

需求分析

项目名称： 互联网新闻分类

项目类别： ☐ 电子商务

☐ 移动终端应用

☒ 大数据分析

☐ 物联网应用

☐ 人机交互应用

☐ 其他()

命题企业： 北京瑞德云网科技有限公司

咨询邮箱： zhangguangjun@ict.ac.cn

2017 年 12 月 1 日

项目需求分析

一、引言

1.1 项目背景

随着经济的发展，互联网已经成为人们生活的一部分，方便了生活的方方面面。互联网上的信息鱼龙混杂，切非常的庞大，想把新闻信息大致浏览是非常困难的，想在那么多的信息中提取自己感兴趣的信息是难上加难。每个人的关注点是不一样的，只关注自己感兴趣的信息变成了一个普遍的需求。有这样的需求，企业会针对这样需求做相应的工作，让用户自己定制自己的需要，有了用户个性化的需求才能对网络大量的新闻数据进行抽取，提取。这样就涌现了互联网新闻主题提取这样的模型分析。提交自己个性化的需要，经过模型的计算，推送给用户关注的信息。

1.2 项目目的

通过真实的互联网新闻主题提取应用案例开发，能够帮助学生快速的掌握大数据应用开发流程，了解当前互联网信息的应用方向，对学生的发展方向做一个指导作用，若发展方向是做互联网，这样的场景会是很多，本场景会对自己有很大的帮助。本案例是企业真实的大数据应用，一些根据用户定制定时推送新闻的应用浓缩。主要在数据

量和数据脱敏方面做了处理，旨在帮助同学掌握真实的大数据应用实现的案例。

二、项目需求

2.1 功能需求

2.1.1 数据录入

互联网新闻数据录入，互联网新闻数据是比较庞大的，新闻数据可以通过网络爬虫批量的爬取大量的数据，新闻数据要求时效比较强，越是实时的越有意义。把大量的新闻数据爬取到本地进行清洗汇总，存储到大数据集群。本项目采用的是已经收集好的互联网新闻数据文本，对整个大数据处理流程做分析的项目，数据已经准备好。上传到大数据集群的 HDFS 文件系统中即可完成数据的录入。

2.1.2 数据处理

对现有的网络新闻数据进行清洗处理，最终处理成结构化的数据储存到 HDFS 文件系统中，提供数据分析提取分析。

2.1.3 数据分析

对清洗过后的新闻数据进行分析，根据现有的数据字段属性情况，可以对互联网新闻数据做主题提取分析，将不同的主题和对应的关键词抽取出来，对新闻正文内容做分词，创建词库，提取主题，做主题

建模，然后对结果数据进行模拟展现。

2.1.4 数据展现

使用 Fire.Fox 交互式的展现提取的主题信息，可以选择查看具体的信息情况，完全根据用户的个性信息展现。

2.2 性能需求

2.2.1 可扩展性

大数据集群可以快速无缝的横向扩展，数据达到一定的规模之后不需要担心数据承载出现问题，加存储或者服务器即可完成集群规模的横向扩展。数据会根据集群的情况合理调整调度，尽可能的合理使用资源。

2.2.2 稳定性

采用的是大数据领域主流的组件进行开发，每个技术都是非常领先。大数据处理平台采用的是高可用，数据副本机制保障数据的安全。并且在行业已经运用到各行各业中，数据的规模也是在 PB 级别，社区活跃度非常高，在运维和开发成本上相对较低，可持续性较强。

2.3 任务要求

2.3.1 大数据平台搭建

能通过安装文档个人完整的把 CRH 技术平台搭建完成，并且正常

运行，为后续的案例开发实验做基础。

2.3.2 数据接入

独立把数据上传至 HDFS 分布式文件系统，提供给 Dataiku 数据分析工具进行数据分析使用。

2.3.3 数据分析

使用 Dataiku 直接读取 HDFS 文件系统中的数据，抽取数据，对新闻数据进行分词入库，提取主题信息分析，把抽取的新闻主题做交互式展现。

三、运行环境

3.1 软件环境

服务器操作系统：RedHat 或者 CentOS（英文版）

服务器操作系统版本：RedHat7 或者 CentOS7

JDK 版本：Oracle1.8

CRH 版本：CRH5.1

分析工具：Dataiku

3.2 硬件环境

推荐测试环境：

内存：8G

存储：100G

CPU：双核处理器

推荐生产环境：

内存：128G 或者 256G

存储：服务器满配每块盘 3T 或者 4T

CPU：48 核

3.2 网络环境

每台服务器或者操作系统之间能相互连通，有时间同步服务器，终端能连接上即可。

四、实现过程

4.1 实现思路

- 数据接入存储为原始数据
- 原始数据提供数据分析使用
- 对分词词库做主题提取分析
- 对提取结果住交互式展现

4.2 实现技术

4.2.1 HDFS

大数据分布式文件存储，实现数据的存储保证数据的安全，HDFS

文件系统的容量可以横向的扩展。

4.2.2 DATAIKU

企业级客户提供基于云技术的大数据服务分析平台，数据分析工程师可以很简单的完成数据的收集，分析，展现。

4.3 实现计划

4.3.1 CRH 环境搭建

使用 CRH5.1 搭建大数据基础平台，搭建过程详见 CRH 安装手册

4.3.2 数据录入

把本地测试数据 PUT 到 CRH 大数据集群的 HDFS 文件系统中。

4.3.3 数据分析

使用 Dataiku 连接 CRH 大数据平台 HDFS 文件系统中，抽取数据，使用 LDA 算法进行海量文本主题建模分析。

4.3.4 数据展现

对抽取出的主题及对应关键词做交互式可视化展现。