易爬（iAvatar）

智能大数据爬取平台

产品白皮书

中网数据（北京）股份有限公司

www.isinonet.com

目 录

1 引言

随着现代信息技术的迅猛发展，一个大规模产生、分享和应用数据的时代正在徐徐开启。近几十年，我们通过互联网产生的累计信息量是人类过去几千年的总和。Web 作为一个巨大的资源宝库，里面有大量有价值的信息，面对类型繁多、数量巨大以及碎片化的网络信息数据，如何有效、准确、全面采集已成为企业大数据战略的首要任务。

图 1 2015-2020 年中国大数据产业规模及预测图

1.1 数据获取的挑战

1) 数据需求范围广，难以全面采集

业务部门的数据需求往往需要采集全网全面的数据，在有限的时间和成本内，批量深度爬取全网，尤其是 FB、Twitter 等社交媒体和特殊渠道的信息来支撑业务数据需求的难度越来越大。

2) 数据获取时间长，难以保证时效性

业务部门数据需求量大，导致获取时间过长，难以将数据实时的流转并供给业务分析应用。数据产生的时间过长，导致数据的时效价值被严重降低。

3) 海量数据中精准获取所需数据难度大

互联网是一个巨大的数据宝库，存在着海量的数据，但是具体的业务需求只需要采集其中和业务最紧密联系的数据。在如此多的数据中，剔除无效数据，精准找到所需数据的难度

大。

4) 大数据防爬技术加大采集数据的难度

越来越多的网站具有大数据防爬技术，并不断更新增强反爬策略，以及各国加大对隐私信息的保护，这些措施都在不断加大数据采集的难度。

2 产品概述

2.1 产品简介

易爬（iAvatar）是一款大规模分布式云化智能爬取软件，可对新闻、论坛、博客、社交、深网等数据源进行深度爬取。平台凝结了中网数据专业的数据爬取攻防技术和爬虫团队多年的最佳实践经验，支持在开源情报、舆情监测、科学研究、市场分析等应用场景下，为用户提供数据爬取服务。

易爬（iAvatar）智能大数据爬取平台目前已支持对 1000+数据源的采集，积累了超过 7T 的数据量，并以日增 5G 的速度持续积累数据。多个政府和企业，包括中国人民解放军某部、重庆发改委、航天科工二院、国家专利局等都在使用易爬（iAvatar）智能大数据爬取平台爬取数据。

2.2 适用范围

易爬（iAvatar）作为数据爬取产品，在目前大规模产生、分享和应用数据的情况下，具有广泛的应用场景。无论是公示到互联网的数据，还是存在于暗网或者具有反爬保护机制的数据，都可以使用易爬（iAvatar）进行数据爬取。具体的使用场景如下：

- 政府开源情报决策分析；
- 政府互联网舆情监测；
- 企业/公司品牌及产品口碑监测；
- 电商行业市场分析；
- 科研机构科学研究。

2.3 客户收益

1) 全面采集国内外数据，助力政府对开源情报的决策分析

开源情报是从公开的媒体挖掘有用的信息，为政府、企业提供所需的情报。易爬（iAvatar）能够全面采集国内外社交媒体及新闻媒体中的新闻资讯、人物动态、机构动向等数据信息，为开源情报的分析提供实时、准确、高效的情报数据，助力情报分析人员决策分析。

2) 实时采集互联网信息，助力政府进行舆情监测并及时预警

舆情监测需要实时汇集国内外信息数据，掌握所关注的动态信息以进行舆情分析，能够及时对不利或危情信息进行预警。易爬（iAvatar）实时提供指定目标源的数据信息，并通过数据挖掘提供社交媒体互动效果分析、关联关系挖掘、传播路径分析、话题事件分析、传播效果评估等功能。

3) 精准采集商业信息，助力企业提升市场营销能力

企业市场分析需要实时而准确地采集竞争对手的新闻、产品、价格、用户反馈等商业情报信息，易爬（iAvatar）精准采集企业品牌以及竞争对手的市场消费的需求，助力企业快速根据市场趋势找到商业机会，提高公司的市场营销能力。

4) 高效采集科研数据，满足科学研究的数据需求

科研机构在研究特定方向时需要实时准确的国内外科技信息与新闻，易爬（iAvatar）可以实时采集分布在各个网站网页上的科研数据，从而全面满足科研人员对于融合信息的收集需求，快速从公开的可信来源轻松获取科学研究的相关数据，为科研人员节约时间与精力。

2.4 产品架构

易爬（iAvatar）智能大数据爬取平台主要由 web 子系统、爬虫生成服务、分布式爬虫子系统和数据存储子系统四部分组成。爬虫生成服务根据 web 子系统下发的爬虫配置信息生成网络爬虫，并把生成后的网络爬虫上传到 web 子系统中；分布式爬虫子系统下载 web 子系统中管理的爬虫，根据爬虫的配置策略执行数据爬取任务，并将爬取的网页数据存储到数据库中。平台架构如图 2 所示。

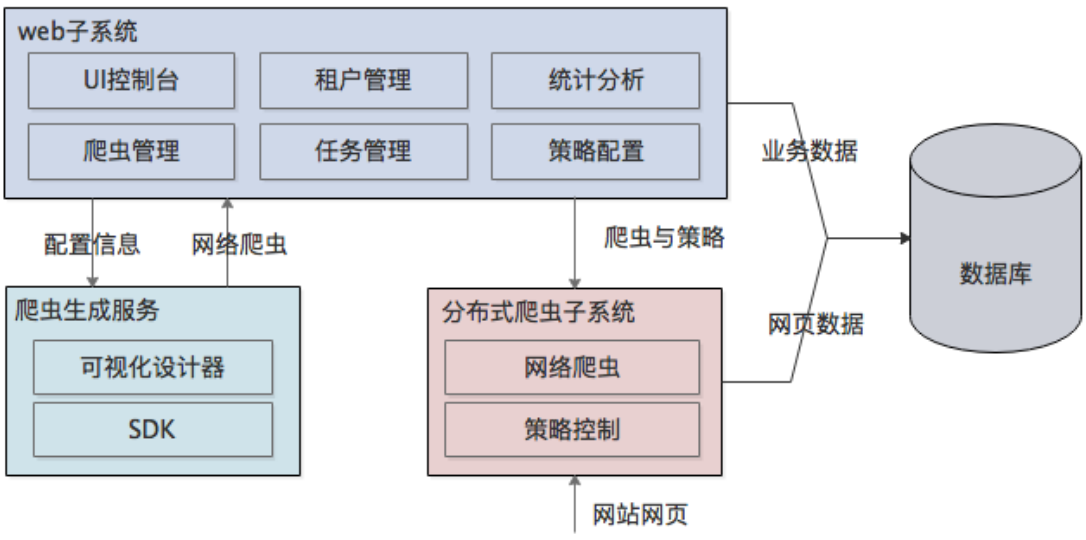

图 2 易爬（iAvatar）智能大数据爬取平台架构

3 产品亮点

易爬（iAvatar）智能大数据爬取平台作为一款全网数据采集、解析、结构化存储的服务平台，其数据采集服务具有安全、高效、准确的特点。

3.1 采集无盲点，数据覆盖率高

易爬（iAvatar）智能大数据爬取平台可爬取新闻、论坛、博客、社交、深网数据。平台现已支持 99%境内外新闻网站的数据爬取，积累了十亿条以上社交媒体数据量、千万条以上论坛数据、千万条以上博客数据。平台还支持批量采集脸书（Facebook）等社交媒体的全维度数据，不仅能爬取脸书账号基本信息，而且能对其帖子、评论信息进行深度采集。

3.2 海量采集资源，提升爬取效率

易爬（iAvatar）智能大数据爬取平台可提供 1,000+境内外 VPN、10,000+境内外 IP 资源、50,000+虚拟账号及 1,000+全球采集节点，轻松应对反爬策略，保证 7*24 小时高并发采集，提升采集效率。

3.3 机器学习，智能精准爬取数据

易爬（iAvatar）智能大数据爬取平台内置爬取引擎，采用机器学习技术有效规避反爬策略。当网站反爬措施更新时，平台可以自动选择最优的反反爬策略来自主规划爬取任务，实现网站数据的持续爬取，用户无需自己研究反爬措施和手动调整爬取策略。

3.4 反反爬引擎，解决难爬痛点

易爬（iAvatar）的反反爬引擎和中网自主研发的金睛反爬防御体系形成爬取-防爬的攻防一体能力，具备爬取+防爬实战化部署和攻防靶场能力。金睛反爬防御体系具备成熟的反爬策略，易爬（iAvatar）与其红蓝对抗，二者相互促进，不断提升易爬（iAvatar）智能大数据爬取平台的反反爬能力。

3.5 零基础，零代码

易爬（iAvatar）智能大数据爬取平台内置海量爬虫模板，涵盖新闻媒体、电子商务、社交网络、生活服务、金融征信、休闲旅游、汽车交通等热门分类；平台提供的可视化设计

器简单易用、全流程可视化操作，用户无需任何爬虫设计经验和编程基础，通过鼠标点击即可制作爬虫、运行任务和爬取数据。

3.6 身份伪装，实现安全爬取

易爬（iAvatar）智能大数据爬取平台具备强大的抗溯源、防追踪、身份隐藏的伪装能力。

能力	手段	作用
抗溯源	从 IP 地址、账号归属地、设备地理位置、语言操作系统等属性的各方面设置境外账号信息。	使他人的追踪溯源终止在境外。
防追踪	培育大量账号模拟用户行为。	抵抗运营商对账号的异常检测，防止他人对培育账号的追踪。
身份隐藏	通过多跳代理的形式获取中转服务器的信息，在可控网络环境下执行上传与下载。	极大程度防护自身人员和组织。

易爬（iAvatar）智能大数据爬取平台还可通过隐传（iTrans）高安全数据回传平台，利用消息加密隐写、数据多通道拆分传输等技术，对数据的回传进行保护。

4 产品功能

本章对产品的功能进行介绍说明，帮助您全面了解产品各个功能模块。

4.1 爬虫制作

平台提供可视化设计器和 SDK 两种方式制作爬虫：

- 可视化设计器：用户需通过可视化设计器，输入起始网站 URL 确定爬取范围，并点击界面标注感兴趣的网站元素，发布后即可完成爬虫设计。可视化设计器界面如图 3 所示。
- SDK 制作爬虫：用户通过界面下载爬虫代码模板，在本地完成爬虫代码编写并上传至平台，发布后即可完成爬虫设计。

SDK 方式制作爬虫可满足用户所有爬虫设计需求；当用户没有编程基础时，可采用可视化设计器制作爬虫。

图 3 可视化设计器界面

4.2 爬取管理

4.2.1 任务调度

任务调度模块实现数据爬取任务的分布式任务调度，包括创建、执行、监控、停止等功能，系统能够自动根据任务优先级和资源状态进行任务分配和任务调整，在数据爬取任务发

生异常时重新分配任务。

图 4 新建任务窗口

4.2.2 资源管理

资源管理是对 VPN 资源、IP 资源、虚拟身份和采集节点等与采集相关的资源信息的集中管理。

- **VPN 资源管理**

VPN 资源管理模块支持统一管理境内和境外的 VPN 账号，提供对 VPN 账号进行添加、编辑、删除、查询等操作。用户可以通过启动和禁用功能实现 VPN 动态的创建和销毁，并对 VPN 资源状态和使用情况进行实时监控。

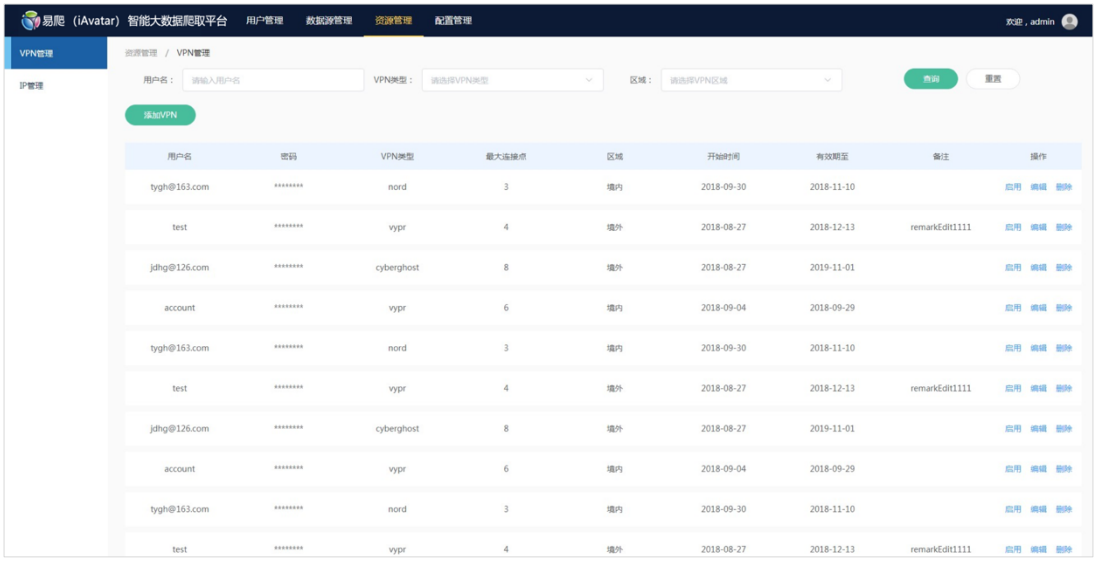

图 5 VPN 资源管理界面

- **IP 资源管理**

IP 资源管理模块支持统一管理境内和境外的 IP 资源，提供对 IP 资源进行添加、编辑、删除、查询等操作。用户可以通过此功能管理购买的 IP 资源或开源 IP 资源，并实时查看 IP 资源状态和使用情况。

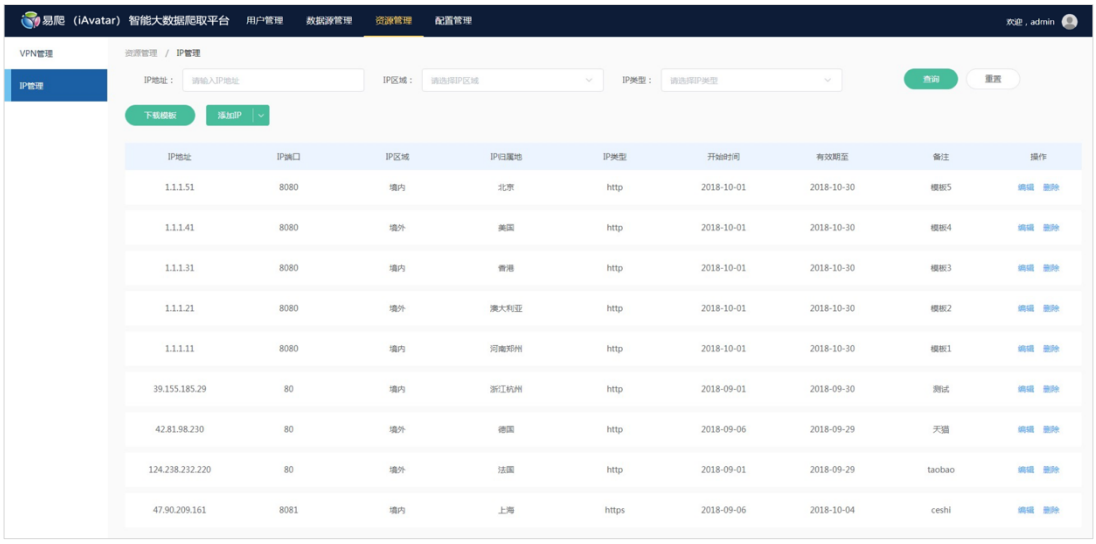

图 6 IP 资源管理界面

- **虚拟身份管理**

虚拟身份管理模块支持统一管理社交媒体类虚拟账号的创建、培育、使用等，提供对虚拟账号创建器生成的或自有导入到平台的虚拟身份进行添加、编辑、删除、查询等操作，并

通过模拟人为社交操作（例如加好友、发帖、点赞、评论等）对虚拟账号进行培育。虚拟身份主要包括账号信息、状态信息、cookie 信息。

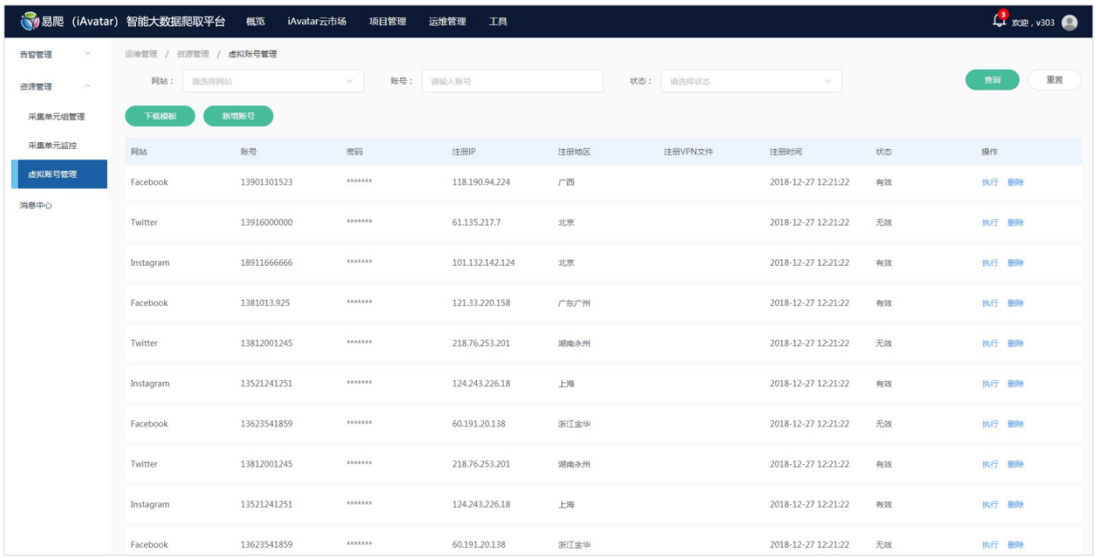

图 7 虚拟身份管理界面

- **采集节点管理**

采集节点管理模块支持管理分布式采集节点，提供对采集节点增加、删除、编辑、查询、释放的操作。该功能具备采集节点线性扩展能力，支持采集器的负载均衡。采集节点提供 CPU、内存的监控能力。

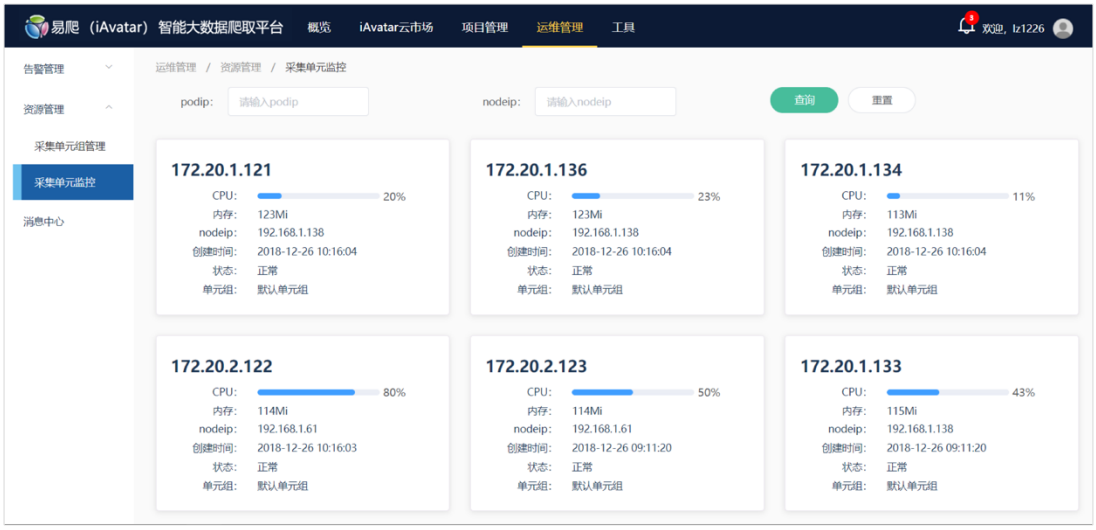

图 8 采集节点管理界面

4.2.3 状态监测

状态监测模块提供对网页页面改版、网页反爬策略、节点运行状态和数据产量等进行告警的功能，并以通知的方式实时推送到 web 前端。

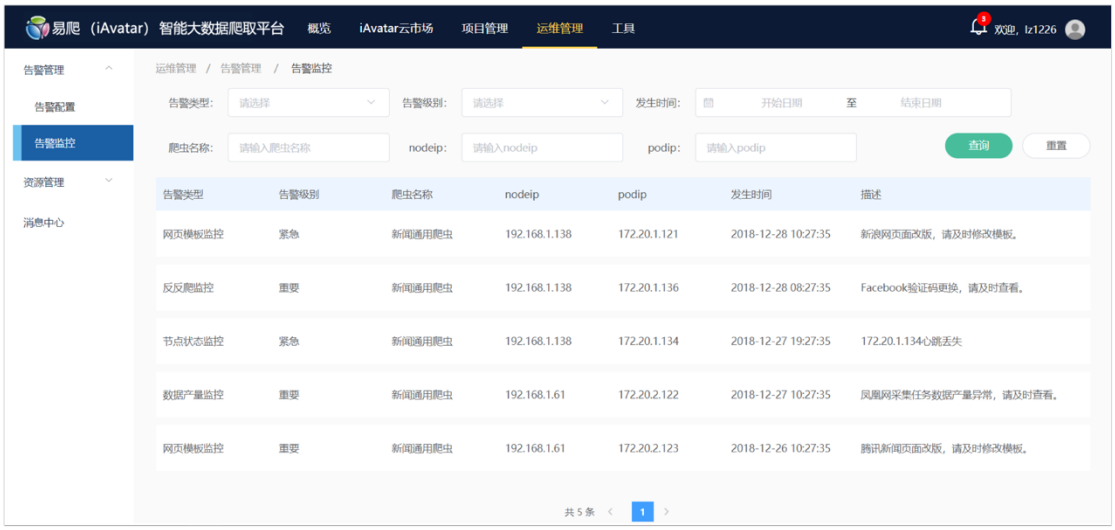

图 9 状态监测界面

4.3 数据爬取

易爬（iAvatar）基于容器技术实现分布式数据爬取运行环境，内置爬取引擎和反反爬引擎，自主规划爬取任务并自动适应反反爬策略，实现智能化数据爬取，主要通过以下几种手段来合理规避反爬策略：

- **虚拟账号创建器**

虚拟账号创建器，可识别注册、登录等表单信息，当遇到目标表单时，会调用相关资源进行账号创建、验证、打标签、保存账号信息等操作。

- **验证码识别器**

验证码识别器集成多种验证码识别能力：极验证码验证、基于 OCR 的文字识别验证、图像验证码，爬虫通过调用验证码识别接口并上传验证码信息，实现验证码类型匹配与识别。

- **IP 扩展器**

IP 扩展器定时获取指定网页源上的 IP 列表并存入 IP 资源库，定时检验 IP 的有效性，剔除无效 IP，自动根据 IP 质量的高低进行分类。

- **反反爬引擎**

反反爬引擎会根据数据爬取情况调整爬取频率，当发现爬取失败率较高时，反反爬引擎会自动限速，实现数据的可持续爬取。

4.4 数据输出

数据输出主要包含数据检索、数据导出、数据分类、数据发布等功能。数据检索提供基于爬取字段的数据信息检索功能；数据导出功能支持多种导出格式，可以按爬取字段自定义导出数据；数据分类功能提供以打标签的形式对数据进行分类；数据发布功能支持将爬取数据发布到云市场共享。

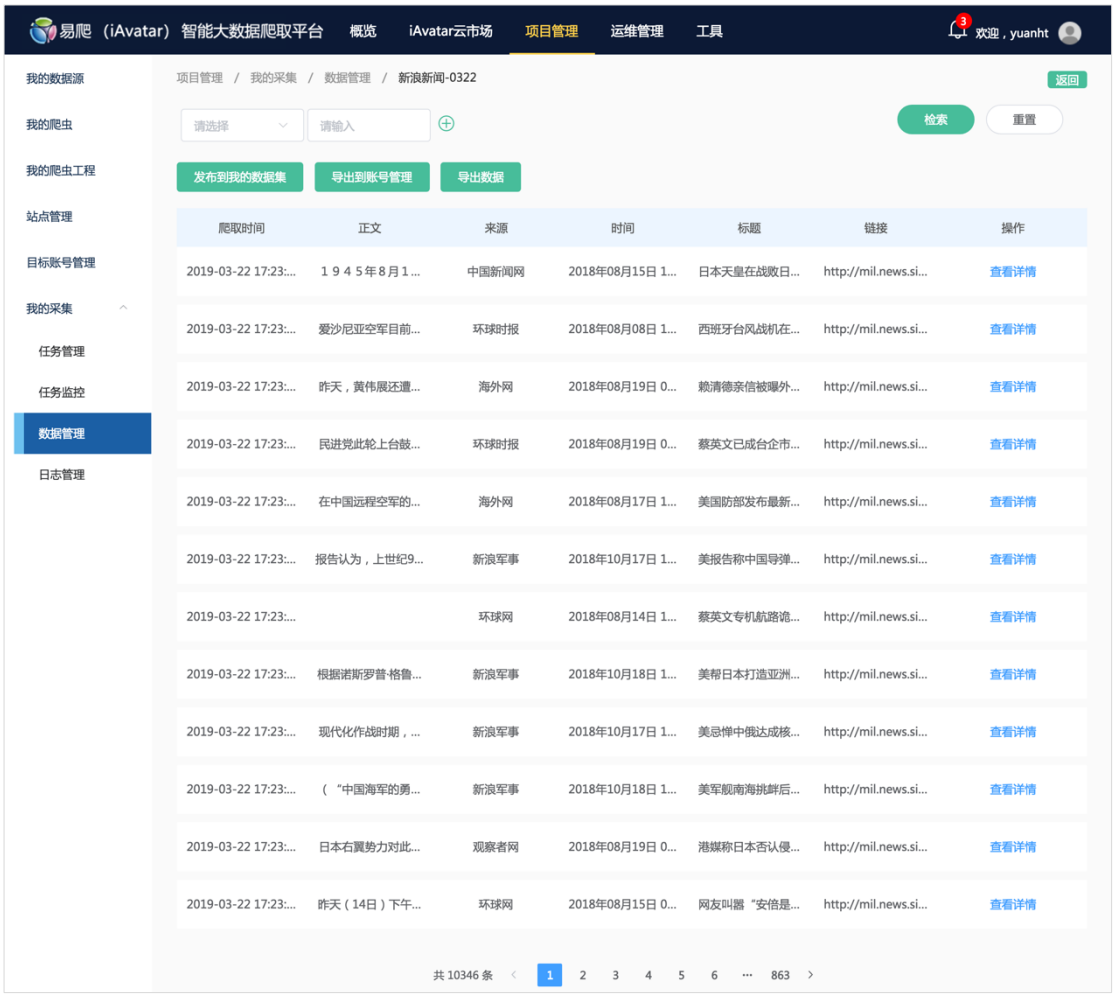

图 10 数据输出界面

5 合作模式

本章节介绍产品的合作模式，帮助您了解合作相关的基本信息。

5.1 License 服务

如果用户对数据爬取有严格的保密要求，平台支持本地部署，将爬取数据存储在用户局域网内的服务器或服务器集群中，防止数据泄露。

用户购买易爬（iAvatar）智能大数据爬取平台软件后，在本地服务器完成安装，使用管理员账号登录平台，添加分配子账号后即可开始使用。

5.2 SaaS 服务

易爬（iAvatar）提供 SaaS 服务，免去用户购买硬件及安装维护的费用，用户只需登录租户账号即可开始使用，大大降低各项支出成本。

5.3 中网云数据服务

易爬（iAvatar）作为中网云数据服务的一部分，主要提供数据采集的服务。

用户提供目标网站的网址，和感兴趣的内容，我们会据此分析网站的结构和数据抓取的复杂度，确定目标网站是否可供挖掘；然后针对性地为抓取网页上的数据而设计相应的爬虫，用来分析网页、下载数据、解析数据、结构化存储等等；最后我们会以用户需要的格式交付。

6 应用案例

6.1 某军开源情报分析系统

某军需要监测某方向的境内外主流新闻媒体和主要军政媒要员的社交媒体信息，融合第三方线下情报进行主要目标人物的全维度分析。需要爬取 1000 家以上的境内外主流新闻媒体源；某方向的 2 万多军政媒要员以及 8 万多扩展账号的脸书(Facebook)全量数据，和重点目标人物需要 30 分钟更新一次增量数据。

面临的挑战主要有：

- 实时获取脸书（ Facebook ）全量数据是业界最难之一。
- 50%的网站有各种各样的反爬措施。
- 60%的网站结构不一致。

iAvatar 系统上线后，部署采集节点 10,000 个以上，采集账号 100,000 个以上，每日采集数据量达到 100,000,000 条以上。

图 11 某军开源情报分析采集节点分布图

6.2 重庆城市信用监测项目

重庆城市信用监测项目是通过全面获取国内主流媒体网站、区县信用网站、区县政府

网站、政务外网的新闻数据，评价各区县信用建设水平和成效，定期披露区县城市信用监测排名，推进重庆市城市信用体系建设工作，提升城市软实力。需要爬取 1951 个数据源，数据覆盖率要达到 95%，3 个月完成全量数据采集。

面临的挑战主要有：

- 35%的网站有 IP 限制、流量限速、验证码等反爬措施。
- 60%的网站结构各异，无法统一适配。
- 任务重，工期紧。

iAvatar 系统上线后，45 天完成所有数据采集任务，工期提前 50%；数据覆盖率高达 99%，全量采集 10,025,165 条数据。

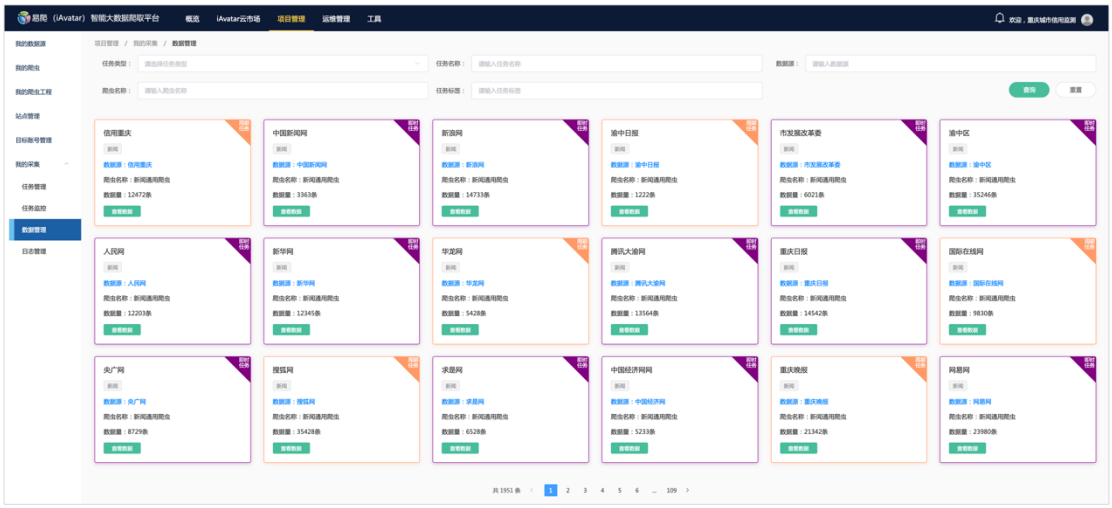

图 12 重庆城市信用监测项目爬取数据列表

6.3 国家专利局大数据采集项目

国家专利局大数据采集项目是采集印度专利网站(http://ipindiaservices.gov.in) 已公布的专利数据。主要爬取该网站 1998 年 1 月 1 日-2018 年 12 月 31 日的专利数据，数据覆盖率要达到 98%，1 个月完成全量数据采集。

面临的挑战主要有：

- 网站有 IP 限流、验证码等反爬措施。
- 预计高达 5T 的专利附件，采集量巨大。

- 网站经常改版，爬取稳定性差。
- 全英文网站。

iAvatar 系统上线后，18 天完成所有数据采集任务，工期提前 40%；数据覆盖率高达 100%，采集了 60 万的专利数据，文件大小达 4.9T 。

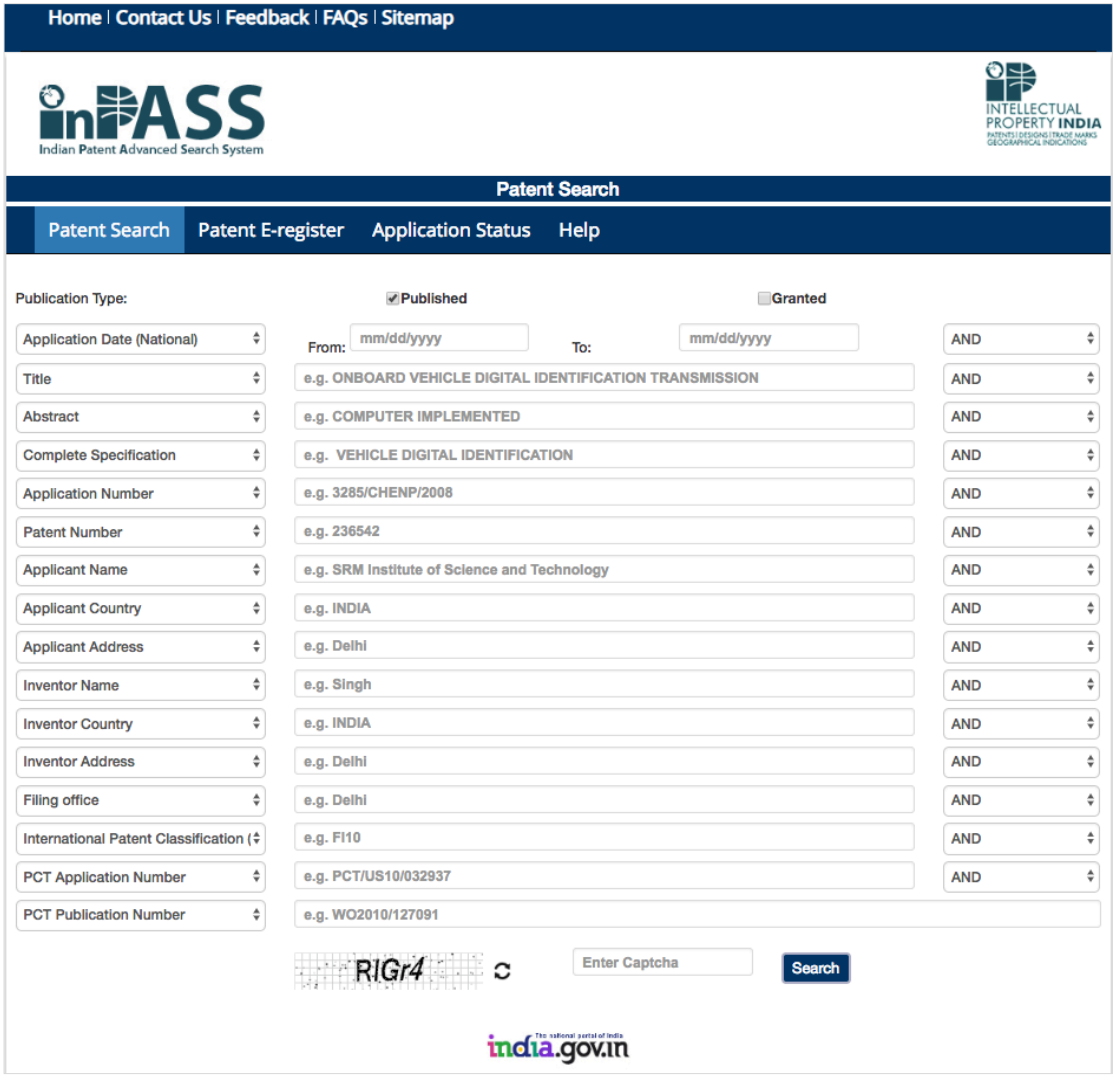

图 13 国家专利局大数据采集项目网站截图

地址：北京市海淀区丰秀中路 3 号院 7 号楼中网数据大楼 100094

电话：+86（10）58251700　　传真：+86（10）58251701

E-Mail：contact@isinonet.com

产品版本号：V3.3　　资料版本号：01