Computer Science and Application 计算机科学与应用, 2019, 9(6), 1134-1141
Published Online June 2019 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2019.96128

Hans汉斯

Design and Development of University Post Bar Public Opinion Analysis System Based on Big Data

Hanqing Cao[1], Quanbin Li[2*]

[1]School of Science and Literature, Jiangsu Normal University, Xuzhou Jiangsu
[2]College of Physics and Electronic Engineering, Jiangsu Normal University, Xuzhou Jiangsu
Email: *liqbzy@163.com

Received: Jun. 6th, 2019; accepted: Jun. 20th, 2019; published: Jun. 28th, 2019

Abstract

With the improvement of computer storage capacity and the development of complex algorithms, data volume has developed exponentially in recent years. Decision-making in all walks of life is changing from "business-driven" to "data-driven". We should make use of the massive processing and intelligent analysis ability of large data to accurately grasp the hot data of the times. Baidu Post Bar is owned by most universities in the country. Under this column, the data of Post Bar is large and scattered. Nowadays, there is an urgent need for an efficient and accurate public opinion analysis system. On this basis, this paper focuses on the analysis of the data design and architecture design of the public opinion analysis system based on the large data environment, as well as the technical lines in the design process. Different from hot search, Baidu Index and other platforms specializing in analyzing user behavior data, this system is completely innovative, more technical and more professional.

Keywords

Big Data, Baidu Post Bar, Public Opinion Analysis System

基于大数据的高校贴吧舆情分析系统设计与开发

曹汉清[1]，李全彬[2*]

[1]江苏师范大学科文学院，江苏 徐州
[2]江苏师范大学泉山校区物电学院，江苏 徐州
Email: *liqbzy@163.com
*通讯作者。

文章引用：曹汉清，李全彬．基于大数据的高校贴吧舆情分析系统设计与开发[J]．计算机科学与应用，2019，9(6): 1134-1141. DOI: 10.12677/csa.2019.96128

收稿日期：2019年6月6日；录用日期：2019年6月20日；发布日期：2019年6月28日

摘　要

随着计算机存储能力的提升和复杂算法的发展，近年来数据量呈指数级发展。各行各业的决策正从“业务驱动”转变为“数据驱动”。我们应该利用大数据的海量处理与智能分析能力，准确抓取时代热点数据[1]。百度贴吧为全国大多数高校所拥有，在此栏目下，贴吧数据信息量大且比较分散。当今，迫切需要一种高效率、高准确率的舆情分析系统。在此基础上本文重点分析了基于大数据环境下的高校贴吧舆情分析系统的数据流程设计和架构设计，以及设计过程中的技术线路。区别于热搜、百度指数等专职于分析用户行为数据的平台，本系统完全创新，技术性更强，专业性更高。

关键词

大数据，百度贴吧，舆情分析系统

1. 引言

大数据背景下，热点数据抓取问题推动着经济的发展，企业、政府部门以及事业单位都需要获取与自己相关的热点信息，分析用户行为数据，进而抓住市场痛点，走向发展前沿。例如百度指数[2]，通过研究关键词关注趋势、洞察网民需求变化、监测媒体舆情趋势、定位数字消费者特征，以百度海量网民行为数据为基础的数据分析平台，是当前互联网乃至整个数据时代最重要的统计分析平台之一，自发布之日便成为众多企业营销决策的重要依据。王海燕等人[3]构建了集数据采集、处理、分析、评价、反馈于一体的大学生网络行为的分析模型，实现了个性化的分析。刘滢[4]基于 Hadoop 云计算平台和 Map Reduce 编程框架设计出四层结构模式智能交通系统，云处理层包含多种模型，实现任务并行化处理；算法层最优化算法，分析和查看某一时刻路段中的车流状况以及每一条道路的分流状况；业务处理层处理业务逻辑，实现业务控制管理和调度；界面管理层管理用户需求信息。赵佳钐等人[5]提出混合架构下的多源异构数据集成方案，针对不同类型的数据分别采用 OldSQL、NewSQL 和 Hadoop 技术进行数据的清洗集成，设计出高校多源异构数据集成系统。李斌等人[6]设计一种大数据高效能平台，以算粒为基本研究对象，深入剖析大数据应用算法的特征，合理划分各计算子任务；其次，构造体系结构匹配矩阵，并将子任务分配到合理的处理部件上；最后，利用动态电压/频率调节技术和数据布局算法，实现非关键任务的电压控制，并优化关键任务的结构布局。

本系统区别于热搜、百度指数等专职于分析用户行为数据的平台，属于一种全新的舆情分析系统，同时基于 Hdoop 平台和 MapReduce 编程框架精准定位用户需求，利于传统架构与大数据架构结合，数据抓取分析效率高，用户更便捷地获取所需要的数据信息，易于扩展，数据量高并发的情况下，具有较好的解决能力。

2. 功能介绍

本系统设计五大功能模块，通过分析关键词可得到较为准确的统计结果，以下为各模块详细介绍(如图 1)：

Figure 1. Detailed description of each module
图 1. 各模块详细介绍

2.1. “高校舆论热点关键词”功能模块

在数据分析阶段，系统利用 hadoop mapreduce、IKAnalyzer 技术对贴吧各高校舆论热点进行分词提取(图 2)，统计后输出于该模块。本模块数据应用于其他模块，更加直观的展示统计的热点信息。

1	有没有	5944
2	学长	5292
3	学姐	5259
4	考研	4391
5	学校	3998
6	一下	3989
7	请问	3472
8	专业	2941
9	可以	2584
10	同学	1912
11	研究生	1794
12	各位	1670
13	想问	1548
14	知道	1488
15	学院	1468
16	一起	1403

Figure 2. College public opinion hotspot keyword interface diagram
图 2. 高校舆论热点关键词界面图

2.2. “高校舆论热点关键词占比分析”功能模块

本模块形成主题“热门 TOP100”，对关键词所占比例以扇形统计图形式进行输出(图 3)。通过对关键词专业的分析，有利于了解大学生就业发展方向，学校增加开设热门专业等。

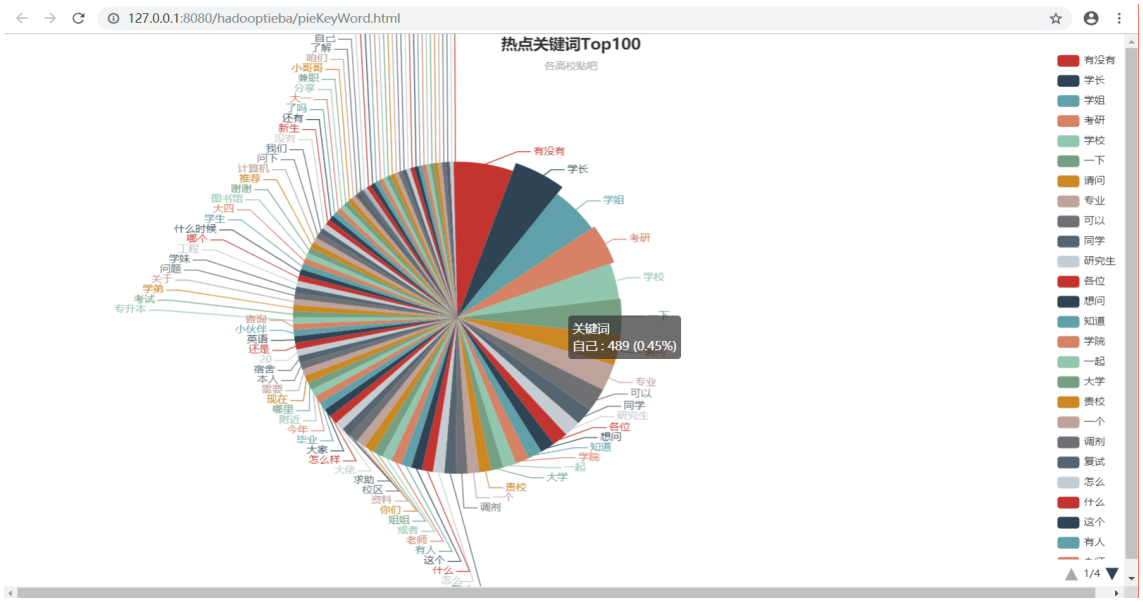

Figure 3. College public opinion hotspot keyword proportion analysis interface diagram
图 3. 高校舆论热点关键词占比分析界面图

2.3. “高校舆论热点关键词男女关注分析”功能模块

本模块键入关键词检索，饼状图形式输出男女生对词关注热度，有利于学校进一步了解大学生男生女生之间思想动态与关注重点的差异，并进行更深层次的探索(图 4)。根据实情，采取不同措施，有针对性的解决舆论热点所带来的负面影响[7]。

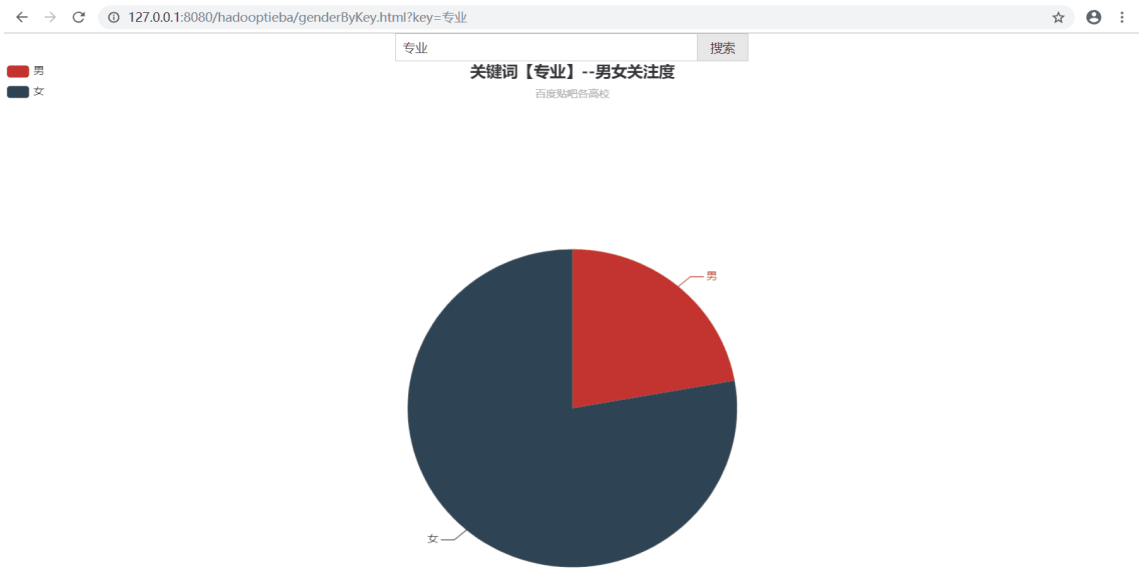

Figure 4. College public opinion hot keywords keyword analysis interface for men and women
图 4. 高校舆论热点关键词男女关注分析界面图

2.4. “高校舆论热点关键词区域分布热度”功能模块

本模块键入关键词检索，中国地图形象输出词的地域分布情况(图 5)，系统的展现了区域文化差异对

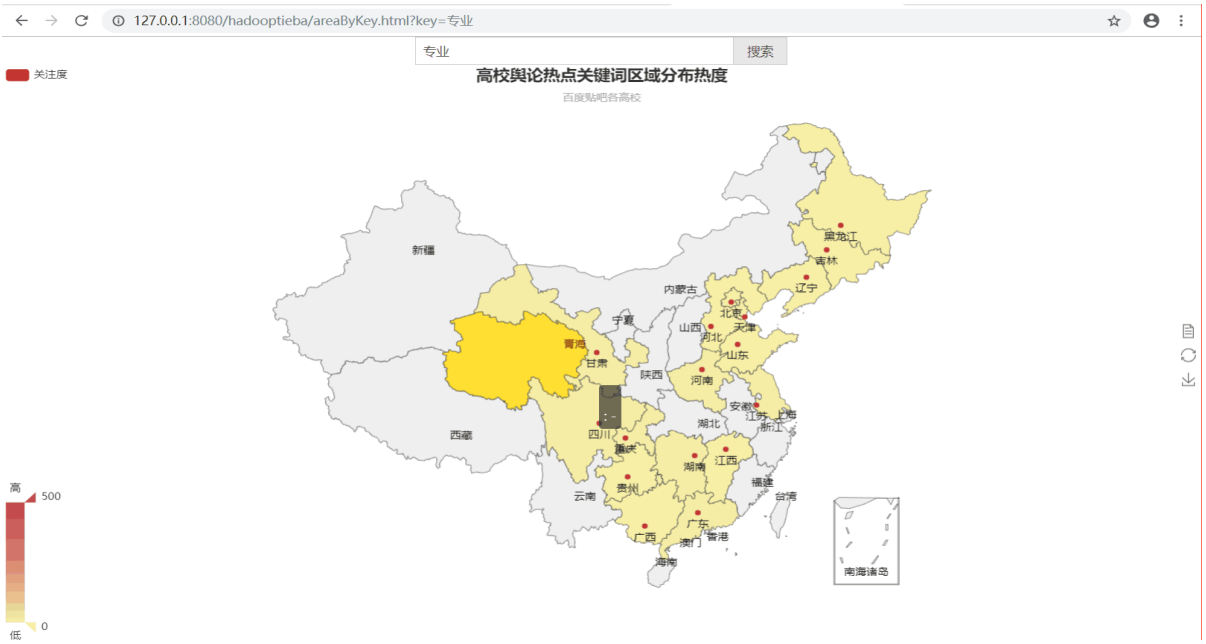

Figure 5. College public opinion hotspot keyword regional distribution hot interface diagram
图 5. 高校舆论热点关键词区域分布热度界面图

舆论热点的影响。

2.5. “高校舆论热点关键词高校分布”功能模块

本模块键入关键词检索，条形图统计输出高校对词关注热度(图 6)。详细地呈现各所高校之间舆论的差异性，高校管理部门结合大学生实际情况采取针对性措施，完善高校网络思想政治教育工作，提高高校思想教育质量[8]。

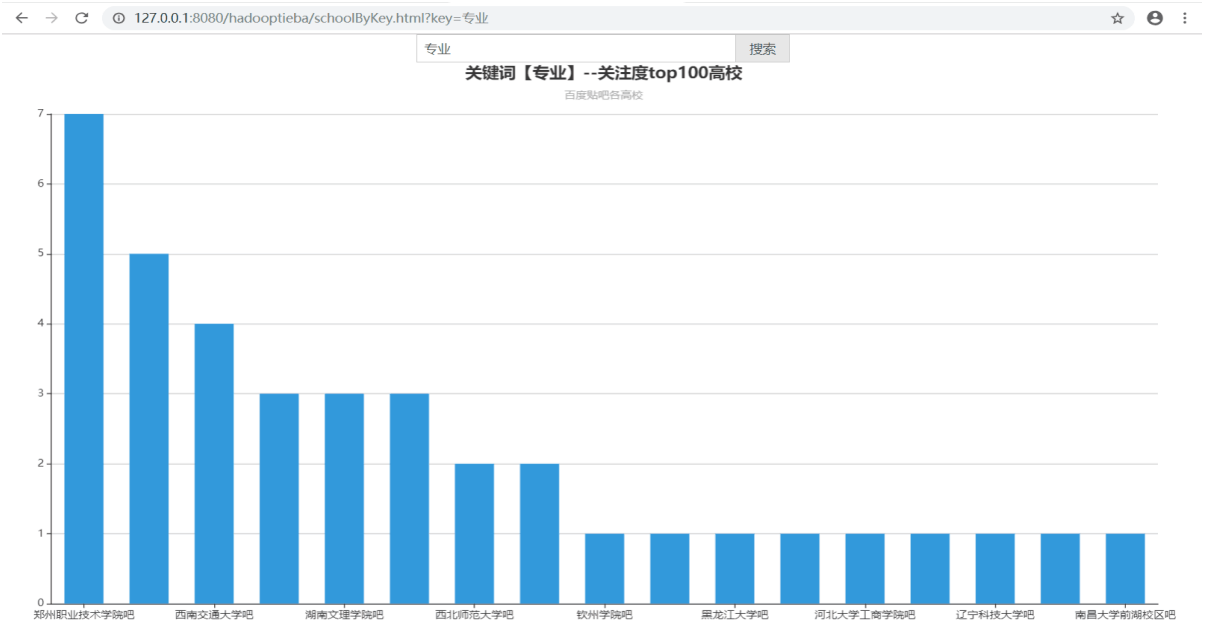

Figure 6. College public opinion hotspot keywords colleges and universities distribution interface map
图 6. 高校舆论热点关键词高校分布界面图

3. 技术线路

系统开发分为四个阶段(图 7)，数据抓取和提取、数据存储、数据分析以及数据可视化。首先，利用 httpclient、json、jsoup、sqoop 技术抓取高校贴吧数据信息，分类写入到 mysql 存储，并将数据导入到 hadoop 空间的 hdfs 中存储。再利用 hadoop mapreduce、IKAnalyzer 进行数据分析和提取。最后利用 spring mvc、mybatis、quartz、amazeui、echarts 技术进行数据可视化展示。

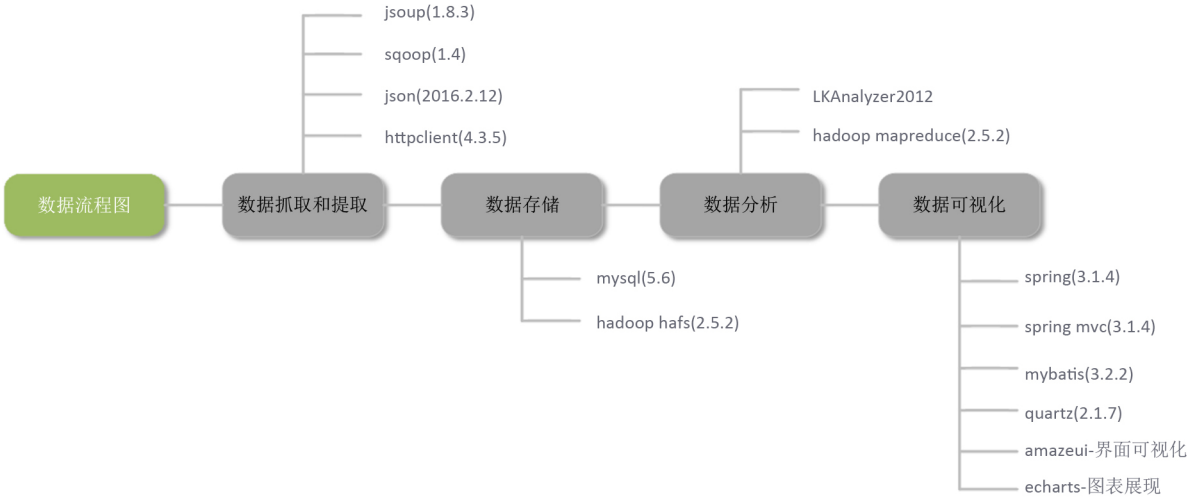

Figure 7. Technical circuit diagram
图 7. 技术线路图

4. 数据流程设计

系统利用 Java 爬虫，htmlparser 与 httpelient 技术抓取高校贴吧数据写入到 mysql 存储，再利用 sqoop 技术将 mysql 数据导入 hadoop 空间的数据存储系统 hdfs 中，并利用 mapreduce 模型进行数据分析、提取，再将数据导入到 Hadoop 空间的数据存储系统 hdfs 中进行存储，最后利用 Java 提取数据写入到 mysql 中存储(图 8)。

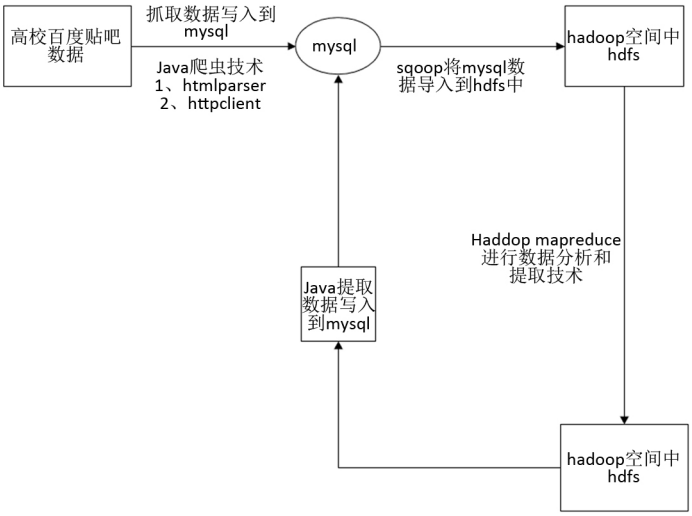

Figure 8. Overall design flow chart
图 8. 总体设计流程图

5. 架构设计

系统采用两套架构(图 9)，传统架构和大数据架构，SSM 做外部抓取与展示技术专业，分布式大数据处理热点信息挖掘效率高。利用 Java 爬虫抓取各高校贴吧数据，分类写入 Mysql 存储，再利用 Sqoop 技术将数据导入 hdfs 空间中存储，并利用 Mapreduce 模型进行数据分析、提取，最后利用 Java 导入 mysql 中存储，通过 webserver 提供网上信息浏览。

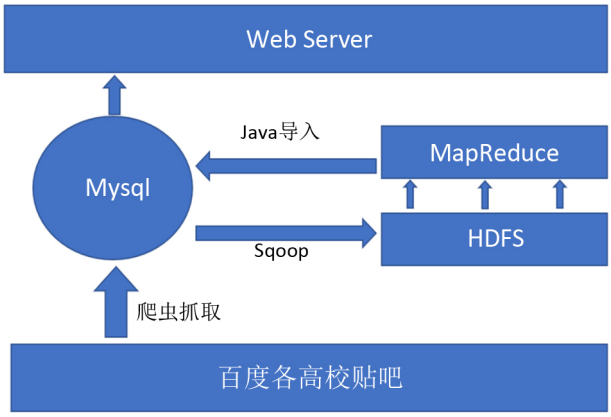

Figure 9. Architecture design
图 9. 架构设计图

6. 结论

基于大数据的高校贴吧舆情分析系统，结合传统架构与大数据架构，SSM 抓取与展示更专业，热点数据挖掘效率更高。基于 Hadoop MapReduce 与 IKAnalyzer 技术实现的数据分析与统计，并采用定时分析，减少了重复劳动，对于数据量高并发的情况下，具有较好的解决能力，更精准地定位用户关注的舆情信息。基于 spring mvc、mybatis、quartz、amazeui、echarts 技术，通过扇形统计图、条形统计图以及饼状图等多种样式界面可视化输出，达到了直观、交互性良好的效果，可对一些有价值的数据进行深入分析与采集，有利于企业抓住市场痛点，推动经济发展。在大数据高速发展的背景下，对海量数据进行抓取分析仍然是一个重要的研究方向。

基金项目

2016 年江苏省教育科学“十三五”规划课题(C-a/2016/01/09)；江苏省高校自然科学基金项目(BK20171166)。

参考文献

[1] GUYOL8888. 企业利用大数据的重要性是什么? [Z/OL]. https://zhidao.baidu.com/question/1929428779751702587.html, 2017-11-06.

[2] 百度. 百度指数[Z/OL]. https://baike.baidu.com/item/%E7%99%BE%E5%BA%A6%E6%8C%87%E6%95%B0/106226?fr=aladdin, 2019-03-27.

[3] 王海燕, 桑晓斐, 赵可云. 基于大数据的大学生网络行为分析研究[J]. 中国教育信息化, 2017(7): 6-10.

[4] 刘滢. 基于大数据平台的智能交通系统架构及功能设计[J]. 综合运输, 2018, 40(9): 86-90.

[5] 赵佳钐, 李坤伦, 徐江, 李院春. 基于混合架构的高校多源异构数据集成系统[J]. 电子技术与软件工程, 2019(7): 160-162.

[6] 李斌, 周清雷, 斯雪明, 聂凯. 基于拟态计算的大数据高效能平台设计方法[J/OL]. 计算机应用研究, 2019(8):

1-9. http://kns.cnki.net/kcms/detail/51.1196.TP.20180424.1023.046.html, 2019-05-07.

[7] 许吉团. 论网络舆论视域下高校思想政治教育的创新[J]. 新西部, 2018(36): 129-131.

[8] 谢继华. 大数据视阈下高校网络思想政治教育创新研究[D]: [博士学位论文]. 成都: 电子科技大学, 2018.