开放获取资源元数据采集系统需求分析报告

1. 引言
   1. 编写目的

明确开放获取资源元数据采集系统的开发目的、功能需求，进一步指导开发设计。

* 1. 背景

毕业论文。

* 1. 名词解释

开放获取期刊：“开放获取”，即可以在公共网络上被免费获取，允许任何用户对该文献的全文信息进行阅读、下载、复制、分发、打印、检索、超链接，支持爬行器收割并建立本地索引、用作软件的输入数据、用于其他任何法律允许的用途。

* 1. 参考资料

1. 任务概述

由于开放获取期刊网页结果多变，抽取粒度较细，具体特点如下：

开放获取资源的描述粒度细。《NSTL开放资源描述元数据标准规范》中对开放获取元数据的描述定义了7类元素集：开放资源集合、开放资源作品、责任者、机构、开放资源使用权益、获取来源、管理信息，而这7类元素集又包含了众多元数据字段，如题名、其他题名、关键词、英文关键词、摘要、英文摘要、作者、机构、唯一标识号、总页数、起止页、参考文献数量、年、卷、期、期刊名称、全文下载链接等。而一般的网络信息的描述，如新闻，只需要采集标题、正文、作者、时间等少量的几个元数据字段便可满足需求。

开放获取资源元数据描述复杂。开放获取资源元数据描述复杂主要表现在两个方面：一是网页中元数据展示多样，即不同开放获取资源的平台网站和期刊网页所揭示的元数据薄厚不均，二是描述元数据的网页结构多变，如在一个期刊网页中，存在网站建立初期所采取的网页模板与现在不一致的情况。而一般的网络信息，如论坛信息，某一个论坛网站在展示信息时采用的是同一个网页模板，不存在不同时期网页结构不一致的情况。因此开放获取资源元数据描述更加复杂。

基于以上特点，使得当前主流的信息抽取工具无法满足抽取需求，主要表现为，无法检测网页结构的变化，使得既定的抽取规则无法覆盖到所有的记录，造成记录的丢失。因此设计开发一套专门适用于开放获取期刊采集的系统，增加网页结构变化检测，优化抽取时的交互设计，提高开放获取期刊抽取准确度和易使用度。

1. 数据描述
   1. 静态数据（系统运行前已有数据）
   2. 动态数据（系统输入数据和运行输出数据）
2. 功能需求
   1. 业务流程图



* 1. 数据流图



* 1. 功能划分

开放获取资源元数据采集系统提供浏览器给用户进行交互。系统主要分为对首页、中间页、详情页的处理，数据库中主要有抽取规则表和元数据表。

首页处理：系统提供一个输入框，供用户输入需要采集网站的URL。输入之后转入到对应的网站。

详情页处理：详情页指的是所需采集的元数据所在的页面，是采集过程中的最后一层。用户在该页面上选择所需采集到元数据字段，并标注该字段的名称。系统根据用户的选择形成抽取规则，每个字段对应一条抽取规则，使用Xpath表示。在调用爬虫程序抽取数据之前，会查询抽取规则库，如果当前规则已经存在，说明之前已经抽取过，则直接将元数据库中的数据返回给用户。如果当前规则不存在，则检查当前抽取规则是否能适用于其他详情页，即检查网页结构是否发生变化，如果变化，则返回给用户结构变化的网页，让用户重新选择抽取字段，并将新的抽取规则添加到抽取规则表中。

中间页处理：中间页指的是所需采集信息详情页的上一层页面，需要点击此类页面中的多个链接，进入到详情页中。该类页面的处理是根据Xpath自动识别出当前页面的同类链接，并存入待访问列表中，以此进行访问。

爬虫程序：根据抽取规则对对应的信息进行抽取，并保存。

* 1. 功能描述
  2. 数据与功能对应关系（矩阵图，功能模块的输入数据和输出数据）

1. 性能需求
   1. 时间要求
   2. 准确性要求
2. 运行环境描述
3. 其他需求