

基于 Hadoop 的云笔记设计与实现

黄 伟 程 浩 张 樊

武昌工学院, 中国·湖北 武汉 430065

【摘 要】近几年,云服务成为了全球科技巨头增长最快的领域,大部分企业也意识到使用云服务可以节省高昂的IT投入费用。本项目分为两期开发项目,两期开发主要完成云笔记的用户模块、笔记模块和基于云分布式的存储和计算,以及在云笔记的使用过程中对用户的行为数据进行记录和分析的过程。项目的Web应用服务采用Python Flask框架完成编写,WebUI部分采用Bootstrap、editor.md、jQuery等前端框架进行编写,开发过程中使用MySQL进行数据存储,在基于云分布式框架的二次开发时,采用分布式架构进行半结构化或非结构化数据的存储。

【关键词】云笔记;大数据;Hadoop

【基金项目】湖北省大学生创新创业训练计划项目成果,课题名称:起航软件工作室——软件服务外包(201713241020)。

引言

随着技术和生活水平的提高,智能设备已成为人们日常生活中最不可或缺的存在。然而,这种不可或缺的存在也会带来很多不便,目前最大问题是数据同步问题。当同时使用多台智能设备时,会发现每台设备系统中都会有自带的同步功能,而这种同步功能却仅限于本设备产品之间的数据同步,这使得人们不得不更换同品牌设备或被迫下载第三方附带同步功能的软件,而这种问题却只是冰山一角。当在设备上寻找或整理凌乱的文档时,可能又会遇到文件资料散落在不同的终端设备或存储设备中、文件版本过多导致无法找到需要的版本或误删重要文件的问题。为了解决这些问题,本项目开发一款用于存储和处理大量文件数据的云笔记项目,能够实现在线文档编辑,数据自动同步,文档数据恢复的功能。

1 相关技术选择

1.1 云笔记Web应用技术选择

云笔记的整套业务实现,采用Python语言进行编写。在后台业务中,将通过pip库,下载pymysql、happybase、impala、Flask以及复杂的依赖关系等模块,完成Web后端各业务逻辑部分功能编写。

1.2 Web应用框架的选择

云笔记的整个Web应用将使用Flask框架进行开发,由于Flask使用的Jinja2模板引擎,在编写Web页面与视图控制器之间交互时显得非常简便。在常用的Flask会话存储方式中,本项目选择redis数据库进行会话存储,它不仅有着性能上的优势,并且在负载均衡和多进程项目部署时,可达到会话统一的功效。

1.3 云计算框架选择

云计算平台的分布式基础架构选用主流的Hadoop框架。在基于Hadoop的云笔记中,采用HBase对用户的笔记信息进行非结构化存储。使用Impala+Hive+Hbase实现大量数据快速查询。

2 平台需求分析

根据需求分析,基于Hadoop的云笔记主要是通过后台逻辑业务,处理和存储用户在前端进行的所有操作,在数据库设计方面会更多倾向于逻辑数据。通过对云笔记平台的分析得到,用户在非登录状态下仅能访问用户共享数据的预览,其大部分功能都需要用户在登录后才能进行操作。

2.1 用户注册:在访客状态下,通过用户登录/注册页面填写电子邮箱地址或手机号码等信息注册新的用户,也可以通过微信扫码,直接跳过页面注册步骤。

2.2 用户登录:访客在已注册账号的情况下,可以通过页面输入邮箱地址或手机号码以及登录密码或动态密码的方式进行登录,也可以通过微信扫码登录个人账户。

2.3 个人中心:成功进行登录的用户可以通过云笔记首页,单击“个人中心”图标或按钮进入个人中心页面。

2.4 文件管理:用户在登录的状态下可以对自己的笔记或文档数据进行删除、移动、回收、重命名、新建、共享等操作。

2.5 支持文档操作:用户在登录的状态下,通过网页或小程序在线查看、编辑文档内容。

2.6 数据共享管理:用户在登录状态下可以通过个人中心页面,对当前用户的所有共享信息进行查看、取消共享等操作,提供文档共享的URL链接与二维码。

2.7 数据历史版本管理:用户在登录状态下可以通过右键选择文档,单击“查看历史版本”选项,查看当前文档的历史版本,并可以进行删除或恢复等操作。

3 平台总体设计

3.1 平台总体架构

基于Hadoop的云笔记平台使用CentOS 7, JDK 1.8.0, CDH 5.7.2, Python3.6.0进行搭建,云笔记平台架构由MariaDB 5.5.44、aliyun轻量应用服务器、HDFS集群组成。服务器采用Python常见的web部署搭配gunicorn+engine。gunicorn是Python的WSGI-HTTP服务器,engine是一款淘宝基于Nginx,针对大访问量网站的需求开发添加了很多高级功能和特性的服务器。

3.2 功能模块规划设计

通过对现有的企业级云笔记平台提供的功能模块进行的调查与分析,云笔记并不仅限于进行笔记文件的操作。实际上,它可以支持管理更多的文档文件类型。基于Hadoop的云笔记可以设计为前台和后台两大模块,主要功能如下。

3.2.1 登录注册模块:用户通过此模块可以选择多种登录或注册方式。

3.2.2 个人中心模块:此模块的主要功能是对用户当前所绑定的手机号、电子邮箱地址、微信账号以及用户基础信息和文档分享状态等提供查看和修改的功能。

3.2.3 消息通知模块:此模块是通过第三方的短信平台提供的短信接口和pip库提供的smtplib模块以及微信小程序提供的消息推送服务进行封装,主要是对用户的重要消息进行实时推送,包括用户找回密码和绑定验证等消息。

3.2.4 文件系统模块:为每个用户提供独立的文件系统,使用户在操作云笔记文件和目录时就像操作移动盘一样简单。

3.2.5 在线编辑模块:此模块是通过在Web端,通过前端框架与Flask控制器之间数据交互,实现笔记与多种类文档在网页或小程序中能够进行在线编辑与查看,以及文档自动保存和自动创建历史版本功能。

3.2.6 文档版本管理模块:用户在编辑文档时,可以通过此模块创建文档快照。

3.2.7 笔记文档共享模块:此模块主要对用户个人文档或文件目录进行分享的状态进行查询和修改,所有用户所共享的目录、

(下转234页)

(4) 衰减阶段

随着该舆情事件进入衰减阶段,发量逐渐递减,但该事件引发的话题逐渐表现出多样性。话题主要涉及“对悲剧本身的哀悼”“身心压力排解”“研究生教育”等方面,因此其情感倾向表现为中性与负面之间的激烈震荡。

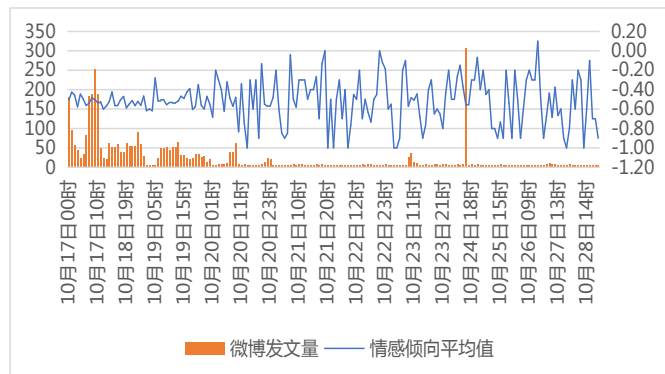


图7 衰减阶段发量及情感倾向平均极性

4 结论

本文以“大连理工大学研究生自杀”事件为例,对高校网络舆情事件情感倾向演化机制及影响因素进行分析。从本次事件的生命周期图可看出,事件从产生到高潮仅经历了2天时间。相较于一般网络舆情事件,高校网络舆情事件的传播速度更快,传播面更广。因此,为了有效引导舆情事件发展,相关部门的主动积极应对是非常必要的。高校等相关部门如果不能在第一时间给予回应,将错过对网民情感引导的最佳时间。所幸,在本次事件中,大连理工大学在事件发生后24小时之内发布了事件通报,不但避免了人们在信息量不足的情况下妄加猜测,更对网民情绪的因势利导起到了重要作用。其次,主流媒体的话题引导对网络舆情情感倾向演化起到至关重要的作用。由于高校网络舆情事件的个人主体主要为高校在校生,所以,在本次突发

事件中,他们有感而发,表达了对逝者的哀悼,为生命的逝去感到痛心。该事件负面情感倾向占比较大,网络情绪主要集中在“悲伤”“难过”“惋惜”等方面。此时主流媒体的介入不但将事件推向高潮,更将话题引向“教育”“成长”“如何克服压力”,使得网络舆情情感倾向极性向中性演进。

总之,由于高校网络舆情主要参与主体是高校学生,他们本身所具有的易情绪化、群体集聚性强等特征使得高校网络舆情事件表现为一旦产生即扩散迅速,且情感极性强。因此,情感倾向演化不但受制于主体本身固有情感特征,还会受到相关部门的危机应对举措,主流媒体话题的影响。

参考文献:

- [1] 张一文, 齐佳音, 方滨兴, 等. 非常规突发事件网络舆情热度评价体系研究[J]. 情报科学, 2011, 029 (009): 1418-1424.
- [2] 郑岩. 高校网络舆情监测系统的目标定位、评判依据与运行保障研究[J]. 情报科学, 2015, 033 (006): 81-85.
- [3] 魏巍. 高校网络舆情特点及对思想政治教育的影响研究[J]. 沈阳工程学院学报(社会科学版) (1期): 140-144.
- [4] 赵扬. 高校网络舆情引导研究[D]. 东北师范大学, 2016.
- [5] 蒋知义, 马王荣, 邹凯, 李黎. 基于情感倾向性分析的网络舆情情感演化特征研究[J]. 现代情报, 2018, 38 (04): 50-57.
- [6] 唐超. 网络情绪演进的实证研究[J]. 情报杂志, 2012, 10 (31): 48-52.
- [7] 陈忆金, 曹树金, 陈桂鸿. 网络舆情意见挖掘: 用户评论情感倾向分析研究[J]. 图书情报知识, 2013 (6): 90-96.
- [8] 陈福集, 黄江玲. 我国网络舆情演变文献研究综述[J]. 情报杂志, 2013 (07): 54-58.

作者简介:

段楠楠 (1981-), 女, 汉族, 湖南省祁阳县, 博士, 副教授, 研究方向: 网络舆情。

(上接231页)

文档、是否有密码, 过期时间等信息都由此模块进行管理。

3.2.8 文档智能整理模块: 此模块是通过云计算, 快速将用户所选择文件夹内所有文件进行分类, 即使是超大量的文件与文件夹结构。

3.2.9 笔记收藏模块: 此模块主要功能是对用户查看他人共享的笔记或文档进行收藏和拷贝。

4 系统实现

4.1 文件系统设计

文件系统以Windows资源管理器的“小图标”样式进行设计, 样式简洁, 操作方便, 考虑到手机用户和平板用户的体验。文件系统的后端通过对数据库的增查改进行逻辑操作。

4.2 文档在线编辑

本平台的云笔记主要的笔记格式并非HTML富文本编辑器, 而是选择使用Markdown文本编辑器, Markdown(.md)是一种普通文本编辑器编写的标记语言, 这种类型的文档便于编辑和查看, 可以通过框架很自动转换为HTML页面进行展示。

4.3 文档在线预览

Office是微软公司开发的一套办公软件套装, 常用组件

有Word、Excel、PowerPoint。设计系统, 通过兼容这三中文件类型, 实现在线预览部分office文档。

5 总结

基于Hadoop的云笔记的开发, 是通过网络传输的方式, 将大量用户的大量文档数据通过平台设计的文件系统进行存储和集中处理, 并为用户提供文件在线处理的Web服务, 给用户带来更加便捷的文件存储和编辑方式。

本项目通过将Hadoop云计算平台整合到Web应用项目中进行研究与设计, 取得相应的设计成果。成功搭建Hadoop分布式计算平台, 使云笔记项目在平台完成部署和测试, 并完成对基于Hadoop的云笔记进行设计和实现。

参考文献:

- [1] 陈彩云, 李鹏, 温武, 郭四稳. 微信二维码授权登录网站的实现[J]. 现代计算机(专业版), 2017 (01): 70-72.
- [2] 陈建尧. 一种基于Hive数据仓库的调度优化研究[J]. 科技通报, 2018, 34 (08): 113-117.