NBA 新闻搜索引擎的设计与实现

赵东杰，曲建波

（华北计算技术研究所 北京 北京 100000）

摘要：快速的响应时间、简洁友好的界面、精确的匹配度等是一个搜索引擎最重要的几个指标，对于专业领域的搜索引擎更是如此。基于 Apache Lucene 的 ElasticSearch 通过倒排索引来提供高效快速的搜索，并提供分词、建议等高级功能，适合作为搜索引擎的后台存储。Redis 提供了高效的内存存储可以加快相应速度，Django 作为一个成熟的 Python 开发框架能够快速搭建项目，专注业务逻辑。基于上述解决方案，通过整合虎扑新闻资源，并借鉴 Google 搜索的设计理念，设计实现一个针对 NBA 新闻的搜索引擎，为用户提供良好的搜索体验。

关键词：虎扑新闻； 搜索引擎； ElasticSearch； Django

引言

在信息爆炸的今天，如何能够快速高效的获取到自己所需要的信息对人们的工作效率至关重要。搜索引擎作为目前最重要的搜索手段，通过一定的策略从网上爬取众多的互联网资源，并通过对信息进行索引、分词等组织和处理方法，为用户提供高效，丰富的搜索体验。根据本学期的信息检索的前沿研究课程设计要求，实现一个专业领域的搜索引擎。

随着 NBA 的薪金空间越来越高，与各个电视、网络平台的合作也越来越紧密，NBA 球迷的数量也是日益增多，因此对 NBA 新闻的需求的日益增多。目前常见的 NBA 新闻查看途径包括腾讯新闻等门户网站、百度等搜索引擎、虎扑等专业领域的应用。其中百度等通用搜索引擎全面性更广，但是在时效性有所欠缺，且内容为了全面性不免宽泛，多了更多无用的信息，不能专注于 NBA 新闻本身。而腾讯新闻、UC 新闻等门户网站，虽然时效性较好，且专业度也足够满足用户的需求，但是过于专注网络流量，新闻标题有时为了吸人眼球，过于骇人听闻，真实性过低。而虎扑作为最专业的篮球论坛，不管是时效性、真实性，都是 NBA 新闻方面的佼佼者，因此将虎扑新闻的 NBA 板块作为我们的数据源。

ElasticSearch 作为基于 lucene 的搜索服务器，提供了一个分布式的全文搜索引擎，并且提供了 RESTFul web 接口，能够快速方便的调用相关功能，为很多大型网站提供了搜索服务，同时也可为个人用户快速集成搜索服务。

Django 作为一个开放源代码的 web 应用框架，可以快速搭建网站，使用户专注于项

目本身的业务逻辑，对于快速搭建本项目的环境是个很好的选择。

Google 作为最成功的搜索引擎，有着众多的经典的设计可以值得在项目界面设计的过程中进行借鉴。

基于上述内容，本文提出了基于 ElasticSearch 的 NBA 新闻搜索引擎，通过借鉴 Google 搜索结果界面的经典设计，为用户提供一个界面友好，功能齐全的搜索引擎。

1 相关技术

1.1 Python 和 Django

Python 作为目前势头最猛的编程语言,大有超越 Java 的趋势。Python 是一种解释型、面向对象、动态数据类型的高级程序设计语言，与传统的 Java 相比，第三方开发库丰富，在网站开发、数据挖掘等领域都有着广泛的应用。Python 简明的语法，以及 Python 中的 ElasticSearch 的相关开发库，能够方便的实现后端的业务逻辑，将搜索到的数据存入到 ElasticSearch 服务器中,为项目提供搜索数据。同时在开发过程中可以设置虚拟开发环境，更有利于项目的独立开发和部署。

Django 项目是一个 Python 定制框架，最初源于劳伦斯出版集团的在线新闻 web 站点,在 2005 年以开源的方式释放出来,它鼓励快速开发,并遵循 MVC 的设计模式。Django 的主要目的是简便快捷的开发数据库驱动的网站。它强调代码复用，把控制层进行了进一步的封装，在编写程序的时候，只需要调用相应的方法即可。同时使用了 URL 正则表达式匹配的方法，能够更快的开发项目。

Django 的业务流程原理，浏览器通过 URL 请求，Django 接收到 URL 请求之后，通过路由系统分发到相应的视图函数，视图函数处理相应的请求，包括数据库的增删改查，并将结果渲染到框架中的模板页面。

1.2 ElasticSearch

ElasticSearch 是一个分布式可扩展的实时搜索和分析引擎,一个建立在全文搜索引擎 Apache Lucene(TM) 基础上的搜索引擎.当然 ElasticSearch 并不仅仅是 Lucene 那么简单，它不仅包括了全文搜索功能，还包括分布式实时文件存储，并将每一个字段都编入索引，使其可以被搜索；实时分析的分布式搜索引擎；可以扩展到上百台服务器，处理 PB 级别的结构化或非结构化数据。

ElasticSearch 是面向文档型数据库，一条数据即为一个文档，使用 JSON 作为文档序列化的格式。其中主要概念与关系型数据库对比如表 1 所示。

一个 ElasticSearch 集群可以包括多个索引，每个索引中可以包含多个类型，而在每个类型中包含了很多的文档，每个文档中又包含了很多的字段。

表 1 数据库概念对比

关系型数据库	数据库	表	行	列
ElasticSearch	索引	类型	文档	字段

1.3 Scrapy

Scrapy 作为一个开源和协作的数据爬取框架，最初的目的就是为了网络抓取而设计的，并且可以方便的进行扩展和定制，从网站中提取到项目所需要的数据。Scrapy 是基于 twisted 框架开发的，这样就保证了它的异步处理，整体架构图如图 1 所示

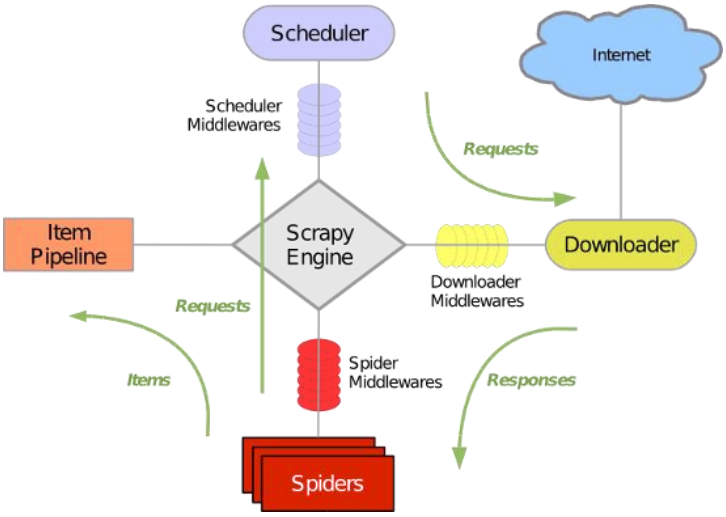


图 1 Scrapy 架构图

Scrapy 主要工作流程是：引擎打开网站，找到处理该网站的 spider 并请求第一个初始 URL，然后引擎把 URL 交给 Scheduler 进行调度。调度器将下一个要抓取的 URL 传给引擎，然后引擎将 URL 通过下载中间件交给下载器 Downloader 进行下载。下载完毕，下载器会生成一个 Response 并返回给引擎，引擎再通过 Spider 中间件将 Response 传递给 Spider，Spider 将处理 Response 得到的 Item 和 Request 分别通过引擎交给 Item Pipeline 和调度器 Scheduler，然后不断重复这个过程，知道调度器 Scheduler 中没有没更多的 Request，结束整个爬取过程。

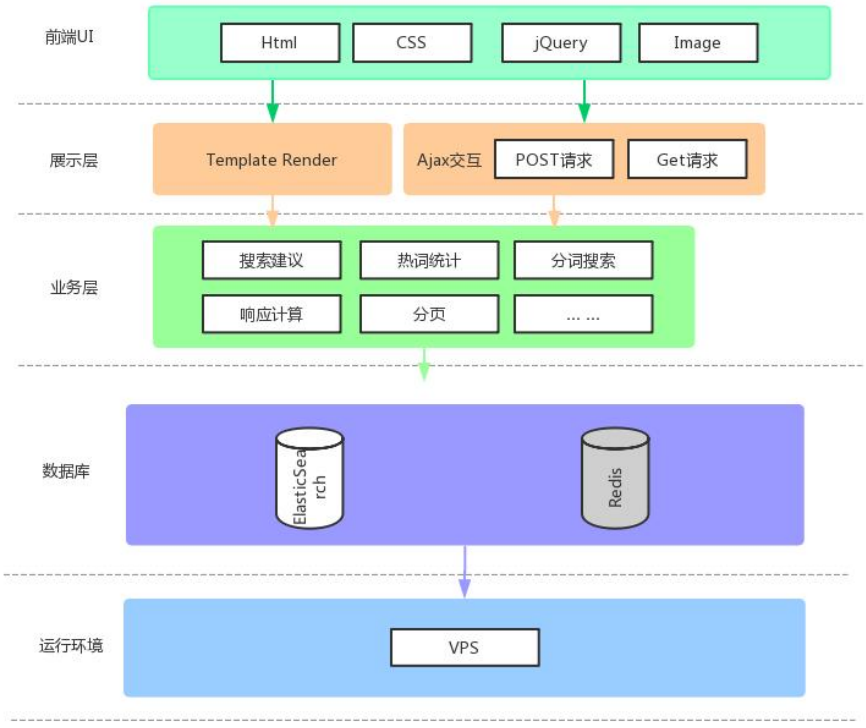
1.4 Redis

Redis 作为一个开源的 key-value 数据库，提供高性能的内存存储。与其他的内存数据库相比，Redis 支持数据的持久化，可以将内存中的数据写入到磁盘，并在 Redis 服务启动时再次加载。Redis 不仅仅支持 key-value，同时也支持 list 等高级数据结构的存储，

并且 Redis 的数据类型都是基于　　基本数据结构，对程序员透明，不需要进行额外的抽象。

2 系统架构设计

搜索引擎的架构设计图如图二所示。



图二　系统架构图

用户通过浏览器访问网站并提交相应请求，后台解析 URL 之后映射到相应的视图函数，视图函数处理相应的业务逻辑并与数据库交互，获取数据后，通过渲染前端页面返回数据。整个服务部署在租用的 VPS 上。

3 系统实现

本系统通过 Ubuntu 18.04 LTS 进行开发，主要开发平台使用 Pycharm，并通过 virtualenv 以及 virtualenvwrapper 两个包来创建独立的虚拟开发环境。前端通过 HTML、CSS、JavaScript 以及 JQuery 框架实现相关界面以及操作。后端通过 redis，elasticsearch-dsl 等库来实现相关的后台数据操作。

3.1 系统环境

开发平台：Ubuntu18.04 LTS；

开发语言：后端 Python，前端 HTML 等；

开发平台：Pycharm；

开发环境：Python3.6,、elasticsearch-rtf、redis；

运行环境：Ubuntu18.04 VPS

3.2 搜索建议

搜索引擎一个比较重要，也是比较人性化，用户友好的功能就是在用户输入时，会给出相应的提示，为用户提供有用的信息，给出搜索的建议。

搜索建议的功能实现主要通过 JavaScript 函数来实现。当输入框内容发生变化是，在 JavaScript 函数中提交一个 Ajax，将输入框的内容提交到相应的搜索建议视图函数，然后在函数中通过 elasticsearch_dsl 库提供的 suggest 函数从之前爬取的存入到 es 中的数据中获取相应的建议数据，并将数据渲染到前端的界面中，即可实现相应的搜索功能。

3.3 搜索记录

一个优秀的搜索引擎还应该对用户的搜索内容进行记录,以便用户后续能够查看自己的搜索。

搜索记录的功能主要通过前端 js 的 LocalStorage 实现。通过在 js 保存一个数组，在每次搜索时更新该数组，同时进行去重处理。然后在每次加载页面的时候将数组读取出来并进行显示。

3.4 热门搜索

搜索引擎还应该为用户进行内容推荐，过于复杂的推荐算法不是本课程的专注点，因此不必考虑数据挖掘等内容，只需根据查询次数对访问词进行排序，按照访问量推荐给用户即可。

为了响应速度，如果将访问内容存入数据库，在每次界面都进行数据库的读操作，耗时太长，不符合系统的响应时间的要求。而 Redis 作为内存数据库，可以提供高效的访问操作，最大限度的提高响应时间，因此我们通过 Redis 来实现热门搜索的功能。

在用户将搜索内容提交到相应的视图函数中，在进行搜索功能的实现之前，首先将搜索词加入到 redis 中，通过 python 中的 redis 中的相应函数，对搜索词的次数递加，并根据搜索次数对搜索词进行排序并返回到前段进行渲染。

3.5 搜索功能

对于搜索引擎来说，最重要的还是搜索结果的响应时间，内容的匹配度等功能。

在搜索功能的实现中，与前面的功能一样，也是通过 URL 映射功能映射到视图中搜索

处理函数中。

搜索时间可以通过python中的datetime模块分别记录搜索前的时间戳以及搜索后的时间戳，然后通过计算差值并转化为秒即可。

分页功能通过 JQuery 提供的分页组件快速实现。

搜索功能通过 Python 中的第三方库 elasticsearch 中的 suggest()方法来实现,并且可在查询方法中设置关键词标红等功能，以及实现模糊搜索，多字段搜索等功能。获取到搜索的数据后，通过对数据进行组织，将数据渲染到搜索结果页面中。

4 项目部署以及测试

通过租用第三方 vps 作为服务器，然后使用 ssh 进行连接，并通过 scp 命令将项目上传到服务器上，之后将本地的虚拟开发环境打包上传到服务器上快速部署。

通过提前发布测试版本，让同学们对系统的功能进行测试，并对其中的 bug 进行修复。

5 总结与未来期望

本文提出了一个基于 elasticsearch 的搜索引擎的解决方案，设计实现了针对 NBA 新闻的搜索引擎，并着重介绍了本人所专注的网站开发以及部署的工作。数据爬取的具体工作由曲建波同学详细介绍。

目前项目功能还过于简陋，前端界面也过于简洁，功能不够丰富，且服务器主要功能都靠手工启动和管理，过程过于繁杂。并且为了快速开发，采用了插件丰富的基于 elasticsearch 5.1.1 的版本，而不是最新的 elasticsearch7 的版本，新版本的功能无法实现。因此服务器的自动化运行以及自动更新、日志管理，前端界面的丰富以及 elasticsearch 的版本迁移是未来的主要工作。

参 考 文 献

[1] 徐伟杰，王挺，薛婉婷．基于 ElasticSearch 的搜索引擎设计与实现[J]. 智库时代,2019(23):228+240.

[2] 严慧,彭绪富,朱小婉,熊旭辉,董叶豪.基于 Scrapy-Redis 分布式数据采集平台的设计与实现[J].湖北师范大学学报(自然科学版),2019,39(01):19-25.

[3] 王芳,张睿,宫海瑞.基于 Scrapy 框架的分布式爬虫设计与实现[J].信息技术,2019(03):96-101.

[4] 张翠丽,孟小艳,杨抒.基于 Django 框架的管理系统的设计与开发[J/OL].计算机技术与发展 ,2019(11):1-9[2019-06-19].http://kns.cnki.net/kcms/detail/61.1450.tp.20190320.1520.002.html.

[5] 何军,陈贵民,黄惠海,郑汉军,陈思德.关于 RESTful 架构的设计[J].网络安全技术与应用,2019(01):37.

[6] 王伟,魏乐,刘文清,舒红平.基于 ElasticSearch 的分布式全文搜索系统[J].电子科技,2018,31(08):56-59+65.

[7] 周璐. 基于 Web 前端的 localStorage 性能研究与改进[D].吉林大学,2014.

[8] 王洪九.运用 jQuery 和 ajax 实现数据库数据的提取和分页[J].信息与电脑(理论版),2012(09):97-98.