浙江大学

Techhub 后端知识库搜索引擎系统

系统设计计划

Group 08

黄亦非、傅诤哲、胡瀚丹、陈鑫

2019/07/09

修订表：

编号	生成版本	修订人	修订章节与内容	修订日期
1	1.0	G08 小组所有成员	所有章节	2019/07/09
2	2.0	G08 小组所有成员	系统验收标准	2019/07/09
3				
4				
5				

审批记录

版本	审批人	审批意见	审批日期
1.0	黄亦非	通过	2019/07/09
2.0	黄亦非	通过	2019/07/09

目录

1 引言

1.1 编写目的

为了让项目得以按计划进行，避免成员间出现协作不畅，本计划书的编写是以作奠基之意。预期目标是保质保量地完成甲方的需求，严格按照项目工程的要求，达到一个卓有成效和执行力的团队所应有的水平。

本文档的预期读者为客户、项目指导老师、项目组的开发人员及其他项目利益共同者。

1.2 背景

针对现在互联网高速发展，但是网上信息获取鱼龙混杂，尤其是在学习新技术的领域当中，各种无用的信息只会浪费我们的时间，消磨学习新技术的热情，所以我们小组计划开发一款后端知识库的垂直搜索引擎。

Techhub 后端知识库搜索引擎系统主要是面向在校大学生用户，致力于在互联网上学习后端的相关技术，能够帮助用户快速的搜索相关领域的技术文档，包括该技术的描述、使用手册、教学视频、使用当中的相关问题等方面。对于用户而言，可以根据输入的问题或者关键字，准确的定位用户的意图，并且返回最准确的结果给用户。同时用户也可以根据提供的几个过滤条件对返回的结果进行对

应的过滤，实现更加精确的结果定位。

同时系统对于数据的来源和处理也非常的重视，致力于提供最准确、最完整、最精确的知识库架构。对于数据我们会进行相关的过滤、去重、结构化信息的提取等等数据处理的动作，保证系统数据的稳定和准确。

1.3 定义

UML：统一建模语言。UML 是一套用来设计软件蓝图的标准建模语言，是一种从软件分析、设计到编写程序规范的标准化建模语言。

Solr：一种搜索引擎框架

React：Web 开发框架

SpringBoot：后端微服务框架

1.4 参考资料

- 《项目描述》（课程资料）

  提供者：课程教学小组

- 《Software Requirements》（课本）

  作者：Karl E. Wiegers (美）

  译者：刘伟琴 刘洪涛

  出版社：清华大学出版社

- 《UML 用户指南》

  作者：Grady Booch 等

  出版社：人民邮电出版社

- 《计算机软件产品开发文件编制指南 GB8567-88》

  制定年份：2006 年

1.5 标准、条件和约定

本项目遵从以下标准：

GB/T 13702-1992 计算机软件分类与代码

GB/T 20918-2007 信息技术

GB/T 19003-2008 软件工程

GB/T 5538-1995 软件工程标准分类法

GB/T 9386-2008 计算机富安居测试文档编制

GB/T 9385-2008 计算机软件需求规格说明

GB/T 5532-2008 计算机软件测试规范

GB/T 18221-2000 信息技术程序设计语言

GB/T 11457-2006 信息技术 软件工程

GB/T 8567-2006 计算机软件文档编制规范

2 总体设计

2.1 功能设计

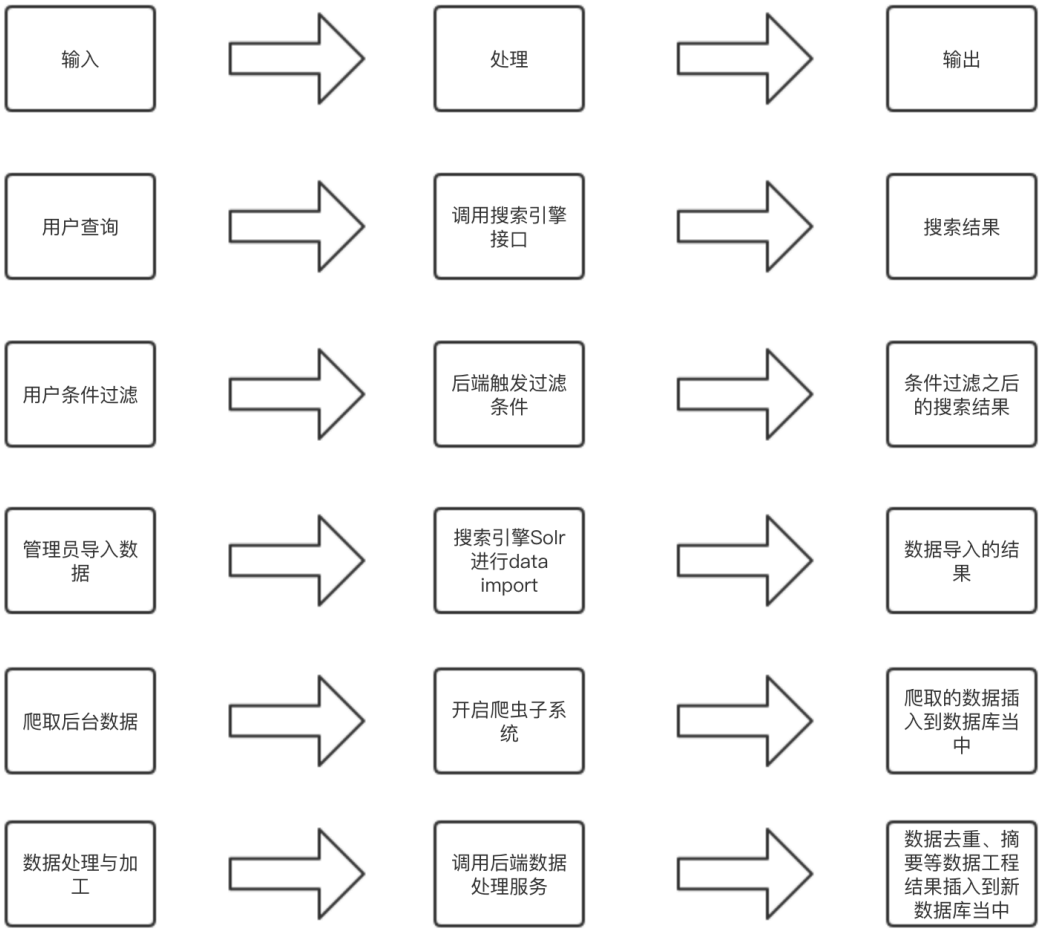
功能模块	功能
前端子系统	查询搜索内容
	筛选过滤条件
搜索服务器子系统	Data Import
	Search 接口
后端 Web 子系统	前端搜索接口
爬虫子系统	信息爬取
	自动滤重
	自动分类
	自动爬取

2.2 用户类型及用户特征

用户类	特征与说明
网站用户	1. 主要用户 2. 同时使用该网站的用户数目可能较多，主要服务的对象为想要找到特定领域的专业技术文档或者是相关问题解答。 3. 要求能够返回足够精确和足够深度的结果 4. 能够根据用户的喜好进行返回数据的过滤 5. 支持根据用户的输入进入搜索引擎搜索，并且返回相关的内容到网站首页。
系统管理员	1. 次要用户 2. 但是权限比较高，具有数据库的更新和审核等管理权限 3. 面向的用户数目比较多，要求能够提供方便并且快速的管理员接口进行管理 4. 操作频率相对较低，但是每一次的操作对于系统造成的影响比较大。

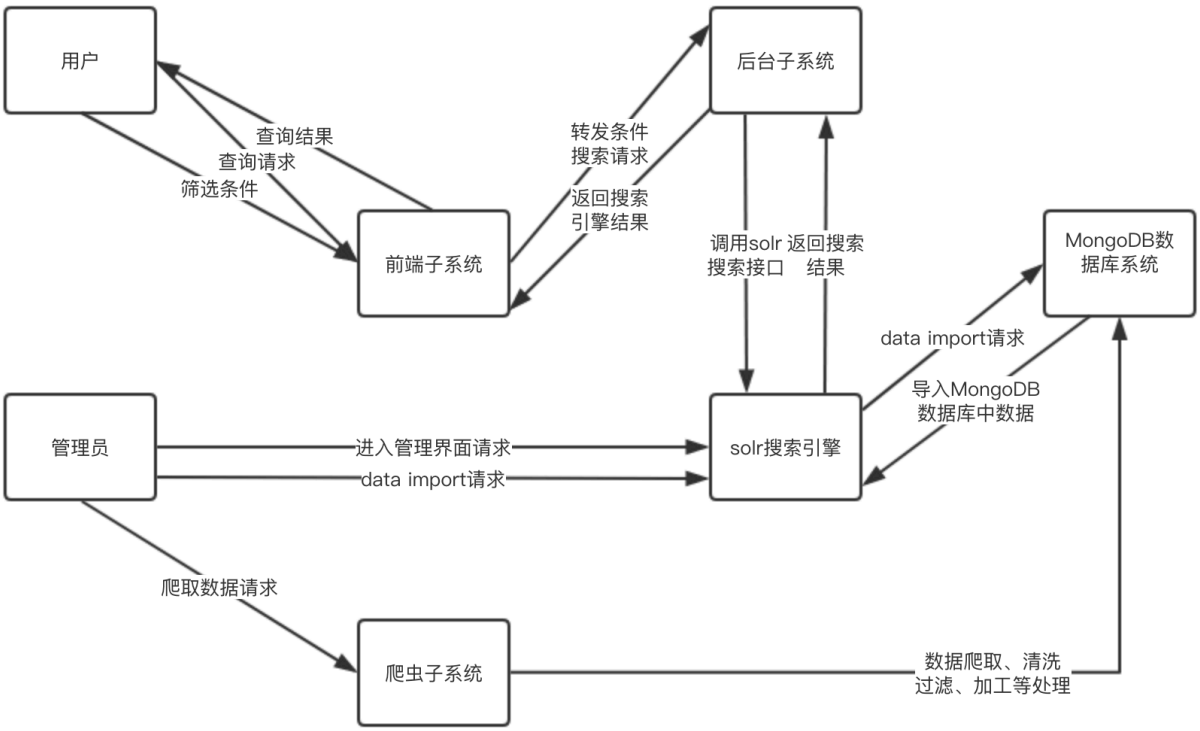
2.3 总体数据设计

2.3.1 IPO 图



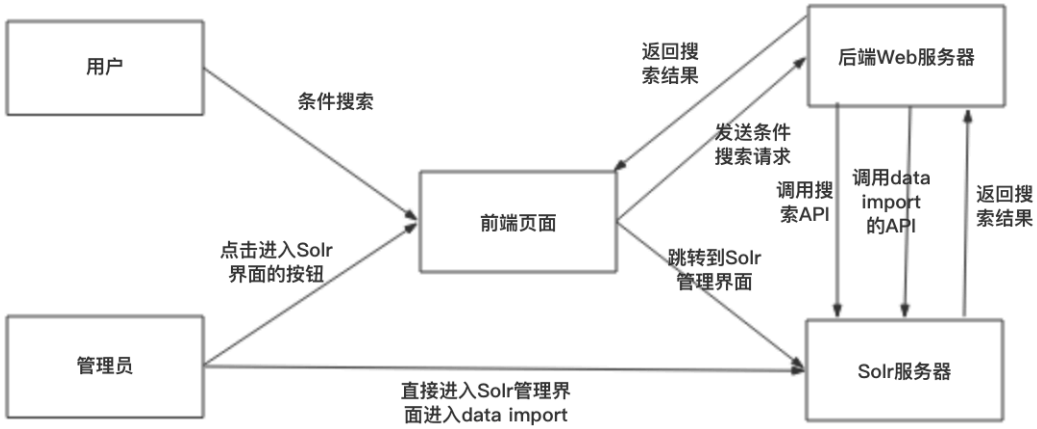
2.3.2 数据流图

环境层数据流图

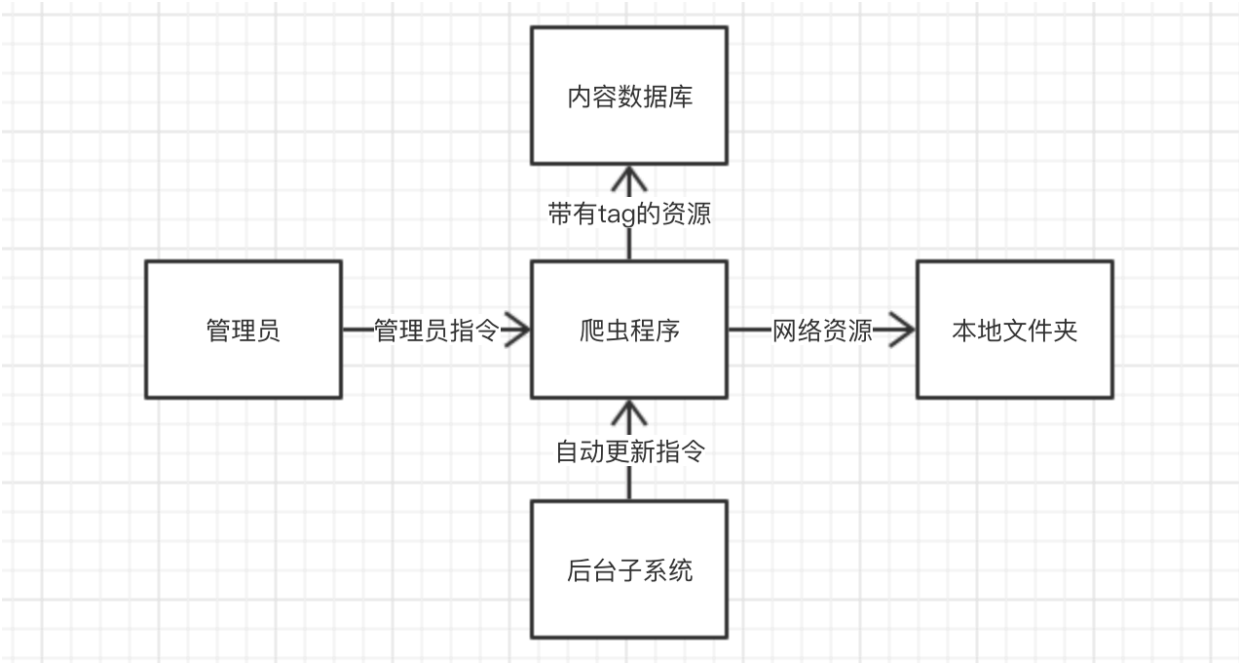


前端子系统数据流图：

后端子系统数据流图：



爬虫子系统数据流图：



2.4 运行环境

OE-1：Techhub 后端知识库搜索引擎系统可以在如下网页浏览器中正常使用：Windows Internet Explore 版本 8、9、10；火狐版本 12 至 26；谷歌 Chrome（全部版本）；苹果 Safari 版本 4.0 至 8.0.

OE-2：数据库平台为 MongoDB，Javascript 版本：Javascript5.0

OE-3：基本硬件配置：

CPU：≥2.0 GHz ；

内存：≥2 GB

硬盘：≥20 GB

硬盘转速：≥5400 rpm

网卡：100M

OE-4：Techhub 后端知识库搜索引擎系统将允许用户使用 Android、iOS、Windows 智能手机和平板电脑进行访问。

3 模块设计计划

3.1 前端子系统模块设计计划

功能点 ID	功能点名称	功能点描述	负责人	截止时间
frontend_01	查询搜索内容	用户在主页搜索栏中输入想要搜索的内容，点击搜索图标按钮后，前端会向后端发送搜索请求，并将返回的结果显示在搜索结果界面中。	胡瀚丹	7.12
frontend _02	筛选搜索条件	用户可以选择搜索来源网站、搜索范围（使用手册、项目实战、问答等），确定筛选后显示符合筛选条件的搜索结果。	胡瀚丹	7.13

3.2 搜索服务器子系统模块设计计划

功能点 ID	功能点名称	功能点描述	负责人	截止时间
Solr_01	Data Import	管理员可以随时手动将 mongoDB 的数据导入到 Solr 中	陈鑫	7.11
Solr_02	Search	向后端 Web 服务器提供搜索的 Http API	陈鑫	7.11

3.3 后端 Web 服务器子系统模块设计计划

功能点 ID	功能点名称	功能点描述	负责人	截止时间
Back_End_01	Search	根据用户输入的关键词、选择的分类调用 Solr 搜索 Api 进行搜索，并对搜索结果进行封装、去重。	陈鑫	7.11

3.4 爬虫子系统模块设计计划

功能点 ID	功能点名称	功能点描述	负责人	截止时间
Scrapy_text_01	信息爬取	系统在服务用户的过程中一般情况下不会实时爬取来自网络的数据，而是通过读取数据库中经过提炼和整合的数据来更好得服务用户。为了保证系统的信息具有一定的实时性，管理员必须隔一段时间爬取一次来自网络的各种信息。	傅诤哲	7.11
Scrapy_text_02	自动滤重	在管理人员爬取大量网络文本资源后，需要对一些相似性很高的文本进行过滤重复或者是合并，由于网络上的资源有很多重复的内容，给用户提供多个重复内容的情况在各类搜索引擎中并不少见，因此对文本数据做滤重十分重要。	傅诤哲	7.11

Scrapy_text_03	自动分类	技术博客一般都具有一些明显的标签来表面他是属于什么技术的、以及介绍了技术什么反面的知识，因此，如果对他们做一定的分类，可以有效得提高搜索的效率。	傅诤哲	7.11
Scrapy_text_04	自动爬取	当查询得到的数据不够多不能满足用户的要求，后台会启动爬虫去一些著名搜索引擎爬去数据，并和己方数据库中的内容做比对来去重和分类，最后将数据插入数据库供以后查询所用。	傅诤哲	7.11

4 系统验收标准

整个研发过程，必须按照《软件开发过程规范》进行，对于项目的评审按照《项目评审 规范》执行，对于项目组的考核按照《软件研发部考核办法》执行。整个项目研发的生命周期按照 CMMI 等级 2 的要求实施。前期上交相关文档有系统总体计划报告、可行性研究报告、系统开发计划书、系统分析说明书、系统设计说明书。后期上交程序设计报告，系统检测计划与测试报告，系统使用说明与维护手册，系统评价报告，系统开发月报告与系统开发总结报告。具体考核检查由邢卫老师和邵健老师负责。

5 小组成员

组长：黄亦非

组员：黄亦非、陈鑫、傅诤哲、胡瀚丹

本文档撰写：黄亦非、陈鑫、傅诤哲、胡瀚丹