**叶峰**

****

手 机：+86-15000475769

电子邮件：1102824396@qq.com

Github： https://github.com/iYefeng

住 址：上海市灵石路735弄14楼101室，200072

籍 贯：上海

健康状况：健康

* **工作经历**

2012-3 ~ 2012-6： 北京联想研究院 实习生

|  |  |
| --- | --- |
| 1 | 开发基于curse的应用程序 |
|  | 根据需求，需要开发一个shell下的**可视化应用程序**，需要有选择按钮，输入框，进度条，滚动条，翻页等功能。采用的是curse库来开发。**该项目的难点是**：curse库目前已经很少有人使用，它的API说明网上很难找到，所以只能通过man一条一条的查询。 |

2015-2 ~ 至今 上海晶赞科技发展有限公司 hadoop数据分析师

|  |  |
| --- | --- |
| 1 | ID图谱 —— 数据分析(采用hadoop)；负责serving（负责id mapping线上服务） |
|  | * 分析运行商和RTB数据（大约8T/天），根据多维度（是否在同一个ip，出现的同时性，浏览网站的相似度，ID间的自身关联），通过一套概率打分机制，将不同的用户ID进行关联。**该项目的难点在于**，数据量巨大，需要同时分析一周或更长时间段内的数据，所以，需要很好的对mr程序进行优化。例如，数据的前期清洗，存储中间数据，按小时均匀分摊任务的负载，以及如何避免outofmemory的问题。 * 关联好的ID关系，需要推送到ID关联服务上。服务采用C++开发，使用thrift协议，单机qps 30k/s，可分布式运行。 |
| 2 | 知识库 —— 垂直行业网站爬虫开发与维护；数据抽取，清洗和入库程序开发 |
|  | * 根据需求，需要爬取的行业多样（初步选择8个行业，电商，酒店，旅游，汽车，视频，房产，poi，保险），爬取网站丰富，内容差异巨大。所以最终选择使用pyspider作为基础开发。在pyspider基础上，增加了代理池和redis爬取任务队列功能。 * 抽水机开发：开发了基于java的**可快速配置**的对爬取后数据进行“按需抽取字段，**清洗**，和最终入库”的通用定时程序。**该任务的一个创新点是**，清洗采用了openrefine[[1]](#footnote-1)的规则，用户可以先在openrefine中对小部分数据进行清洗，清洗结果满意以后，可以从openrefine中导出清洗规则，放到抽水机配置中，抽水机可以根据清洗规则，清洗数据。 |
| 3 | zampSDK —— 主要负责数据分析，例如，归因分析，反作弊，行为分析等 |
|  | * 对zampsdk[[2]](#footnote-2)进行维护与开发。归因分析，反作弊，行为分析等模块全部采用spark重构，将原来复杂的逻辑简化，同时根据需求，增加了多维度(增加了8个维度：地域，运营商，版本，操作系统，设备，推广频道，事件，子事件)的关联分析。**该该任务的难点在于**：原先数据分析采用mr程序，逻辑极其复杂，并且很多程序中采用了多路输出，导致代码阅读困难，修改更困难。**该任务的创新点**：反作弊模块，采用了自动机实现。保证了以后可以很容易地增加规则。 |
| 4 | 电信研究院“悦me网关数据监控”项目 |
|  | * 对网关数据（url，cookie， user-agent，refer）中的有用信息进行提取。提取维度有（ID，电商，汽车，酒店，视频）。采用C++开发了提取程序内核，只要输入提取规则，就可以从url，cookie， user-agent，refer中提取信息，并实现了python，java的api调用，单核 qps 27k/s。**该任务的难点**：提取规则量大（大于3万条规则），并且多采用正则表达式，所以对每次提取，需要对规则进行合理的筛选，不能采用枚举。考虑到提取信息中可能会有脏数据，所以，提取规则中要可以设置数据清洗过滤功能，相应的，程序中也要实现这个功能。User-agent提取采用ac自动机实现。 * 提取数据的多维度统计分析。采用mr和spark实现，需要按照合同要求，实现多维度的关联分析，并将统计结果写入mysql。 |

* **掌握技能**
  + 熟悉mapreduce，spark，hive，hbase
  + 熟悉mysql，mongodb，redis
  + 熟悉python，C++，java，scala
  + 熟悉protobuf，thrift
  + 熟悉 lr，svm，pca，ac自动机，tire树，tf-idf，cf，random forest等算法（https://github.com/iYefeng/traits/tree/master/algorithm）
  + 熟悉pyspider，scrapy
* **教育经历**

2012-9 ~ 2015-3 北京航空航天大学 硕士学位（**管理科学与工程）**

* **全年级综合第1名** GPA 3.5/4
* 运筹学，统计学，计算机编程(C/C++, python)

2008-9 ~ 2012-6 北京航空航天大学 学士学位（**工业工程）**

* **全班第2名** GPA 3.6/4

2005-9 ~ 2008-6 浙江省慈溪中学

* **获得奖励**

**2014-12 北京市优秀毕业生 / 2013-10 研究生国家奖学金** / **2013-9 一等综合奖学金** / 2014-5 北航研究生十佳提名奖 / 2012-9 优秀志愿者称号 / 2012-7 北航优秀毕业生 / **2010-11 全国大学生数学建模竞赛全国一等奖** / 2009-6 全国大学生数学竞赛三等奖 / 2010-11 北京航空航天大学学科竞赛奖学金一等奖 / 2009-11 北京航空航天大学优秀生称号 / 2010-4 北京航空航天大学物理竞赛一等奖 / 2009-3 北京航空航天大学数学竞赛二等奖 / 2010-10 北京航空航天大学学习优秀奖学金 / 2010-12北京航空航天大学校三好学生的称号

1. 一个 google开源的数据清洗工具 [↑](#footnote-ref-1)
2. 一个类似于talkingdata的移动app追踪监测服务。 [↑](#footnote-ref-2)