**软件可行性分析报告**

**文档名称：单页应用搜索引擎优化系统可行性分析报告**

**项目名称：单页应用搜索引擎优化系统**

**编写人：赖信涛**

目录

[单页应用搜索引擎优化系统可行性分析报告 1](#_Toc484353294)

[一、 概述 1](#_Toc484353295)

[1.1项目名称 1](#_Toc484353296)

[1.2项目背景 1](#_Toc484353297)

[1.3项目产品的主要用途与性能 2](#_Toc484353298)

[二、 可行性研究的前提 2](#_Toc484353299)

[2.1 要求 2](#_Toc484353300)

[2.2 系统的主要功能 3](#_Toc484353301)

[2.3 系统运行的软硬件环境 4](#_Toc484353302)

[2.4技术可行性分析 5](#_Toc484353303)

[2.5操作可行性分析 5](#_Toc484353304)

[2.6法律可行性分析 5](#_Toc484353305)

[2.7 评价尺度 5](#_Toc484353306)

[三、对现有系统的分析 6](#_Toc484353307)

[3.1架构图 6](#_Toc484353308)

[3.2功能图 6](#_Toc484353309)

[四、所建议的系统 7](#_Toc484353310)

[4.1 影响 7](#_Toc484353311)

[4.2 局限性 7](#_Toc484353312)

[4.3 技术条件方面的可行性 8](#_Toc484353313)

[五、投资及效益分析 8](#_Toc484353314)

[5.1 投资成本 8](#_Toc484353315)

[5.2 收益 8](#_Toc484353316)

[5.3 收益/投资比 8](#_Toc484353317)

[七、结论 8](#_Toc484353318)

[八、参考资料 8](#_Toc484353319)

单页应用搜索引擎优化系统可行性分析报告

# 一、 概述

## 1.1项目名称

单页应用搜索引擎优化系统

## 1.2项目背景

近年来，兴起了很多单页面的网页开发框架。它们单页应用大量采用Ajax（Asynchronous JavaScript and XML, 异步的[JavaScript](https://zh.wikipedia.org/wiki/JavaScript" \o "JavaScript)与[XML](https://zh.wikipedia.org/wiki/XML" \o "XML)技术）技术，在不刷新网页的情况下就可以更新网页的内容。给用户体验带来了很大的提升。

但是随之也产生很多问题，用单页应用开发网页最大的痛点就是被搜索引擎的收录较差。由于网页的内容是需要执行页面内的JavaScript脚本来生成的。所以当搜索引擎的爬虫请求页面内容的时候，单页应用响应的只有一包含JavaScript脚本的HTML文档。在JavaScript脚本之前，网页内并没有实质的内容。

本文提出，使用基于无界面浏览器技术的分布式爬虫对自身网页进行解析，填充内容，对搜索引擎提交完整的网页。以此来解决搜索引擎的收录问题。本文提出的系统采用无界面浏览器技术，可以模拟真实用户打开网页，把用户实际看到的网页保存下来。系统由爬虫模块负责任务调度和抓取网页，然后使用无界面浏览器对网页解析，之后由存储模块将相应的信息保存到文件系统和数据库。当搜索引擎发来请求的时候，Web服务器模块就可以根据保存的信息做出响应。利用分布式的爬虫，使用任务队列分配网页的采集任务，任务调度灵活，扩展性高。针对搜索引擎的爬虫单独设置网络服务器，SEO（search engine optimization，搜索引擎优化）针对性高。

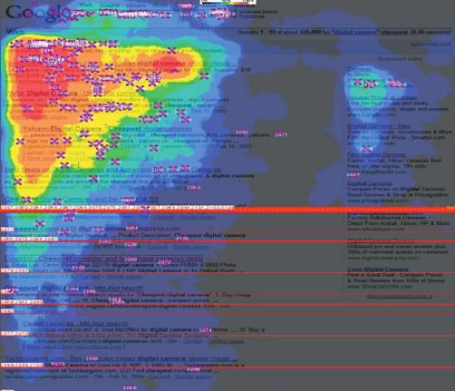
## 1.3项目产品的主要用途与性能

针对本文要解决的问题，即让搜索引擎正常收录单页应用，需要将原网站的网页解析并保存下来，还要能在搜索引擎的网页爬虫发来请求的时候，返回解析后的网页作为响应。所以，需要爬虫模块来抓取网页，phantomjs模块解析网页，存储模块保存网页，web服务器模块响应网页，这几个模块相互配合。

# 二、 可行性研究的前提

## 2.1 要求

对于信息展示类的网站来说，大约有70%的流量来自于搜索引擎的结果引入。搜索引擎对于网站的收录，对网站来说是至关重要的实验数据显示约85 % 的用户只翻看搜索引擎 返回结果的前10 个结果,即返回结果页面的第一页。这个用户行为决定了尽管搜索引擎返回的结果 数目十分庞大,但真正可能被绝大部分用户所浏览的,只有排在最前面的很小一部分而已。下图展示了用户的浏览习惯，基本上靠前的一小部分结果占据了搜索引擎的大部分流量，而互联网的主要流量来自于这里。要想提高网站在搜索引擎展示结果中的数量和排名（排名越靠前，被点击的概率越大），首先要保证网站的内容能正常并且全面地被搜索引擎收录。



用户浏览习惯金三角

如果单页应用没有针对搜索引擎优化的话，是很难被正常收录的。因为单页面应用只有一个页面，搜索引擎只是下载页面的HTML文档，并不会执行文档中包含的脚本。这样，就造成了搜索引擎收录页面数量少、收录到的页面内容空洞、价值不高、网页的页面之间结构化差，从而会进一步导致网站在搜索引擎的结果中展示的频率低、排名靠后。会损失很多宝贵的搜索流量来源。

所以单页应用不同于一般的网页，除了常规的搜索引擎优化工作之外，还需要一些特殊的搜索引擎优化方案，来保证单页应用所有的页面能够被正常收录。本文的目的便是解决这个问题，如果能让单页应用被搜索引擎收录，就可以带来很多宝贵的搜索流量，提升网站的价值。目前流行的一些其他搜索引擎优化方案，也可以基于本文做跟深一步的优化。

## 2.2 系统的主要功能

将网页抓取下来，是系统的基础功能。能通过页面中的链接连续不断地自主发现新的页面，自己分配工作，是理想中的爬虫的工作方式。作为一个爬虫模块，任务调度必不可少，必须使用一种合理的任务队列，使用调度算法分配任务，才能使爬虫稳定、有序地执行。除此之外，还需要解析页面中的链接，将新的链接放回到队列中。这个过程应该注意发现链接的效率，和内存占用，以及判断是否是本站链接，如果不是，不应该作为任务执行。如果提供sitemap，爬虫可以根据sitemap进行更全面地抓取。

对HTML文档的解析是系统的重点功能，有一个专门的模块负责解析的任务。由于这部分解析JavaScript脚本需要使用JavaScript语言来编写程序，系统的其余模块需要使用Python编写，所以为了使这两个模块之间能互相配合，便将解析JavaScript的模块以服务的形式开发。运行的时候监听某一个端口，python程序通过往这个端口发送请求，此模块回应请求，以这种方式交互。

存储模块需要实现的是将保存下来的网页与新抓取的网页进行快速对比，判断网页是否已经更新。系统可以将网页提交给搜索引擎，被搜索引擎正常收录。其中对于文档的对比是一个技术难点，页面的对比必须满足程序执行速度块，占用内存少，占用存储空间少的几个要求。此外，存储使用的数据库可能要面对非结构化的数据，可以灵活扩展与管理。

Web服务器模块的功能比较简单，只是将已经保存的网页返回便可以。但是有一种特殊的情况，就是有请求的目标是尚未被索引的网页。这种情况下应该马上去处理这个页面，处理完之后将页面反回给浏览器。也就是说，Web服务器应该有往任务队列中放高优先级任务的机制。

此外，稳定性与安全性是本系统的目标。所以应该使用容器机制和合适的进程管理界面。应该提供直观地观察系统运行状态的方式，能够快速重启或维护系统各个组件。维护方面应该使用授权机制，保证基本的安全性。

## 2.3 系统运行的软硬件环境

操作系统：Linux

安装有： Vim, Phantomjs, Python ,zsh

## 2.4技术可行性分析

本文提出，使用分布式爬虫的技术通过构建动态的镜像网站来解决但页面应用的SEO问题。通常，爬虫技术用于数据采集，搜索收录等场景[2]。但是配合无界面浏览器技术，可以打破爬虫只能下载页面内容的局限性，能够使用JavaScript解释器引擎执行页面内的JavaScript代码，从而获得真实的、有内容的页面，而不仅仅是无意义的HTML文档内容。

本文提出的系统采用容器的方式部署，屏蔽了不同操作系统之间底层的不同，从而使部署变得简单。组件与组件之间通过网络地址加端口的方式通讯，非常灵活，采用分布式的爬虫和任务队列，扩展性高。

所有组件的设计方式，保证了本系统的高可用性，适用于各种不同规模的网站。对于小型网站，可以采用单台甚至与原网站使用同一台服务器的方式。对于大型网站，可以使用两台到几十台的服务器集群，专门解决SEO的问题。

## 2.5操作可行性分析

使用tmux部署，非常简单易用。

## 2.6法律可行性分析

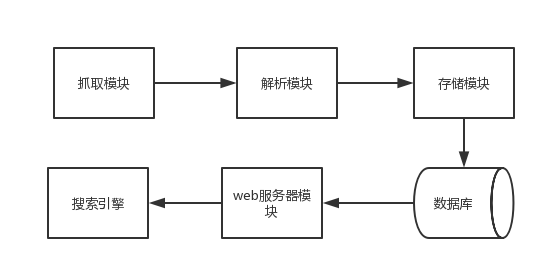
所使用工具和技术及数据信息不违反法律.

## 2.7 评价尺度

本系统比较简单而且只在本机使用，暂不考虑网络，因而开发费用小。系统开发人员为1人，时间大概在60天左右（2017年4月1日 至 2017年5月31日）。

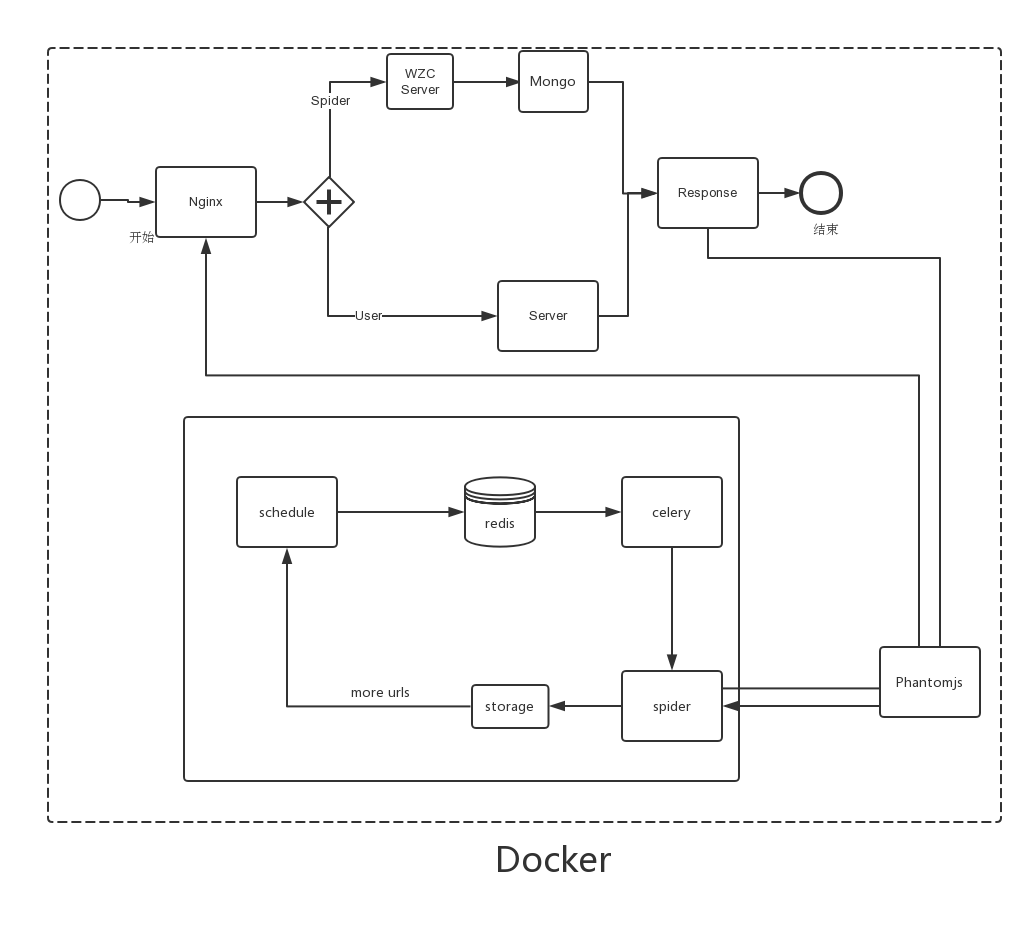
# 三、对现有系统的分析

## 3.1架构图



## 3.2功能图

系统的每一层用一个程序集来实现



# 四、所建议的系统

本系统为单页应用的SEO优化而设计，整个系统主要分为两部分。第一部分是爬虫引擎，主要负责从目标网站抓取页面内容，将爬取的任务信息存到数据库。第二部分是网页服务器部分，主要负责给搜索引擎返还真实的页面内容。

## 4.1 影响

系统应该有足够的存储空间。

## 4.2 局限性

系统可能无法在短时间内快速索引所有网站。

## 4.3 技术条件方面的可行性

本系统是一个基于Linux的系统,现有技术（开发环境和使用的架构）已较为成熟,利用现有技术完全可以实现系统开发目标.同时,开发期限较为宽裕,预计可以在规定期限内完成开发任务.

# 五、投资及效益分析

## 5.1 投资成本

出人力开发外无其他成本。

## 5.2 收益

作为开源软件发布，没有收益

## 5.3 收益/投资比

无。

# 七、结论

本项开源，免费。开发完成之后能为社区带来很大收益，能简化单页应用的SEO，非常可行。

# 八、参考资料

1. 谢平华，MD5 算法在重复邮件识别方面的研究和实现，科技传播, 2011
2. 余慧佳, 刘奕群, 张敏, 茹立云, 马少平，基于大规模日志分析的搜索引擎用户行为分析，中文信息学报，第 21 卷 第 1 期 2007 年 1 月
3. 王霜，修保新，肖卫东，Web 服务器集群的负载均衡算法研究，计算机工程与应用，2004