

视采采集器

产品白皮书

caijiqi.net 敬上

日期	版本	作者	修改内容
2006-11-13	0.1	caijiqi.net	新版

目 录

- 1 概述
 - 1.1 目的
 - 1.2 产品简介
 - 1.3 市场分析
 - 1.3.1 互联网应用
 - 1.3.2 信息搜索
 - 1.3.3 资料录入
 - 1.4 需求概述
 - 1.4.1 网站采集
 - 1.4.2 信息采集
 - 1.4.3 数据结构化
- 2 用户特点
 - 2.1 网站管理员
 - 2.2 信息采集用户
 - 2.3 数据结构化用户
- 3 运行环境
- 4 运行体系
- 5 系统特性
 - 5.1 I/O体系
 - 5.2 容器体系
 - 5.3 缓存体系
 - 5.4 插件体系
- 6 功能说明
 - 6.1 结构化采集
 - 6.2 可视化元数据定义
 - 6.3 插件支持
 - 6.4 客户端环境模拟
 - 6.5 多线程采集
 - 6.6 全局发布
 - 6.7 分页采集
 - 6.8 关联文件下载
 - 6.9 规则保存
 - 6.10 模板修饰
 - 6.11 结果过滤、替换
 - 6.12 重复过滤
- 7 支持信息

1 概述

1.1 目的

本文从技术的角度向用户介绍采集系统的系统结构、系统特性及功能特征。并分析采集系统所面对的市场行情和当前的用户需求。

本文适合用户以及技术人员阅读和参考。

1.2 产品简介

视采采集器是一款所见即所得的采集别人网站数据和论坛文章帖子的数据采集软件。它使用先进的数据结构化分析技术，通用性强，所见即所得，简单易用。系统提供可视化定义规则，即时结果预览，即时帮助向导，网站自动采集，论坛自动跟贴等先进功能。系统模拟各种浏览器特性，突破多种防采限制。可作为论坛采集器、新闻采集系统、CMS 采集器等网站数据采集工具使用。

1.3 市场分析

1.3.1 互联网应用

随着互联网的发展及普及，互联网用户迅速增长，上网已成为人们生活中的日常内容，人们通过网站阅读，发表，搜索，交流，购物等，所有这些上网行为，由点到线，都将汇聚庞大的商业价值。因此，互联网成为众多人的梦想帝国，淘金之地。不管您是腰缠万贯，还是身无分文，这里只谈信息为王，服务至上。因此信息的创造、收集、组织和再加工是网站的生存基础。信息采集系统可以通过网站管理员指定的网站地址和预定义的抓取规则，自动获取网页内容，自动按照自身网站系统的数据结构抽取数据，并发布到网站系统中，让您不花丝毫心血和金钱，就可以使您的网站一夜之间网罗天下。

1.3.2 信息搜索

由于各种用户群体的网络连接，使得互联网成为一个包罗万象的信息库，商业的、学术的、个体的等等信息都可以在互联网上发布和获取，因此，企业可以通过互联网获取客户资源、市场行情、商业信息等。但在这茫茫的信息大海，我们常常缺少一种工具来发现我们所关心的内容，并有效的组织和储备它们，使之成为企业的内部资源。信息采集系统可以根据数据模式，自动通过搜索引擎检索数据，将匹配的信息显示在您的桌面上。

1.3.3 资料录入

企业管理系统，企业信息管理系统、客户服务系统等各种信息处理系统，它们只能处理结构化的数据，如学生信息包括用户名、性别、年龄等属性，它们必须保存在预定义的结构里。但系统外界会有大量的非结构化数据，如客户提交的材料、公司内部文档等。而这些数据通常是人工统计和人工输入各类信息处理系统中。信息采集系统它能将一篇文档按信息系统的数据结构自动抽取成多个字段，并自动将这些字段导入到企业的各类信息处理系统中。

1.4 需求概述

网站管理员最大的心愿是提供最丰富的网站内容，吸引更多访问量；市场营销人员每当通过蛛丝马迹而获取到隐藏的客户资源而兴奋不已；企业后勤人员做梦都想远离这些枯燥无味的文字录入。采集系统好比一双慧眼，让您看得更远，获得更多。

1.4.1 网站采集

网站管理员希望将别人的整站数据下载到自己的网站里或者将别人网站的一些内容保存到自己的服务器上。从内容中抽取相关的字段，发布到自己的网站系统中。有时需要将网页相关的文件也保存到本地，如图片、附件等。

网站管理员会定时从同一网站上抓取内容，希望已经抓取的内容不要再发布到网站系统中。对于一些网站，需要登陆才能获取页面。网站管理员希望能够通过一个内容列表页面获取所有的相关内容，包括内容列表的其它分页。当第二次抓取相同网站时，希望不要再重复第一次的设置。

1.4.2 信息采集

网站管理员从互联网中收集各类图片、笑话、新闻、技术等各类信息，然后分类、编辑，发布到自己的网站系统中。网站管理员一般通过搜索引擎搜索各类关键字获取目标网址，然后再提取网页中的内容。关键字的组织决定获取内容的准确性和数量。由于内容来自不同的网站，所以提取内容的方法也各不相同。对于某一类的信息，发布到网站系统的数据结构是相同的。

网站管理员对站内进行搜索，将相关的内容在首页上进行编排和索引。

企业从互联网上搜索 email 和电话号码，并且能够查看该信息的相关信息，以便了解该对象的基本情况。企业希望能够搜索某一类别的客户信息，如这个客户属于女性，年龄为 20 到 30 岁等。并且能够将采集到的对象信息保存到企业内部的客户管理系统中。

企业需要了解某一产品的信息，希望得到该类产品的报价、厂商等，以及这些信息的对比情况。并且能够得到报价、厂商的进一步信息。这些信息希望能够保存到企业的内部的 ERP

系统或其它系统中。

1.4.3 数据结构化

企业办公产生的电子文档，客户提交的客户资料等这些数据，一般需要大量的人力手工输入到企业的 ERP 系统或信息系统中，企业希望能够通过软件从这些文档中抽取相关的数据自动导入到系统中。这些数据一般都有固定的模板格式，并且同一类文档的模板格式是相同的。如客户的家庭信息，客户 1 和客户 2 的模板格式是一样的，只是内容不一样。

2 用户特点

2.1 网站管理员

系统的最终用户群包含网站的管理员们，对一些给目标地址做了隐藏的网站，可能会使他们操作失败，特别有些网站对网页内容作了扰乱处理，使得他们更难准确地定义规则。对于这些问题，系统提供一些范例和匹配通配符，告诉他们的应对方法。并且官方网站提供交流论坛，共享和学习相互之间的经验。

2.2 信息采集用户

对于信息采集的用户，系统提供丰富的模板模式供用户使用，如 email 匹配模式、电话号码匹配模式等，只需要用户选择一个模板，就可以获得他们想要的信息。当然官方网站提供丰富模板资源供以下载。

2.3 数据结构化用户

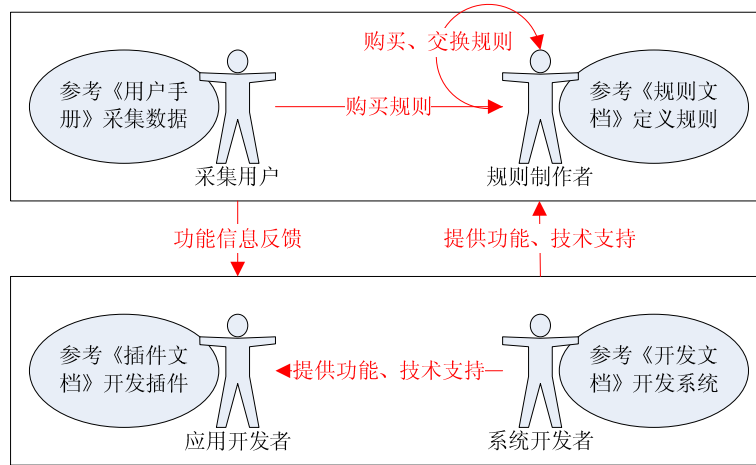
对于数据结构化的应用，会有第三方的技术人员提供支持。

系统预留了输入输出的编程接口，一些用户扩展这些接口，使系统应用到更多的场合下。针对这些用户，系统提供详细的接口说明，扩展示例代码。我们提供开发包，并描述每个类和每个方法的使用方法和功能。

还有一类用户属于商人的范畴，他们仅从事规则的制作，在网上交换或出售自己的规则。他们更关心网站的搜索和内容的质量，他们分两种类型，一种是猎人，他们能够发现各种各样的信息，能够满足各类网管的内容需求，他们从数量上获取大把金钱。当然，对于个别稀有的内容，价格就像黄金一样了。另一种属于黑客，他们精通 web 技术，机智并执着，在他们手里，都是一些很难发掘到的精品，当然价格都是高昂的。

由于采集系统属于开源软件，会有很多人去分析和使用采集系统中的组件，扩展和完善采集系统。他们使用的级别不仅是系统的界面上，而是深入到代码的内部上，他们需要参考采集系统的各类技术文档，所以系统除了用户手册，其它的开发文档也是必不可少的。

协作关系：



3 运行环境

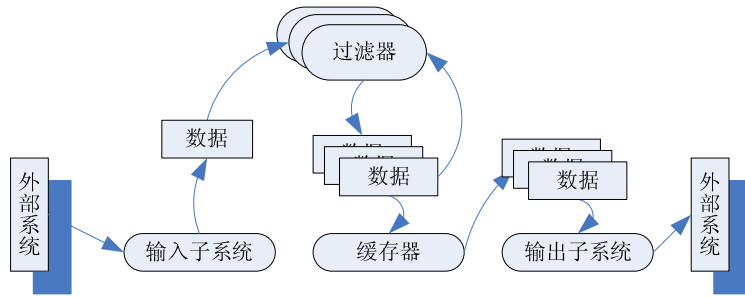
为了使采集系统适应多种运行环境，系统采用多种体系结构和多种语言版本。

采集系统分单机版和 web 版。Web 版又分多个不同语言的实现版，如 java 版、php 版、.net 版等。

软件结构	编程语言	操作系统	数据库	运行环境
单机版	vc	window	access	window
	java	window/unix	mysql	Jdk
web 版	java	window/unix	mysql/mssql/oracle	servlet 容器+jdk
	php	window/unix	mysql	php 容器
	.net	window	mssql	iis 服务器

4 运行体系

采集系统基本组件包含输入子系统，缓存器、输出子系统。数据通过多个过滤器多深度的提取下，被保存在缓存器中。示意图如下：



5 系统特性

5.1 I/O 体系

系统使用统一的输入输出接口对各类外部目标进行读取和发布数据。透明的支持现在和未来各类外部系统的交互要求。

5.2 容器体系

容器管理体系，使系统运行更加高效，并且提供更高的用户交互能力。特性如下：

1. 控制过滤器的并发数，适应不同的目标限制。
2. 过滤器的状态报告，时刻了解内容的采集过程。
3. 采用复用和调度策略，并发更加高效。

5.3 缓存体系

缓存区为输出子系统提供全局的数据索引，使输出子系统具备以下几种能力：

1. 可以在全局范围内对数据进行校验和再加工。
2. 可以跨层次地关联单元数据，发布采集的中间数据，

5.4 插件体系

采集系统支持丰富的插件类型，插件管理器负责加载插件和索引插件。插件分以下几种类型：输入插件、输出插件和过滤器插件，功能如下：

1. 输入插件支持不同的外部对象读取。如 http 服务器、ftp 服务器、文件系统等。
2. 采集插件支持不同的数据格式采集以及特殊的信息采集。如网页采集、word 采集、电子邮件地址采集等。
3. 输出插件支持各类系统的发布，如 bbs 系统、信息系统等。

6 功能说明

6.1 结构化采集

系统对半结构化数据进行语义分析，根据语义规则智能提取数据。

6.2 可视化元数据定义

用户在可视化的目标界面上指定所要采集的内容。

6.3 插件支持

系统拥有丰富的插件功能，支持各类目标的采集和各类系统的发布。如 ftp 采集，http 采集以及数据库发布，文件发布。

6.4 客户端环境模拟

模拟客户端环境，支持客户端和服务端的基本会话功能。如浏览器的 session 机制、cookie 机制。支持用户登录。

6.5 多线程采集

系统多任务并发，多线程采集。支持线程的并发控制和状态监视。

6.6 全局发布

系统提供上下文关联的全局缓存区，发布模块可以联合不同层次的单元数据。用户可以检查和编辑缓存区中的单元数据。

6.7 分页采集

根据页码规则，自动采集内容的下一页。

6.8 关联文件下载

系统可以根据设置自动下载页面包含的其它文件。如 flash、图片等。

6.9 规则保存

采集对象、过滤规则、发布目标等信息保存在规则文件里，用户可以导入导出规则文件，与其它人共享或交换规则文件。系统提供友好的向导页面供用户配置规则文件。

6.10 模板修饰

可以按预定义的模板结构发布数据。

6.11 结果过滤、替换

自动过滤数据的格式及语法，如过滤 html 语言，word 格式等。支持常量替换和环境变量替换。

6.12 重复过滤

自动删除采集结果中的重复数据。

7 支持信息

资源	说明
www.caijiqi.net	项目官方网站，发布项目文档，提供系统下载。
QQ:107175884	
Mail:hotheartboy@gmail.com	