**SINA 科技热词可视化展示**

**一 设计目标**

1 服务器上设置爬虫定时任务，借助jsoup工具从新浪科技专题爬取当前页面的相关内容，将每个链接对应的内容以文本文件的格式保存到数据库表中。

2 服务器上设置分词定时任务，读取数据库中爬虫记录，使用分词工具ansj对文本内容进行切分，汇总每个词的出现次数，生成<名词，出现次数>键值对，生成的<名词，出现次数>键值对插入到数据表当中。

3 服务器端设置访问接口，按照出现次数多少对数据表中分词进行排序，选择出top N的热词，返回json数据。

4使用Ajax和Echarts，访问服务器，可视化显示词频。

**二 工作流图**

**三 设计与编码**

**1 本项目分成两个部分完成。**

第一部分：网页内容的爬取，分词，入库和热词接口提供。

第二部分：借助ajax和echarts进行可视化展示。

**2 实现**（具体参见代码）：

CrawService.java：使用jsoup对新浪科技（<http://tech.sina.com.cn/>）首页进行爬取，设置一定的爬取规则，将爬取的内容存数据库。

LoadDbService.java：读取数据库中爬虫信息，分词处理，生成键值对，并插入到MySQL数据表中。

WordCountController.java：服务器访问接口，查询数据库中词频信息，并在返回json数据。

index.html：ajax访问数据服务端，echarts可视化的数据。

**四 MySQL**

**1 创建词频表：**

CREATE TABLE `word\_count` (

`id` bigint(20) NOT NULL AUTO\_INCREMENT,

`word` varchar(255) NOT NULL,

`count` int(11) NOT NULL DEFAULT '0',

`create\_time` timestamp NOT NULL DEFAULT CURRENT\_TIMESTAMP,

PRIMARY KEY (`id`),

KEY `uk\_word\_time` (`word`,`create\_time`) USING BTREE COMMENT '单词和时间处理',

KEY `idx\_count` (`count`) COMMENT '词频索引',

KEY `idx\_create\_time` (`create\_time`) USING BTREE COMMENT '创建时间添加索引方便查当天数据'

) ENGINE=InnoDB DEFAULT CHARSET=utf8;

**2 选择热词**

select word, count from word\_count where date(create\_time)=#{day} ORDER BY count desc limit 0,50;

**五 效果展示**

