



<搜象 OCR
材料结构化内容管理平台>
产品介绍

搜象（厦门）信息技术有限公司

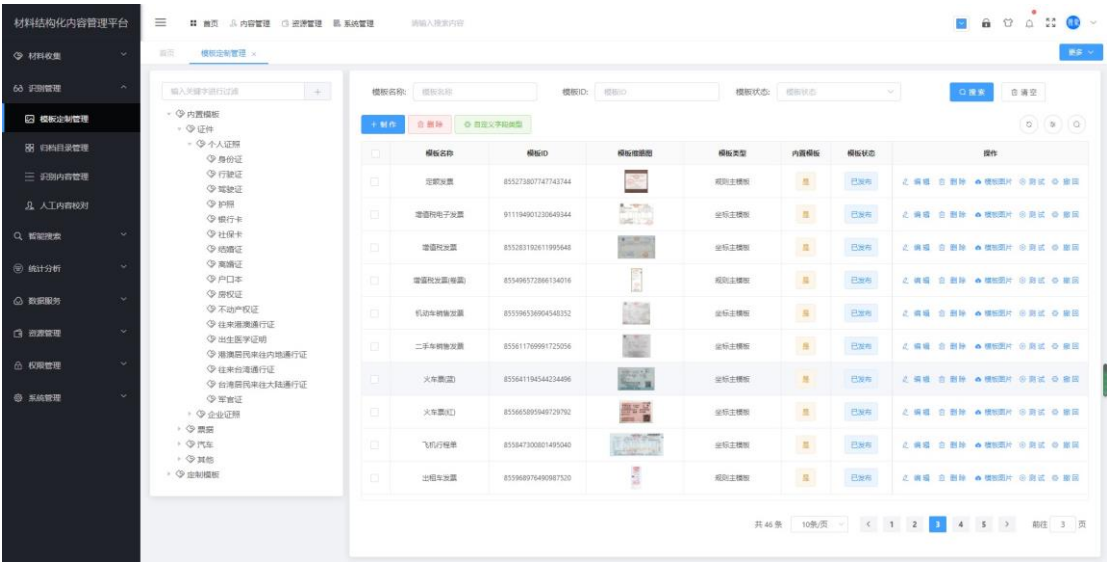
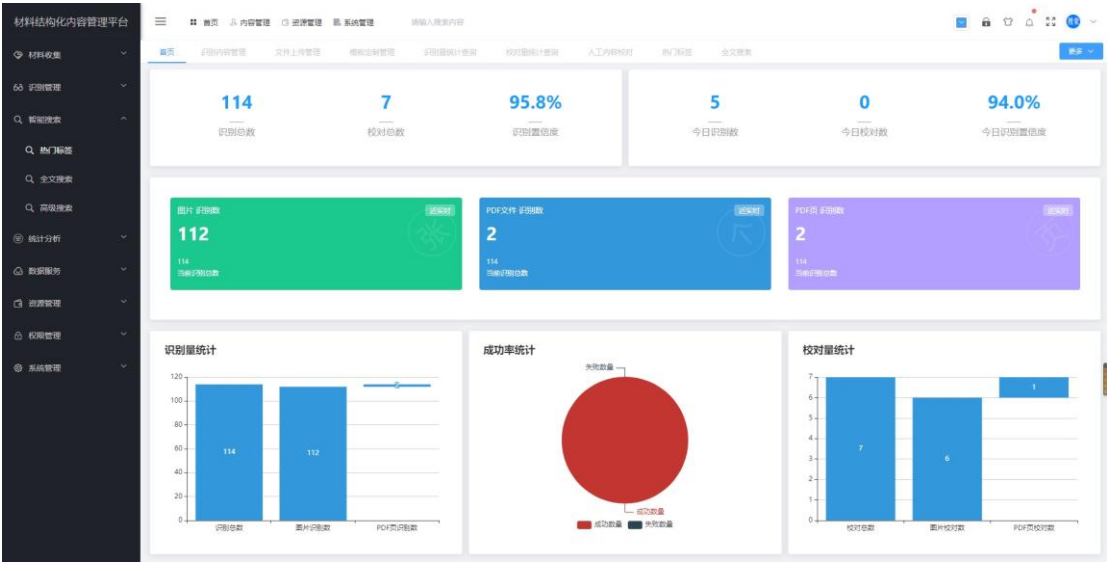
2021 年 9 月 25 日

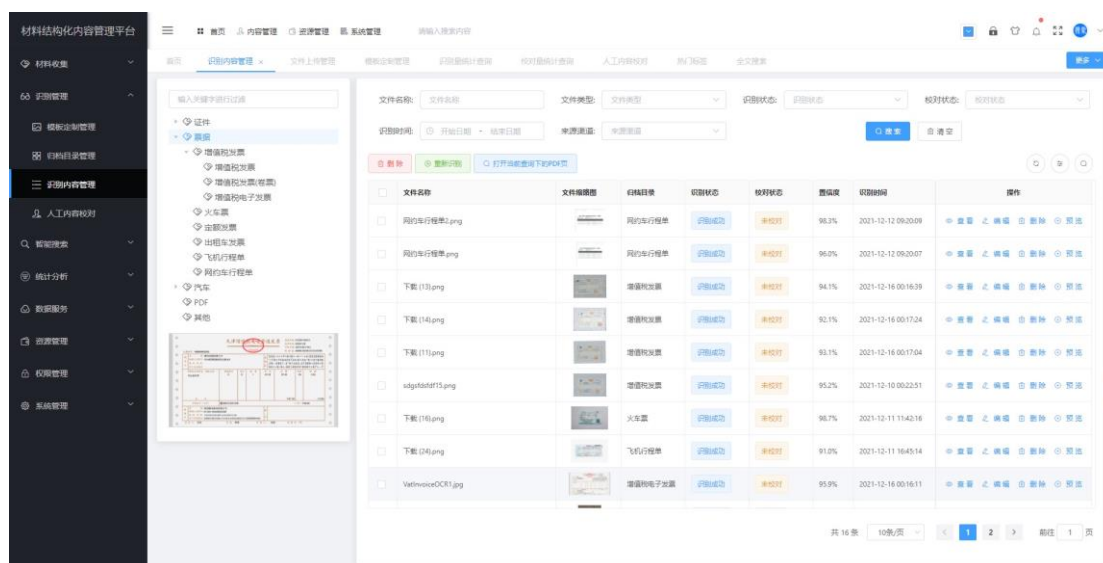
目 录

1.1	产品概述.....	3
1.2	产品特点.....	4
1.2.1	优秀的用户体验.....	4
1.2.2	自定义 OCR 模板，结构化文字识别.....	4
1.2.3	支持图片、PDF 和压缩文件上传，支持多应用渠道接入	4
1.2.4	图像非失真压缩、旋正、自动归档.....	5
1.2.5	分布式对象存储.....	5
1.2.6	内容标签聚类、热门置顶.....	5
1.2.7	全文搜索、组合式条件搜索.....	5
1.2.8	自定义数据统一查询服务.....	5
1.2.9	异步总线方式，提高系统并发能力和稳定性.....	5
1.2.10	工业级的安全认证及权限管控.....	5
1.2.11	大数据支撑.....	6
1.2.12	容器化运维.....	6
1.2.13	支持多租户模式.....	6
1.3	功能模块	6
1.3.1	材料收集.....	6
1.3.2	识别管理.....	6
1.3.3	智能搜索.....	6
1.3.4	统计分析.....	6
1.3.5	数据服务.....	7
1.3.6	资源管理.....	7
1.3.7	权限管理.....	7
1.3.8	系统管理.....	7

1.1 产品概述

本系统主要针对企业中一些需要进行结构化内容提取的材料资源，例如申请材料、审批材料、人事档案以及各种证件发票等非结构化纸质或文件资源，系统进行资源的收集，结构化 OCR 模板定义，结构化文字识别以及内容的管理，形成并沉淀资源的结构化、半结构化以及非结构化数据，对外提供数据服务。





材料结构化内容管理平台-识别内容管理

1.2 产品特点

1.2.1 优秀的用户体验

有赖于我们前期对架构方案的技术选型以及在设计开发中对细节的专注。

1.2.2 自定义 OCR 模板，结构化文字识别

依托专业的模板定制组件，可对各行业业务所需的图像资料进行按需模板定制、字段结构化识别。

1.2.3 支持图片、PDF 和压缩文件上传，支持多应用渠道接入

图片支持但不限于 jpg、jpeg、png、bmp、gif、tif、tiff、jfif、heic、jpe 及 dib 等格式；压缩文件支持兼容 Window 和 Linux 的 ZIP 格式，自动提取压缩文件中的图片和 PDF 进行识别。

支持平台本地上传，支持多应用渠道接入上传。

1.2.4 图像非失真压缩、旋正、自动归档

几 M 大小图片可非失真压缩至一百多到两百多 K，适合大规模存储，可开启自动旋正系统参数，对于非正向图片进行旋正存储，基于模板匹配技术，可对图像进行自动归档。

1.2.5 分布式对象存储

非结构化的图片和 PDF 数据进行底层的对象存储，高性能写入和读出，大规模数据场景下组件可设置多节点分布式存储。

1.2.6 内容标签聚类、热门置顶

默认文字识别结果自动标签生成，可按需自定义，内容标签自动聚合，热门标签置顶排序，点击标签可检索出相关图片和 PDF 页。

1.2.7 全文搜索、组合式条件搜索

集成 Elasticsearch 的强大检索能力，支持中文分词搜索、精确搜索、模糊搜索以及字符串匹配搜索等多条件组合下的复杂搜索。

1.2.8 自定义数据统一查询服务

通过简单编写类 SQL 脚本，可自定义平台的数据 API，提供给外部应用渠道调用。

1.2.9 异步总线方式，提高系统并发能力和稳定性

基于 kafka 消息队列的总线模式，可以有效降低处理数据时对数据库的依赖，有效缓存削峰，提高并发处理能力，让系统稳健运行。

1.2.10 工业级的安全认证及权限管控

采用 OAuth2 协议标准对用户的访问进行安全认证，独立的权限管理服务将用户的权限粒度控制在菜单/按钮权限、数据权限以及接口权限。

1.2.11 大数据支撑

在大数据存储及计算方面，平台相关组件支持分布式、支撑横向水平扩展，例如分布式文件存储、NewSql 数据库、分布式搜索引擎以及分布式消息总线等。

1.2.12 容器化运维

平台采用容器化技术部署运维。

1.2.13 支持多租户模式

平台支持多租户模式，根据客户需求进行定制部署。

1.3 功能模块

1.3.1 材料收集

主要对文件上传的管理，可上传图片、PDF、压缩文件，图片支持但不限于 jpg、jpeg、png、bmp、gif、tif、tiff、jfif、heic、jpe 及 dib 等格式；压缩文件支持兼容 Window 和 Linux 的 ZIP 格式，自动提取压缩文件中的图片和 PDF 进行识别。

1.3.2 识别管理

可定制 OCR 识别模板，创建文件归档目录，对文件 OCR 识别结果进行人工校对编辑等管理。

1.3.3 智能搜索

支持热门标签置顶排序、标签检索，全文检索以及组合条件检索，并对检索结果进行内容查阅。

1.3.4 统计分析

对识别量和校对量进行统计，支持识别量统计查询，校对量统计查询。

1.3.5 数据服务

通过简单编写类 SQL 脚本，可自定义平台的数据 API，提供给外部应用渠道调用。

1.3.6 资源管理

对任务作业进行调度管理，包括首页数据缓存、标签统计缓存、识别量统计、校对量统计以及 ES 批量索引等作业。

1.3.7 权限管理

系统角色权限管理，包括菜单权限、数据权限、接口权限。

1.3.8 系统管理

包括用户管理、机构管理、岗位管理、系统字典管理、业务字典管理、顶部菜单管理、参数管理以及应用管理。