

**毕业设计开题报告**

题 目： 网络爬虫大数据分析系统

专 业： 计算机科学技术

院 系： 信息工程学院

年 级： 2013级

学 号： 13150078

姓 名： 袁澄

指导教师： 周刚

职 称： 教授

湖北经济学院教务处 制

**一、背景：**

**随着互联网越来越快速的发展，人与人的交流也越来越紧密而且方便，互联网技术的进步给我们的生活带来了丰富的体验。比如娱乐方面：我们可以在网络上搜索电视剧电影进行观看，社交方面：各种聊天工具，通过互联网将越来越多的人都纳入到互联网这张巨网中，生活方面：网上购物，不仅仅是衣物等生活用品，更可以直接买到要用的生鲜食材，互联网影响到人类的方方面面。随之而来，我们存留在互联网的数据量也是越来越大，比如说网上购物时下的订单信息，和人聊天的聊天记录和图片视频等等，还有其他的各种娱乐资讯新闻等等，互联网上的数据量，正在以一种无法想象的速度开始迅速增长着，据不完全统计：信息大爆炸以来，人们可以明显感受到大数据的来势凶猛。资料显示，目前每天全球互联网流量累计达1EB(即10亿GB或1000PB)，这意味着每天产生的信息量可刻满1.88亿张DVD光盘。其中，人们一天上传的照片数量相当于柯达发明胶卷后拍摄的图片总和。人们不知不觉中留下的“数据碎片”已经渗透到社会经济、大众日常生活的各个角落。**

**于是有人又问了，这么多的数据，我怎么才能从中找到自己想要的数据呢，我们的日常生活中有问题第一时间想要获取知识的来源途径，不再像以前没有互联网那样去问自己的长辈亲友，而是去百度一下，谷歌一下，搜搜一下等等，通过各种搜索引擎，我们输入关键字搜索引擎帮我们来完成数据的筛选。**

**而我要做的课题是关于搜索引擎中的一个小的部分，就是数据的收集和分析和一些删除，添加还有分析推荐的功能。**

**一些概念知识：**

**Big Data(大数据)**：或称巨量资料，指的是所涉及的资料量规模巨大到无法通过目前主流软件工具，在合理时间内达到撷取、管理、处理、并整理成为帮助企业经营决策更积极目的的资讯。

**网络爬虫**：为搜索引擎从万维网下载网页。一般分为传统爬虫和聚焦爬虫。

传统爬虫从一个或若干初始网页的URL开始，获得初始网页上的URL，在抓取网页的过程中，不断从当前页面上抽取新的URL放入队列，直到满足系统的一定停止条件。通俗的讲，也就是通过源码解析来获得想要的内容。

聚焦爬虫的工作流程较为复杂，需要根据一定的网页分析算法过滤与主题无关的链接，保留有用的链接并将其放入等待抓取的URL队列。然后，它将根据一定的搜索策略从队列中选择下一步要抓取的网页URL，并重复上述过程，直到达到系统的某一条件时停止。另外，所有被爬虫抓取的网页将会被系统存贮，进行一定的分析、过滤，并建立索引，以便之后的查询和检索；对于聚焦爬虫来说，这一过程所得到的分析结果还可能对以后的抓取过程给出反馈和指导。

**Docker**: 是一个开源的应用容器引擎，让开发者可以打包他们的应用以及依赖包到一个可移植的容器中，然后发布到任何流行的 Linux 机器上，也可以实现虚拟化。容器是完全使用沙箱机制，相互之间不会有任何接口。

**Hadoop:** Hadoop实现了一个分布式文件系统（Hadoop Distributed File System），简称HDFS。HDFS有高容错性的特点，并且设计用来部署在低廉的（low-cost）硬件上；而且它提供高吞吐量（high throughput）来访问应用程序的数据，适合那些有着超大数据集（large data set）的应用程序。HDFS放宽了（relax）POSIX的要求，可以以流的形式访问（streaming access）文件系统中的数据。

**YARN:**Apache Hadoop YARN （Yet Another Resource Negotiator，另一种资源协调者）是一种新的 Hadoop 资源管理器，它是一个通用资源管理系统，可为上层应用提供统一的资源管理和调度，它的引入为集群在利用率、资源统一管理和数据共享等方面带来了巨大好处。

**HDFS:** Hadoop分布式文件系统(HDFS)被设计成适合运行在通用硬件(commodity hardware)上的分布式文件系统。它和现有的分布式文件系统有很多共同点。但同时，它和其他的分布式文件系统的区别也是很明显的。HDFS是一个高度容错性的系统，适合部署在廉价的机器上。HDFS能提供高吞吐量的数据访问，非常适合大规模数据集上的应用。HDFS放宽了一部分POSIX约束，来实现流式读取文件系统数据的目的。HDFS在最开始是作为Apache Nutch搜索引擎项目的基础架构而开发的。HDFS是Apache Hadoop Core项目的一部分。

**HIVE:** hive是基于Hadoop的一个数据仓库工具，可以将结构化的数据文件映射为一张数据库表，并提供简单的sql查询功能，可以将sql语句转换为MapReduce任务进行运行。 其优点是学习成本低，可以通过类SQL语句快速实现简单的MapReduce统计，不必开发专门的MapReduce应用，十分适合数据仓库的统计分析。

**二、实现目标：**

爬虫引用创建的docker container，包括的模块有scrapy, mongo, celery, rabbitmq等，实现关键词删选，排序，分析，推荐等功能。

**三、体系框架：**

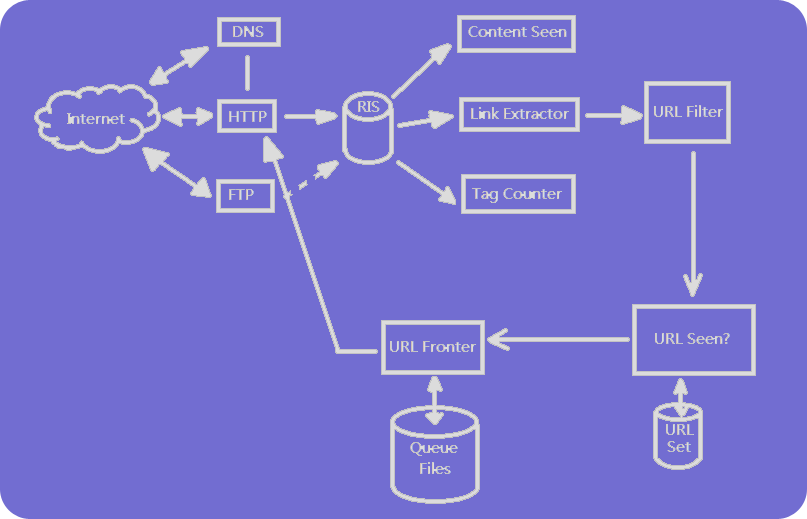
python -> crawler 用python实现爬虫（非应用爬虫）

crawler + docker -> crawler-node 将爬虫装入docker形成节点，并暴露一些端口口用于外部读写文件和与其他节点进行通讯

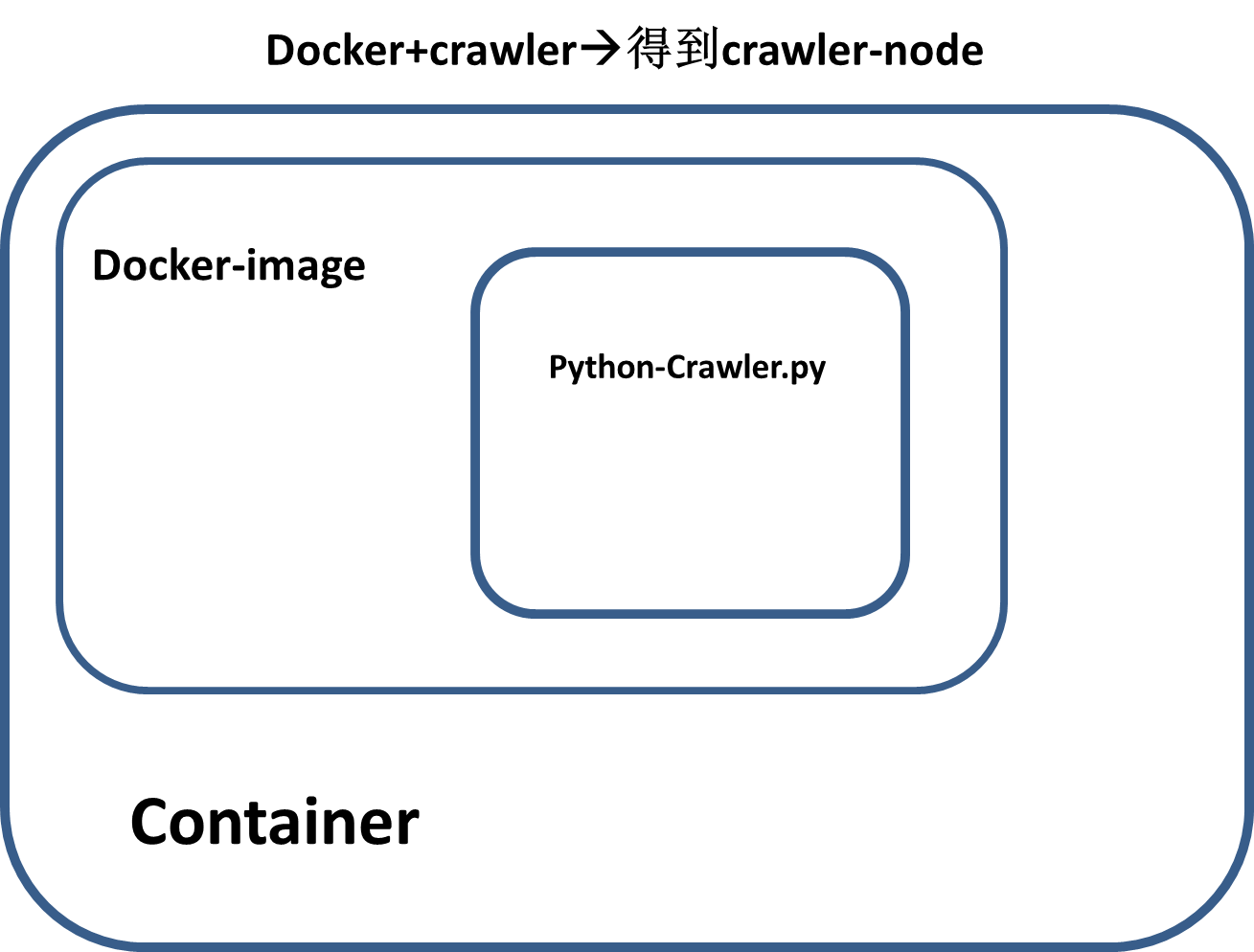
crawler-node + hdfs/hive(hadoop/spark/yarn) -> resSet 将爬虫得到的结果集合大数据平台的一些技术手段形成分析前的数据库

resSet + analysis -> action performance 数据分析后可以得到相类似的行为或者是类似的关键字结果集合，并依此形成推荐

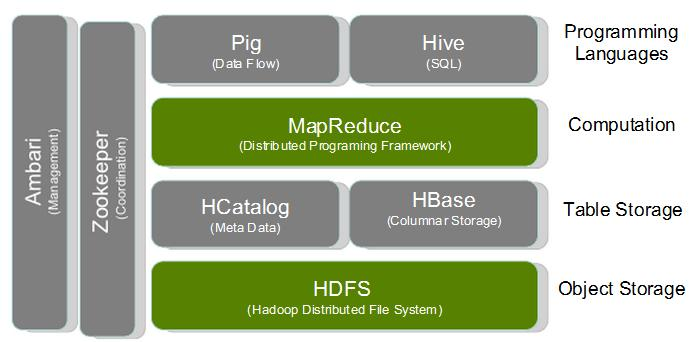
**四、实现路线图：**



Python实现**crawler**



节点部署概念图：



**五、参考文献：**

[1] docker与爬虫， https://github.com/LiuRoy/spider\_docker

[2] spark与爬虫, https://github.com/USCDataScience/sparkler

[3] Sanjay Ghemawat · Howard Gobioff · Shuntak Leung,The Google file system

[4]Jeon, Hocheol. "Lexicon-based context-sensitive reference comments crawler." Journal of Information Science 41.3 (2015): 342-353

[5] http://java-source.net/open-source/crawlers Open Source Crawlers in Java<br>

[6] http://en.wikipedia.org/wiki/Web\_crawler#Open-source\_crawlers Open-source crawlers