

基于 OpenGauss 的图片关键字智慧检索系统

蒋新辉 覃镇锋 彭石燕

(柳州铁道职业技术学院, 广西 柳州 545616)

摘要: 近些年, 随着信息技术高速发展, 信息爆炸时代的到来, 基于 OpenGauss 的图片关键字智慧检索系统运用而生。该系统采用深度学习技术, 不仅能对图片进行自动分类, 还可以提取图片特征, 并利用 OpenGauss 数据库管理系统实现高效的数据存储和检索。通过实验验证, 该系统能够快速准确地检索相关图片, 具有较高的准确率和实用性, 而且该系统的实现可以为图片检索提供全新的解决方案, 可广泛应用于各个领域。对此, 本文基于 OpenGauss 对图片关键字智慧检索系统展开研究, 以供参考。

关键词: OpenGauss; 图片; 智慧检索系统

随着移动手机设备以及物联网设备的爆炸式增长, 越来越多图片数据需要被存储到数据库中。海量非结构化的数字图像中存储了诸如密码, 数据表格, 身份证件等重要信息, 这些信息通常被存储在云数据库中。这就使得每一个云数据库用户都需要从中快速准确地搜索所需图片信息, 从而对围绕数据库的图片检索系统提出了更高的挑战和要求。因此, 针对如何快速有效地从云数据库系统中提取图像关键字信息, 本团队提出一种基于 OpenGauss 的图片关键字检索系统。而且结果证明, 围绕 OpenGauss 数据库的关键字图像检索系统有效提高了图片检索准确性和效率。

伴随着数字摄影技术和计算机视觉技术的不断发展, 图片检索技术逐渐成为了一个热门的研究方向。而且图片检索系统的应用领域非常普遍, 例如, 旅游, 餐饮等。在表 1 中按照不同应用领域列举了使用图片检索技术的一些代表性公司。图片检索系统主要包括以下三个部分: 首先是采集和保存所需的图片资源, 通过抽取图片特征或者进行标注将资源进行存储的存储系统, 然后根据用户提出的图片检索需求, 分析需求并将图片转化为用于检索数据库系统的访问请求, 最后根据相应的检索模型和算法, 从数据库系统中检索出与目标相似度较高的图片并返回给用户。

表 1 应用图片检索的相关领域

领域	代表应用
电商	Amazon, eBay, 淘宝, 京东
旅游	马蜂窝, 携程
搜索	Google, 百度, 搜狗
学习	作业帮, 学霸君, 搜题

近些年来, 作为图片检索系统的重要组成部分的图像内容检索也越来越受到广泛关注。随着数字时代的发展和无纸化办公, 大量的纸质书籍和各种公文文件开始被扫描成文档图片。再加之, 现今的智能移动设备上普遍配置摄像头, 大家早已养成了随身拍、随手拍的拍照习惯。用户除了单纯的上传外, 也有着对图片进行归档、以供日后查找的需求。

为了满足用户等对图片检索的需求, 本系统基于华为 OpenGauss 数据库存储技术, 提出一种图片关键字智能检索系统。系统的后端采用了 PP-OCR 智能算法取代人工图片标注提取图片文本信息(包括图片的文字信息和位置信息), 并将检索的图片信息和路径信息写入 OpenGauss 数据库, 其次, 通过优化 Elasticsearch 工具的检索索引, 提高 OpenGauss 图片检索效率。最后, 在 Centos 以及 OpenEuler 平台上实现了该索引和接口, 充分利用了 PC 端的计算能力和移动端的便捷性, 构成了基于微信小程序

的图片关键字检索系统。

一、图片检索系统应用场景

在当今大数据时代, 文档内容的智能检索识别技术有巨大的应用需求。这些应用场景主要来自生活中的各种设备拍照所形成的图片; 来自生活发票、车票、电子文档、车牌、标识、标牌等图片文本; 网络场景图像和合成文档图像: 互联网(包括移动互联网, 社交媒体如微信、微博等)上有大量用户上传的图像(其中相当一部分有文字)和合成文档图像(如合成文本图像、自然图像中嵌入文字等)都是用户可能需要存储和检索、并管理的图片。然而当用户要找某一张图片时, 以人工方式逐一搜寻文件夹, 耗时耗费精力。本系统则很好地考虑了移动互联网的图像管理和检索要求, 不仅可以用于云端海量图片的存储检索, 也能适用于管理个人设备上的图片内容。

二、OpenGauss 图片关键字检索整体框架技术说明

本系统整体架构主要分为 4 部分: 微信小程序界面模块, 主要功能是图片上传、字符串检索、呈现检索结果, 实现服务器和移动设备的交互; flask 服务器模块, 该模块主要是实现 PP-OCR 模块的图片信息提取、接受移动端发出的检索指令、发送检索结果等功能。Elasticsearch 模块, 采用 Elasticsearch 工具实现文本去冗余、排序功能、优化搜索方式等功能; OpenGauss 数据库模块, 是整个系统核心模块, 在其中除了在核心系统中承担基础图片文本数据管理之外, 在数据库的智能管控扮演着重要角色, 本系统基于 OpenGauss 数据库, 采用深度学习算法实现智能图片提取, 智能检索能力, 对于提升 PC 端、手机移动端系统的数据库图片智能检索的平稳运行和高性能数据处理起着关键作用。

(一) paddleOCR 技术介绍

识别系统选用实用的超轻量化 OCR 系统, 即 PP-OCR。PP-OCR 的整体模型尺寸仅为 3.5M 和 2.8M 分别用于识别 6622 个汉字和 63 个字母数字符号。该模型引入了一系列策略来增强模型能力或减小模型大小。并同时采用预训练的中文和英文识别模型, 包括文本检测器(97K 图像)、方向分类器(600K 图像)和文本识别器(17.9M 图像)。此外, 该 PP-OCR 还在法语、韩语、日语和德语等其他语言识别任务中得到验证, 可以满足任何图文字提取的需要。

本 OpenGauss 智能检索文本识别系统采用超轻量化 PP-OCR 系统, 该系统由文本检测、检测框校正和文本识别三部分组成。图像文本检测的目的是对图像中的文本区域进行定位。在 PP-OCR 中, 使用可微二值化(DB)作为文本检测器, 它基于简单的分割网络。简单的 DB 后处理使得它非常高效, 该模型采用了以

下 6 种策略：轻主干、轻头、去除 SE 模块、余弦学习率衰减、学习率热身和 FPGM 剪枝进一步提高其有效性和效率。最后，检测框检测到的文本在识别之前，需要将文本框转换为水平矩形框，以便进行后续的文本识别，检测帧由四个点组成，易于通过几何变换实现。另外，在 PP-OCR 中，使用 CRNN 作为文本识别器，它在文本识别中有着广泛的应用和实用性。CRNN 集特征提取和序列建模于一体。它采用了连接时态分类（Connectionist Temporal Classification，CTC）损失来避免预测与标注之间的不一致。为了增强文本识别器的模型能力和减小模型大小，可以轻主干、数据增强、余弦学习率衰减、特征映射解析、正则化参数、学习率预热、光头、预训练模型和 PACT 量化等策略。由于视角、缩放、弯曲等多种因素的影响，图片文本通常会发生巨大的变化。本模型采用结构化对图像文本结果进行结构化来解决高密度和长文本带来的不同问题需要，以获取更高精度图片文本识别模型。

（二）Elasticsearch 检索技术介绍

针对 OpenGauss 数据库检索无法满足海量文件资源数据高速分布式检索的问题，采用了基于 Elasticsearch 的关键字分布式搜索引擎的构建方法，并以 Elasticsearch 技术为核心，结合 LogStash、Redis 等多门技术，实现了 openGauss 图片关键字检索系统。系统使用 LogStash 将数据库中的标准化数据同步到 Elasticsearch，再通过 JDBC 读取 openGauss 后端数据，微信小程序展示搜索结果，也可通过访问 Redis 缓存的文本信息和文本位置信息，以便更快地访问信息。与传统数据库检索相比，该系统的检索响应时间明显缩短，检索结果全面，能够满足各类平台图片检索的需要。

数据检索框架如下：结合数据规模、性能和功能需求，基于 Elasticsearch 的关键字检索系统分为数据导入、数据索引、数据检索、缓存四个模块。首先用 Logstash 将 PP-OCR 识别出来的数据导入到 Redis 进行缓存库中，利用 Elasticsearch 对缓存数据进行处理，排序去除冗余，加快存储效率、减少文本存储空间，再将文本信息和位置、图片路径存储到 OpenGauss 中，当用户发送搜索请求时通过 Elasticsearch 官方提供的 JavaAPI 调用 Elasticsearch 实现数据的检索，最后用微信小程序对搜索结果进行前端展示。

（三）OpenGauss 数据库模块技术生态介绍

作为项目核心组成部分，OpenGauss 数据库相关技术的运用是必不可少的环节。基于 OpenGauss 数据库内核，OpenGauss 官方开放了数据同步接口，并基于数据接口开发了 JDBC 驱动使得开发者得以方便快速地接入 OpenGauss 数据库；同时，基于其原生的运行视图、监控运维接口及工具集、备份恢复接口及工具集，OpenGauss 还为开发者提供了方便快捷的数据库客户端 DataStudio，使得开发者得以及时准确地对存储的数据进行查看、添加、修改和删除。

在本项目中，我们安装配置了 CentOS 系统，在 CentOS 系统上安装系统软件依赖，执行 OpenGauss 安装，配置好用户后，通过 DataStudio 客户端进行数据库连接，建立了所需数据库及数据表。然后基于 JDBC 驱动，通过在 Elasticsearch 中进行相关配置，完成了 OpenGauss 和 Elasticsearch 的数据同步。在此基础上，开发 flask 服务器，通过 http 访问 Elasticsearch 的 api 接口，实现了服务器和 OpenGauss 的数据交互。基于 javascript 语言，开发微信小程序，实现与 flask 服务器的通信，从而完成了小程序前端和数据库的交互访问。

三、检索系统在微信小程序上实验和分析

为了测试检索系统的效率，在实验硬件环境为 Intel（R）Core（TM）17-67003.40GHz 处理器、8GB 内存、操作系统为 win10 专业版 64 位操作系统下进行测试。采用了车票、电子邮件、图片文本、PPT 图片等 64 张 JPG、JPEG、PNG 等格式图片制作成测试数据集。

在系统健壮性测试中，我们从系统的容错性、稳定性、兼容性、检索速度四个方面来测试，容错性指相关模块的独立运行能力，稳定性指是否存在内存泄漏问题，兼容性指对输入图片格式的兼容性。在检索速度测试中，只要是测试从 64 张图片中检索每张图片的速度。

根据表 2 测试结果，可以看到，图片检索系统对于常见的图片格式都够兼容，而且系统的容错性和稳定性、精度均可满足用户需求。

表 2 图片系统测试结果

测试类型	容错性	稳定性	兼容性	检索速度（每张图片 /s）
结果	通过	通过	通过	0.56

为了更直观的了解本系统功能，该系统通过微信小程序交互可以实现上传图片到任意文件夹，并在上传过程中通过 PP-OCR 实现图片文本信息的提取，并将提取的结果和图片路径经过 Elasticsearch 技术排序和去除冗余后上传至 OpenGauss 数据库中。

用户将关键字输入至微信端，随后传入服务器进行解析，通过 Elasticsearch 工具搜索方式搜索 OpenGauss 数据库中的文本信息，根据搜索关键字将含关键字的图片路径、文本信息位置传输至微信小程序结果展示端。

四、检索系统总结与展望

当前，文档图片检索不仅具有重要的学术研究意义，而且还拥有着良好的发展前景和商业价值。利用 PP-OCR、Elasticsearch、Logstash 等技术搭建 OpenGauss 图片智慧检索系统，推动 OpenGauss 生态技术发展，改善传统图片检索方式，并在手机端和 PC 进行了实验，测试了系统的容错性、稳定性、兼容性、检索速度，实验结果证明，该系统可以非常高效的检索所需图片，并具有一定的容错性、稳定性、兼容性。

但该系仍存在一些可改进和推广的工作将在后续进行，具体如下：

1. 基于 OpenGauss 开发更全面图片存储和检索功能，实现并推进其高并发、高性能、易运维的特点，并尝试将其嵌入到浏览器和 app 中，优化其图片检索方式，加快 OpenGauss 图片数据检索技术落地。

2. 利用 OpenGauss 数据库并行查询能力，结合本系统实现的智能模型，实现在数据库内检索图像内容关键字信息。

3. 在系统模型上，可以基于特定场景（例：板书、学生试卷、车票等），收集数据集训练，提高特定场景下 OCR 的精度，并融入 NLP 技术，实现系统的关键字和关键短语的提取，进一步优化 OpenGauss 的储存方式和检索方式，以更好地代替传统存储图片打标签的方式。

参考文献：

[1] 赵朋悦，王翔，刘守训. 基于 VGG 的高准确率翻拍广告图片检索系统[J]. 无线通信技术，2022，31（02）：32-36.

项目信息：中青年课题基金项目（2022KY1414）