

**Geeking搜索引擎项目报告**

团队名称：Geek

团队成员：

陈新荃 2014E8018461048

高妍

林裕杰

肖卡飞

2014年12月14日

目录

[目录 2](#_Toc405806037)

[**一．项目总体介绍** 3](#_Toc405806038)

[**二．设计方案** 3](#_Toc405806039)

[1. 网页爬虫 3](#_Toc405806040)

[1.1 爬取策略 3](#_Toc405806041)

[1.2 具体实现 3](#_Toc405806042)

[2. 索引构建 3](#_Toc405806043)

[2.1 数据结构 3](#_Toc405806044)

[2.2 构建流程 4](#_Toc405806045)

[2.3 网页过滤 4](#_Toc405806046)

[2.4 文本分词 4](#_Toc405806047)

[3. 检索策略 4](#_Toc405806048)

[3.1 检索流程 4](#_Toc405806049)

[3.2 结果排序 4](#_Toc405806050)

[3.3 结果聚类 4](#_Toc405806051)

[4. 前端处理 5](#_Toc405806052)

[4.1 页面元素 5](#_Toc405806053)

[4.2 自动补齐 5](#_Toc405806054)

[4.3 搜索词推荐 5](#_Toc405806055)

[4.4 摘要快照及高亮 5](#_Toc405806056)

[**三．测试与评估** 5](#_Toc405806057)

[1. 测试环境 6](#_Toc405806058)

[2. 测试结果 6](#_Toc405806059)

[2.1 功能测试 6](#_Toc405806060)

[2.2 性能测试 6](#_Toc405806061)

[**四．创新点** 6](#_Toc405806062)

[1. 标题权重 6](#_Toc405806063)

[2. 中间字自动补全 6](#_Toc405806064)

[3. 【还有什么尽管添加】 6](#_Toc405806065)

[**五．经验与总结** 7](#_Toc405806066)

[1. 个人经验总结 7](#_Toc405806067)

[2. 组长总结 7](#_Toc405806068)

# **一．项目总体介绍**

🡪陈新荃

采用什么方法，实现了什么功能，基于什么语料，创新点

时间、代码量、github结果等

# **二．设计方案**

## 1. 网页爬虫

### 1.1 爬取策略

🡪咖啡

### 1.2 具体实现

🡪咖啡

第三方jar包等

## 2. 索引构建

🡪裕杰

总体介绍

### 2.1 数据结构

🡪裕杰

数据库、索引在内存中的结构等

### 2.2 构建流程

🡪裕杰

### 2.3 网页过滤

🡪陈新荃

### 2.4 文本分词

🡪陈新荃

## 3. 检索策略

🡪裕杰

总体介绍

### 3.1 检索流程

🡪裕杰

流程图

### 3.2 结果排序

🡪陈新荃

合并、排序等

### 3.3 结果聚类

🡪裕杰

聚类探究与算法实现

## 4. 前端处理

🡪高妍

总体介绍，功能实现

### 4.1 页面元素

🡪高妍

截图解释

### 4.2 自动补齐

🡪高妍

### 4.3 搜索词推荐

🡪高妍

### 4.4 摘要快照及高亮

🡪高妍

# **三．测试与评估**

🡪陈新荃

总体介绍，测试内容等

## 1. 测试环境

🡪陈新荃

软件、硬件、语料集、

## 2. 测试结果

### 2.1 功能测试

🡪高妍

排序、聚类、自动补齐、相关搜索推荐、snippet、快照等

### 2.2 性能测试

🡪陈新荃

启动速度、响应速度

# **四．创新点**

## 1. 标题权重

🡪陈新荃

## 2. 中间字自动补全

🡪高妍

## 3. 【还有什么尽管添加】

# **五．经验与总结**

## 1. 个人经验总结

陈新荃：

高妍：

林裕杰：

肖卡飞：

## 2. 组长总结

Balabala…