TBSGET 网络文献采集软件

功能特点

北京金信桥信息技术有限公司

Http://www.tbs.com.cn

第一章 软件结构

1.1 系统架构

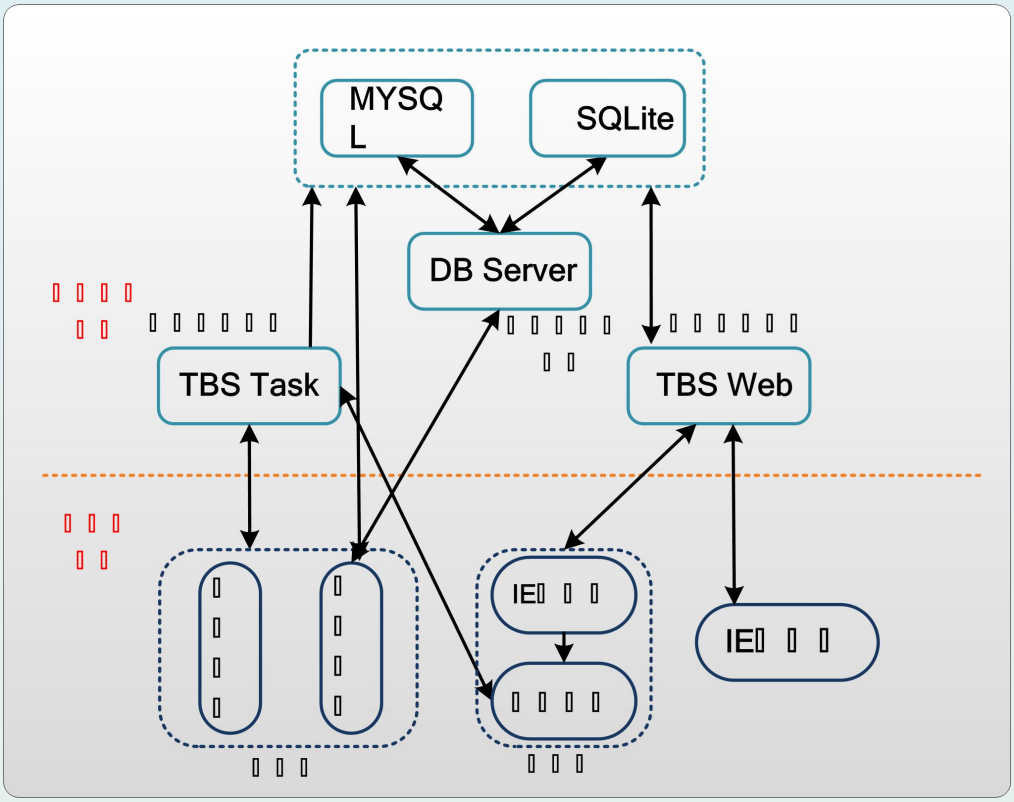


图 2.1 系统整体构架图

1.2 功能架构

TBS 网络信息采集软件是将 web 浏览器中的信息通过容器 IWebBrowser2/IHTMLWindow2 里的接口设置采集规则，并通过调动任务进行采集，最后将采集到的结果进行保存、入库。其软件组成主要包括：任务调度模块（TBSTASK）、信息规则配置模块、信息采集模块、信息入库模块四大模块。如下图：

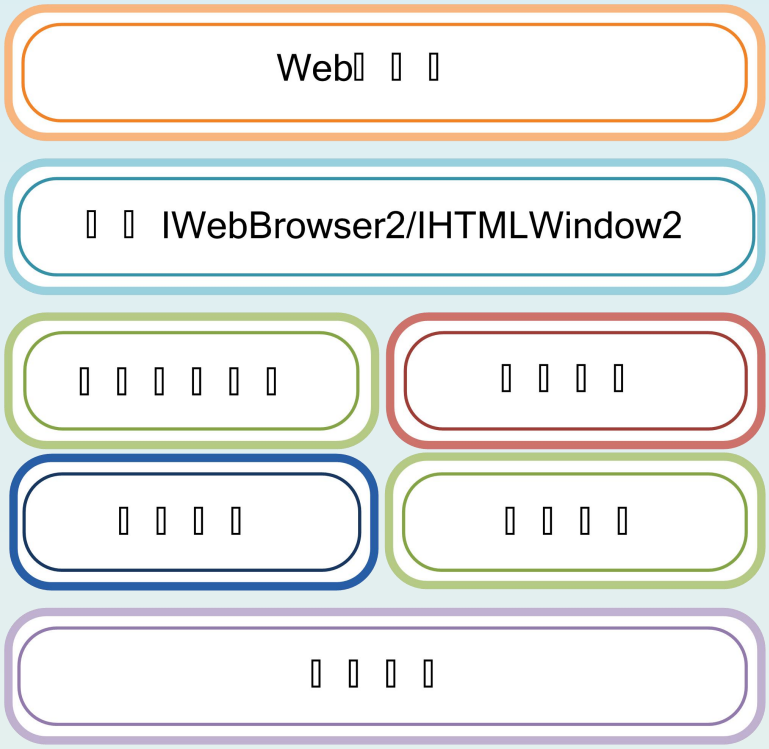


图 2.2 层次构架图

任务调度模块（TBSTASK）是对所有的采集规则进行自动调度，以减少人工操作的繁琐性。

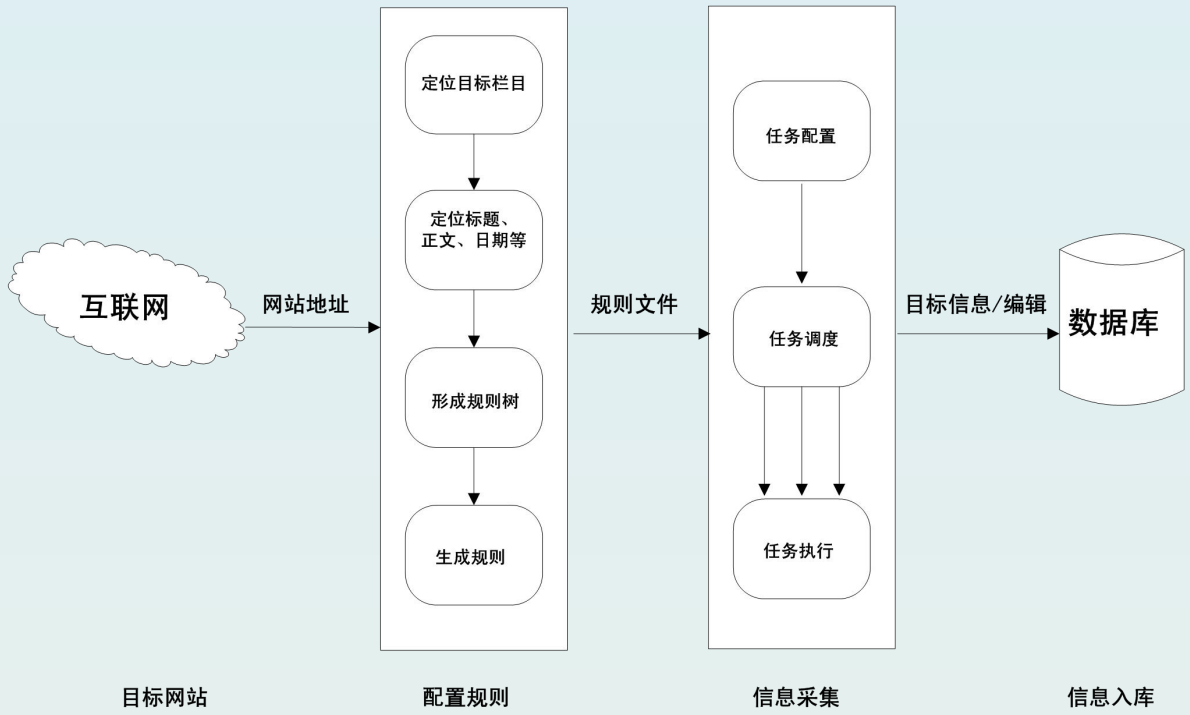
信息规则配置模块主要是对要采集的站点中的相应栏目进行采集规则的配置。

信息采集模块主要是执行所配置的采集规则。

信息入库模块主要是将采集后的信息进行归档入库，以便于发布软件调用。

信息编辑模块主要是对采集信息中无法自动过滤的信息进行人工编辑修改操作.

1.3 工作原理



第二章 功能及特点

TBS 网络信息采集软件采用 C/B/S 结构，用户界面友好，编辑规则容易，稳定性较高，在 Windows 平台（TBSGET 采集服务端以及 TASK 服务端已支持 LINUX 系统），客户端在 IE8.0 以及 IE 的高版本上都可以长时间正常运行。

2.1 软件功能

★ URL 自动去重功能：URL 去重功能是系统默认的排重功能，对于采集过的页面，将其地址存入防重数据库，进行新一轮采集时，首先查询防重数据库，如果已经下载，则自动丢弃该任务。

★ 标题自动去重功能：对于 URL 不同，但标题重复的条目信息，单靠 URL 排重无法满足实际需求，对此，系统提供了标题排重功能。

★ 实时采集实时入库功能：对于采集下来的信息可以实现实时入库发布。

★ 多任务并行处理功能：可以设置多个采集任务，任务调度器自动完成任务调度，分布式处理，采集效率高。

★ 垃圾信息自动过滤功能：对网页中的广告、无用链接等垃圾信息软件可以进行自动过滤、去除，仅提取有用的信息。

★ 正文或标题信息的自动过滤功能：对于写入正文或标题节点中的垃圾信息，可以对其进行自动过滤。

★ 日期格式的自动转换功能：对于日期格式不规范的网站，可以实现日期格式的自动转换，以便于获取新闻的正确发布日期，并对数据库中的日期字段进行索引。

★ 全方位的采集功能：采集的对象包括文字内容、图片、PDF、DOC、flash 动画、视频等等各类网络内容。支持图文混排对象的同时采集。

★ 下载日志管理功能：对于已经下载的网页和下载出错的网页均作日志记录，便于查看和维护。

★ 采集页面自动跳转功能：若在设定的时间范围内无法打开采集页面，系统会自动放弃该页采集，自动跳转到下一页的采集。

★ 进程自动退出功能：若在规定的时间内无法完成采集任务，自动退出采集任务，等待下一轮采集。

★ 弹出窗口自动关闭功能：有些网站在采集过程中总是有规律的弹出一些提示窗口，需要点击一下“确定”按钮，才能继续采集，对于这种情况，系统提供了对该窗口的自动关闭功能。

★ 关键词过滤功能：有些用户并不需要网页中某一栏目的所有信息，而是需要栏目中包含某些关键词的信息，对此，系统提供了关键词采集的功能。用户可以通过设置界面设置要过滤的关键词，这样，系统就会根据设置执行采集任务。关键词过滤含与、或、非三种表达式。

★ 词表过滤功能：可根据提供的词表信息，对采集的信息进行按关键词自动分类。

★ 格式转换功能：可根据需要将采集下来的文本信息，转换成网页格式、UNICODE 格式、GB2312 格式、JSON 格式等。

★ 图片下载控制功能：系统可根据用户的实际需要设置是否下载图片信息。

★ 自动翻页功能：系统可以根据用户实际需求设置概览或正文中的自动翻页，并对翻页的页数进行可控配置。

★ 条目采集时间自动调控功能:有些网站是非常抵制自动信息采集的,因此，在采集时间上进行了相关排查,为此，该软件增加了采集时间人为调控的功能,用户可以根据需要设置采集的总体时间及单条记录采集时间，该时间可以是固定时段，也可以是随机时段。从而最大化的模拟了人工操作。

★ 采集条数的自动调控功能:不同的网站，尤其是文献数据库的网站，允许用户每天下载数据库记录的条数不定，如果靠用户自己记住每天采集的数量，势必影响用户的工作效率，为此，该软件提供了每天下载记录数量的配置功能，用户只要设置好每天采集的数量，便可高枕无忧，一劳永逸。

★ 采集代理 IP 的自动转换调用功能：对于一些有着 IP 限定的网站，既一个 IP 代理只能采集限定数量的网站，如果用户想要在一台机器上采集限定数量以上的的信息，势必要不断的更换 IP 代理，这样无疑会影响工作人员的工作效率，为此，系统提供了 IP 代理自动转换调用的功能，用户只需在设置界面设置好相关的代理 IP，系统就会根据实际情况，反复调用不同的 IP 代理。

★ 采集文本和附件自动上传功能：URL 对于一些需要将采集的信息存放到中转服务器而不是本地服务器的用户，系统提供了信息自动上传的功能。只要在设置界面进行了相关设置，采集信息就会自动上传到指定的目标服务器上。

★ 强悍的抗干扰功能：很多网站都针对采集行为作了各种干扰措施，TBS 网络采集软件利用的是模拟人为操作技术，因此这些反采集的干扰措施对采集基本无效。

★ 自动运行调度功能：任务的运行调度由服务器来完成。主要是根据任务的当前设置，包括任务对应的功能程序、运行有效时间、次数、间隔、运行参数等，从而实现任务周期调度功能。

★ 任务管理功能：任务的管理功能主要由客户端来完成。通过 B/S 客户端操作界面，用户可以增、删、改任务，设定任务的运行有效时间、次数、间隔、运行参数等，也可设定任务是否处于激活状态。同时，允许用户对任务进行归类管理，支持批量任务状态修改。

2.2 软件特点

★ 提供交互式界面，界面友好，操作方便。

★ 采集后的信息的保存方式多样，可满足多种发布软件的需求。

★ 模拟人工的 cookie 功能，具有极强的隐蔽性。

★ 采用任务调度方式，实时采集实时发布，减少人工干预，有效提高用户的工作效率。

★ 支持多任务并发运行，有效提高采集效率。

★ 自动删除重复信息，有效保持记录的唯一性。

★ 规则配置方法多样化，覆盖面广，可最大程度上满足用户的需求。
★ 自动过滤采集页面的垃圾信息，有效提高了信息内容的准确率。
★ 服务端支持 WINDOWS 和 LINUX 双系统。
★ 采集端支持 WINDOWS 和 LINUX 双系统。
★ 支持 JAVASCRIPT 的翻页，图片翻页，大大提高了采集效率。
★ 可通过 JS 编程进行后续规则编辑，以便更精准的提取所需信息，进行信息统计分析。
★ 软件操作简单，采集配置人员懂一些简单 HTML 和 JS 脚本，即可熟练进行采集配置。
★ 系统采用可编程模块化设计，方便用户进行功能扩展。
★ 规则利用率高，一次配置，可长期使用，定时调度，大大减轻了人工重复劳动，有效提高了工作效率。
★ 采集时间动态分配，有效避开防爬监控。
★ 采集数量动态控制，减少人为失误。
★ 采集 IP 动态调用，减少人为操作，有效提高工作效率。
★ 提供了多种采集内核，解决了因浏览器的兼容性而导致采集任务无法进行的问题。
★ 支持自动登录，解决了因登录受阻，而导致信息无法采集的问题。
★ 支持自动翻页并能控制翻页页数，有效提高了用户的工作效率。
★ 支持关键词过滤采集，大大提高了采集信息的精准率。
★ 支持多次跳转的 JS 解析，成功解决了因跳转过多而导致的信息无法采集的难题。
★ 支持采集和存储分布式管理，有利于用户对软件和存储的维护与管理。
★ 支持标题和正文的过滤，有效提高了采集信息的准确性。
★ 支持实时采集实时入库，使用户能在第一时间检索到新采集的内容。
★ 支持 SQL、MYSQL、SQLLITE 等数据存储。
★ 支持网页、论坛、博客、微博、数据库等各类网络数据的采集。
★ 高效的解析、采集速度，配合多线程、多项 目同时运行的功能，可以确保你的下行带宽充分得到利用。
★ JS 解析的自动判断识别，可快速检查被采集的页面是否需要执行 JS 解析，如果不需要的，尽量不使用低效的 JS 解析模式。
★ 系统容错性较强，采集结果数据完整度高，只要规则配置得当，基本不会出现采集结果遗漏的情况。

第三章 联系方式

地址：北京市海淀区中关村东路 66 号世纪科贸大厦 B 座 2306 室
邮编：100190
总机：（010）62670903/62670700
传真：（010）62670877
客户服务：（010）62670903
公司网站：http://www.tbs.com.cn