**Cyberin媒体库**

需求说明书

当前版本号：Cyberin媒体库V1.0

需求确认时间：2019年5月9日

**脉讯在线（北京）信息技术有限公司**

目录

[1．系统概述 2](#_Toc7957085)

[1.1项目背景 2](#_Toc7957086)

[1.2 目的 3](#_Toc7957087)

[2．功能需求描述 3](#_Toc7957088)

[2.1功能需求 3](#_Toc7957089)

1．系统概述

## 1.1项目背景

互联网发展不断加快，数据更新速度也在不断提速，客户对数据的时效性和全面性要求也更高。之前数据抓取的频率按照一天四次更新，更新频率过慢，导致客户对数据的时效性不满，为提升数据抓取效率，达到数据实时抓取，提升客户满意度，因此就有了搭建cyberin媒体库的想法，cbyerin媒体库的搭建是对爬虫业务的一次提升，在抓取速度方面更上一个台阶。

## 1.2 目的

Cyberin媒体库搭建是为了提升数据抓取效率和数据的全面性，保证定制过的网站，数据都可以覆盖到。除提升抓取速度外，还增加了网站地域、类型、来源的整理，保证定制过的网站获取更全面的信息。

# 2．功能需求描述

## 2.1功能需求

|  |  |  |  |
| --- | --- | --- | --- |
|  | | | |
|  | | | |
| **概况** | **脚注** | **名称** | **需求说明** |
|  | A | 查找条件 | 按条件搜索需要的数据 |
|  |  | 抓取状态 | 抓取状态为三种，可按 “全部”、“完成”、“失败”分别进行筛选，默认显示“全部” |
|  |  | 抓取等级 | 抓取等级分三种，可按“全部”、“高”、“中”、“低”分别进行筛选，默认显示“全部” |
|  |  | 作者/互动/原创转载是否处理 | “作者/互动/原创转载是否处理”状态为四种，可按 “全部”、“未处理”、“已处理”、“无”分别进行筛选，默认显示“全部”，“无”表示对应的字段为空 |
|  |  | 是否应用危机APP | 是否应用危机APP筛选分三种，分别为“全部”，“是”，“否”，默认为“全部” |
|  |  | 是否应用搜搜 | 是否应用搜搜筛选分三种，分别为“全部”，“是”，“否”，默认为“全部” |
|  |  | 多条件查找 | 分别按“链接”、“网站”、“二级版面”、“三级版面”、“迅迅APP别称”、搜搜APP别称”、“网站类型”、“地域”选项查找，一个选项可以对应多个条件查找。如果多条件查找，一行输入一个条件，如果网站类型一组含多个关键词，词之间用英文逗号隔开 |
|  |  | 搜索 | 点击“搜索”按钮，分别按照①-⑥条件搜索，并在C处显示结果 |
|  | | | |
|  | B | 批量上传/修改 | 批量添加/修改链接及对应的相关信息，也可以通过“下载模板”查看需要上传的文件格式。上传的文件为csv格式，上传和下载模板字段见附件。点击“选择文件”，选择需要上传的文件，点击“上传”，①若上传的文件均为新增，上传成功后提示“成功上传\*\*\*条，新增\*\*\*条”；②若上传的文件均为修改后数据，上传成功后提示“成功上传\*\*\*条，修改\*\*\*条”；③若上传的文件有新增也有修改，上传成功后提示“成功上传\*\*\*条，新增\*\*\*条，修改\*\*\*条” |
|  | | | |
|  | C | 内容展示 | 对媒体库详细数据的展示 |
|  |  | 删除与导出 | 对勾选的数据删除或者是导出  当有勾选指定数据时，点击“删除”，删除勾选的数据，并提示“确认将\*\*条数据删除？”点“是”则直接删除，点“取消”，则取消删除；点击“导出”，导出的内容是勾选的数据  当不勾选数据，点击“删除”默认不删除任何数据，并提示“请选择需要删除提示的数据”；点击“导出”，默认导出媒体库所有数据  如果有条件查找时，点击“导出”，导出的内容为按条件查找的结果 |
|  |  | ID | 链接对应的id，每个链接对应一个id，id为唯一值，不可重复 |
|  |  | 链接 | 显示添加的链接 |
|  |  | 网站 | 显示链接对应的网站名 |
|  |  | 二级版面 | 显示链接对应网站的二级版面名 |
|  |  | 三级版面 | 显示链接对应网站的三级版面名。最多显示三级版面，如果四级版面，默认都只显示到三级。  网站、二级版面、三级版面需与cyberin平台每条数据的来源对齐，显示格式为“网站-二级版面-三级版面”，若没有三级版面，则显示为“网站-二级版面” |
|  |  | 迅迅别称 | 网站的另一种名称，供迅迅产品线使用，cyberin来源默认取网站、二级版面、三级版面组合 |
|  |  | 搜搜别称 | 网站的另一种名称，供搜搜产品线使用，cyberin来源默认取网站、二级版面、三级版面组合 |
|  |  | 网站类型 | 对网站打标签，一个网站可以对应多个标签，标签之前用英文逗号隔开 |
|  |  | 地域 | 显示链接对应网站所属地域， 如果网站为行业网站或没有明确的地域显示，默认为“全国” |
|  |  | 抓取等级 | 显示对应链接的抓取等级。抓取等级分为高、中、低三档。高：每10分钟抓取一次；中：每半小时抓取一次；低：每一小时抓取一次。前期所以链接都按高优先级抓取 |
|  |  | 昨日抓取量 | 显示对应链接昨天的日抓取量，每天6点更新前一天的抓取量 |
|  |  | 是否有作者/互动/原创转载 | 整理链接时，整理人员会标注对应链接是否有作者，互动，原创，转载信息，如果显示有作者或有互动或有原创转载，对应的“作者/互动/原创转载是否处理”默认显示为“未处理”；若“是否有作者/互动/原创转载”显示“无”，对应“作者/互动/转载是否处理”显示为空 |
|  |  | 添加人 | 显示添加该条信息的人员姓名 |
|  |  | 添加时间 | 显示对应链接添加的时间。上传数据时，默认取上传的时间为添加时间 |
|  |  | 修改时间 | 对应链接的数据有修改内容，显示修改后时间，如果没有做修改，默认为空 |
|  |  | 最新抓取时间 | 显示最近一次抓取完的时间 |
|  |  | 抓取状态 | 指最后一次数据抓取状态，状态分为两种：完成、失败。如果正常抓取完成显示“完成”；如果没有抓取完成，显示“失败” |
|  |  | 作者/互动/转载是否处理 | 状态分为三种：未处理、已处理、空。“是否有作者/互动/原创转载”显示有作者或者是有互动或原创转载时，“作者/互动/原创转载是否处理”显示“未处理”；当技术人员对有作者/互动/原创转载做了处理后，点击“未处理”后可切换为“已处理”状态，“未处理”与“已处理”状态可互相切换； “是否有作者/互动/原创转载”显示无时，作者/互动/转载是否处理”显示为空。“未处理”和“已处理”状态的修改，不影响“是否有作者/互动/原创转载”内容显示 |
|  |  | 是否应用迅迅 | 对应列表的抓取数据如果需要上到迅迅产品线，点击勾选，有勾选，会直接上到迅迅产品线，默认为不勾选 |
|  |  | 是否应用搜搜 | 对应列表的抓取数据如果需要上到搜搜产品线，点击勾选，有勾选，会直接上到搜搜产品线，默认为不勾选 |
|  |  | 保存、删除 | 每个字段都是可编辑的状态，点击“保存”，对修改后的内容直接保存，点击“删除”将删除对应记录的数据 |  |  | 翻页 | 默认按添加时间排序，一页显示100条，可翻页到任何一页查看添加的列表信息 |
|  |  | 备注 | 网站抓取活覆盖有单独说明，在此备注 |

## 2.2整理标准

1、媒体列表包含的网站必须有版块；

2、整理的网站需将网站对应各版块内容都整理出来，包括二级、三级的版块信息；

3、综合门户类网站需将地域各版块信息整理出来；

4、网站如果包含博客、论坛、问答等版块信息，也需要整理在内；

5、整理列表需考虑更新频率，若更新频率过高，需加上第二页列表；

6、电子报、移动端、自媒体属单独定制，不在媒体库列表中。

7、网站类型包括：央媒，综合，地方，财经，数码科技，时尚，汽车，交通（除汽车）母婴，教育，旅游，娱乐，生活服务，投诉，游戏，企业,房产，社交，电商，购物，健康等。