高铁采集器 V9.8

产品白皮书

成都日渊大数据有限公司

2018 年 7 月

目录

一、引言

1.1 文档主题

成都日渊大数据有限公司是国内最早从事互联网数据服务的企业之一，多年来专注于互联网数据采集领域，面向国内外的广大用户提供技术支持与服务。目前拥有超过十万的合作客户，其中包括政府机构和众多知名企业。乐维公司一直秉承着为客户节约成本，提升价值的服务理念，做客户最值得信赖的合作伙伴。

由成都日渊大数据有限公司自主研发推出的核心产品——高铁采集器，是一款能够高效采集网页数据的采集软件。高铁采集器作为国内使用人数最多的网页数据采集产品，曾多次被网易新闻、电脑报、安徽商报等知名媒体报道，备受业界关注。软件通过对网页数据的提取，处理，发布等操作，使网页数据的提取利用变得简单便捷，能够显著提升使用者的工作效率。

本文档就高铁采集器的研发背景、产品概述、功能、应用等各个方面进行了系统的介绍，以帮助本文档的读者快速，全面的了解高铁采集器。

1.2 适用范围

本文档适用于需要全面、系统的了解高铁采集器产品的人群，其他关于高铁采集器的信息咨询，可以联系高铁采集器的客服人员为您解答。

1.3 相关术语

- **采集任务：**采集任务是高铁采集器中对于数据采集和数据发布任务的完整配置，包含采集规则和发布模块。
- **采集规则：**即我们对如何采集和采集什么的问题给出一些设置让采集器按照设置的规则来执行，这个设置可以从高铁采集器里面导出保存为.ljobx 文件，也可以再次导入高铁采集器。
- **发布模块：**在高铁采集器中，发布模块是对“将已经采集到的数据发布到哪里”进行的设置。包括 WEB 在线发布模块和数据库发布模块，其设置分别可以导出保存为.wpm 文件和.dbm 文件，并可以再次导入高铁采集器，多次使用。
- **发布接口：**发布接口是一个小型的页面程序，通常和 WEB 在线发布模块配合使用来满足用户的特定需求。即采集器将采集的数据发送到发布接口文件中，接口文件得到数据，并按照用户特定需求灵活地处理数据。
- **标签：**是指用来提取某项内容信息的一个字段名字，由用户在编辑规则的时候指定，比如标题、手机号、邮件、作者，内容标签采集到的信息在发布模块中就可以通过该标签名对应获取到，格式为[标签：标签名]如[标签：标题]。标签在高铁采集器里面有分为两种：分别为列表页标签和内容页标签，顾名思义列表页标签就是在获取列表页时（即采网址时）就获取到内容信息，

内容页标签是在获取内容页或多页内容时（采内容）才获取内容信息。注：通常还有一种说法为 html 标签，这里的标签是指一些 html 代码里面的属性标识符，如：<a href 里面的 a 标签，里面的 font 标签为 html 标签，该术语在内容处理的 html 标签排除项出现。

二、高铁采集器的研发背景：

2.1 从搜索引擎到网页数据采集

以惊人的速度发展起来的网络，成就了万维网这个拥有着大量信息资源的宝藏，基于万维网信息资源而生的搜索引擎则实现了信息的有效提取和利用；但仍在飞速发展的网络让我们对互联网信息产生了新的需求，不仅要搜索到信息，还要将所需数据信息快速地收集到目标中去，这个目标可能是一家网站，一个数据库，一间网店，一篇文档……所有需要数据的地方，正是这种对数据利用的强烈需求催生了网页数据采集。

2.2 从手动采集到软件采集

对网页数据的采集需求最初通过人工手动采集来实现，我们把需要的数据复制下来，再粘贴到目标中去，就完成了一个最简单的采集过程。手动采集可以满足少量的采集需求，但网页数据是海量的，我们需要的数据往往也是大量而又复杂的，传统手动采集会耗费更多时间更多精力，因此我们需要一个高效的采集工具，来帮助我们快速完成采集，在这种需求之下，高铁采集器应运而生。

高铁采集器实现了将数据从采集到处理到发布的一系列智能操作，能够快速稳定地应对大量的数据采集需求，取代手动采集模拟人工操作，大幅提升工作效率。

三、高铁采集器综述：

3.1 高铁采集器简述

高铁采集器是一款专业的网页数据抓取、处理、分析，挖掘软件。软件凭借灵活的配置，可以轻松迅速地抓取网页上散乱分布的文本、图片等文件，并通过数据清洗、过滤、去噪等预处理后进行整合聚集存储，再进行数据的分析挖掘，最终将可用数据呈现。

3.2 功能详述

高铁采集器主要包含网址采集、内容采集、数据处理、数据发布、多任务多

线程运行、HTTP 二级代理服务器、计划任务管理器、任务运行日志管理和插件拓展九大特色功能，下面对九大主要功能进行详细说明。

3.2.1 网址采集

高铁采集器可以通过网址采集规则的设定，快速采集到所需的网址信息。可手动输入、批量添加或直接从文本导入网址，并能自动筛选去除重复的网址信息。

支持多级页面网址的采集，多级网址采集可以使用页面分析自动得到地址、手动填写规则两种方式。应对多级分页中内容不同，但地址相同的页面网址采集，高铁采集器设置了 GET，POST 和 ASPXPOST 三种 HTTP 请求方式。

高铁采集器支持网址采集测试，可以验证操作的正确性，避免操作有误导致采集结果不准确。

3.2.2 内容采集

高铁采集器可以通过分析网页源代码，设定内容采集规则，精准采集到网页中散乱分布的内容数据，并支持多级多页等复杂页面中的内容采集。

通过定义标签，能够将数据进行分类采集，比如将文章内容的标题与正文分开采集。高铁采集器配置了三种内容提取的方式：前后截取、正则提取、正文提取。可选性强，用户可以按照使用需求进行选择。

内容采集同样支持测试功能，可选用一个典型页面来测试内容采集的正确性，以便及时更正和进行下一步数据处理。

3.2.3 数据处理

对于采集到的信息数据，高铁采集器可以对其进行一系列的智能处理，使采集到的数据更加符合我们的使用标准。主要包括 1）标签过滤：过滤掉内容中不需要的空格，链接等标签；2）替换：支持近义、同义词替换；3）数据转换：支持汉译英、简转繁、转换为拼音等；4）自动摘要、自动分词：支持自动生成摘要和自动分词功能；5）下载选项：高铁采集器支持任意格式的文件探测下载，并能够将相对地址智能补全为绝对地址。

3.2.4 数据发布

高铁采集器将数据采集下来后默认将数据保存在本地数据库（sqlite、mysql、sqlserver），用户可以根据自己的需求选择对数据的后续操作以完成数据发布，支持直接查看数据、在线发布数据和入数据库，并支持用户进行发布接口的使用和开发。

根据数据库类型用相关软件打开可以直接查看数据，配置一个发布模块即可将数据在线发布到网站，可以设置自动登陆网站，获取栏目列表等；如果入到用户自己的数据库中，用户只需写几个 SQL 语句，程序就会按照用户的 SQL 语句导入数据；保存为本地文件时支持本地 SQL 或文本文件（word、excel、html,txt）格式。

3.2.5 多任务多线程运行

高铁采集器可以选择同时运行多个任务，支持不同网站或同一站点下不同栏目的内容同时采集，能够有计划的调度任务。单个任务在采集内容和发布内容时均可以使用多线程运行，提升运行效率。

3.2.6 HTTP 二级代理服务器

高铁采集器可以通过二级代理服务器的功能实现 IP 的更换，避免因 IP 被限制访问而导致的采集无法运行，用户需先获取一些代理 IP，然后将代理 IP 导入采集器中完成设置即可。

3.2.7 计划任务管理器

高铁采集器支持计划任务管理，能够定时自动地进行采集发布，实现自动更新的功能，可对加入计划任务内的任务设置其执行的频率和开始运行的时间，执行频率可以选择每周、每天、每间隔，或根据用户需求自定义 corn 表达式执行。

3.2.8 任务运行日志管理

高铁采集器配置了采集监控系统，任务运行管理器将采集监控模块生成的记录信息组装成日志条目，如果启用了自动运行功能或需要对程序运行状况进行监控，可以查看任务运行日志中某个日期时间段内的运行情况，来做具体的分析。可以具体了解到任务的成功数量、失败数量，重复数量和用时等数据。

3.2.9 插件扩展

高铁采集器支持 PHP 和 C#插件扩展，可以帮助用户对采集的数据进行修改处理，完成用户的更多需求，极大的扩展了采集器的功能。用户可以按照插件开发手册自行开发所需插件，也可以使用高铁采集器官方开发的一些插件资源。

高铁采集器中配置了插件管理器，可对插件列表进行管理和选择插件方法，支持插件测试。

3.3 版本介绍

高铁采集器分为免费版、基础版、旗舰版机器码版、旗舰版自动授权版、企业版尊享版，企业版豪华版六个版本，每个版本对应的功能和权限有所不同，高铁采集器支持低版本向高版本的升级。

- **免费版：**高铁采集器免费版已经具备许多基础功能，比如分页采集，post 获得网址，支持网站登录采集等，能够满足一般用户的使用需求，用户同样可以在论坛和 QQ 群中获取技术支持。
- **基础版：**高铁采集器基础版在免费版的基础上为用户提供更为强大的采集功能，包括：ftp 上传功能，标签自由组合功能，Http 请求功能，定时定量采

集发布，多页采集，Sqlite 数据库存储数据，采集到的数据编辑后再发布等。基础版只能绑定一台电脑或服务器，包含免费更换电脑一次和一年的免费升级年限服务。

- **旗舰版机器码版：**高铁采集器旗舰版机器码版相对于基础版增加了文件自动上传到网站，二级页面缓存，二级代理服务器，图片加水印等。同样只能绑定一台电脑或服务器，包含免费更换电脑一次和一年的免费升级年限，附赠一个采集规则和一个发布模块。
- **旗舰版自动授权版：**高铁采集器旗舰版自动授权版与上述机器码版功能一样。区别在于自动授权版绑定一台机器后可以随时更换机器码，不限次数，同样有一年的免费升级年限，附赠两个采集规则和一个发布模块。
- **企业版尊享版：**企业版尊享版较之其他版本增加了数据发布到 Oracle，Http 接口管理采集器运行的功能，包含两个加密狗和三个机器码，需要绑定三台机器，赠送三次免费更换授权。使用加密狗时无需绑定机器，即企业版尊享版可以同时供五台机器使用，附赠四个采集规则一个发布模块，一年免费升级年限。同时享有一次免费的网络培训（三小时）。
- **企业版豪华版：**豪华版在功能上与尊享版相同，区别在于豪华版包含四个加密狗和六个自动授权，即可以供十台机器同时使用，且六个自动授权绑定机器后可随时更换授权。加密狗同样无需绑定机器，附赠八个采集规则和两个发布模块，以及一个插件，享有一次免费的网络培训（三小时）。

四、高铁采集器的特性

4.1 全网通用

4.1.1 基于 web 结构的采集

高铁采集器的采集原理是基于 web 结构的源代码提取，几乎适用于所有的网页，以及网页中能够看到的所有内容；

4.1.2 扩展性强

高铁采集器支持接口和插件多种扩展延伸，打破操作局限，满足更加多样化的使用需求，使高铁采集器真正做到全网通用。

4.2 功能全面

4.2.1 集采集发布于一体

高铁采集器在每个功能上都做了优化设置，除了最基础的数据采集，更是融入了强大的数据处理和数据发布功能，全面完善了对于数据利用的整个流程。

4.2.2 多元化的功能配置

高铁采集器在许多细节操作中配置多项可选方式。1）多种提取方式：网址和内容的提取均设置了多种方式，网址采集包含手动填写采集规则、页面自动分析，内容提取包含前后截取、正则提取、正文提取，标签组合，用户可根据自己的需要选择不同方式；2）多识别系统：正文识别、任意编码识别等多种智能识别系统；3）图片、压缩文件、视频等任意格式的文件都能轻松下载；4）支持 Access/MySQL/MsSQL/Sqlite/Oracle 五种类型的数据库发布；5）可选择使用加密狗，随时移动更安全。

4.3 高效稳定

4.3.1 分布式高速采集系统

高铁采集器采用分布式高速采集系统，将任务分配至多个服务端同时运行采集，多任务多线程式的运行模式，能够最大化提升运行效率。

4.3.2 占用资源少

任务量得以分解，服务端所占用资源减少，使得高铁采集器的性能更加稳定。

4.4 数据精准

4.4.1 采集监控系统

实时地监控采集，任务运行日志报错统计，及时修复，确保数据不被遗漏。

4.4.2 数据处理准确

多种精细化的数据处理方式，结合测试功能让高铁采集器做到数据采集无误，精准可用。

五、高铁采集器的典型应用

5.1 使用群体及运用

5.1.1 企业人员

收集潜在的客户信息，快速挖掘新客户；通过分析客户行为开展业务，降低风险和预算，洞察竞争对手的业务数据，助力商业决策。

5.1.2 电商运营

按照用户需求定向采集商品信息、商家信息、产品评价，挖掘相关数据背后的潜在价值，进行精准的营销优化，提升运行效率。

5.1.3 政府机关

实时汇集国内外信息数据，掌握所关注的动态信息，进行舆情监控，及时对不利或危情信息进行预警，并通过分析数据指导社会与经济的发展。

5.1.4 网站站长

实现定时采集数据和自动发布数据，采集优质内容加工处理后填充发布到网站，让网站快速拥有强大的内容支撑，轻松提升流量与人气。

5.1.5 个人需求者

批量下载大量的文件，图片等内容，解决个人在学术研究或生活，工作等方面的数据需求，取代手动复制粘贴，提高效率，节省下更多时间。

5.2 应用案例

案例一：地震台网中心

某地震中心通过高铁采集器汇集到各类地震监测数据并加以分析，同时能够实时监控数据动态，及时预警最新地震活动分布范围。

案例二：某品牌保险

数据为保险行业带来四大精准：精准营销、精准定价、精准管理，精准服务。通过采集器抓取、筛选和分析出精算、营销、投保、服务、理赔等各个环节的统计数据，更加科学地设定各种费率；向客户提示保障不足的地方，筛选出最适合的保险产品和服务类型并向其精准推送。

案例三：淘宝店长

电商运营人员用高铁采集器采集到同类商品的属性、评价、价格，市场销量占比，从而进行某商品标题的搜索优化，根据同类经验制造爆款，提升网店的运营水平与效率。

案例四：视频网站

对采集到的视频数据进行流量分析，排序，使得精品内容得以不断涌现，并能将精品快速发布到目标网站中，提升网站流量，助力内容与营销升级。

案例五：著名大学科研人员

高铁采集器帮助科研人员完成大量科研数据的检索、采集，在短时间内快速批量下载大量内容，取代手动采集，节省下更多时间，工作效率快速提升。

六、运行环境

6.1 系统环境

Win10/Win7/Win8/Win2012/Win2010/Win2008/Win2003/Vista/Xp

6.2 框架支持

要求电脑安装.NET4.0 框架支持，下载地址：
http://www.microsoft.com/zh-cn/download/details.aspx?id=17718

七、高铁采集器技术支持

官方 QQ：402248381

官方网站：https://www.highferrum.com

微信：gaotiecaijiqi

邮箱：402248381@qq.com

手机：18707070070

座机：028-86755595

地址：四川省成都市锦江区庆云南街 69 号 1 栋 16 楼 12 号

邮编：610021

联系人：王琪灵

工作时间：非节假日 9:00-18:00