

视采采集器

采集系统技术方案

www.caijiqi.net 敬上

日期	版本	作者	修改内容
2006-11-12	0.1	www.caijiqi.net	新版

目 录

- 1 概述
 - 1.1 目的
 - 1.2 需求概述
 - 1.3 系统需求
 - 1.3.1 多样化的采集目标
 - 1.3.2 多样化的数据格式
 - 1.3.3 分布式海量数据
 - 1.3.4 数据横向和纵向采集
 - 1.3.5 用户操作简单、快捷
 - 1.4 交互目标
 - 1.4.1 采集目标
 - 1.4.2 发布目标
- 2 系统设计
 - 2.1 运行体系
 - 2.2 系统结构
 - 2.2.1 过滤器容器
 - 2.2.2 缓存器
 - 2.2.3 插件管理器
 - 2.2.4 输入输出
 - 2.2.5 过滤器
 - 2.3 策略及设想
 - 2.3.1 体系策略
 - 2.3.2 采集策略
 - 2.4 模块结构
 - 2.5 过滤管道
 - 2.6 规则文件
 - 2.7 单元关联
 - 2.7.1 父子间的关联
 - 2.7.2 兄弟间的关联
 - 2.8 界面设计
 - 2.8.1 主界面结构
 - 2.8.2 采集界面结构
 - 2.8.3 规则定义界面结构
 - 2.8.4 信息发布设置界面
- 3 功能说明
 - 3.1 结构化采集
 - 3.2 可视化元数据定义
 - 3.3 插件支持
 - 3.4 客户端环境模拟
 - 3.5 多线程采集
 - 3.6 全局发布
 - 3.7 分页采集
 - 3.8 关联文件下载

- 3.9 规则保存
- 3.10 模板修饰
- 3.11 结果过滤、替换
- 3.12 重复过滤
- 4 支持信息

1 概述

1.1 目的

本文分析系统需求，说明系统结构和解决方案。

本文适合技术人员阅读和参考。

1.2 需求概述

网站、企业、营销人员都有对信息的需求，不同的信息领域，不同信息使用者，信息的获取方法和获取途径大不相同。采集系统需要满足多样化的采集应用，以及适应未来的需求增长。

1.3 系统需求

1.3.1 多样化的采集目标

信息分布在各种信息存储系统中，各种存储系统有着各自的交互机制，需要采集系统提供多种并可扩展的连接模块。

1.3.2 多样化的数据格式

信息以多种形式存在，如网页、word 文档、pdf 等。这些不同的格式数据需要采用不同的采集机制。

1.3.3 分布式海量数据

由于网络通信的延时和网络带宽的限制，并发多线程通信能够有效地减低延时和抢夺资源。

1.3.4 数据横向和纵向采集

需要系统自动采集数据的下一页；自动采集数据的关联附件；自动根据当前采集结果采集下一数据。

1.3.5 用户操作简单、快捷

多样且复杂的数据格式增加用户的作业难度，用户希望所见及所得，及时提供相应的操作提示信息。

1.4 交互目标

1.4.1 采集目标

采集目标为以下几种：

1. web 系统
2. 文件系统
3. 数据库系统
4. 其它文本数据源

1.4.2 发布目标

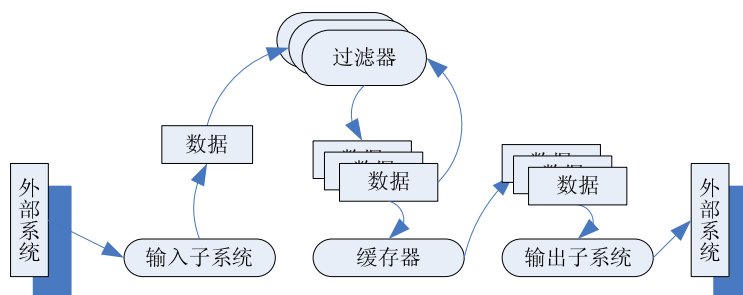
发布目标为以下几种：

1. 文件系统
2. 数据库系统
3. 其它文本数据存储系统或接收设备

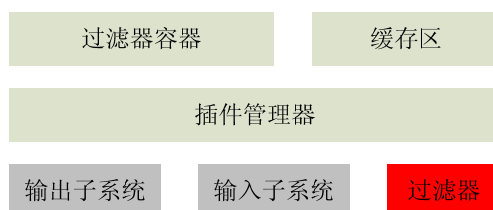
2 系统设计

2.1 运行体系

采集系统基本组件包含输入子系统，混存器、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：



2.2 系统结构

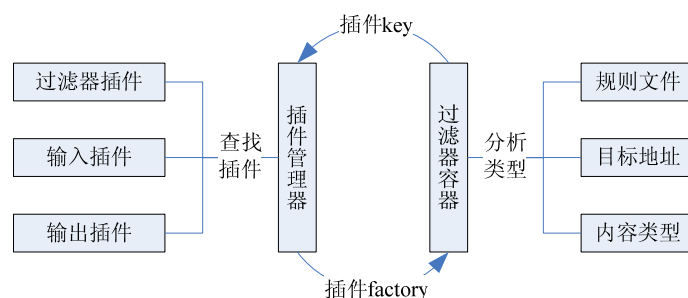


输出子系统、输入子系统、过滤器以插件的方式结合到系统中。过滤器容器通过插件管理器引用插件模块，驱动系统的执行。

2.2.1 过滤器容器

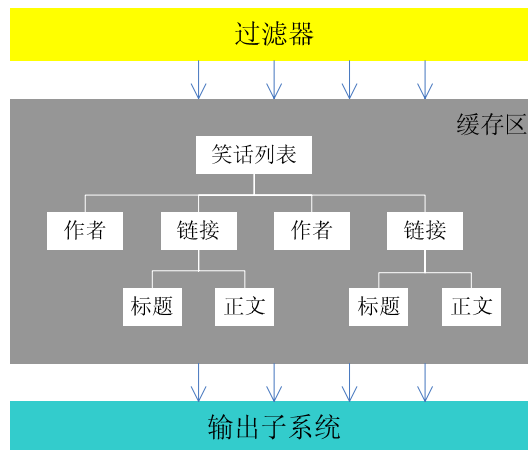
容器创建当前类型的过滤器实例并传递当前的输入输出句柄和全局缓存区句柄。容器控制过滤器的并发数。当所有的过滤器生命结束时，容器将触发输出子系统的执行。

容器通过规则文件和目标地址生成插件关键字，根据关键字查找插件管理器获得当前的过滤器插件和当前的输入输出插件的工厂句柄。



2.2.2 缓存器

过滤器将采集的数据送入缓存区。缓存区维持数据的采集顺序和上下文关系。输出子系统通过单元标识索引该单元及上下文单元。



2.2.3 插件管理器

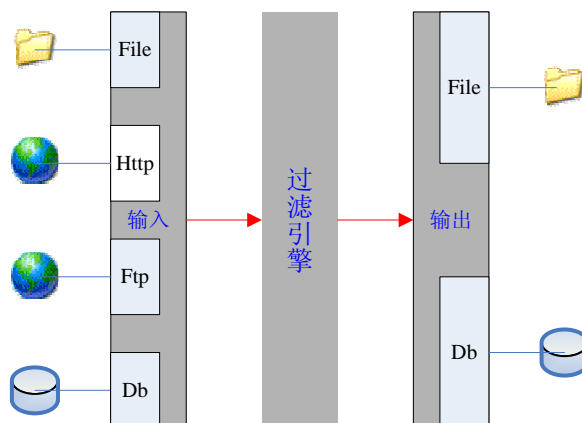
采集系统支持丰富的插件类型, 插件管理器负责加载插件和索引插件。插件分以下几种类型: 输入插件、输出插件和过滤器插件, 功能如下:

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。

插件管理器通过关键字来索引各类插件工厂。

2.2.4 输入输出

采集系统采用统一的输入输出接口与各类外部目标交换数据, 数据交换的过程由特定的模块实现。特定的模块是采集系统和外部目标之间的桥梁, 类似于 **window** 的设备驱动模块, 不同的输入输出机制对应不同的输入输出模块。**I/O** 体系负责管理和调度这些输入输出模块。输入输出模块包括标准的输入输出模块和扩展的输入输出模块。扩展的输入输出模块继承标准的输入输出模块根据外部目标的连接要求进行特定的会话处理。



标准的输入模块包含以下几种类型：

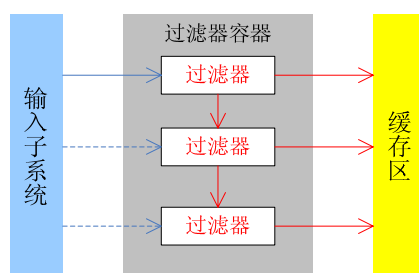
1. ftp 协议输入模块
支持 ftp 服务器的访问
2. http 协议输入模块
支持 web 服务器的访问
3. file 协议输入模块
支持文件的读取
4. jdbc 输入模块
支持数据库以 jdbc 接口的访问
5. odbc 输入模块
支持数据库以 odbc 接口的访问

标准的输入模块包含以下几种类型：

1. file 协议输入模块
支持文件的写入
2. jdbc 输入模块
支持数据库以 jdbc 接口的访问
3. odbc 输入模块
支持数据库以 odbc 接口的访问

2.2.5 过滤器

过滤器句柄由容器创建，并发执行的。过滤器的输出结果输入到下一个过滤器，同时结果也被存储在缓存区，供输出子系统全局引用。



2.3 策略及设想

2.3.1 体系策略

为了适应不同的采集目标和采集机制，采集系统采用插件体系和容器管理体系,用户可以通过安装插件包来支持特殊的应用。采集系统包含三类可扩展的插件模块：输入插件、采集插

件和输出插件。

三类插件在容器的驱动下相互协作。容器根据采集规则文件，创建入口过滤器，然后以多线程的方式启动过滤器，过滤器根据采集地址请求相应的输入模块读取数据，将过滤的结果保存在缓存区中，然后向容器请求它的下一个过滤器，如果返回值不为空，则以多线程的方式启动它们。容器收到过滤器的请求时，如果下一个过滤器为空，则调用输出模块，输出模块从缓存区在全局范围内读取数据，发布采集结果。

2.3.2 采集策略

采集系统对不同的采集目标采用不同的采集机制，对半结构化的数据进行语义分析，智能的抓取数据。对于网页采集，过滤器分析它的 html 标签，然后根据标签的类别和属性抓取指定的数据；对于 word 文档，过滤器分析 word 的文档格式及 word 对象，智能的抓取数据；对于采集特殊信息，如 email、手机号码等信息，过滤器通过模板模式来抓取信息。

2.4 模块结构

个模块如下：

采集规则文件
采集规则定义模块
采集规则解析模块
采集规则管理模块

采集容器管理模块
采集容器

缓存器管理模块
缓存器

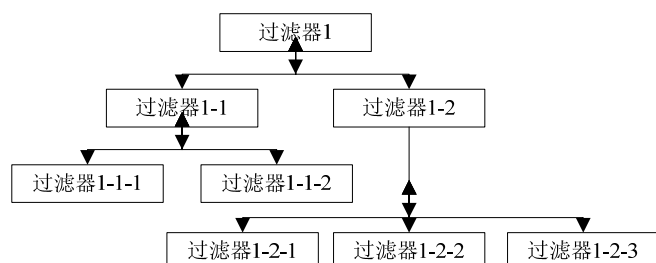
http 输入插件
文件系统输入输出插件
jdbc 输入输出插件
过滤器插件

输入插件管理模块
输出插件管理模块
过滤器插件管理模块
插件管理器

采集单元状态报告模块

2.5 过滤管道

过滤管道是信息流的通道，它是过滤引擎根据过滤器的输入关系形成的树状数据通道。示意图如下：



图中有五个管道：

1==1-1==1-1-1;

1==1-1==1-1-2;

1==1-2==1-2-1;

1==1-2==1-2-2;

1==1-2==1-2-3

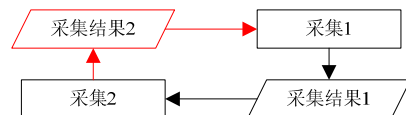
2.6 规则文件

采集规则按照约定的语法来定义。描述语言可以采用 XML，规则逻辑可以使用正则表达式或自定义的脚本语言，或者是这两者的结合。

采集系统对采集逻辑进行的合法性的检查，检查是否有输入输出环状，检查是否有输出的结果没有发布等情况。

案例：

1) 输入输出首尾互联



2) 输出的结果没有发布



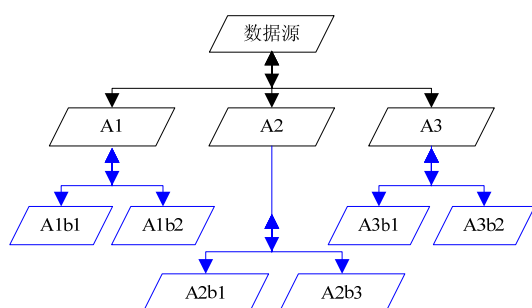
2.7 单元关联

系统在发布过程中，值关联有以下几种形式：

2.7.1 父子间的关联

过滤器 A 有多个匹配结果 a1、a2、a3，值 a1、a2、a3 又作为输入源进行第二次匹配 B，又产生多个匹配结果(a1b1、a1b2)，(a2b1、a2b2)，(a3b1、a3b2)。系统发布时要保证[a1、(a1b1、a1b2)], [a2、(a2b1、a2b2)], [a3、(a3b1、a3b2)]的关联性。

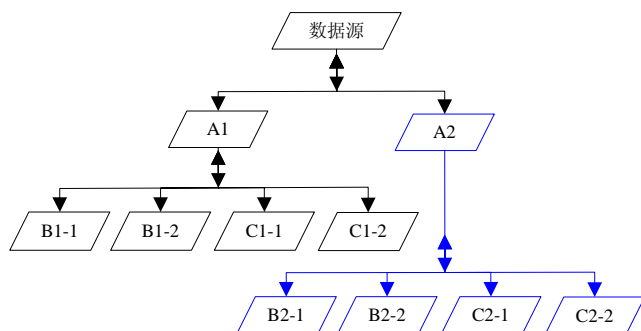
图：



2.7.2 兄弟间的关联

数据 a、b、c 是从数据源过滤的结果。字母的下标表示一个过滤器匹配的多个值。B 和 C 关联时只能是以下形式：[b1-1、c1-1]，[b1-2、c1-2]，[b2-1、c2-1]，[b2-2、c2-2]。我们称之为顺序关联。

图：



采集存储器在存储数据结构上必须能表示上述的数据关系，并且能通过过滤器的下标进行索引和向上查找。

2.8 界面设计

2.8.1 主界面结构

模块列表	
公共信息	
功能列表	工作区
	帮助信息提示
状态条	

系统采用二级导航栏的结构引导用户操作，实时地帮助信息提示，解决用户的疑难问题，让用户轻松、流畅地完成作业。

2.8.2 采集界面结构

规则文件窗口区
采集单元状态窗口区
过滤器状态窗口区

用户可以在规则文件窗口区选择以往的规则文件进行采集。采集单元状态窗口区及时地向用户报告已经采集的单元数据。过滤器状态窗口区向用户报告已经执行结束或正在执行的过滤器状态信息，使用户时刻了解系统的执行状态。

采集界面在主界面的工作区中展示。

2.8.3 规则定义界面结构

规则属性窗口区
采集对象展示区
单元定义区
采集对象源码区
功能按钮区

1. 规则属性窗口区：设置规则的基本属性
2. 采集对象展示区：以可视化方式展示采集对象，用户直接在可视化对象上选取采集对象
3. 单元定义区：在采集对象上指定采集的数据
4. 采集兑现源码区：显示采集对象的源代码

2.8.4 信息发布设置界面

发布目标属性设置区		
目标大单元列表区	目标小单元列表区	采集单元列表区
		变量设置区
	关联按钮区	
	单元关联展示区	
操作按钮区		

用户连接发布目标，在目标单元列表区中分级展示单元列表，指定目标单元和采集单元的对应关系。单元关联展示区展示当前已经关联的单元列表。

3 功能说明

3.1 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。

3.2 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。

3.3 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。

3.4 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。

3.5 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。

3.6 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。

3.7 分页采集

根据页码规则，自动采集内容的下一页。

3.8 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。

3.9 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。

3.10 模板修饰

可以按预定义的模板结构发布数据。

3.11 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。

3.12 重复过滤

自动删除采集结果中的重复数据。

4 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	