

S2JH4Net Nutch-AJAX

李霞

Java/JEE技术

oschina号：xautlx

擅长的领域：全栈

公司的行业背景：企业应用/互联网

开源项目 > WEB应用开发 > Web开发框架

xautlx / s2jh Java LGPL-3.0

捐赠 0 Unwatch 207 Star 337 Fork 242

代码 Issues 10 Pull Requests 0 附件 0 Wiki 29 统计 服务 管理

xautlx / s2jh4net Java LGPL-3.0

捐赠 0 Unwatch 116 Star 131 Fork 98

代码 Issues 17 Pull Requests 0 附件 0 Wiki 1 统计 服务

集结最新主流时尚开源技术的面向互联网Web应用的整合前端门户网站、HTML5移动站点及后端管理系统J2EE相关主流开源技术架构整合及一些企业应用基础通用功能和组件的设计实现的最佳实践和原型参考。

140 Commits 2 Branches 0 Tags 0 Releases

Master	+ Pull Request	+ Issue	文件	挂件
Li Xia	最后提交于 2年前	Upgrade Apache Commons Collections to v3.2.2		
.settings	Li Xia	常规代码优化		
assets	Li Xia	独立运行包配置		
runtime	Li Xia	独立运行包配置		
src	Li Xia	修正Test.ftl中过期的模板代码内容		

项目地址:

<https://gitee.com/xautlx/s2jh>

<https://gitee.com/xautlx/s2jh4net>

框架特色

- 面向主流企业级WEB应用系统的界面和常用基础功能设计实现
- 主体基于主流的 (Spring MVC + Spring3 + Hibernate4/MyBatis3) 架构
- 引入JPA、Spring-Data-JPA提升持久层架构规范性和开发效率
- 基于流行jQuery/Bootstrap等UI框架和插件整合, 良好的浏览器兼容性和移动设备访问支持
- 提供一个基础的代码生成框架, 简化实现快速基本的CRUD功能开发
- 基于Maven的项目和组件依赖管理模式, 便捷高效的与持续集成开发集成

技术架构

- 技术列表 - 框架主要技术(Java/Web/Tool)组件列表介绍
- 技术特性 - 主要技术选型和设计说明
- 异常处理 - 介绍框架的异常处理的策略设计
- 移动支持 - 以Android为例的Web App与Native App整合应用

开发指南

- 开发配置 - 开发基础环境配置说明
- 工程结构 - 对整个项目工程代码结构进行概要性介绍
- 代码生成 - 用于基本CURD框架代码生成的工具
- 基础功能 - 框架已经实现的基础功能介绍说明
- UI组件 - 框架UI组件设计思路和用法演示
- 表格组件 - 功能强大的Grid表格组件扩展增强
- 表单控制 - 介绍Web开发过程最主要的表单处理设计

xautlx / nutch-ajax Java

捐赠 0 Unwatch 49 Star 71 Fork 39

代码

Issues 3

Pull Requests 0

附件 0

Wiki 1

统计

服务

管理

基于Apache Nutch和Solr以及Htmlunit, Selenium WebDriver等组件扩展, 实现对于AJAX加载类型页面的完整页面内容爬取、解析、清洗、持久化、全文检索等处理 -- 编辑

48 Commits

2 Branches

master

+ Pull Request

+ Issue

文件

Li Xia 最后提交于 2年前 . 将oschina maven私服提

apache-nutch-2.3

document

solr-4.8.1

.gitignore

.project

README.md

文档结构图

- 基于 Nutch&Solr 定向采集解析和索引搜索的整合技...
 - 总体介绍
 - 文档内容说明
 - 主要功能特性
 - 主要技术特性
 - 特别说明
 - 基本工具安装配置
 - JDK 版本及安装
 - Eclipse IDE / Spring Tool Suite
 - Windows 版本 MongoDB
 - Eclipse 插件安装配置
 - JDK 设置
 - Apache IvyDE
 - MongoDB 可视化客户端插件: MonjaDB
 - Git 参数配置
 - Eclipse Workspace 参数配置
 - 项目资源库获取说明
 - 项目导入及开发环境配置
 - 项目导入 STS 工作空间
 - 项目运行 Quick Start
 - Ntuch 运行脚本执行过程分析
 - 启动本地 MongoDB 服务和客户端使用
 - 导入项目预置的各 Job Run/Debug 配置
 - 关于 Windows 环境运行出现 hadoop 相关异常处
 - 安装用于 Selenium WebDriver 执行的 FireFox
 - MySQL 解析数据存储
 - Solr 运行及索引
 - Nutch 扩展插件说明
 - 插件列表
 - lib-pinyin: 汉字转拼音组件
 - 用法场景
 - 在 IncomingFilter 调用接口把中文内容转输
 - lib-HttpURLConnection: 基于正则的 URL 请
 - lib-HttpClient: 基于多线程的 Htmlunit 页
 - protocol-s2jh: 基于 Htmlunit, Selenium Web
 - parse-s2jh: 解析特定业务属性并进行持久化
 - index-s2jh: 扩展的 index 索引处理插件
 - 主要配置文件介绍
 - nutch-site.xml
 - gora.properties
 - xxxx-urfilter.txt
 - regex-normalize.xml
 - htmlunit-urfilter.txt
 - log4j.properties

基于 Nutch&Solr 定向采集解析和索引搜索的整合技术指南文档

总体介绍

文档内容说明

文档内容主要涉及 [nutch-ajax](#) 及 [nutch-solr-commerce](#) 项目源码已实现内容 (包含一些 [Nutch](#) 和 [Solr](#) 标准的功能和原理说明但不保证完整性)。具体可详见文档目录列表。

内容不包括 What Is Not

本教程内容以单机运行模式为例, 不涉及 [Nutch](#) 及 [Solr](#) 相关的基于 HDFS、[Hadoop](#)、[HBase](#) 等分布式和集群部署运行等高级特性, 此类相关技术点在本教程不做任何说明, 请自行参考官方资料相关教程。

主要功能特性

常规的 HTML 页面抓取: 对于常规的例如新闻类没有 AJAX 特性的页面可以直接用 [Nutch](#) 自带的 protocol-http 插件抓取。

常规的 AJAX 页面抓取: 对于绝大部分诸如 [jQuery ajax](#) 加载的页面, 可以直接用 [htmlunit](#) 扩展插件抓取。

特殊的 AJAX 请求页面抓取: 诸如淘宝/天猫的页面采用了独特的 [Kissy Javascript](#) 组件, 目前测试 [htmlunit](#) 无法正确解析, 因此退而求其次采用效率低一些的 [Selenium WebDriver](#) 方式实现页面数据抓取。

主要技术特性

基于 Apache [Nutch 2.3](#) 及 [Solr 4.8.1](#) 系列
基于 [Htmlunit](#), [Selenium WebDriver](#) 等扩展实现 AJAX 网站数据爬取
基于 [Nutch](#) 的 [protocol-http](#) 插件实现数据整合进行索引和搜索

特别说明

文档所列举到的工具软件类型及版本, 操作过程等, 皆以实际操作过程涉及相关为参考, 如果你是新手建议一步步按照文档进行, 如果你是熟手你可以参考相关描述直接在现有熟悉的环境自行配置。

[Nutch](#) 的运行涉及 [Hadoop](#), [HBase](#) 等这些组件, 官方的代码和指南基本都是面向或优先定位在 [Unix/Linux](#) 环境进行开发和部署运行的。但是考虑到比较绝大部分开发人员都是熟悉 [Windows](#) 开发环境, 考虑到引入 [Linux](#) 作为开发演示环境会把开发过程进一步复杂化, 因此本教程全部基于 [Windows](#) 系统进行说明, 并且会对其中涉及到一些平台问题进行特殊说明。

除特殊说明以外, 以下教程假定相关软件或项目的操作根目录为: `DEV_DIR=D:\projects\training`

项目地址:

https://gitee.com/xautlx/nutch-ajax

技术功底

立身之本，一切尽在代码中

技术管理

一个人再厉害也干不完所有活
还需要一个善于进取和分享的团队
以及有效的配置和开发过程管理

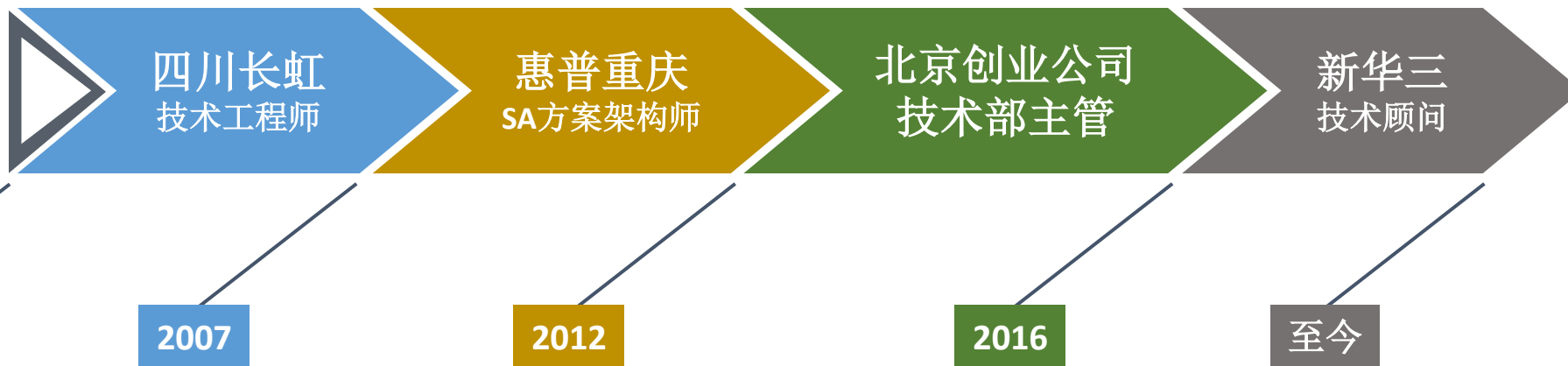
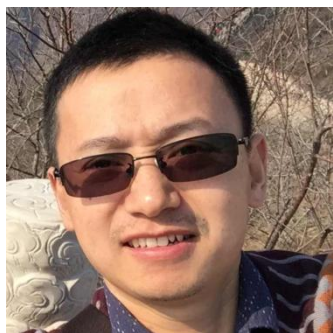


参与开源

工作中提炼和参与开源项目
开源项目提升工作质量和效率

开发运维

敏捷开发，持续集成，自动化运维
有效提升开发、测试、运维的效率和质量

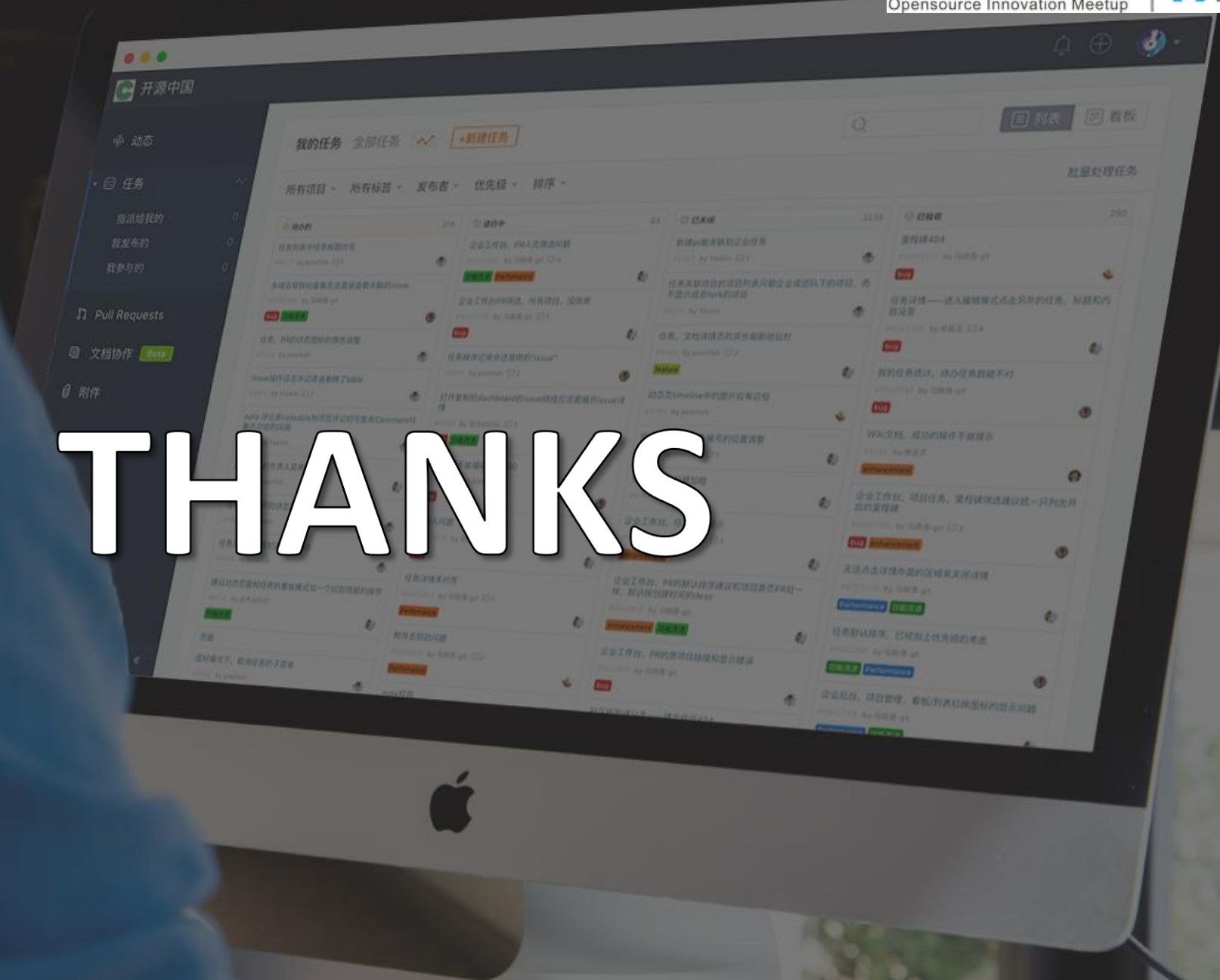


新的征程

右侧是我的项目QQ交流群的二维码，欢迎大家进群一起交流，分享经验。

如果大家觉得我的项目还不错，希望能够给点掌声鼓励一下，谢谢！





THANKS