算法描述和软件设计说明书

### 一、题目分析及解题思路

1. **实现互联网事件天级或更高频次抓取，能够采集政治会议、展会、体育赛事、演唱会、突发异常天气等各类中比较主要的事件，每个类型的事件数据至少来自2个数据源（网站）。**

|  |  |  |
| --- | --- | --- |
| 数据类型 | **数据源（网站）** | |
| 展会 | E展会 | http://www.eshow365.com/ |
| 中国会展门户 | http://www.cnena.com/ |
| 体育赛事  演唱会 | 大麦网 | https://www.damai.cn/ |
| 永乐票务 | http://www.228.com.cn/ |
| 突发异常天气 | 中国天气 | http://www.weather.com.cn/ |
| 政治会议 | 活动家 | https://www.huodongjia.com/ |
| 航班信息 | 飞常准 | http://www.variflight.com/ |

**2.实现事件的去重功能，主要有两个方面，一是不同数据源（网站）的事件去重，二是不同天抓取的事件去重。**

**去重算法：**本项目最初采用相似度阈值去重法，但因大部分展会、演唱会等名称相似但实则不同的信息存在被误删的现象，如：第X届、XX站等。最终采取近义词匹配去重法，结合时间及地理位置信息，有效避免了大量信息误删的情况。

**3.事件画像建模，即事件属性自动化提取。属性需包括主办方级别、影响区域范围、影响力度、影响人群等。**

在“设置>>下载数据”中选择要下载的文件类型，服务器将自动提取事件属性并生成相应文件。

**属性自动化提取算法：**将不同种类事件采用不同规则筛选出部分属性的匹配项。例如，主办方包含“中华人民共和国”或“中国”则为国家政府组织，包含“世界”或 “全球”则为国际政府组织。

**4.可视化展现，可视化展示事件的影响情况，形式不限。**

**主视图**采用百度地图清晰准确的展示热点的地理位置信息，并提供可交互的动态展示详细信息。

**统计视图**采用echarts图表工具展示，包括：

**分时地区**统计视图：根据用户选择的时间动态生成改时间段内各省份的各种数据总量，并展示在小型中国地图中。

**全局日期**统计视图：把数据库中所有今年的数据，按日期分类统计，并生成一个3D视图，其中x轴为日期，y轴为月份，z轴为数量。

**局部日期**统计视图：把该场馆在该时间段内的所有类型数据的开始时间展示到二维日历视图中。

### 二、软件支持

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 开发环境 | | Window 10、MyEclipse 10、Tomcat 6、Java EE 6、MySQL 5.5 | | |
| 实验平台 | | CentOS 6.5、Java JDK 1.8、Tomcat 8、MySQL 5.6 | | |
| 开发语言 | | Java、JavaScript、HTML、CSS | | |
| 第三方代码 | 后端 | Jsoup | 1.8.3 | https://github.com/jhy/jsoup |
| Xsoup | 0.3.1 | https://github.com/code4craft/xsoup |
| Log4j | 1.2.17 | https://github.com/apache/log4j |
| Poi | 3.16 | https://github.com/apache/poi |
| CrawlerSYS | 0.1.0 | https://github.com/zrtzrt/CrawlerSYS |
| 前端 | Bootstrap | 3.3.7 | https://github.com/twbs/bootstrap |
| Echarts | 3 | https://github.com/ecomfe/echarts |
| BaiduMap | 2 | http://map.baidu.com |

### 三、软件结构

**CrawlerSYS**

分布式爬虫相关组件，参见《CrawlerSYS设计文档》。

**Dao**

数据持久化层，用于与数据库通信，分为接口和实例。依次包括：地理位置信息、其他活动信息、展会信息。

**DataSource**

自定义爬虫，从不同数据源中使用不同方法爬取不同结构的数据，具体请参见《数据库结构》。依次包括：飞常准上的航班信息、全部站点爬虫启动器、大麦网上的场馆信息、活动家上的会议信息、永乐票务上的活动信息、大麦网上的活动信息、中国会展门户上的展会信息、E展会上的展会信息、中国天气上的天气预警信息。

**Entity**

各个类间、与数据库间通信使用的实体类。各结构如下：

Location：int id;int type;String name; String city; String province;String detail;Double longitude; Double latitude;

Show：int id;String name;String picture;Date startDate;Date endDate; LocationEntity locat; int MINprice;int MAXprice;int showtype;String info; List<String> url;

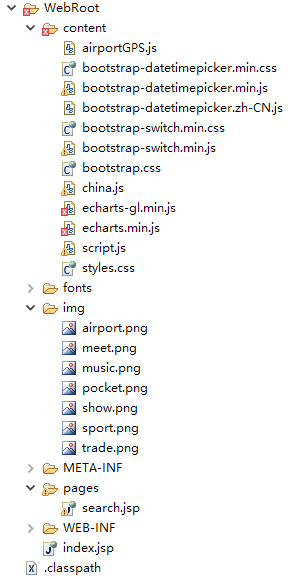
Trade：int id;String name;String en\_name;String picture;Date startDate;Date endDate; LocationEntity locat;String industry;String host;int area;int times;String frequency;String used;List<String> url; int hot;

Out

从数据库中未过期数据自动化提取事件属性，并生成符合要求格式的相应文件。

Servlet

控制爬虫功能及初始化爬虫功能的服务器和查询符合所有条件的地理位置信息服务器。

Content

可视化界面的额外文件，依次包括：所有机场的地理位置信息、选择日期的样式、选择日期的动画和功能、选择日期的中文支持、开关的样式、开关的动画及功能、自定义的bootstrap样式、小型中国地图的数据、3D图表的功能、普通图表的功能、主要脚本文件、主要样式文件。

Img

所有图标文件，依次包括：机场、会议、音乐、娱乐、体育赛事、展会。

Pages

顶部搜索栏JSP页面。

### 四、数据库结构

