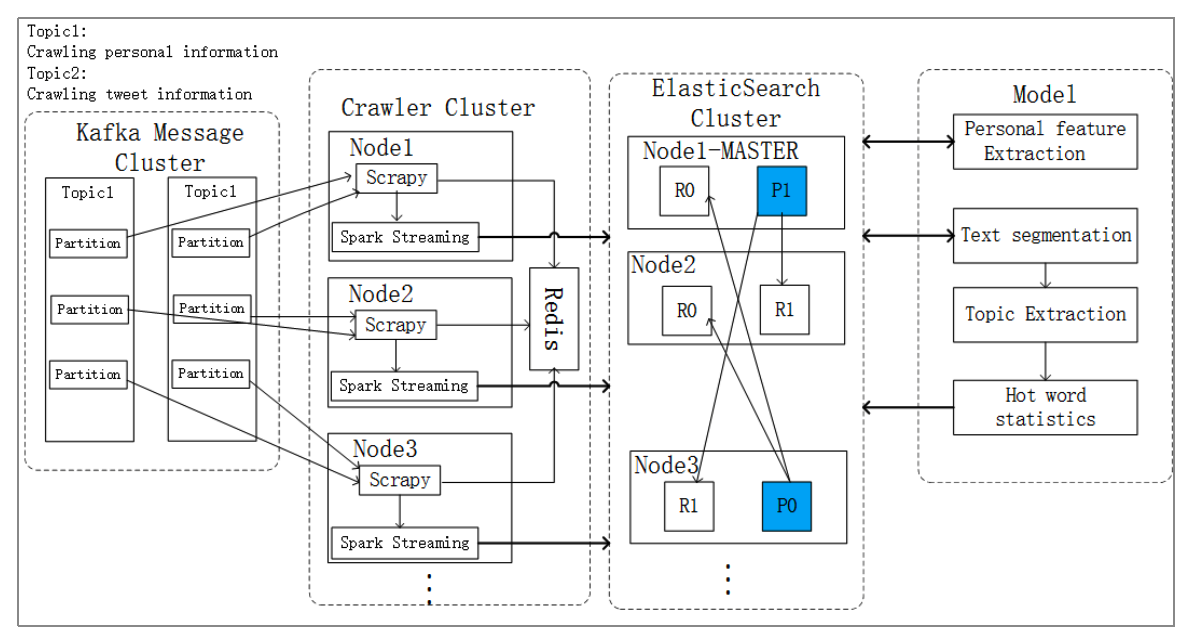


Figure 3: Internal structure diagram of SNES

* 1. Data crawling module

The crawler cluster consists of several nodes, each of which acts as a consumer of message subscription module (Chapter 3.3) and each node contains the crawler framework Scrapy-redis, Spark Streaming framework and Redis database. The frame structure and internal data flow of the data crawling module is shown in Fig. 4. To improve data crawl speed and avoid interruption due to network problems or the restrictions of target websites, we design the top scheduler to communicate with the message subscription module and the brief work flows are as follows:

Step 1: The top scheduler requests a new data crawling task from message subscription module;

Step 2: If the task is not null, go to Step 1, else go to Step 3;

Step 3: The top scheduler issues a data crawling command to spider;

Step 4: The spider gets the first URL from redis database and send to engine;

Step 5: If the URL is already in the crawled queue, go to Step 1, else go to Step 6;

Step 6: The engine requests random cookie from cookie pool and random the cookie and proxy Ip;

Step 7: The spark streaming module processes the obtained data and stores it to massive data storage and extraction module (Chapter 3.4);

Step 8: Put the URL into crawled queue and go to Step 1.

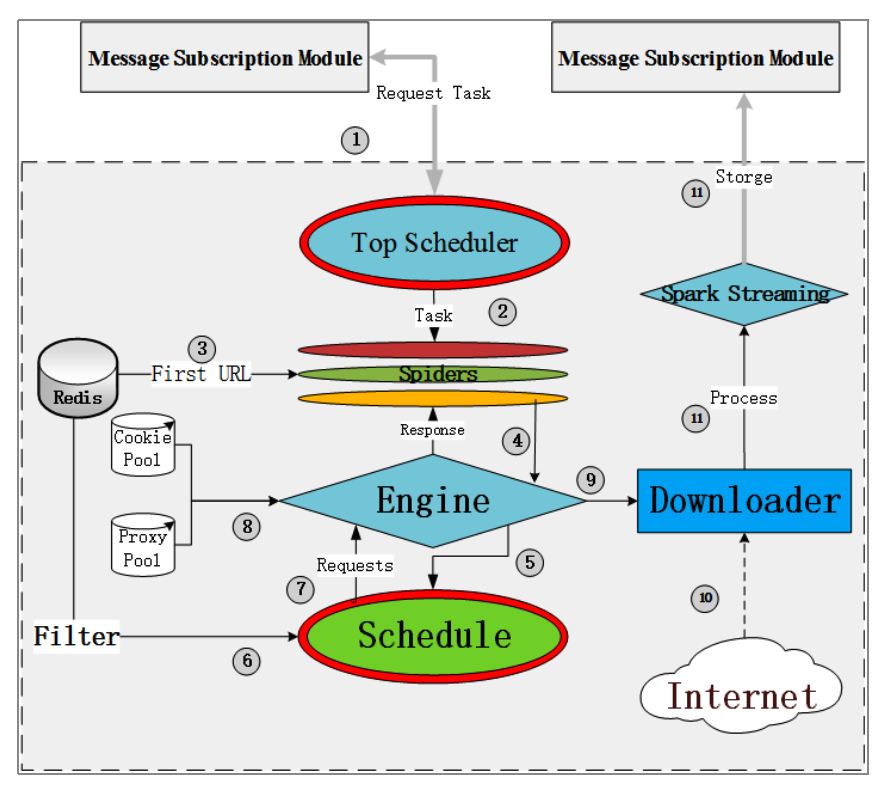


Figure 4: Frame structure and internal data flow of data crawling module

* 1. Message subscription module

To improve the scalability, fault tolerance and decoupling between modules in our platform, we design the message subscription module based on Kafka. The message subscription cluster consists of two working nodes. Each message topic in the figure is divided into three partitions, each consumer node consumes one partition of each topic. After empirical experiment, it is concluded that to get better performance, the number of topic’s partition is preferably set to a multiple of the consumer and it is best to ensure that the number of message partitions that each consumer monitors is equal. Users can subscribe to specific events or topics as they want and the data crawling module in Chapter 3.1 would consume the message and crawl the corresponding data.

For each topic, the message subscription maintains a partitioned message that looks like the left of Fig. 5. Each partition is an ordered, immutable sequence of records that is continually appended to a structured commit message. Each record in the partitions is assigned a sequential ID number called the offset that uniquely identifies each record within the partition. In fact, the only metadata retained on a per-consumer basis is the offset or position of that consumer in the message. This offset is controlled by the consumer: normally a consumer will advance its offset linearly as it reads records, but, in fact, since the position is controlled by the consumer it can consume records in any order it likes. As shown in the right of Fig. 5, a consumer can reset to an older offset to reprocess data from the past or skip ahead to the most recent record and start consuming from now.

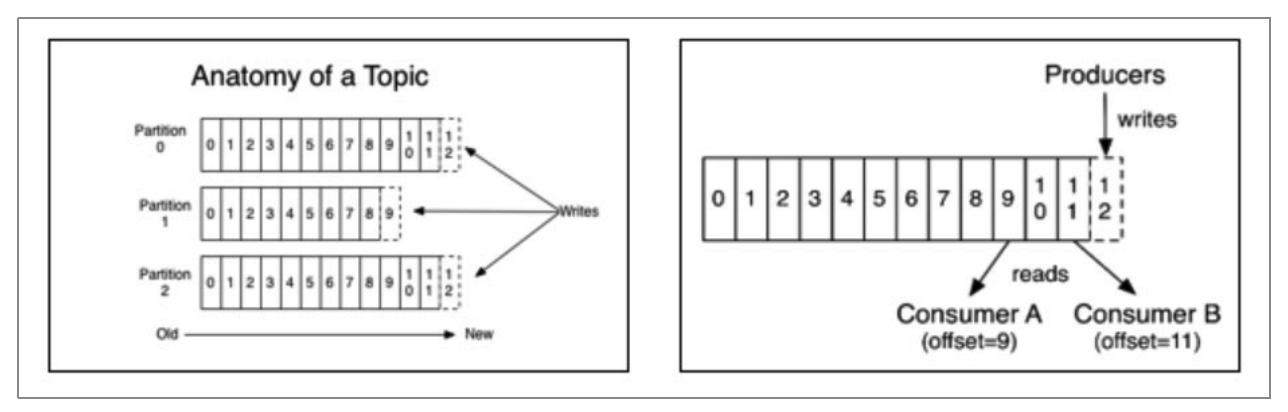


Figure 5: The theory of producing and consuming topics

* 1. Massive data storage and extraction module

Mining social network public opinion and extracting hot topics require efficient storage and retrieval of massive UGC which are unstructured data. We design a massive data storage and extraction module based on Elastic Search. The data structure of the massive data storage and extraction module is shown in Fig. 6. Each Node has three shards, where P is the primary shards and R is the redundant shards. The primary shard P1, P2, and the redundant shard R0 are stored in node 1, the redundant shard R0, R1 and R2 are stored in node 2, the primary shard P0 and the redundant shard R1, R2 are stored in node 3 There are a total of 3 primary shards and 6 redundant shards. At the same time, we also noticed that Node 1 also has a master ID, which means it is a master node. It is more special than other nodes. It has the authority to control the entire cluster, such as resource allocation, node modification.

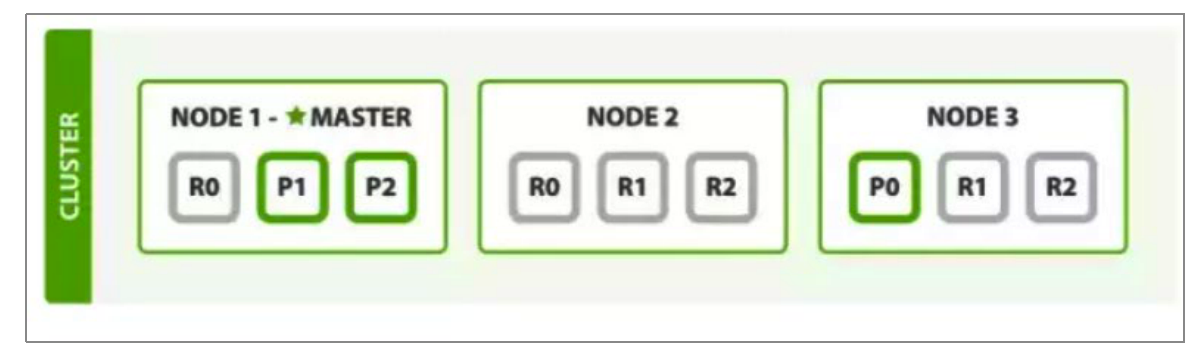


Figure 6: The theory of producing and consuming topics

1. **Experiment**

We evaluate the effects of public opinion monitoring and performance of the SNES by applying it to Sina Weibo real-time data.

* 1. Experiment environment

To ensure the data acquisition, processing and retrieval capability of the plat- form, and consider the scalability and load balance of the system, we constructed the experimental platform environment as shown in Tab. 1. The network bandwidth test of the platform is shown in Fig. 7.

Table 1: The hardware facility and application of experiment platform

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Server | CPU | Memory | Disk | Application |
| Server-1 | 3.60 GHz 2 core 4 threads | 8GB | 500GN | ES-Master |
| Server-2 | 3.60 GHz 2 core 4 threads | 8GB | 500GN | ES-Data |
| Server-3 | 3.60 GHz 2 core 4 threads | 8GB | 500GN | ES-Data |
| Server-4 | 3.60 GHz 2 core 4 threads | 8GB | 500GN | ES-Data |
| Server-5 | 3.60 GHz 2 core 4 threads | 8GB | 250GN | Kafka-Broker |
| Server-6 | 3.60 GHz 2 core 4 threads | 8GB | 500GN | Kafka-Broker |
| Server-7 | 3.60 GHz 2 core 4 threads | 8GB | 250GN | Kafka-Broker |
| Server-8 | 3.60 GHz 2 core 4 threads | 8GB | 500GN | Spider&Spark |
| Server-9 | 3.60 GHz 2 core 4 threads | 8GB | 250GN | Spider&Spark |
| Server-10 | 3.60 GHz 2 core 4 threads | 8GB | 500GN | Spider&Spark |

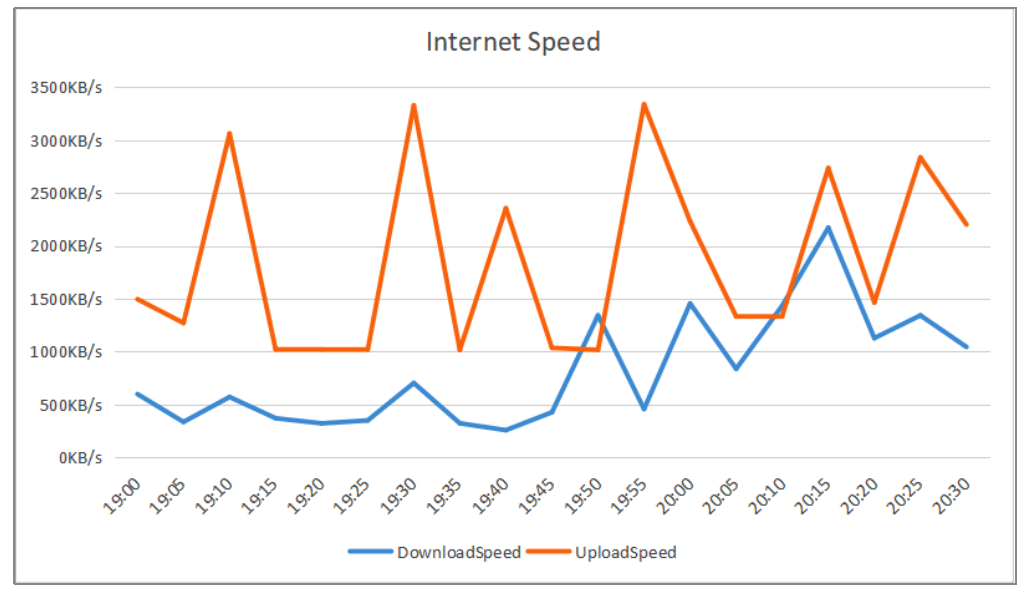


Figure 7: Network bandwidth of the experimental environment

* 1. Experiment data

The experimental data in this paper is collected from Sina Weibo data in real time, including personal information, relationship network, tweets and interactive information (like or dislike, comment), etc. Detailed data information for incident moni-toring and data crawling test is shown in Tab. 2; the data format for searching and writing latency test is shown in Tab. 3.

Table 2: Statistics of the experimental data for incident monitoring

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Data Source | Persons | Relations | Tweets | Interacts |
| Sina Weibo | 130,028 | 71,032 | 1,501,202 | 4,130,187 |

Table 3: Data format for searching and writing

|  |  |  |
| --- | --- | --- |
| Field | Type | Length |
| userId | Integer | 32bit |
| userName | Char(20) | 160bit |
| content | Char(150) | 1200bit |
| date | Long | 64bit |
| platform | Short | 16bit |
| thumb | Integer | 32bit |
| comment | Integer | 32bit |
| forward | Integer | 32bit |

* 1. Experiment result
     1. Performance evaluation

In this paper, the search efficiency, the speed of crawling data and writing efficiency are used as evaluation indicators for this platform. The speed of crawling data. We use the Eq. (1) to measure the speed of crawling data.

(1)

where n is the number of crawling cluster nodes; h is the average CPU frequency, the unit is GHz; t is the total crawling time, the unit is second(s); c is the total number of pages crawled during t seconds; bw represents the average of network bandwidth during the time, the unit is MB/S. Therefore, sp represents the average value of the page crawled by the single-threaded single node during the CPU clock cycle under the unit bandwidth condition. The value of sp can reduce the influences of the number of nodes, CPU threads, CPU frequency, and network bandwidth to some extent. In the experiment conducted in this paper, the crawling time is 48 hours, the total number of pages crawled is 864078, h=3.6, the bw=2 and the result sp=0.07, which is calculated by Eq. (1).

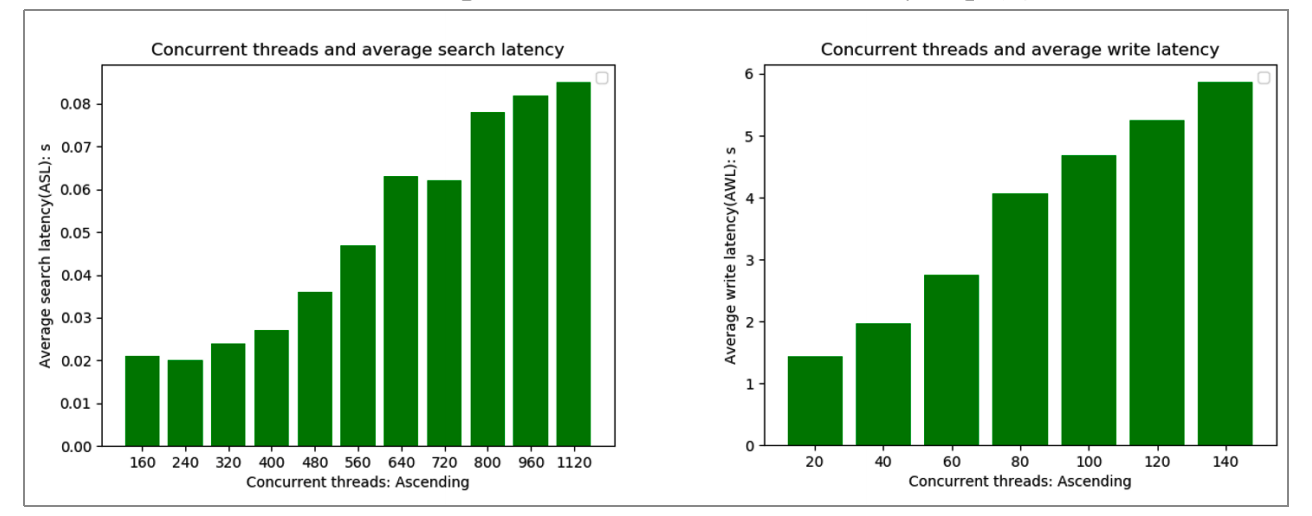


Figure 8: Search and write latency of SNES

We evaluate the system’s search latency and index latency with concurrent read and write scripts, with a test time of 50 minutes. We measure the delay by the Eq. (2)

(2)

where n is the number of experiments, ti is the delay of the ith experiment, the unit is second; ASL is average search latency. The left of Fig. 8 shows the result of our concurrent searching test, we conduct 4000 experiments with each specific number of concurrent requests and get the search latency. The right of Fig. 8 shows the result of our concurrent writing test, we conduct 2000 experiments with each specific number of concurrent requests and get the writing latency. All data is randomly generated and the format is shown in Tab. 3.

* + 1. Incident monitoring

This experiment is aimed at “The Event of HuaZhu Information Leakage on 2018-08-28 in China”, the search keyword is “HuaZhu Leak” and the monitoring time is from August 28, 2018 to September 10, 2018. The left of Fig. 9 shows the trend of the transmission volume monitored by the system about the “The Event of HuaZhu Information Leakage”. It can be seen that the line is rising from August 28 to September 2, representing that the event is heating up. After September 04, it is gradually declining until it is at a lower level. This is consistent with the development of the entire event. The right of Fig. 9 shows the hot word cloud of the event obtained by extracting and filtering the content topic from the crawled microblog. It can be seen that words such as “HuaZhu”, “Information”, “Leakage”, “Data” and “Cybersecurity” have the highest proportion in the word cloud, and these words can properly describe and summarize the content of this event and the discussions and opinions of Weibo users and related media on this incident.

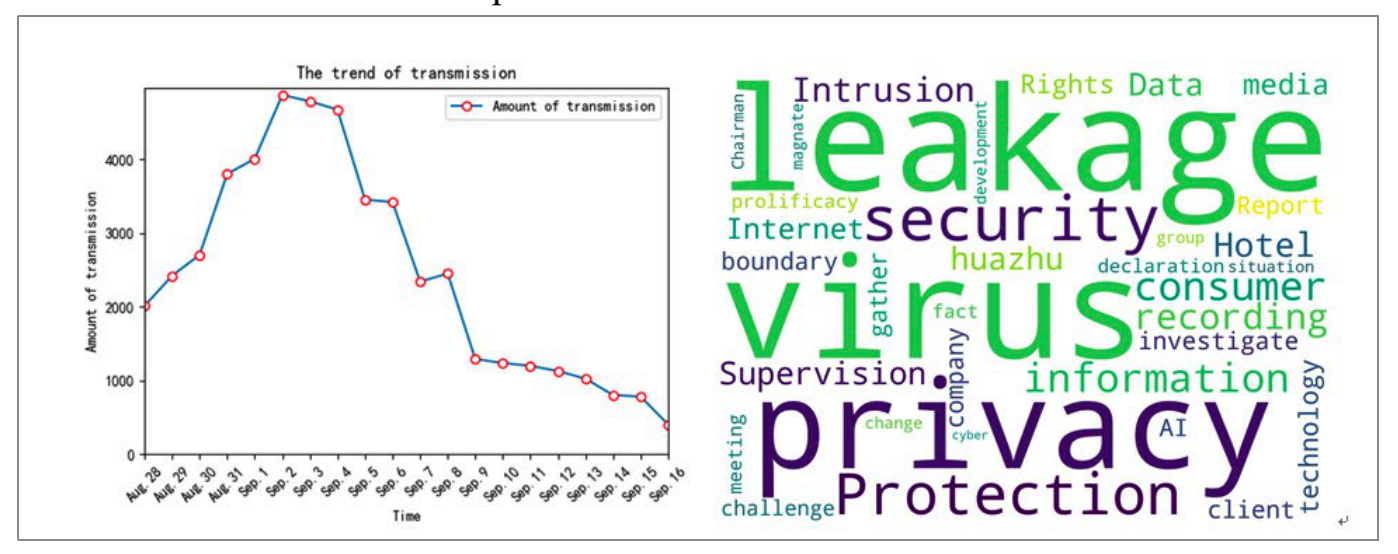


Figure 9: Daily spread and hot words cloud of “Huazhu Info Leakage”

1. **Conclusion and discussion**

SNES we proposed in this paper creatively integrates high-throughput subscription module based on Kafka, distributed crawling framework, real-time streaming computing framework and massive data storage and extraction module based on ElasticSearch. After the above experiment, the SNES architecture has the following advantages:

1. Crawling ever-changing social media data in real time.
2. Storing and retrieving massive data quickly and efficiently.
3. Good scalability and fault tolerance.

Currently, our platform has defects in data security storage and encryption protection, our future work will focus on improving the security of the platform.

**Acknowledgement:** This work is supported by State Grid Science and Technology Project under Grant Nos. 520613180002, 62061318C002, the Fundamental Research Funds for the Central Universities (Grant Nos. HIT.NSRIF.201714),Weihai Science and Technology Development Program (2016DXGJMS15) and Key Research and Development Program in Shandong Provincial (2017GGX90103), Fujian Young and Middle-aged Teacher Education Research Project, Grant No. JAT160466, Jiangsu Polytechnic College of Agriculture and Forestry Key R&D Projects (2018kj11), Study and Development of Smart Agriculture Control System Based on Spark Big Data Decision (2017N0029).

**References**

**Burstein, P.** (2003): The impact of public opinion on public policy: a review and an agenda. Political Research Quarterly, vol. 56, no. 1, pp. 29-40.

**Fan, Y.** (2018): Design and implementation of distributed crawler system based on scrapy. IOP Conference Series: Earth and Environmental Science, vol. 108.

**Hindelang, M. J.** (1974): Public opinion regarding crime, criminal justice, and related topics. Journal of Research in Crime and Delinquency, vol. 11, no. 2, pp. 101-116.

**Jie, D.; Jungang, X.** (2009): IPOMS: an internet public opinion monitoring system. Applications of Digital Information and Web Technologies, pp. 433-437.

**Li, J.; Wei, Q.; Xi, Y.** (2018): Analysis and research on the development of university libraries based on wechat and facebook social network platform. Journal of Sichuan Library Science, no. 6, pp. 69-74.

**Li, R.; Li, B.; Zhang, G.; Jiang, J.; Luo, Y.** (2018): A high-performance and flexible chemical structure&data search engine built on couchdb&elasticsearch. Chinese Journal of Chemical Physics, vol. 31, no. 3, pp. 341-349.

**Lu, S.** (2018): Design and Implementation of a Message Publishing Service Based on Kafka (Ph.D. Thesis). Nanjing University.

**Shen, L.** (2018): Design and Implementation of Weibo Public Opinion Monitoring System Based on Spark (Ph.D. Thesis). University of Electronic Science and Technology.

**Tan, L.; Zhou, J.** (2018): Real-time traffic data processing platform based on spark streaming. Computer System Application, vol. 27, no. 10, pp. 133-139.

**Tang, Y.** (2013): Design and Implementation of Internet Public Opinion Monitoring System (Ph.D. Thesis). Beijing University of Posts and Telecommunications.

**Wang, J.; Guo, Y.** (2012): Scrapy-based crawling and user-behavior characteristics analysis on taobao. Cyber-Enabled Distributed Computing and Knowledge Discovery, pp. 44-52.

**Xie, D. X.; Xia, W. F.** (2014): Design and implementation of the topic-focused crawler based on scrapy. Advanced Materials Research, vol. 850, pp. 487-490.

**Yan, H.; Yu, X.** (2018): Software system design based on Kafka message platform. Electronic Technology and Software Engineering, no. 18, pp. 38.

**Yan, L.; Wang, X.** (2018): Real-time movie recommendation research based on spark streaming computing. Software Guide, pp. 1-5.

**Zhang, X.; Li, Z.; Liu, G.; Xu, J.; Xie, T. et al.** (2018): A spark scheduling strategy for heterogeneous cluster. Computers, Materials & Continua, vol. 55, no. 3, pp. 405-441.

**Zhou, Y.; Han, X.** (2015): Application examples of elasticsearch in e-commerce system. Information Technology and Standardization, vol. 5, pp. 30.

Chuiju You,Dongjie Zhu,2,Yundong Sun,Anshan Ye,Gangshan Wu,Ning Cao,Jinming Qiu,Helen Min Zhou. SNES: Social-Network-Oriented Public Opinion Monitoring Platform Based on ElasticSearch[J]. CMC: Computers, Materials & Continua,2019,61(3).

* 1. SNES：基于ElasticSearch的面向社交网络的舆情监测平台

崔菊友，朱东杰，孙云东，叶鞍山，吴刚山，宁草、金明秋、周海伦

**摘要：**随着社交网络的快速发展，基于社交网络的舆论监督变得越来越重要。许多平台在舆论监督方面都取得了一些成功。但是，这些平台在可伸缩性，容错性和实时性能方面不能表现良好。在本文中，我们提出了一个基于ElasticSearch（SNES）的新颖的面向社交网络的公众舆论监控平台。首先，SNES集成了分布式爬虫集群模块，可提供实时社交媒体数据访问。其次，SNES集成了ElasticSearch，它可以近乎实时地存储和检索大量的非结构化数据。最后，我们设计了基于Apache Kafka的订阅模块，以消息推送和消费的形式将平台的模块连接在一起，从而提高了消息吞吐量和动态水平扩展的能力。大量的经验实验证明，该平台能够很好地适应具有高度实时数据的社交网络，并在公众舆论监控中具有良好的性能。

**关键词：**社交网络，公众舆论监控，ElasticSearch，scrapy-redis。

1. **简介**

随着互联网技术的飞速发展，网民数量迅速增加。据统计，2018年上半年，中国微博用户数达到3.37亿。以微信为代表的即时通讯用户数达到7.56亿，使用率达到94.3％[李，魏，席 (2018)]。社交网络已进入一个强调用户参与度和体验的时代，用户可以随时随地在Web上创建内容，共享内容并彼此交互。因此，社交网络上有大量的用户数据。这些海量社会媒体数据蕴含着巨大的价值，并反映了整个社会的舆论导向，吸引了越来越多的研究人员对网络舆论和热点话题进行研究和探索。Hindelang [Hindelang (1974)]发现，刑事司法与公众舆论之间的联系是紧密的。Burstein [Burstein (2003)]发现，即使考虑政治组织和精英的活动，公众舆论的影响仍然很强。首先，社交网络上的数据具有即时性和爆炸性的特点，需要实时从社交平台上抓取数据，以发现公众意见并及时准确地提取热门话题。

其次，UGC（User generated content，用户生成的内容）是庞大且非结构化的，快速存储和检索非结构化数据对于社交网络民意监控平台至关重要。最后，平台不是静态的，而是可以随着数据规模和性能要求的不断扩大而不断扩展和调整的，因此，平台的可扩展性和模块之间的低耦合是必不可少的。研究人员在在线舆论监测平台的设计上取得了一些进展。Shen [Shen (2018)]提出的基于Spark的微博舆情监测系统框架在预测微博内容源的爆炸方面取得了良好的效果，但是该系统不能很好地适应微博内容源的实时性。社交媒体数据，因为缺少数据获取模块，并且由于缺少搜索引擎而无法快速检索目标信息。唐等[唐 (2013); Jie和Jungang (2009)]在他们的平台中添加了一个Web搜寻器模块，该模块可以实时地搜寻社交网络上的最新数据。但是，其平台中直接通过相互调用来执行功能的模块是高度耦合的，不具有良好的可扩展性。

在本文中，我们提出了一个基于Elasticsearch（开源分布式全文检索系统）的新颖平台“ SNES”，并设计了基于Apache Kafka（高吞吐量分布式发布和订阅消息系统）和Spark Streaming（实时流计算框架）的订阅模块。大量的经验实验证明，该平台可以很好地适应社交媒体数据，并在公众舆论监控中获得良好的表现。

SNES具有以下优点：

1. 实时抓取不断变化的社交媒体数据；
2. 快速高效地存储和检索海量数据；
3. 卓越的性能、可扩展性和容错性。
4. **相关工作**

首先，在社交网络上挖掘舆论并提取热门话题需要大量且多来源的社交媒体数据，而社交媒体数据的爆炸性和即时性决定了无法以传统方式获取和处理数据。Scrapy是用于爬网网站的应用程序框架，可用于各种有用的应用程序，例如数据挖掘，信息处理或历史档案。Wang等人[Wang and Guo (2012)]使用Scrapy抓取架构抓取淘宝共享平台的数据，并分析了网页的格式。Scrapy-redis是一个基于Scrapy的分布式搜寻器组件。它使用Redis来存储和调度请求，并存储由爬网生成的项目以进行后续处理。Fan [Fan (2018)]设计并实现了基于scrapy-redis的分布式爬网系统。Xie等人[Xie and Xia (2014)]设计了一个以主题为中心的网络爬虫，该爬虫可以尽快爬网和收集与主题相关的网页。抓取数据后，我们需要处理数据，例如过滤掉非法数据和转换数据。通过使用高级算法（例如map，reduce，join等），Spark Streaming可以处理来自多个数据源（包括Kafka，Flume，Twitter，ZeroMQ，Kinesis和TCP套接字）的实时流数据，并具有高吞吐量，容错能力。最后，处理结果可以存储在文件系统，数据库等中。Tan等人[Tan and Zhou (2018)]设计了一个基于Spark Streaming的实时流量处理平台，并验证了该系统可以实时应用于大规模高并发数据流。Zhang等人[Zhang，Li，Liu等 (2018)]提出了一种更有效的计划任务的策略，并将其添加到Spark的源代码中。Yan 等人 [Yan and Wang (2018)]设计了一个基于Spark流计算的实时电影推荐系统，可以更好地满足用户的实时需求。因此，本文提出的SNES体系结构包含一个分布式爬虫集群模块，该模块使该平台能够实时抓取不断变化的社交媒体数据。

其次，挖掘社交网络舆论和提取热门话题需要有效存储和检索非结构化数据的大量UGC。Elasticsearch是一个开源的，分布式的全文本搜索引擎，支持RESTful接口。它可以在很短的时间内存储，搜索和分析大量数据，通常用作具有复杂搜索方案的核心引擎。我们可以设置相同的群集名称以在同一网段中形成一个分布式群集，并通过向群集添加更多节点以进行水平扩展来平衡负载。ES支持全文搜索，可以为每个字段建立索引，并且可以搜索每个字段的数据。ES采用非集中式集群设计，集群中的中心节点发生故障后，ES会自动选择一个新节点作为中心节点，并自动迁移数据片段，以确保用户数据的安全性和访问性。它的许多优点使其在解决与大数据有关的问题方面具有良好的性能，并且是企业大数据解决方案中的首选工具。周等[Zhou and Han (2015)]为电子商务系统设计并实现了基于ES的弹性搜索集群，该集群大大降低了CPU使用率，内存使用率，数据库服务器维护成本，并大大提高了搜索效率和系统稳定性。Li等[李，李，张等 (2018)]设计了一种名为DCAIKU的高性能化学结构和数据搜索引擎，该引擎建立在CouchDB和ElasticSearch引擎的基础上，该引擎可以处理关键字搜索和结构搜索，从而以高精度和低延迟对数百万条记录进行处理。因此，本文提出的SNES体系结构结合了开源的高分布分布式全文本搜索引擎，该引擎可以近乎实时地存储和检索大量的非结构化数据。

最后，平台应具有良好的可伸缩性，容错能力以及模块之间的解耦。Apache Kafka是一个分布式流处理平台，可以以高及时性和容错性发布，订阅和保留数据，因此可以用作消息中间件。与其他主流消息传递中间件（包括ActiveMQ和RabbitMQ）相比，Kafka具有以下优势：

1. 操作简单直观；
2. 出色的消息吞吐量；
3. 支持负载平衡和动态水平扩展与一个完全分布式的体系结构。

Kafka的这些优点对于构建高度可用和高效的消息传递订阅服务是必需的。Lu [Lu (2018)]设计和实现了一个基于Kafka的消息订阅服务，称为MPS，它扩展了微服务之间的通信方式，并通过解决许多问题（例如服务之间的高度耦合的HTTP和RPC调用）来提供服务之间的异步解耦。在平台内，容易阻塞的请求，未缓冲的请求。Yan等[Yan and Yu (2018)]使用Kafka作为消息平台来完成海量数据的传输和并发实时处理。因此，我们设计了基于Apache Kafka的订阅模块，该模块以推送和使用消息的形式将平台的模块连接在一起，以提高消息吞吐量和动态水平缩放的能力。

1. **SNES的设计**
   1. 总体设计

在多源数据获取方面，SNES通过将分布式爬网框架与基于Kafka消息订阅服务的订阅模块相结合，提高了数据爬网速度和爬网任务的可调度性。在海量数据的存储和检索方面，我们设计并添加了以ES为中心的全文存储和搜索引擎，以优化整个系统平台的存储和检索效率。通过组合不同模块的方式，设计了基于Kafka的消息订阅服务，任务和指令以消息的形式在模块之间传输，大大减少了模块之间的耦合，提高了平台的可扩展性。该架构的整体拓扑如图1所示。

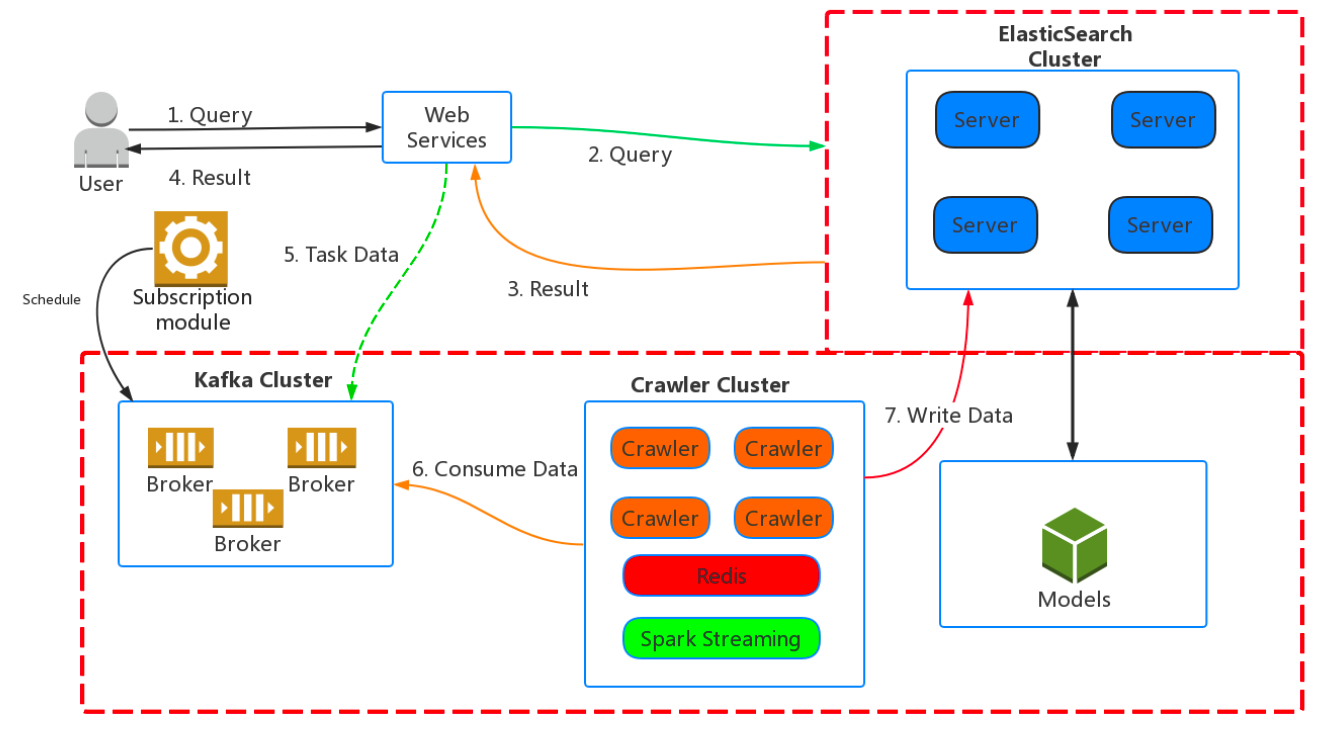


图1：SNES的整体拓扑

考虑到社交网络上舆论和热点话题的数据大小，性能和功能要求，我们添加了数据存储和检索模块，消息发布订阅服务模块，Web服务逻辑处理模块以及多源分布式信息， 流处理模块进入SNES体系结构。工作步骤可以总结如下：

步骤1：用户通过Web服务搜索指定的内容；

步骤2：如果ES中有命中信息，则转到步骤7，否则转到步骤3。

步骤3：服务通过消息订阅服务模块发出消息爬行命令；

步骤4：数据抓取模块接收相关命令，对指定的数据进行抓取；

步骤5：流处理模块处理抓取的数据；

步骤6：ES存储处理后的数据；

步骤7：将匹配信息返回到Web服务。

整个平台的网络拓扑如图2所示。共有四个集群，包括多源数据爬网集群，存储和搜索引擎，基于Kafka的消息传递服务和基于Spark的数据处理集群。每个模块通过内部网交换机连接，这减少了耦合并提高了通信速率。以上部分均在内部LAN环境中，为了提高安全性，与外部网络的数据通信必须通过防火墙。为了避免目标网站的IP访问频率的限制，我们使用代理IP服务来攀登新浪微博和微信等国内社交网站，并采用虚拟云托管服务来攀登Twitter和Twitter等国外社交网站。Facebook。图3描绘了系统的内部结构。它包括四个部分：消息发布订阅服务，社交网络信息爬网服务，数据存储和检索集群以及民意信息处理模型。

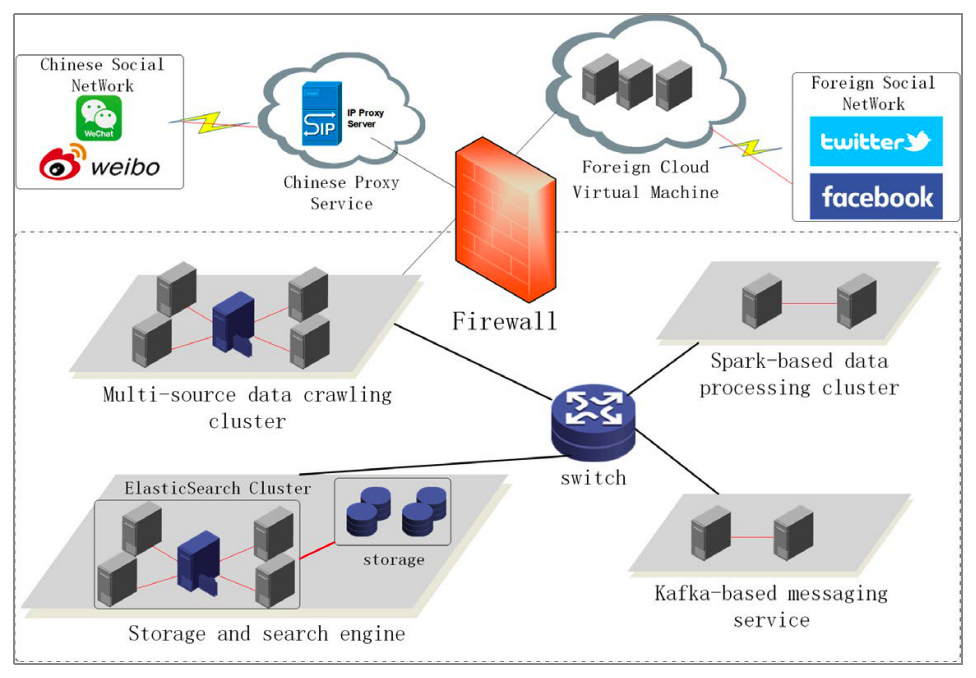


图2：SNES的网络拓扑

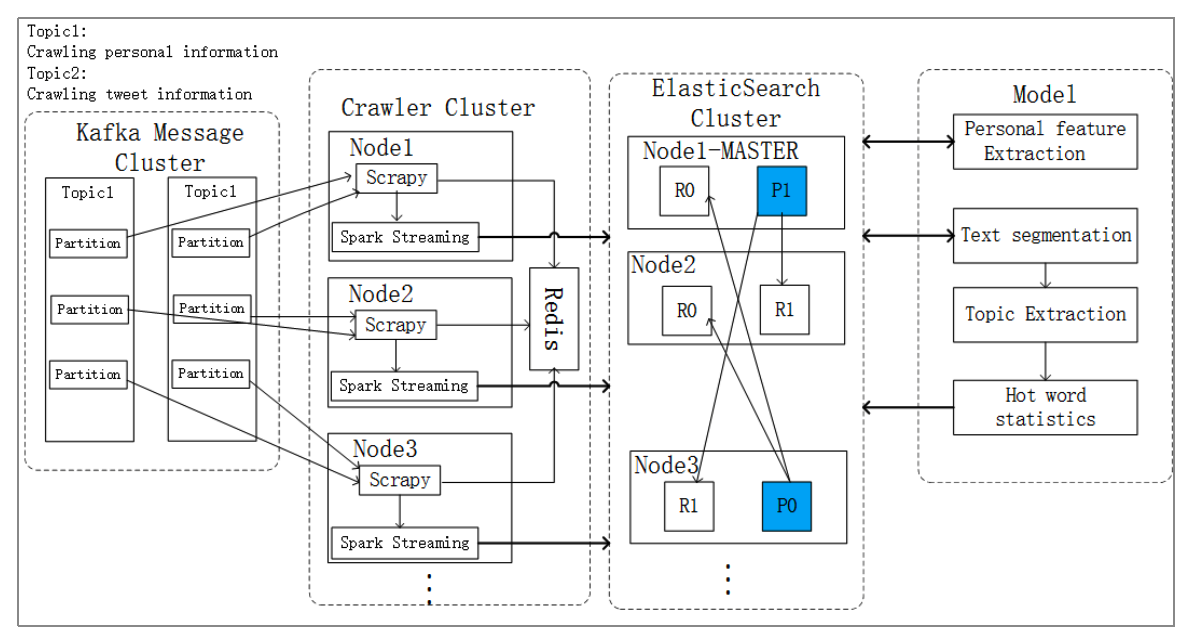


图3：SNES的内部结构图

* 1. 数据抓取模块

搜寻器集群由几个节点组成，每个节点都充当消息订阅模块的使用者（第3.3章），每个节点都包含搜寻器框架Scrapy-redis，Spark Streaming框架和Redis数据库。数据爬网模块的框架结构和内部数据流如图4所示。为了提高数据爬网速度并避免由于网络问题或目标网站的限制而中断，我们设计了与消息订阅模块进行通信的顶级调度程序简要的工作流程如下：

步骤1：顶层调度程序从消息订阅模块请求新的数据抓取任务；

步骤2：如果任务不为空，则执行步骤1，否则执行步骤3；否则，执行步骤3。

步骤3：顶级调度程序向Spider发出数据爬网命令；

步骤4：Spider从redis数据库获取第一个URL并发送给引擎；

步骤5：如果URL已在爬网队列中，请转到步骤1，否则转到步骤6；否则，请转到步骤6。

步骤6：引擎从cookie池中请求随机cookie，并对cookie和代理Ip进行随机化；

步骤7：火花流处理模块处理获取的数据并将其存储到海量数据存储和提取模块中（第3.4章）；

步骤8：将URL放入爬网队列，然后转到步骤1。

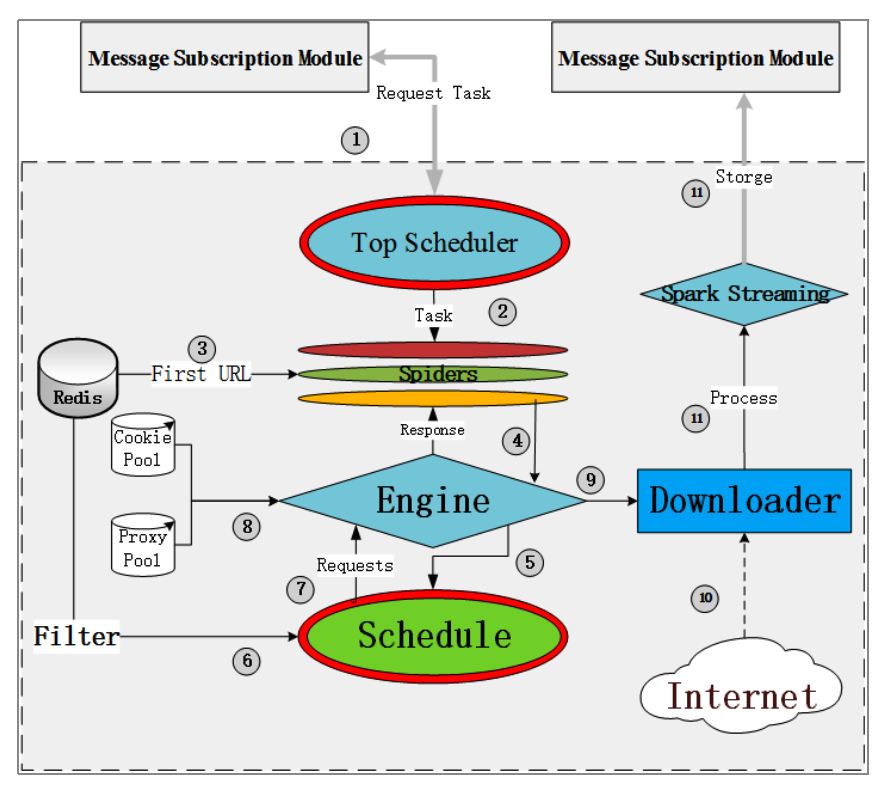


图4：数据爬网模块的帧结构和内部数据流

* 1. 消息订阅模块

为了提高平台中模块之间的可伸缩性，容错性和去耦性，我们设计了基于Kafka的消息订阅模块。消息订阅群集由两个工作节点组成。图中的每个消息主题都分为三个分区，每个使用者节点使用每个主题的一个分区。经过经验实验后得出的结论是，为了获得更好的性能，主题分区的数量最好设置为使用者的倍数，并且最好确保每个使用者监视的消息分区的数量相等。用户可以根据需要订阅特定的事件或主题，第3.1章中的数据爬网模块将使用该消息并爬网相应的数据。

对于每个主题，消息订阅都维护一个看起来像图5左侧的分区消息。每个分区都是有序的，不可变的记录序列，该记录序列被连续附加到结构化提交消息中。分区中的每个记录都分配有一个称为偏移的顺序ID号，该ID唯一地标识分区中的每个记录。实际上，基于每个消费者保留的唯一元数据是该消费者在消息中的偏移量或位置。此偏移量由使用者控制：通常，使用者在读取记录时会线性地推进其偏移量，但实际上，由于位置是由使用者控制的，因此它可以按喜欢的任何顺序使用记录。如图5的右侧所示，使用者可以重置到较早的偏移量以重新处理过去的数据，也可以跳到最近的记录并从现在开始使用。

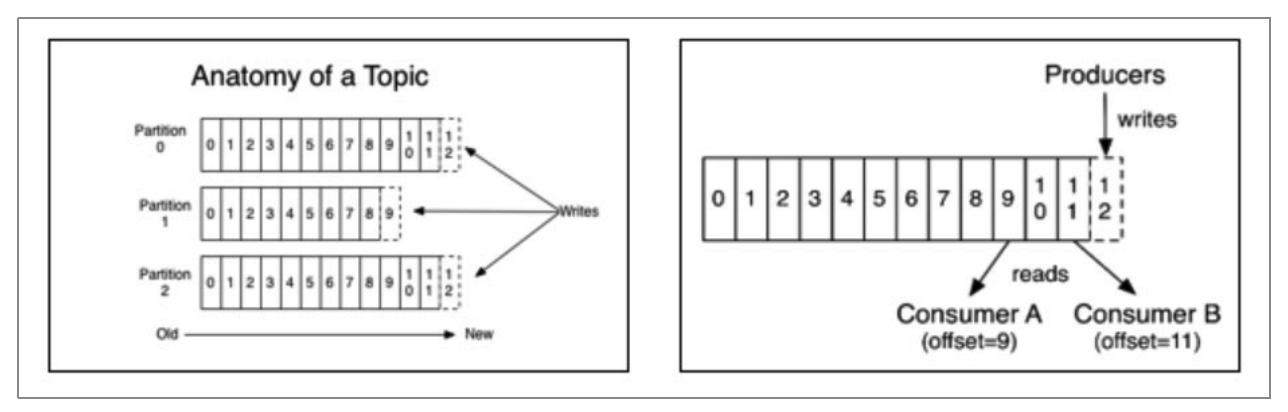


图5：主题产生和消费的理论

* 1. 海量数据存储和提取模块

挖掘社交网络舆论和提取热门话题需要有效地存储和检索非结构化数据的大量UGC。我们基于Elastic Search设计了一个庞大的数据存储和提取模块。海量数据存储和提取模块的数据结构如图6所示。每个节点具有三个分片，其中P是主要分片，R是冗余分片。主分片P1，P2和冗余分片R0存储在节点1中，冗余分片R0，R1和R2存储在节点2中，主分片P0和冗余分片R1，R2存储在节点3中。总共3个主要分片和6个冗余分片。同时，我们还注意到节点1也有一个主ID，这意味着它是一个主节点。它比其他节点更特别。它具有控制整个群集的权限，例如资源分配，节点修改。

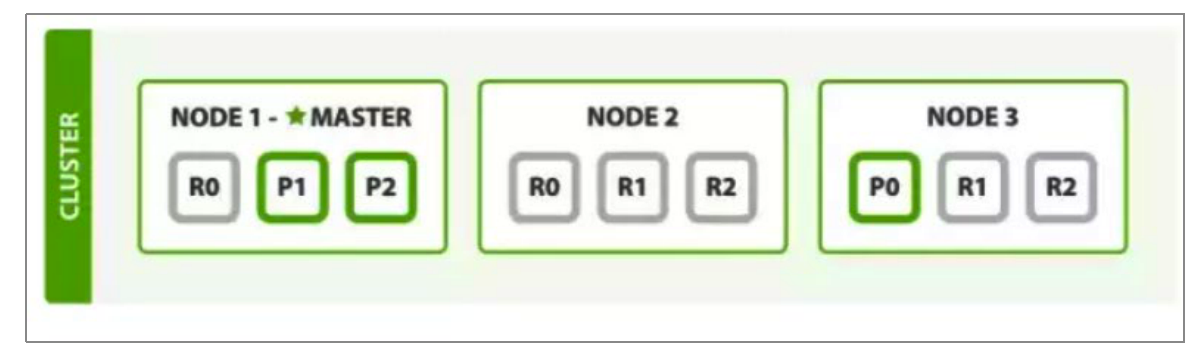


图6：主题产生和消费的理论

1. **实验**

通过将其应用于新浪微博实时数据，我们评估了舆论监督和SNES绩效的影响。

* 1. 实验环境

为了确保平台的数据采集，处理和检索功能，并考虑系统的可扩展性和负载平衡，我们构建了实验平台环境，如表1所示。该平台的网络带宽测试如图7所示。

表1：实验平台的硬件设施和应用

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 服务器 | CPU | 内存 | 磁盘 | 应用 |
| Server-1 | 3.60 GHz 2 core 4 threads | 8GB | 500GN | ES-Master |
| Server-2 | 3.60 GHz 2 core 4 threads | 8GB | 500GN | ES-Data |
| Server-3 | 3.60 GHz 2 core 4 threads | 8GB | 500GN | ES-Data |
| Server-4 | 3.60 GHz 2 core 4 threads | 8GB | 500GN | ES-Data |
| Server-5 | 3.60 GHz 2 core 4 threads | 8GB | 250GN | Kafka-Broker |
| Server-6 | 3.60 GHz 2 core 4 threads | 8GB | 500GN | Kafka-Broker |
| Server-7 | 3.60 GHz 2 core 4 threads | 8GB | 250GN | Kafka-Broker |
| Server-8 | 3.60 GHz 2 core 4 threads | 8GB | 500GN | Spider&Spark |
| Server-9 | 3.60 GHz 2 core 4 threads | 8GB | 250GN | Spider&Spark |
| Server-10 | 3.60 GHz 2 core 4 threads | 8GB | 500GN | Spider&Spark |

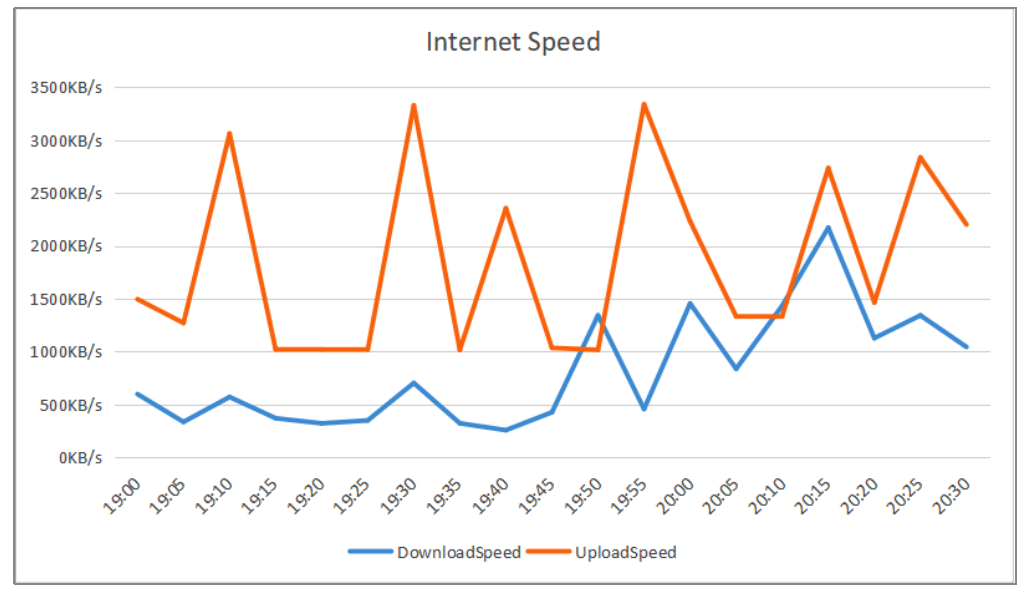


图7：实验环境的网络带宽

* 1. 实验数据

本文的实验数据是从新浪微博数据中实时收集的，包括个人信息，关系网，推文和交互信息（喜欢或不喜欢，评论）等。如表2所示； 表3显示了用于搜索和写入延迟测试的数据格式。

表2：用于事件监视的实验数据的统计信息

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 数据源 | 人数 | 关系 | 推文 | 互动 |
| 新浪微博 | 130,028 | 71,032 | 1,501,202 | 4,130,187 |

表3：用于搜索和写入的数据格式

|  |  |  |
| --- | --- | --- |
| Field | Type | Length |
| userId | Integer | 32bit |
| userName | Char(20) | 160bit |
| content | Char(150) | 1200bit |
| date | Long | 64bit |
| platform | Short | 16bit |
| thumb | Integer | 32bit |
| comment | Integer | 32bit |
| forward | Integer | 32bit |

* 1. 实验结果
     1. 性能评估

本文将搜索效率，爬网速度和写入效率用作该平台的评估指标。爬行数据的速度。我们使用等式（1）来衡量爬网数据的速度。

(1)

其中n是爬网群集节点的数量； h为平均CPU频率，单位为GHz； t是总爬行时间，单位是秒； c是t秒内抓取的页面总数； bw表示该时间段内网络带宽的平均值，单位为MB / S。因此，sp表示单位带宽条件下在CPU时钟周期内单线程单节点爬网的页面的平均值。sp的值可以在某种程度上减少节点数，CPU线程，CPU频率和网络带宽的影响。在本文进行的实验中，爬网时间为48小时，爬网的总页数为864078，h = 3.6，bw = 2，结果sp = 0.07，由等式（1）计算得出。

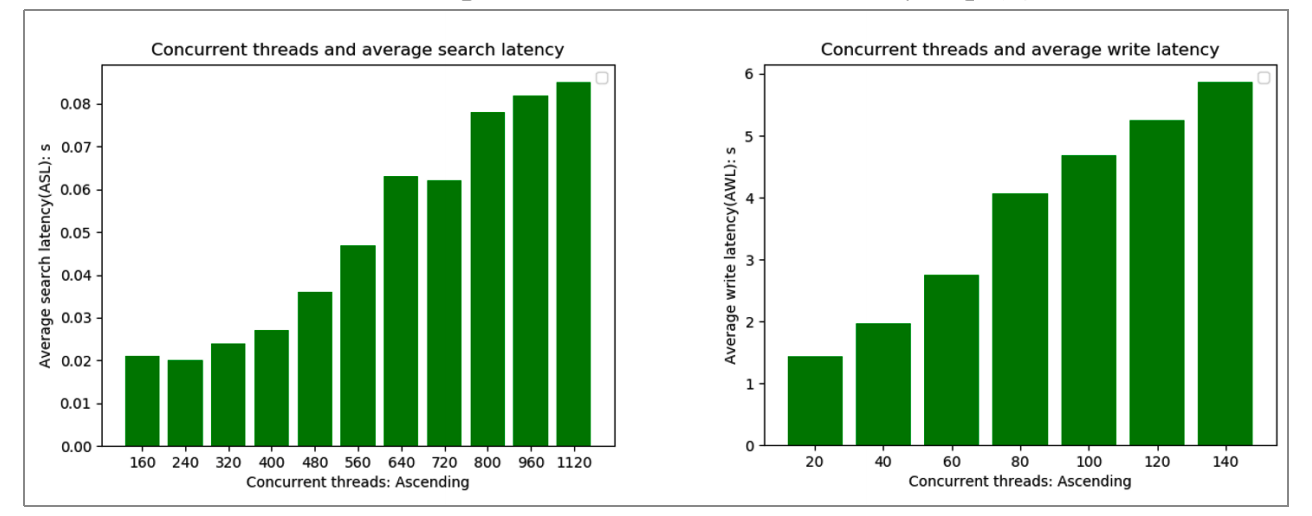


图8：SNES的搜索和写入延迟

我们使用并发读写脚本评估系统的搜索延迟和索引延迟，测试时间为50分钟。我们用等式（2）来衡量延迟

(2)

其中n是实验次数，ti是第i个实验的延迟，单位是秒； ASL是平均搜索延迟。图8的左侧显示了并发搜索测试的结果，我们对每个特定数量的并发请求进行了4000次实验，并获得了搜索等待时间。图8的右侧显示了我们的并发写入测试的结果，我们对每个特定数量的并发请求进行了2000次实验，并获得了写入延迟。所有数据都是随机生成的，格式如表3所示。

* + 1. 事件监控

本实验针对“2018年8月28日中国华住信息泄露事件”，搜索关键词为“华住泄漏”，监控时间为2018年8月28日至2018年9月10日。图9显示了有关“华珠信息泄漏事件”的系统监控的传输量趋势。可以看出，从8月28日到9月2日，生产线正在上升，表示事件正在升温。9月4日之后，它逐渐下降，直到达到较低水平。这与整个事件的发展是一致的。图9的右侧显示了通过从爬网的微博中提取和过滤内容主题而获得的事件的热词云。可以看出，诸如“ HuaZhu”，“ Information”，“ Leakage”，“ Data”和“ Cyber security”之类的词在词云中所占比例最高，这些词可以恰当地描述和总结该事件的内容以及微博用户及相关媒体对此事件的讨论和看法。

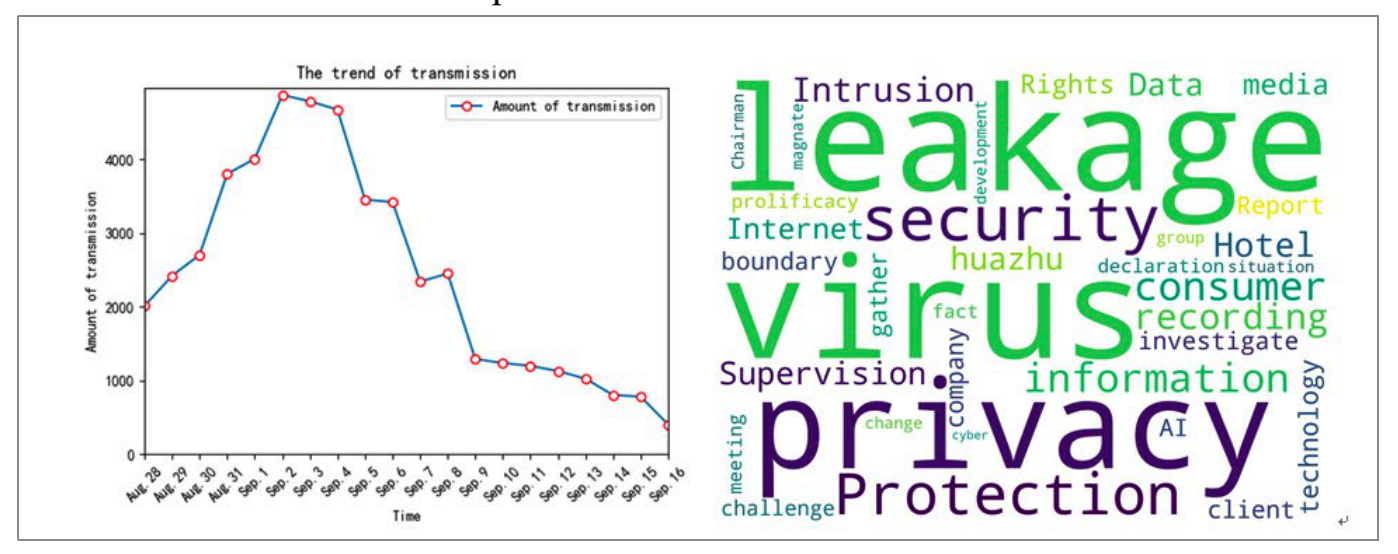


图9：“华住信息泄漏”的每日传播和热词云

1. **结论与讨论**

我们在本文中提出的SNES创造性地集成了基于Kafka的高吞吐量订阅模块，分布式爬网框架，实时流计算框架以及基于ElasticSearch的海量数据存储和提取模块。经过以上实验，SNES体系结构具有以下优点：

1. 实时爬行不断变化的社交媒体数据。
2. 快速有效地存储和检索海量数据。
3. 良好的可扩展性和容错性。

当前，我们的平台在数据安全性存储和加密保护方面存在缺陷，我们未来的工作将集中在提高平台的安全性上。

**致谢：**该工作得到国家电网科技计划项目的资助，资助号为520613180002、62061318C002，中央大学基础研究基金（资助号为HIT.NSRIF.201714），威海科技发展计划（2016DXGJMS15）和 山东省重点研究开发计划（2017GGX90103），福建省中青年教师教育研究项目，授权号JAT160466，江苏农林科技职业技术学院重点研究开发项目（2018kj11），智能农业控制系统研究与开发 基于Spark大数据决策（2017N0029）。

**参考文献**

**Burstein, P.** (2003): The impact of public opinion on public policy: a review and an agenda. Political Research Quarterly, vol. 56, no. 1, pp. 29-40.

**Fan, Y.** (2018): Design and implementation of distributed crawler system based on scrapy. IOP Conference Series: Earth and Environmental Science, vol. 108.

**Hindelang, M. J.** (1974): Public opinion regarding crime, criminal justice, and related topics. Journal of Research in Crime and Delinquency, vol. 11, no. 2, pp. 101-116.

**Jie, D.; Jungang, X.** (2009): IPOMS: an internet public opinion monitoring system. Applications of Digital Information and Web Technologies, pp. 433-437.

**Li, J.; Wei, Q.; Xi, Y.** (2018): Analysis and research on the development of university libraries based on wechat and facebook social network platform. Journal of Sichuan Library Science, no. 6, pp. 69-74.

**Li, R.; Li, B.; Zhang, G.; Jiang, J.; Luo, Y.** (2018): A high-performance and flexible chemical structure&data search engine built on couchdb&elasticsearch. Chinese Journal of Chemical Physics, vol. 31, no. 3, pp. 341-349.

**Lu, S.** (2018): Design and Implementation of a Message Publishing Service Based on Kafka (Ph.D. Thesis). Nanjing University.

**Shen, L.** (2018): Design and Implementation of Weibo Public Opinion Monitoring System Based on Spark (Ph.D. Thesis). University of Electronic Science and Technology.

**Tan, L.; Zhou, J.** (2018): Real-time traffic data processing platform based on spark streaming. Computer System Application, vol. 27, no. 10, pp. 133-139.

**Tang, Y.** (2013): Design and Implementation of Internet Public Opinion Monitoring System (Ph.D. Thesis). Beijing University of Posts and Telecommunications.

**Wang, J.; Guo, Y.** (2012): Scrapy-based crawling and user-behavior characteristics analysis on taobao. Cyber-Enabled Distributed Computing and Knowledge Discovery, pp. 44-52.

**Xie, D. X.; Xia, W. F.** (2014): Design and implementation of the topic-focused crawler based on scrapy. Advanced Materials Research, vol. 850, pp. 487-490.

**Yan, H.; Yu, X.** (2018): Software system design based on Kafka message platform. Electronic Technology and Software Engineering, no. 18, pp. 38.

**Yan, L.; Wang, X.** (2018): Real-time movie recommendation research based on spark streaming computing. Software Guide, pp. 1-5.

**Zhang, X.; Li, Z.; Liu, G.; Xu, J.; Xie, T. et al.** (2018): A spark scheduling strategy for heterogeneous cluster. Computers, Materials & Continua, vol. 55, no. 3, pp. 405-441.

**Zhou, Y.; Han, X.** (2015): Application examples of elasticsearch in e-commerce system. Information Technology and Standardization, vol. 5, pp. 30.

Chuiju You,Dongjie Zhu,2,Yundong Sun,Anshan Ye,Gangshan Wu,Ning Cao,Jinming Qiu,Helen Min Zhou. SNES: Social-Network-Oriented Public Opinion Monitoring Platform Based on ElasticSearch[J]. CMC: Computers, Materials & Continua,2019,61(3).

1. 专业阅读书目
   1. 书目一

内容摘要：

正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文（ABS）

* 1. 书目一

内容摘要：

正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文（ABS）

封底

1. 对论文正文中某一特定内容的进一步解释或补充说明性的注释，置于本页地脚，前面用圈码标识。 [↑](#footnote-ref-0)