基于贵州交通行业网络舆情监测平台

开发计划书

2016年8月

修改日志

|  |  |  |
| --- | --- | --- |
| 修改日期 | 修改人 | 修改内容 |
| 2016年8月16日 | 张杰 | 建立文档 |

目录

[项目背景 4](#_Toc459216529)

[项目愿景 4](#_Toc459216530)

[本文档的预期读者 4](#_Toc459216531)

[系统总体架构 5](#_Toc459216532)

[系统工作量预估与目前完成情况 6](#_Toc459216533)

[项目面临的困境 7](#_Toc459216534)

# 项目背景

贵州交通部门为了能够实现对交通行业的舆论进行分析、监控、引导以及跟踪而建立贵州交通行业舆情监控系统(后称系统)。系统主要依托目前的与计算以及大数据分析技术，采用爬虫技术作为基础，提供可展示，可配置，可配置、高性能以及安全健壮的舆情监控服务、舆情分析服务。

# 项目愿景

系统主要是为贵州省交通的主管部门、交通行业的运营者以及参与提供对各类与交通行业相关的网络舆情进行采集、分析、预防、跟踪为一体立体服务，实现从采集到展示的整个过程。系统通过对各类消息在网络中传播的特征、模式以及途径进行分析，建立跟分析相关的采集任务，通过对采集的数据进行清洗以及入库，使用大数据的手段对清洗后的数据进行聚类分析于关联分析，提取出当前贵州的交通行业相关的企业和主管部门在网络上的影响。从而为贵州交通行业的主管部门和企业决策者提供可参考的权威信息和知识，促始贵州交通行业健康快速地发展。

# 本文档的预期读者

本文档的预期读者主要有:公司的领导、网络舆情项目组、以及其他相关的人员。

# 系统总体架构

系统的总体架构主要采取分层架构的方式进行展示。系统的各层的组件之间主要采取数据共享的方式来完成数据的流转和分析过程。

舆情分析结果展示服务

舆情分析规则管理服务

分析算法管理服务

采集规则管理服务

分析数据展示服务

分析数据展示平台

分析数据管理

网络舆情数据分析平台

数据分析算法管理

分析数据入库管理

数据采集任务管理

分析任务调度管理

舆情分析规则管理

数据采集规则管理

分析数据装载管理

网络爬虫

采集调度

采集规则

库

数据清洗入库

网页解析

采集数据处理

根据系统的层级系统架构，可以看出，系统包括从数据采集到数据分析以及展示的全部模块。系统的展示成通过配置分析规则、数据采集规则以及进行任务调度，数据分析平台通过数据采集任务调度通知采集层进行相应的数据采集调度，采集层采集数据之后进行网页的解析，数据解析完成后按照数据采集规则进行数据处理，数据处理完成之后进行清洗入库。数据分析平台根据数据舆情数据仓库现有的数据，数据分析规则以及分析算法对采集的数据进行数据的加工处理，最后将分析结果进行入库，展示成通过数据查询根据UI需求进行数据展示，并提供相应的数据分析规则、采集规则以及算法等一系列的管理服务。

# 系统工作量预估与目前完成情况

根据对系统关键需求的分析和数据分析业务模式，系统在业务层次上主要划分为三个层次：

1. 数据采集层
2. 数据分析层
3. 数据展示层

各层的主要模块可参考系统的总体架构。

目前系统的总体开发进度如下表

|  |  |  |
| --- | --- | --- |
| 层级名称 | 开发进度 | 备注 |
| 数据采集层 | 目前完成网络爬虫部分的开发，能够从网络上下载相关的业务，但是对网页的解析和规则管理以及数据的入库没有完成，数据的初步清洗没有启动，整个层级的完成的工作量大约在20%左右。 |  |
| 数据分析层 | 没有启动 |  |
| 数据展示层 | 没有启动 |  |

系统的总体工作量预估如下表

|  |  |  |  |
| --- | --- | --- | --- |
| 层级名称 | 预估工作量 | 预估代码量 | 备注 |
| 数据采集层 | 6个人月 | 3万到5万行 |  |
| 数据分析层 | 20个人月 | 5万到8万行 |  |
| 数据展示层 | 4个人月 | 3万到5万行 |  |

系统目前人力资源配置

|  |  |  |  |
| --- | --- | --- | --- |
| 层级名称 | 人员配置 | 模块划分 | 备注 |
| 数据采集层 | 张杰 | 规则管理，数据清洗 |  |
| 杨城 | 网页解析，采集规则库管理 |
| 马兴江 | 网络爬虫，采集任务调度 |
| 数据分析层 | 无 |  |  |
| 数据展示层 | 无 |  |  |

项目发布版本规划

第一版本发布：2016年10月份

# 项目面临的困境

1. 数据分析层技术选型问题。
2. 数据展示层的技术选型问题。
3. 各层之间的数据共享问题。
4. 数据分析平台选择问题。