 DOI: https://doi.org/10.12349/iser.v6i6.6024

Design and practice of unified service platform for full-text retrieval

Jianqin Yang

CNOOC Research Institute Co., Ltd., Beijing, 100028, China

Abstract

By expanding the functions of the full-text search service management module, the platform achieves permission management, monitoring management, and data management; the remote service component is fully functional, comprehensively supporting the extended application of full-text search services; the full-text search service interface has been optimized to provide standard internal and external interface services for the service platform. The index data loader adapter function has been extended as a supporting tool for the full-text search service platform, enabling convenient and rapid storage of application system data into the full-text search service platform. The file management system has been comprehensively upgraded, achieving a new optimization and upgrade of the file management system along with the expansion and application of full-text search functions.

Keywords

full text retrieval; archival service; service platform

档案全文检索统一服务平台设计与实践

杨建钦

中海油研究总院有限责任公司，中国 · 北京 100028

摘　要

通过全文检索服务管理模块的功能扩展，实现平台的权限管理、监控管理和数据管理等功能；远程服务组件功能完善，全面支持全文检索服务的拓展应用；全文检索服务接口优化，为服务平台提供对内、对外的标准接口服务。索引数据加载适配器功能扩展，作为全文检索服务平台的配套工具，实现应用系统数据便捷、快速存储至全文检索服务平台。文件管理系统的全面升级，实现文件管理系统的全新优化升级与数据全文检索功能扩展及应用。

关键词

全文检索；档案服务；服务平台

1 引言

随着智能油田、智慧油田的建设，越来越多档案研究更加重视海量数据的检索以及高效利用，并且都进行过相关的技术探索及建设。但受限于档案课题研究范围的束缚以及数据的信息壁垒，数据的检索引擎存在重复建设的问题，缺乏统一的管理机制。同时，不同的检索引擎，学习成本高，会导致项目建设周期延长，建设成本增加等问题。通过对 ElasticSearch 的二次封装开发，打造独立的产品服务，以平台的模式为其他应用系统提供全文检索服务 [1]。

当前档案多数的应用系统都进行过全文检索研究及建设，但由于缺少统一建设、统一管理的全文检索服务，各系统在全文检索方面的建设中都存在重复工作量。应用系统建设过程中，以往只能通过简单的关键词查询来实现检索的目的，系统无法基于检索关键词热度的增长以及用户的使用习惯而进行自主优化和统计分析。

针对当前存在的问题，本文提出了相关的技术解决思路，在充分利用现有的建设资源的基础上，打破信息孤岛，遵循“统一规划、统一标准、统一建设、统一管理”的原则，建立档案统一的全文检索服务平台，通过平台功能扩展，实现档案全文检索服务的统一提供、统一管理和统一维护，系统的全新优化升级与功能扩展。

2 研究现状

全文检索服务平台存在问题如下：

①系统重复建设问题：当前多数的应用系统都进行过全文检索研究及建设，但由于缺少统一建设、统一管理的全文检索服务，各系统在全文检索方面的建设中都存在重复工作量。

②全文检索建设成本高：应用系统进行全文检索功能建设时，需要进行相关专业知识的储备，集群部署及授权较

【作者简介】杨建钦（1982–），中国山东郓城人，硕士，高级工程师，从事GIS、信息系统集成研究。

为繁琐，学习开发语言需要的时间长、难度大，项目建设周期被无形拉长。

③缺乏统一的管理机制：应用系统进行全文检索功能建设时，仅局限在本应用系统，总院暂无统一有效的监管手段，缺乏对全文检索服务引擎及整体数据应用情况的监控，无法统计分析并加以利用。

④缺乏全文检索机制：应用系统建设过程中，以往重视的主要是结构化数据的检索，缺乏对非结构化数据的处理及检索。即使支持全文检索，也无法根据实际的需求自定义检索效果。

⑤缺乏知识的深度管理：应用系统建设过程中，以往只能通过简单的关键词查询来实现检索的目的，系统无法基于检索关键词热度的增长以及用户的使用习惯而进行自主优化和统计分析。

3 设计方案

全文检索服务平台依托 ElasticSearch 全文检索服务引擎，结合 Docker 技术、RestFul 标准接口等，使用 SpringBoot+VUE 框架，打造一体化平台，以平台模式为其他应用系统提供全文检索服务和监控管理等功能 [2][3]。

图 1 全文检索服务总体架构

全文检索统一服务平台包含四大主要模块：

①服务管理平台：全文检索服务的管理平台，针对不同用户角色，可通过该平台进行账户角色管理、申请检索服务使用、处理申请、ElasticSearch 服务的资源分配、ElasticSearch 服务器相关指标的监控、索引管理和监控、数据管理和监控等。

② Docker 镜像：Docker 是一种应用容器，在 Docker 内预装 ElasticSearch、Java、检索服务接口、依赖组件包等，为服务平台提取全文检索服务。Docker 镜像封装目的是傻瓜式、便捷式部署在宿主机中。

③检索服务接口：预装在 Docker 内部，为整个应用服务平台提供 RElasticSearchtFul 模式的 API 接口。接口包括：索引管理、索引重建、数据插入、数据更新、数据删除、数据检索等。

④远程服务组件：远程服务组件部署在宿主服务器，作为服务平台的扩展，协助平台将多类别配置文件（ElasticSearch 配置文件、ElasticSearchSP 配置文件、IK 配置文件等），一键部署式部署到宿主服务器。

全文检索服务的数据来源主要包含以下三类：适配器：实现扫描文档、图片等文本提取（非结构化数据）和关系型数据库数据提取（结构化数据）后写入服务。服务 API：通过整理原始数据为服务写入数据 API 对应的 JSON 格式，调用接口写入服务。数据迁移：将已有 ElasticSearch 服务的数据拷贝至全文检索服务挂载的数据目录下，重建索引即可将数据导入服务。以往使用 ElasticSearch 服务时，用户需要耗费大量时间学习 ElasticSearch 相关知识，例如服务的集群部署、初始化授权、分词器安装、开发语言的学习等，操作繁琐，学习时间长、难度大，项目建设周期被无形拉长。而全文检索服务平台使用一键化部署，无需学习 ElasticSearch 集群部署及授权配置，系统管理员只需分配宿主机并点击确认按钮，平台会自行生成 ES 集群、ESSP 服务、IK 分词等配置信息、调用远程组件在宿主机创建配置文件、自动化部署

全文检索服务集群、初始化授权信息、集成分词工具、创建分词库，将原本复杂的部署过程缩短至数秒内，实现傻瓜式的快捷部署。

索引适配器主要实现提取，转储和数据应用功能。提取：实现扫描文档、图片等的文本提取（非结构化数据）和关系型数据库（结构化数据）的提取过程。转储：索引数据适配器会根据表结构或预设信息创建模型，调用数据写入 API 实现结构化和非结构化数据的存储。应用：检索接口会调用底层 ElasticSearch 服务，应用通过调用检索服务接口即可实现全文数据的查询。

全文检索服务部署架构采用前后端分离模式开发。中间件采用 Nginx 和 Docker。平台部分：前端采用 Nginx 分布式集群云部署；后台依赖 MySQL 数据库。将两部分部署至平台服务器。服务部分：ElasticSearchSP 全文检索服务依赖 ElasticSearch 环境，将两部分打包至 Docker 镜像内，与远程服务组件一同部署至各宿主机。调用方式：在创建全文检索服务时，平台通过远程服务组件生成配置文件并创建 Docker 容器，从而启动全文检索服务。

全文检索服务平台经过多轮选型验证，最终采用 ElasticSearch 全文检索服务引擎。ElasticSearch 一个开源的基于 Lucene 的分布式、高扩展、高实时的搜索与数据分析引擎。它可以很方便地使大量数据具有搜索、分析和探索的能力。充分利用 Elasticsearch 的水平伸缩性，能使数据在生产环境变得更有价值 [4][5]。

4 应用效果

经过对全文检索服务平台、文件管理系统多轮的功能测试，系统功能完善，参与测试的模块达到了测试方案中所规定的要求，测试功能点覆盖系统的 99%，无遗留问题，系统满足需求规格书中的要求，可以上线运行。全文检索服务平台目前处于试运行阶段，全文检索服务平台功能扩展及文件管理系统升级建设课题组常驻现场进行开发工作，课题组的开发人员与技术支持人员对全文检索服务平台的试运行进行了持续的支持和维护。鉴于驻场开发的便捷性，用户可随时与全文检索服务平台的运维人员进行问题的反馈及优化整改意见的对接，同时课题组也积极收集试运行阶段的问题以及功能优化修正意见，并经过项目组的讨论确定，由系统支持人员完成全文检索服务平台的升级及优化工作。

全文检索服务平台的主要运维工作包括：应用系统的安装及部署、应用系统的升级、应用系统的数据文件的定期备份、数据库的日常运维等。效果与意义如下：

①未来其他应用系统使用该服务平台，不需要进行二次开发，只需要通过指定的 API 接口服务，即可以实现数据资料信息的入库及资料检索。因此，对其他应用系统来说，全文检索服务的使用及引入，不需要额外的技术学习及储备的工作量，只需要熟悉平台操作流程及有限的 RElasticSearchtful 标准的 API 接口的即可。

②平台为管理人员提供了便捷的管理界面，可通过平台管理多套 ElasticSearch 服务。用户通过界面的交互配置，即可实现 ElasticSearch 的本地 / 远程配置（如集群信息、节点信息、数据存放路径、日志存放路径等）、远程控制（启动 / 停止检索服务、数据的管理等）等，无需学习复杂的 ElasticSearch 集群部署和授权，只需一键完成分配，平台将结合远程组件，自动完成配置文件创建、容器创建、多节点集群搭建、授权初始化的过程；同时服务平台也实现了注册的应用系统的便捷管理，可实时启用 / 禁用其他应用系统对服务的使用。

5 结语

全文检索服务平台作为一个通用产品平台，满足其他应用系统的全文检索服务需求，未来拥有良好的应用前景。目前已提供工程智能设计平台进行试用，并获得了客户的高度认可。

ElasticSearch 服务在使用过程中，会产生相应的日志文件，同时会占用一定的磁盘容量，建议及时通过监控查看日志占用情况，如有必要需及时对日志进行清理。

参考文献

[1] 张建中,黄建飞,熊拥军.基于ElasticSearch的数字图书馆检索系统[J].计算机与现代化,2015(6):69-73.

[2] 张云,许江淳,李玉惠,等.基于Nginx服务器负载均衡技术的研究与改进[J].软件,2017,38(8):6-12.

[3] 马豫星.Redis数据库特性分析[J].物联网技术,2015(3):105-106.

[4] 饶琛琳.ELKStack权威指南[M].2版.北京:机械工业出版社,2014:16-30.

[5] CROFTWB,METZLERD,STROHMANT.搜索引擎———信息检索实践[M].刘挺,秦兵,张宇,译.北京:机械工业出版社,2010:76-93.

[6] MCCANDLESSM,HATCHERE,GOSPODNETICO.Lucene实战[M].牛长流,肖宇,译.2版.北京:人民邮电出版社,2016:81-83.