舆情站点爬取

1. 类别：

建立字典表，记录所有的类别和网站的编号（如图1.1所示）。

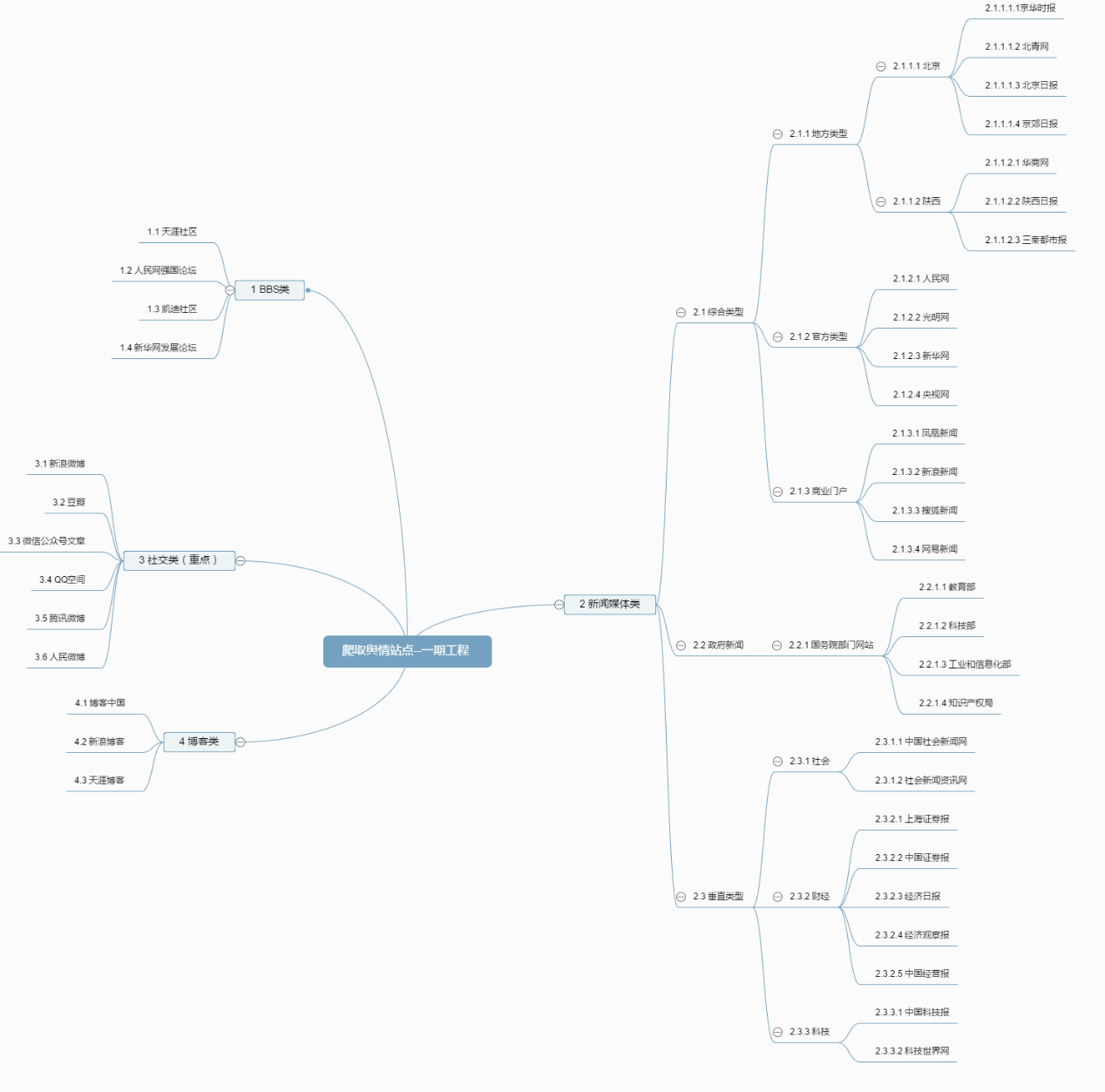


图1.1

1. 每个类别代表网站分析：
2. BBS类：天涯网（略）
3. 新闻类：
   1. 综合类型：
      1. 地方类型：三秦报
4. 爬取内容：

新闻内容：

|  |  |  |
| --- | --- | --- |
| 序号 | 名称 | 字段 |
| 1 | url | url |
| 2 | 一级类别 | first\_category |
| 3 | 二级类别 | second\_category |
| 4 | 标题 | title |
| 5 | 发表时间 | publish\_time |
| 6 | 摘要 | abstract |
| 7 | 来源 | source |
| 8 | 图片 | pics |
| 9 | 内容 | content |
| 10 | 标签 | tags |

图片：

|  |  |  |
| --- | --- | --- |
| 序号 | 名称 | 字段 |
| 1 | 图片ID | pic\_id |
| 2 | 图片URL | Pic\_url |
| 3 | 图片 | picture |
| 4 | 图片描述 | pic\_desc |

专题：

|  |  |  |
| --- | --- | --- |
| 序号 | 名称 | 字段 |
|  | 专题名称 | special\_name |
|  | 专题描述 | special\_desc |
|  | 专题发布时间 | special\_publish\_time |
|  | 新闻 | news |
|  | 图片 | pics |

1. 爬取流程：

一般新闻：



专题新闻：



* + 1. 官方类型：新华网

1. 爬取内容：

普通新闻

|  |  |  |
| --- | --- | --- |
| 序号 | 名称 | 字段 |
| 1 | URL | url |
| 2 | 新闻类别 | news\_category |
| 3 | 新闻标题 | news\_title |
| 4 | 发表时间 | publish\_time |
| 5 | 来源 | source |
| 6 | 内容 | content |
| 7 | 图片 | picture |
| 8 | 视频 | video |
| 9 | 责任编辑 | editor |
| 10 | 评论 | comment |

评论：

|  |  |  |
| --- | --- | --- |
| 序号 | 名称 | 字段 |
| 1 | 用户ID | comment\_userID |
| 2 | 用户昵称 | comment\_nick |
| 3 | 地点 | comment\_address |
| 4 | 时间 | comment\_time |
| 5 | 内容 | Comment\_content |
| 6 | 顶的数量 | ding\_num |

专题：

|  |  |  |
| --- | --- | --- |
| 序号 | 名称 | 字段 |
| 1 | 专题名 | special\_name |
| 2 | 介绍 | special\_desc |
| 3 | 时间 | special\_time |
| 4 | 图片 | special\_pic |
| 5 | 视频 | special\_video |
| 6 | 新闻 | special\_news |

1. 爬取流程：

普通新闻：



专题新闻：



* 1. 政府国务院网站：教育部

1. 爬取内容：

文件公告：

|  |  |  |
| --- | --- | --- |
| 序号 | 名称 | 字段 |
| 1 | 爬取路径 | Spider\_path |
| 2 | 网页位置 | position |
| 3 | 信息名称 | info\_title |
| 4 | 信息索引 | info\_index |
| 5 | 生成日期 | info\_date |
| 6 | 发文机构 | info\_institution |
| 7 | 发文字号 | document\_num |
| 8 | 信息类别 | info\_category |
| 9 | 内容概述 | info\_abstract |
| 10 | 内容 | Info\_content |
| 11 | 附件 | file |
| 12 | 二维码 | QRCode |
|  | 文章URL | url |

新闻动态：

|  |  |  |
| --- | --- | --- |
| 序号 | 名称 | 字段 |
| 1 | 爬取路径 | path |
| 2 | 当前位置 | position |
| 3 | 标题 | news\_title |
| 4 | 时间 | News\_time |
| 5 | 来源 | News\_source |
| 6 | 责任编辑 | News\_editor |
| 7 | 内容 | News\_content |
| 8 | 图片 | Picture |
| 10 | 附件 | file |
| 11 | 二维码 | QRCode |

数据：

|  |  |  |
| --- | --- | --- |
| 序号 | 名称 | 字段 |
| 1 | 爬取路径 | path |
| 2 | 当前位置 | position |
| 3 | 标题 | data\_title |
| 4 | 数据原始页面 | Data\_content |
| 7 | 备注 | remarks |
| 8 | 责任编辑 | editor |
| 9 | 附件 | file |
| 10 | 二维码 | QRCode |

1. 爬取策略：

从主页（http://www.moe.gov.cn/）开始爬取，爬取页面中的所有链接，如果该链接是非含有以上三种类型信息的网页（网页不以.html结尾或者以index[\_]?[\s]\*.html结尾），则记录该连接的文字信息，并跟进，若是含有以上三种类型信息的网站，则根据页面结构去爬取相应的信息。



* 1. 垂直类型：中国社会新闻网

1. 爬取内容：

|  |  |  |
| --- | --- | --- |
| 序号 | 名称 | 字段 |
|  | 新闻类别 | category |
|  | 新闻标题 | Title |
|  | 发布时间 | publish\_time |
|  | 来源 | source |
|  | 内容 | content |
|  | 编辑 | editor |
|  | 图片 | picture |

1. 爬取规则



1. 社交类：新浪微博
2. 爬取内容：

用户：

|  |  |  |
| --- | --- | --- |
| 序号 | 名称 | 字段 |
| 1 | 用户id | user\_id |
| 2 | 用户头像 | user\_image |
| 3 | 昵称 | nickname |
| 4 | 是否加V以及加V类型 | is\_v |
| 5 | 地址 | address |
| 6 | 性别 | sex |
| 7 | 生日 | birthday |
| 8 | 关注用户的数量 | concern\_num |
| 9 | 粉丝的数量 | fans\_num |
| 10 | 博客的数量 | blog\_num |
| 11 | 等级 | level |
| 12 | 简介 | desc |
| 13 | 标签 | tags |

博客表：

|  |  |  |
| --- | --- | --- |
| 序号 | 名称 | 字段 |
| 1 | 博客ID(mid) | mid |
| 2 | 用户ID | user\_id |
| 3 | 内容 | content |
| 4 | 图片 | picture |
| 5 | 链接 | links |
| 6 | 视频链接 | video\_links |
| 7 | 发表时间 | time |
| 8 | 话题 | topic |
| 9 | @的用户 | at\_user |
| 10 | 赞的数量 | praise\_num |
| 11 | 评论的数量 | comment\_num |
| 12 | 转发的数量 | forward\_num |
| 13 | 转发的微博id | forward\_id |

评论表：

|  |  |  |
| --- | --- | --- |
| 序号 | 名称 | 字段 |
| 1 | 评论ID | comment\_id |
| 2 | 博客ID | blog\_id |
| 3 | 评论用户ID | comment\_user\_id |
| 4 | 内容 | comment |
| 5 | 图片 | comment\_picture |
| 6 | 赞的数量 | comment\_praise\_num |
| 7 | 回复的用户 | replay\_user |
| 8 | 时间 | comment\_time |

粉丝关注表

|  |  |  |
| --- | --- | --- |
| 序号 | 名称 | 字段 |
| 1 | 关注的人 | focuse\_id |
| 2 | 粉丝 | fans\_id |

1. 爬取规则：



1. 博客类：博客中国
2. 爬取内容：

博客：

|  |  |  |
| --- | --- | --- |
| 序号 | 名称 | 字段 |
| 1 | 标题 | title |
|  | 子标题 | sub\_title |
| 2 | 作者 | author |
| 3 | 作者ID | author\_ID |
| 4 | 发表时间 | publish\_time |
| 5 | 所属栏目 | category |
| 6 | 阅读量 | read\_num |
| 7 | 评论量 | comment\_num |
| 8 | 点赞量 | hand\_up\_num |
| 9 | 点差量 | hand\_down\_num |
| 10 | 赏的数量 | admire\_num |
| 11 | 文本内容 | content |
| 12 | 图片 | picture |
| 13 | 链接 | link |
| 14 | 文章URL | URL |

作者：

|  |  |  |
| --- | --- | --- |
| 序号 | 名称 | 字段 |
| 1 | 作者名 | author\_name |
| 2 | 作者的博客名 | author\_blog\_name |
| 3 | 简介 | introduce |
| 4 | 头像 | image |
| 5 | 文章数 | article\_num |
| 6 | 浏览数 | read\_num |
| 7 | 粉丝数 | fans\_num |
| 8 | 主页URL | main\_page\_URL |

评论：

|  |  |  |
| --- | --- | --- |
| 序号 | 名称 | 字段 |
| 1 | 用户ID | comment\_user\_ID |
| 2 | 博客ID | comment\_blog\_ID |
| 3 | 评论时间 | comment\_time |
| 4 | 评论内容 | comment\_content |
| 5 | 赞的数量 | praise\_num |
| 6 | 评论回复ID | reply\_ID |

粉丝关注：

|  |  |  |
| --- | --- | --- |
| 序号 | 名称 | 字段 |
| 1 | 关注 | focuse |
| 2 | 粉丝 | fans |

专题：

|  |  |  |
| --- | --- | --- |
| 序号 | 名称 | 字段 |
|  | 专题名 | special\_name |
|  | 专题相关信息介绍 | special\_info |
|  | 视频 | special\_video |
|  | 图片 | picture |
|  | 帖子 | blogs |

1. 爬取规则：

单个博客爬取：



博客专题爬取：



在进行单个博客爬取更新时，只需要根据时间从最新文章中爬取更新，若该文章的作者以前没有被怕取过，增加该作者的信息。专题的爬取更新只能根据年份进行更新。