技术报告：

基于主页标题文字的
热点新闻跟踪与分析、可视化展示

开发小组

2014 年 6 月

技术报告

1 需求分析

目前新闻网站内部的新闻有比较好的分类，可以比较容易的收集关于同一个事件的相关新闻。但是收集不同门户网站的相关新闻就显得比较困难。而用户希望得到多个网站相关的新闻信息。所以将门户网站的新闻进行聚类整理就显得十分重要。

同时，如果不同门户网站同时出现了相关新闻，也能反映出当前社会的热点。对门户网站新闻进行聚类也有利于社会热点的研究，如果有大量历史数据的数据库，则在社会研究上也有较好的附加效益。

2 模块介绍

实现这些功能需要的步骤如下：

- 从各种门户网站抓取标题数据
- 在数据库存储有关数据
- 用户提交查询的需求
- 算法对相关标题数据进行聚类
- 将结果在界面上展示

2.1 网页抓取

本部分集成了 html 和 json 解析, 能够适应大部分的 web 页面和 ajax 请求数据解析, 尽可能多的支持各个新闻网站. 所有 Spider 继承自 SpiderBase, 并可选继承 GumboBasedSpider 或 JsonBasedSpider 以获取 HTML 及 JSON 解析功能.

主要类的设计如下:

- SpiderBase
  所有 Spider 的基类, 提供网页抓取的基础功能 (如 HTTP 请求, 编码转换)
- GumboBasedSpider
  提供基于第三方库 gumbo 的 html dom 树解析, 类中声明了两个纯虚的回调函数, 留给子类实现, 以适应不同网站的不同网页结构. 其中 searchNodeCallback 在递归遍历 dom 节点时调用, 由子类判断节点是否有用, 是否需要深入遍历;resultCallback 在遍历先前存储的节点时, 有子类进一步过滤出需要保留的信息.
- JsonBasedSpider
  提供对于 json 格式数据的解析, 基于第三方库 JSON++.

- SpiderFor163,SpiderForQQ,SpiderForSina
  具体 Spider 实现, 提供对于相应门户网站新闻主页的抓取
- SpiderFor163,SpiderForSinaRoll,SpiderForQQRoll
  具体 Spider 实现, 提供对于相应门户网站滚动新闻的抓取

2.2 数据存储

为了提供历史新闻跟踪分析功能, 需要存储历史抓取数据, 因此设计了数据存储部分. 为了快速查找, 这部分使用数据库作为后端存储. 目前选用的轻量级的关系型数据库 sqlite.

主要类的设计如下:

- StorageBase
  提供与数据库引擎无关的数据操作接口, 使主程序与数据库解耦. 目前, 仅实现了插入和按日期查找的接口.
- SqliteDatabaseStorage
  继承自 StorageBase, 实现接口数据类型向 sqlite 数据库的类型转换, 以及 sql 语句构造等功能.

2.3 核心算法

核心算法部分涉及 include/tracker/和 src/tracker/目录下的所有文件。为了方便中文处理，核心算法部分全部使用 C++11 中的 u32string。

- title.h 声明了标题类 title 和聚类标题集 titlebundle，用于与外部通信。
- tracker.h 声明了聚类器的基类 tracker，其中 trackFocus 的纯虚函数以 vector<title> 作为输入，以 vector<titlebundle> 作为输出。
- trackerForce.h, trackerForce.cpp 声明并定义了一个基于单个字的聚类器 trackerForce，继承自 tracker。
- trackerCluster.h, trackerCluster.cpp 声明并定义了一个基于分词和 K-means 算法的分类器 trackerCluster，继承自 tracker。
- trie.h 声明并定义了一个字典树的模板 trie，trie<u32string> 在 trackerCluster 中被使用。
- segment.h segment.cpp 声明并定义了一个基于双向最大匹配的分词算法，在 trackerCluster 中被使用，输入是 u32string，输出是 vector<u32string>。

2.4 用户交互

UI 部分采用 HTML+JS 方式开发, 并在程序中使用 libmicrohttpd 提供内置 http 服务器.

前端采用 jQuery 框架, 并使用 jquery.awesomeCloud 插件显示词云.

3 目录结构

```
├———3rdparty 第三方库目录
|       ├———JSON++
|       └———gumbo-parser
├———bin 目标文件
├———documentation 文档目录
├———include 所有头文件
|       ├———spider
|       ├———storage
|       ├———tracker
|       └———web
├———scripts 工具脚本
├———src 源码目录
|       ├———spider 爬虫部分
|       ├———storage 存储部分
|       ├———tracker 核心算法
|       └———web 网页服务器
├———testdata 算法测试数据
└———tests 单元测试程序
```

4 编译方法

在 Mac OS X 系统下, 安装如下依赖包:

brew install curl libmicrohttpd pcre sqlite cmake autoconf automake bison flex libtool pkg-config

进入顶层目录编译:

make