TBSCIS 网络资源采编软件

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 系统结构

TBSCIS 网络资源采编软件整体的通讯与数据交换采用 C/S 结构来构建。C/S 结构可以充分利用两端硬件环境的优势，将任务合理分配到 Client 端和 Server 端来实现，降低了系统的通讯开销。C/S 结构的基本原则是将计算机应用任务分解成多个子任务，由多台计算机分工完成，即采用“功能分布”原则。

TBSCIS 网络资源采编软件主要构成：服务器端软件、管理端软件、操作端软件。

1.1 系统架构

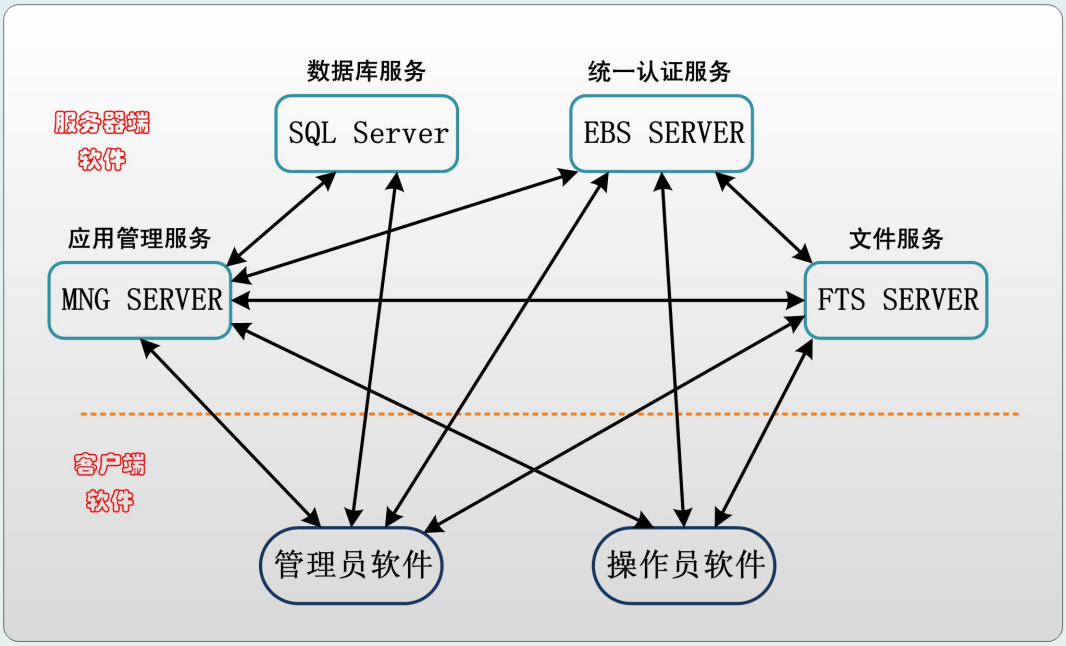


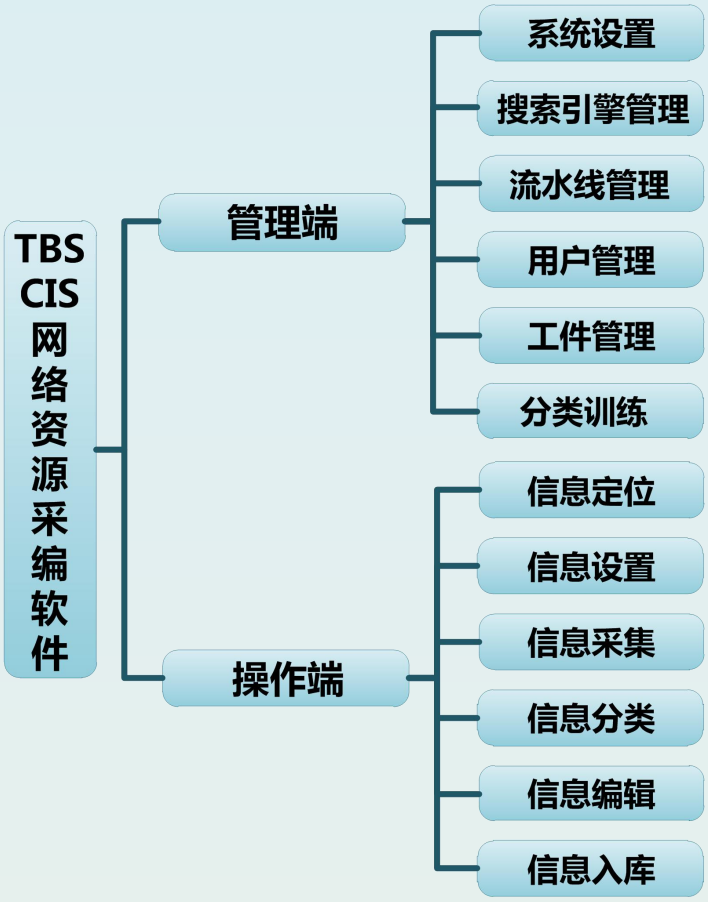
图 2.1 系统架构图

1.2 功能架构

在 TBSCIS 网络资源采编软件的 C/S 架构的客户端中，根据客户端功能以及操作对象的不同，划分成管理端和操作端两个模块。

管理端软件是对整个网络采编软件的各项功能进行设置和管理，通过设置后操作端才可以进行正常的采编处理操作。

操作端软件是对网络资源采编过程进行管理，通过操作端软件的各个工序的实现采编处理操作。



1.3 系统软硬件部署

TBSCIS 网络资源采编软件主要包括五个部分： EBS 统一认证服务器、FTS 文件服务器、MNG 调度服务器、MSSQL 数据库服务器、TBSCIS 管理软件、TBSCIS 客户软件。

EBS 统一认证服务器、FTS 文件服务器、MNG 应用管理服务器、MS-SQL 数据库服务器可以根据实际情况部署在一台服务器上，也可以部署在不同的机器上，由服务端自动平衡负载，系统也可以挂接多个 FTS 文件服务软件。另外，TBSCIS 管理软件、TBSCIS 客户软件可以安装在同一台个人电脑上。

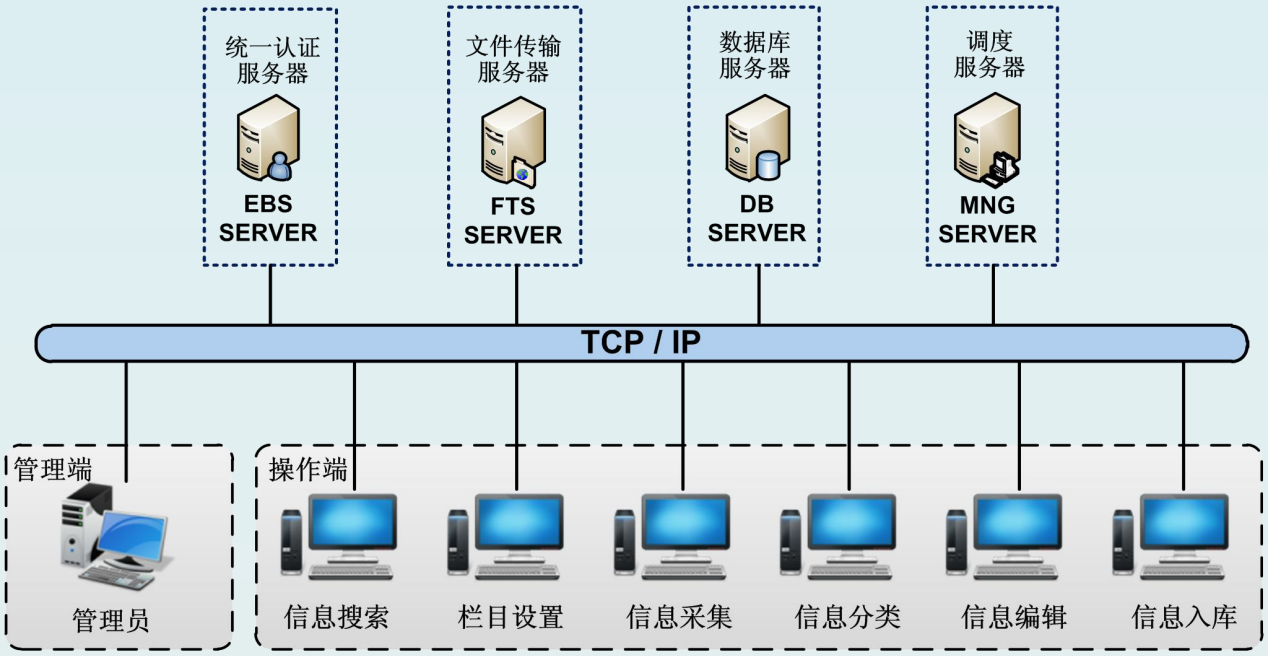


图 2.3 系统网络部署

1.4 技术参数

采编软件中操作端负责信息搜索、信息设置、信息采集、信息分类聚类、信息入库等操作，对于大量的网站的监控和采集就需要多个操作端来协同操作，并发地获取数据，这一方面就对客户端数量提出了较高的要求。

当客户的数据采集完毕，都要上传至服务器里作为后续工序的“加工材料”，这样大量的数据会被不定时的汇总到服务端，因此需要服务要有大量的存储空间和快速的响应能力。

对于网页采集的过程，在 2M 宽带情况下，抓取的速度以 200kb/s 进行，单个网页平均长度 20kb，那么采集一个网页则需要 0.1s 的时间；网页处理的过程在进行关键字查询，链接，时间提取，大约需占用 0.5s。处理一个网页的总时间为 0.6s。则一个小时约可以抓取并分析 6000 多个页面，相当于 2 个网站的页面（只是针对企业网站，媒体类的可能有近百万的页面需要处理）。按 10 个左右并发数，那么一台机器一天可以采集和更新 400 个网站。

第二章 功能及特点

TBSCIS 网络资源采编软件主要包含服务管理端软件和操作端软件两部分。管理端软件负责流水线的建立、工序的设置、操作员的分配、基本信息设置等等。操作端软件负责信息搜索、信息设置、信息采集、信息分类聚类、信息入库等操作，其中信息采集、信息分类、信息入库操作都可以长时间自动运行。

2.1 系统功能

2.1.1 管理端软件

■ **系统设置**

设置数据库服务器、FTS 服务器、MNG 服务器的 IP 地址及端口信息，使得软件中各个服务程序能够有机的连接起来，使系统能够正常的运转。

该模块中涉及功能属于 TBSCIS 软件的初始化操作，软件部署完成后需首先通过系统配置功能对软件

进行初始化处理。

- **系统设置**

用于设置软件运行时所依赖的 SQL-SERVER 数据库的信息，同时创建软件数据库文件。

- **数据库设置**

用于在软件数据库中初始化相关操作工序，以及对软件数据库进行备份和恢复操作。

- **文件服务器设置**

设置加工过程中产生的工件在服务器上的存放路径。

■ **搜索引擎管理**

搜索引擎管理模块主要对操作端信息定位模块中“按关键字搜索”功能调用的搜索引擎进行管理。

- **增加新引擎**

添加一个新的搜索引擎，操作端信息定位模块中会相应增加一个新的搜索引擎供调用。

- **更新引擎内容**

对已有的搜索引擎的配置信息进行更新。

- **更改引擎名**

对已有的搜索引擎的名称进行更改。

- **删除引擎**

删除软件中已有的搜索引擎。

■ **流水线管理**

在网络资源采编软件中，所有对网络资源的采编操作处理都是通过操作软件的流水线进行处理。通过流水线管理，用户可以根据不同的采集需求对多条流水线进行管理。

- **新建流水线**

新建网络资源采编流水线，并对流水线的相关信息进行设置。

- **编辑流水线**

对软件中已有的流水线相关信息进行编辑和修改。

- **删除流水线**

对软件中已有的流水线进行删除操作。

- **配置工序**

软件中的采编流水线由工序组成，通过配置工序功能可对软件中管理的流水线根据不同的采编需求来配置不同的工序。

- **配置员工**

通过配置员工功能可以将不同的员工配置到不同流水线的不同工序中。

同一流水线的同一工序中可以包含多个员工。

同一员工可以隶属于不同流水线的不同工序中。

■ **用户管理**

用户即流水线管理中提到的操作员，通过用户管理可以用用户组的管理方式对用户进行同一管理和流水线工序分配。

- **添加新组**

新建一个用户组，可以通过用户组对具有相同属性的用户进行管理。

- **编辑组名**

对软件中已有的用户组名称进行编辑。

- **删除组**

对软件中已有的用户组进行删除。

当用户组下包含用户时，该用户组不能删除。

➢ **添加新用户**

添加一个新的用户，同时设置该用户的账号、密码及所属组别。

➢ **修改用户**

对软件中已有用户的中文名称和所属组别 ID 进行修改。

➢ **删除用户**

对软件中已有的指定用户进行删除。

➢ **添加流水线**

将指定用户添加到软件管理的现有流水线中。

➢ **添加工序**

将指定用户添加到指定流水线的指定工序中。

同一用户可以添加到流水线的不同工序中。

■ **工件管理**

显示采集流水线中各工序中的工件状态，便于管理员掌握各流水线加工进度。

➢ **编辑工件**

对工件的所在工序、工件运行状态、操作状态进行修改。

➢ **删除工件**

将指定工件从流水线中删除。

➢ **批量处理**

同时对多个工件批量设置所在工序、工件操作状态及运行状态。

■ **分类训练**

对操作软件“信息分类”模块的信息分类功能所调用的分类模型进行管理。

➢ **训练设置**

通过训练设置可以新建分类训练模型，同时对分类训练模型的相关信息进行设置。

➢ **修改模型**

对软件中现有的分类训练模型进行修改。

➢ **分类训练**

利用文本数据对指定模型进行训练，训练文本的内容越贴近模型，训练次数越多，模型在进行分类处理的时的准确率就越高。

➢ **添加新类**

用户可以为指定模型添加分类。

➢ **编辑类名**

可对软件中管理的指定分类名称进行编辑。

➢ **删除类**

对软件中指定的分类进行删除。

➢ **添加训练样本**

对模型下面的指定分类添加训练样本用于分类训练。

➢ **利用已有类**

利用现有的类中的分类训练信息对其他分类进行训练。

➢ **保存训练文本到类**

将指定的训练文本添加到指定类中。

2.1.2 客户端软件

■ **信息定位**

- 允许用户选择输入某个目标资源进行检索。
- 支持简单检索和复合检索两种检索方式。
- 支持自动检查链接的有效性。
- 支持不同目标资源的特定检索条件。
- 提供多种检索结果输出格式。
- 利用 Cache 技术保存检索结果，提高检索访问速度，有效利用网络资源。
- 支持并发检索，对于并发的同一个检索条件共用检索线程资源，最大程度地利用网络和系统资源。

信息设置

- 支持按栏目进行设置
- 支持按关键字进行设置
- 允许设置采集时间间隔
- 允许用户人工更改采集状态
- 允许用户对采集过程的各种情况进行设置

信息采集

- 支持各种标准格式信息资源的采编，如 HTML 页面、文本信息、表格、图片、声音、视频等。
- 实现对网页与内联图片的统一采集。
- 支持繁体页面（BIG5 码）的采编，并自动转换为标准的简体码（GB 码），支持 Unicode 码集。
- 支持由程序自动生成的页面内容的采集，如由 JavaScript 生成的页面。
- 能方便抓取由数据库自动生成或者需要身份验证的网站内容。
- 支持单篇网页及网站历史数据的批量下载。
- 高效的采编技术和更新策略，采用多线程并发搜索技术，采集过程高效准确，且提供高效的更新手段，已经采编过的信息不会重复采集，更新时只获取前次采集后更新的网页。
- 整合新版的信息采集，增加了对采集页面的控制，如框架、脚本、路径、视频、控件下载等的控制
- 对采集配置的记忆功能
- 可远程上传采集的内容

信息分类

- 高效的垃圾信息过滤。软件可对网页进行内容分析和过滤，自动去除广告、版权、栏目等无用信息，精确获取目标内容主体。
- 智能化信息自动分类技术。采用 TBS 基于内容的自动分类技术，可对采集的网页进行基于内容的自动分类，不需人工干预。自动分类的准确率基本可以满足信息粗加工或大多数应用的实用要求。同时系统提供分类训练工具，允许用户自行根据自己的分类需求和数据特点设定分类结构和生成特征模板，适应不同行业的需求。
- 聚类技术是按照某种相似度值将一个集合划分成若干个子集，使得子集内部的元素之间有较大的相似度，而子集之间的元素的相似度值较小。对于文本聚类，就是将一个文档集合按照相似性划分成若干个类，使得每个类中的文档相似，而不同类中的元素之间不相似。聚类过程也包含三个阶段：文本预处理、文本特征提取、聚类。前两个过程和分类是相同的，在生成待聚类文本的特征向量集合之后，就可以用聚类算法对文本集进行聚类了。由于聚类依赖于文本之间的相似度，因此，计算文本之间的相似度是聚类的基础。
- 基于内容相似度计算的自动去重。不是利用简单的规则判断，而是利用内容的相似性进行排重判断，准确性高。
- 利用文摘框架通过文本分析技术自动提取文摘。

信息编辑

➢ 可以方便地查看已采集的各种信息，可对其进行修改和删除等操作，达到用户对采集到信息的直接控制。

信息入库

➢ 可以根据用户需要，经采集到的信息，自动或人工的导入 SQLSever 数据库（关系数据库）或 TBS 数据库（用于发布及检索）、或者将其另存备份。

2.2 系统特点

2.2.1 加工流程方案完全由用户定制

由于采编的信息资源的来源广泛、特点各异以及不同用户的需求不一致，导致了信息资源加工过程的复杂性。但是整个信息资源加工大致可以分为信息搜索、信息定位、信息采集、信息去重、信息分类聚类、信息编辑审核、信息入库、信息发布等工序。通过完善的接口设计和流程分析，系统提供用户任意确定工艺流程操作个数和顺序，实现单机信息资源加工和机群间高效率的协同作业。

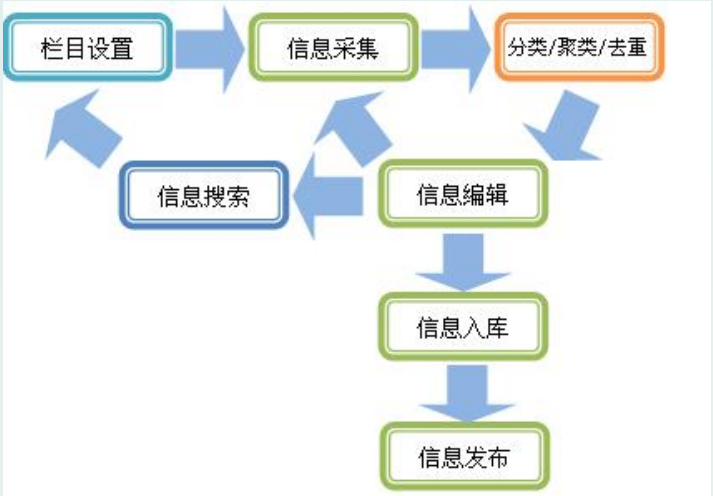


图 3.1 信息加工流程图

2.2.2 高效的信息定位功能

目前 Internet 上已经存在多个功能强大的搜索引擎，超级检索引擎已经随时将网上产生的新闻等建立了索引，可以直接为我们所用。无须再使用自己的 ROBOT 到网络上抓取。TBS 元搜索引擎直接利用这些超级搜索引擎，可以从多个搜索引擎上获取检索信息，对结果进行合并去重处理，然后将结果返回给用户，方便用户进行资源的查找定位。

2.2.3 先进的信息采编技术

系统给用户提供了功能强大的可视化的采集规则的配置界面，极大的提高了用户进行规则的配置的效率、降低了对操作人员的计算机知识的要求。

适应网站内容格式的多变性，能完整地获取需要采集的页面，遗漏少。

能方便地将网页中的信息提取出来，如日期，标题，作者，栏目等内容；过滤网页中的无用信息。

系统通过多线程处理技术，可以同时启动多个搜索器，快速高效地对目标站点或栏目进行信息采集。

2.2.4 开放性好，和其他信息服务系统有机集成

被采集到的信息可以根据用户系统环境需要，存储到 TBS 全文数据库、SQL Server、Oracle 等关系数据库中，使得其他信息服务系统可以方便地利用，从而系统和其他系统的有机集成。

2.2.5 智能更新策略

提供高效的更新手段，已经采编过的信息不会重复采集。

第三章 工作流程

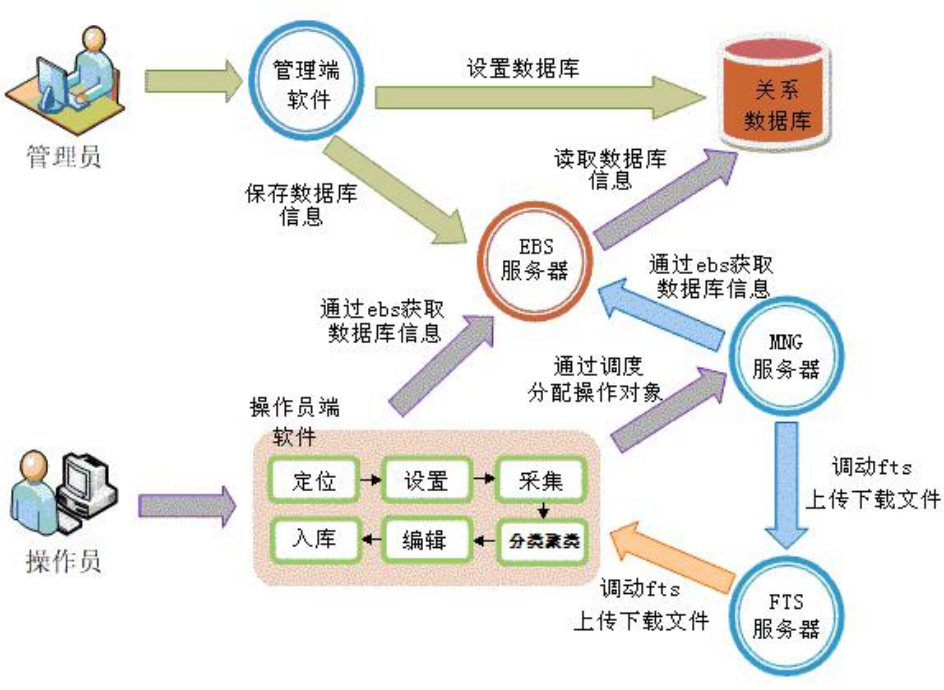


图 4.1 工作流程图

第四章 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