Software Engineering and Applications 软件工程与应用, 2020, 9(6), 516-522
Published Online December 2020 in Hans. http://www.hanspub.org/journal/sea
https://doi.org/10.12677/sea.2020.96059

Hans汉斯

政策大数据聚合及分发平台关键技术研究与示范应用

涂著刚

贵阳高新数通信息有限公司，贵州 贵阳
Email: 58424862@qq.com

收稿日期：2020年11月21日；录用日期：2020年12月20日；发布日期：2020年12月28日

摘　要

本文结合网页搜索引擎、数据自动分级分类、公共服务平台三个关键点，开展基于垂直搜索引擎的数据采集、基于CNN的数据自动分级分类、基于微服务架构的高并发云服务的关键技术路线研究。针对传统长短期记忆神经网络和卷积神经网络分类识别率不高的问题，提出融合CNN和注意力机制的长短时记忆来提高政策数据分类识别精确度。首先通过word2vec获取政策数据文本形成词向量矩阵，然后输入到传统机器学习分类模型中，在此基础上使用模型融合技术融合单一模型中分类效果较好的模型，最后得到融合模型和单一模型的分类结果并进行对比。结果表明，本文提出的技术路线有效解决了信息及时性不强、信息完整性不足、信息推送精准度不够等关键问题，实现政策扶持落地转换能力和效果的有效提升。

关键词

数据聚合，垂直搜索引擎，信息推送，微服务架构

Key Technology Research and Demonstration Application of Policy Big Data Aggregation and Distribution Platform

Zhugang Tu

Guiyang Hi-Tech Data Communication Co., Ltd., Guiyang Guizhou
Email: 58424862@qq.com

Received: Nov. 21st, 2020; accepted: Dec. 20th, 2020; published: Dec. 28th, 2020

文章引用: 涂著刚. 政策大数据聚合及分发平台关键技术研究与示范应用[J]. 软件工程与应用, 2020, 9(6): 516-522.
DOI: 10.12677/sea.2020.96059

Abstract

Based on the three key points of web search engine, automatic data classification and classification, and public service platform, this paper studies the key technical routes of data collection based on vertical search engine, automatic classification and classification of data based on CNN, and high concurrency cloud service based on microservice architecture. Aiming at the problem that the recognition rate of traditional long-term and short-term memory neural network and convolution neural network is not high, this paper proposes a long-term and short-term memory fusion of CNN and attention mechanism to improve the classification and recognition accuracy of policy data. The first mock exam is to get the first mock exam vector matrix from word2vec data, and then input it into the traditional machine learning classification model. Based on this, we use the model fusion technology to fuse the model with better classification effect in the single model, and finally get the classification results of the fusion model and single model and compare them. The results show that the technical route proposed in this paper effectively solves the key problems such as weak timeliness of information, lack of information integrity, and insufficient accuracy of information push, so as to effectively improve the landing conversion ability and effect of policy support.

Keywords

Data Aggregation, Vertical Search Engine, Information Push, Microservice Architecture

1. 引言

自 2015 年，李克强总理在政府工作报告中提出“大众创业、万众创新”以来，全国各地大大小小的创新型公司如雨后春笋般成立；中共十八大后，各地在优化营商环境方面实招频出、创新不断，出台实施约 2 万亿元规模减税降费举措，打出债券、信贷、股权“三支箭”缓解融资难题，各级企业尤其是中小微企业是最直接的受益者。要强化实现政策“落地感”，首要必须确保广泛宣传、高效执行和对应落实到企业。而随着互联网+技术的推广、智能终端的普及，大数据时代信息的快速获取渠道逐步完善丰富，数据的收集、挖掘、分析也给政策宣贯和落地转化奠定了良好的技术基础。

为了给企业更好的指向性，需要识别出企业真实需要的政策信息数据，因此文本分类中的传统机器学习、自然语言处理和深度学习技术可实现文本数据的分类挖掘及信息检索，已经在新闻公告分类、社会舆情分析、垃圾信息过滤等方面广泛应用[1]。

本文重点研究如何从庞大的政策类数据文本中提取出价值信息，并实现自动化的检索、分类汇总和分发推送，因此利用文本分类技术对政策信息进行分类识别是近年学术界的研究热点[2]。要解决这些问题，本文对数据的获取、分析、挖掘、归类开展深入研究，通过融合卷积神经网络 CNN 和长短时记忆注意力机制模型的引入，重点解决营销新闻文本识别分类的问题，并提出融合政策全面覆盖、企业需求精准定位、项目资金申报服务于一体的综合移动服务云平台设计路线，研究并实现面向中小企业的数据分发综合云服务平台。

2. 政策宣贯服务平台存在不足析

伴随着互联网+技术的推广、智能终端的普及，APP 信息推送和短信事务提醒进入了大家的生活，也为大家的工作带来了便利[3]。迈入大数据时代后，信息的快速获取渠道逐步完善丰富，数据的收集、挖掘、分析也给行业针对性服务奠定了良好基础。虽然网上现有大量各种的数据资源，为中小企业在政策指引下的发展奠定了信息数据基础。但由于不同行业、不同区域、不同发展阶段、不同类型的企业所涉及到的政策信息多样，数据来源广泛、结构复杂、动态实时，在有效利用的过程中急需解决以下问题：

2.1. 政策推送存在一定覆盖盲区

公司机构想要了解最新的政策支持和项目信息，需要主动去获取，带来人员和资金上更多投入。其中，中小微企业在成立之初缺乏技术资金资产，及时了解政策导向并获取扶持是公司能生存下来关键因素之一；而新创业者绝大多数没有运营经验，都是以技术为基础，对政策情况基本不了解，不能及时获取项目申报信息，或者知道项目申报信息但不知道如何申报。

政府部门多数通过官方网站发布新政策及扶持信息，无法确保辖区或行业内的企业都能收到并学习；“政策是否落实、企业是否及时获取到亟需的政策信息”等问题不能得到及时反馈，在政策覆盖面上形成了一个较大的盲目区。上述因素导致企业和政府部门之间产生沟通障碍，导致扶持政策落地最后一公里问题突出。

2.2. 海量信息致针对性服务缺乏

虽然网上现有大量各种的数据资源，为中小企业在政策指引下的发展奠定了信息数据基础；但由于不同行业、不同区域、不同发展阶段、不同类型的企业所涉及到的政策信息多样，数据来源广泛、结构复杂、动态实时，导致各类支持中小微企业发展的政策，存在一定的零散现象，不够系统。即使是结构化数据，由于适用的范围和应用服务对象的不同而存在巨大差异，尤其是跨行业的数据之间无法互通，为数据的融合性分析和针对性服务带来巨大的困难。例如，税务部门出台的企业减税降费政策与金融部门解决企业融资难的政策契合点较少，零散的举措难以达到总体效果。

2.3. 政策落地专业配套服务缺乏

主管部门重点关注于发布各类惠企政策，不注重为企业详细解读政策，没有较好的帮助企业提升谋发展、谋改革、谋创新的积极性和主动性。同时，不注重分析政策的落实与效果，并未考虑到本地的实际情况，没有较好地追踪政策落实的效果、发挥的作用。

中小微型企业，对各类政策资源、政策服务流程、项目申报要求等了解还不系统全面，不同政府部门之间、政府和企业之间都还存在较大的时空信息不对称，政策之间没有形成聚合力量，政策落地服务质量和效率有待进一步提升，对政策研究、政策咨询和政策服务的需求剧增，企业需要人才服务、金融服务、项目咨询公司以及法律、财务专业事务所提供在线服务等，从而能让众多小微企业能将全部精力放在公司的运营以及产品的发展上。鉴于此，需要通过整合资源、联动服务、构筑合力，最大限度地释放产业经济政策效能。

3. 平台关键技术路线研究

要构建平台实现全国各部门的产业扶持政策信息采集和推动，帮助企业及时了解掌握涵盖国家、企业、申报、科技、创业、金融、税务的各级政策信息，需要开展三方面的关键技术研究：1) 实现对聚焦、实时和可管理的网页信息采集，实现非结构化内容到结构化数据的数据解析，达到实现精准、全面的全

文索引和联合检索[4]；2) 构建不同行业、地域的数据挖掘及文本分类模型，实现对文本词向量的轨迹分类，实现不同用户数据的精准分发推送；3) 实现多线程高并发的分布式处理优化，实现企业级容器并发管理，实现应用发布及更新管理。平台技术路线如图 1 所示。

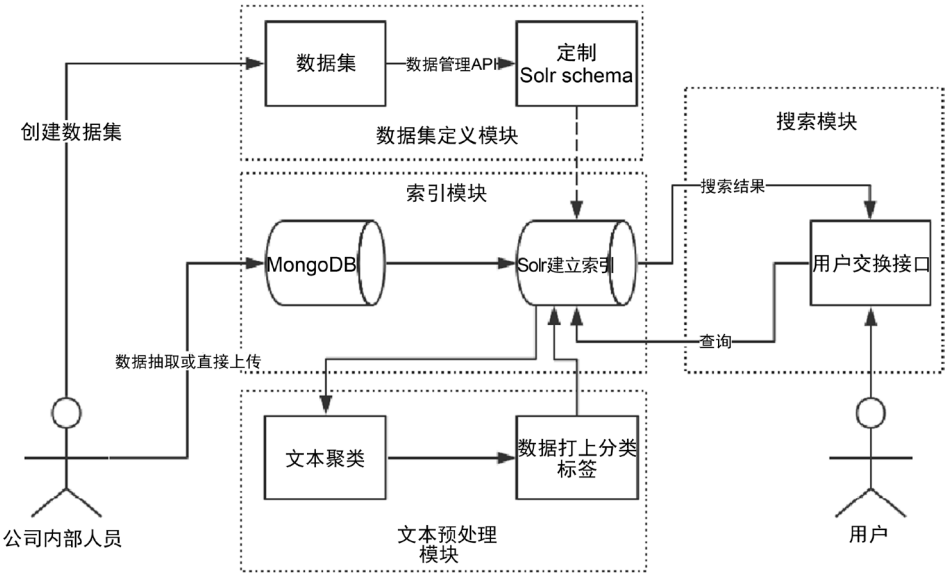

Figure 1. Platform technology roadmap
图 1. 平台技术路线设计图

3.1. 基于垂直搜索引擎的数据采集

通过垂直搜索引擎技术，构建非结构化数据实时采集及解析，高精准的全文广泛索引和联合检索，实现对聚焦、实时和可管理的网页信息采集，实现非结构化内容到结构化数据的数据解析，达到实现精准、全面的全文索引和联合检索。基于大数据的垂直搜索引擎，由信息采集、索引、查询和展示四个部分构成，包括 Solr (Search On Lucene Replication)、索引器、检索器、中文分词模块、用户接口模块，通过将 Solr 与索引器连接，索引器与检索器相互连接，检索器与中文分词模块连接，用户接口模块与中文分词模块连接，实现全文检索。其中 Solr 索引器负责对原始数据库的文档构造索引，并且存储在索引数据库中。检索器利用索引数据库中的索引来查找与用户查询相匹配的文档，计算各个文档和查询关键词的相关度，并将相关度大于阈值的文档按照相关度递减的顺序排列，返回给用户；中文分词模块使用全二分最大匹配快速分词算法；用户接口模块为可视化的查询输入和结果输出界面。

具体过程中，1) 通过采用基于垂直搜索引擎的网页采集技术，按需实现控制采集目标和范围、按需支持深度采集及按需支持复杂的动态网页采集，达到更加聚焦、纵深和可管控的需求，并且网页信息更新周期也更短，获取信息更及时；2) 通过采用全二分最大匹配快速分词算法，最大限度地提高分词效率；3) 通过垂直搜索支持全文检索和精确检索，并按需提供多种结果排序方式，支持结构化和非结构化数据联合检索，满足作者、内容、分类的组合检索。

3.2. 基于卷积神经网络的文本分类

文本分类的主要过程包括文本预处理，特征提取，文本表示和分类器训练四部分。2017 年夏从零等人提出了基于事件卷积特征的新闻文本分类，通过卷积神经网络从新闻文本中提取特征文本，以对文本进行分类，但通常会忽略上下文且语义准确度不高[5]。对于捕获单词的语义，文本中信息的顺序仍然不

能令人满意。同时，特征提取的方法存在数据稀疏、维度爆炸等问题，降低了训练模型的泛化能力[6]。

基于此，本文通过结合卷积神经网络和注意力机制的长短时记忆网络结构，实现基于卷积神经网络(Convolutional Neural Network, CNN)的数据挖掘及分类，通过构建基于邻接矩阵构建文本分类模型，实现对文本词向量的轨迹分类，并通过最优次模式分配距离方法进行验证测试，提升分类的精确度实现政策类文件分类的准确率提升[7]。

具体过程中，1) 借助卷积神经网络结构和群结构的相似性，使用卷积神经网络分类文本的词向量和个数，在获得文本词向量分类基础上，进一步分类群结构，获取群协作分类关系；2) 通过距离敏感性参数调整最优次模式对于真实点集和估计点集之间距离误差的权重；3) 通过关联敏感性参数调整最优次模式距离对于真实点集和估计点集之间数量不同的关联误差权重。

3.3. 基于微服务架构的高并发云服务

以微服务架构为基础构建云服务平台，实现多线程高并发的分布式处理优化，通过 Rancher 实现企业级容器并发管理，通过 Docker 实现应用发布及更新管理，通过 Redis 分布式锁建立及特定服务保障，通过多线程的用户数据分发实现精准推送。集群管理架构如图 2 所示。

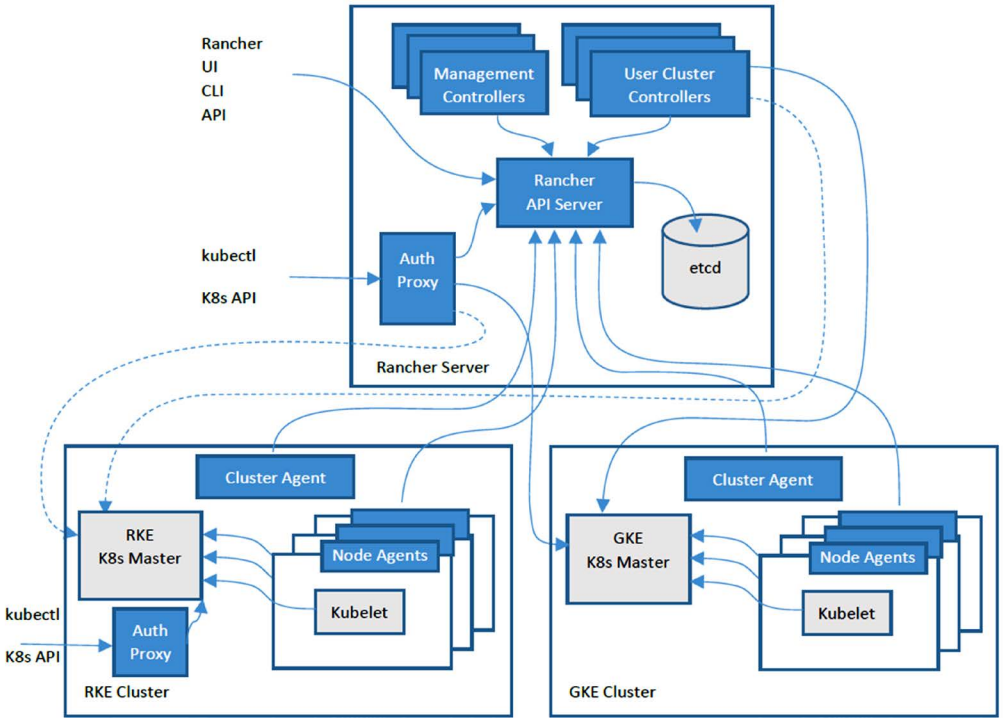

Figure 2. Rancher cluster management architecture
图 2. Rancher 集群管理架构图

具体过程中，首先通过 Rancher 为容器提供一揽子基础架构服务：包括 CNI 兼容的网络服务、存储服务、主机管理、负载均衡、防护墙；跨越公有云、私有云、虚拟机、物理机环境运行，实现一键式应用部署和管理；与各种 CI/CD 工具协同工作，实现开发、测试、预生产和生产环境的自动部署，提供整体可视化的主机、容器、网络及存储管理，简化运维人员故障排除和生产部署的工作量；并通过企业应用服务目录实现企业数据中心的应用管理部署。其次，通过 Docker 镜像提供容器运行时所需的程序、库、资源、配置等文件外，包含了一些为运行时准备的一些配置参数(如匿名卷、环境变量、用户等)。镜像不

包含任何动态数据，其内容在构建之后也不会被改变，在当某些相同的层已经存在的时候，完全不需要重新传输，大大提高镜像在网络上的传输效率。最后，通过研究 Redis 构建合适的分布式锁，在高并发下消除选择竞争、保持数据一致性，即客户端在执行连贯的命令时上锁的数据不会被别的客户端的更改而发生错误，保证特定服务在多个容器中只有一个运行，同时保证命令执行的成功率。对于需要进行政策信息分析推送的用户进行查询分发、批量处理，以遍历的方式进行用户推送，保证用户不会漏推或推送重复；信息推送采用先集中查询信息再对多用户进行集中分析多线程处理的方式，保证信息推送的顺序性和完整性。

4. 平台研发与应用

基于上述技术路线，研发出《政策快报》综合服务云平台。《政策快报》是一款集政策信息、政策解读于一体的国家扶持企业政策申报信息服务平台，是企业“及时、全面获取政府政策类资金、资源”的利器。平台收录全国各个部分的涉企政策项目信息，为企业挖掘有价值的政策信息、项目申报指南等信息内容，帮助中小企业用好政策福利，平台主要内容如图 3 所示。提供申报信息、国家政策、政策解读、金融服务，另外提供专业的政策申报咨询服务机构，高效完成信息收集、项目申报，为企业提供一站式服务。《政策快报》综合服务平台的主要子系统，包括数据采集系统、元数据管理系统、智能推动系统、权限管理系统、用户反馈系统等，主要功能模块如图 4 所示。

Figure 3. Cloud platform of policy express
图 3. 《政策快报》综合服务云平台

Figure 4. Main function module diagram of policy express
图 4. 《政策快报》主体功能模块图

《政策快报》综合服务平台主要特点包括：及时全面的获取政府每年数十万至数千万的政策类资金、资源，更高优先级展示项目申报信息，高效稳定；聚合发改、工信委、工商、国税、地税、科技、财政、人社、农委、环保、质检、商务、国土、住建、规划、交通、水利、安检等数万个各级政府部门网站精选内容，支持按区域、按发布机构、按时间筛选呈现信息。覆盖项目申报信息、国家政策、政策解读、金融服务、人才服务信息，只要用户关注的信息，都能搜到、关注，也可按条件订阅，避免价值信息遗漏，信息精准、及时，智能呈现全面政策信息和招商引资渠道，第一时间通达企业决策人；并提供咨询服务、金融服务，形成整体解决能力。

《政策快报》综合服务云平台存储的网页快照和相关附件有近两千万条左右，能高效的对数据库进行存储及查询进行各种优化，保证高并发下同一个数据的二十万次上锁执行释放锁的操作，平台的稳定运行稳定，反响热烈，截止 2020 年 7 月，累计安装超过 350 万次。APP 主要功能截图如图 5 所示。

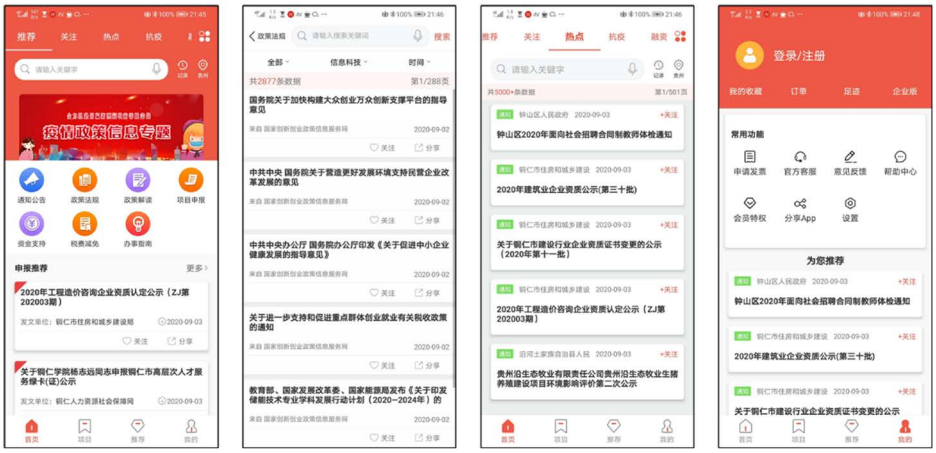

Figure 5. Example of comprehensive service cloud platform of policy express
图 5. 《政策快报》功能示例图

5. 总结

《政策快报》平台的出现，打通了企业项目申报及资金需求扶持的通道，打通了企业技术提升及科技成果转化的渠道，释放企业精力降低各项成本投入，有效减少政策覆盖存在盲区，提升政策转换能力和落实效果。平台面向中小企业用户以及政府部门服务，持续优化金融、人才、法律、财务、专利和商标等增值业务，解决了企业及时获取有效信息的刚性需求，并保持完善的后续服务。

参考文献

[1] 刘逸琛, 孙华志, 马春梅, 姜丽芬, 钟长鸿. 一种基于高层特征融合的网络商品分类[J/OL]. 北京邮电大学学报: 1-7[2020-12-03].

[2] 叶蔚, 常青, 杨芳. 基于虚拟化的移动应用架构研究和设计[J]. 计算机工程与设计, 2019, 40(2): 292-297.

[3] 刘鹏程, 孙林夫, 张常有, 王波. 基于交互注意力机制网络模型的故障文本分类[J]. 计算机集成制造系统: 1-27[2020-12-03].

[4] 何力, 谭霜, 项凤涛, 吴建宅, 谭林. 基于深度学习的文本分类技术研究进展[J]. 计算机工程: 1-15[2020-12-03].

[5] 吴汉瑜, 严江, 黄少滨, 李熔盛, 姜梦奇. 用于文本分类的 CNN_BiLSTM_Attention 混合模型[J]. 计算机科学, 2020, 47(S2): 23-27+34.

[6] 张宇昂, 贾云鹏, 刘家鹏. 一种多特征融合的长文本分类方法[J]. 中国电子科学研究院学报, 2020, 15(9): 910-916.

[7] 刘高军, 王小宾. 基于 CNN + LSTMAttention 的营销新闻文本分类[J]. 计算机技术与发展, 2020, 30(11): 59-63.