# 本期项目建设方案

## 信息资源规划和数据库建设方案

### 信息资源规划

#### 信息资源分类

##### 原始数据分类

请填写《附表1 数据资源分类》

##### 特征数据分类

请填写《附表1 数据资源分类》

##### 领域知识数据分类

请填写《附表1 数据资源分类》

##### 业务应用数据分类

请填写《附表1 数据资源分类》

##### 支撑数据分类

请填写《附表1 数据资源分类》

#### 信息资源结构设计

##### 原始数据分类

###### 网上采集数据分类

信息资源分类

**表 1.1‑1网上采集数据分类表**

| **序号** | **信息资源** | **说明** |
| --- | --- | --- |
| 1 | 采集门户网站信息资源 | 门户网站主体信息  门户网站评论信息 |
| 2 | 采集新闻网站信息资源 | 新闻网站主体信息  新闻网站评论信息  新网站闻弹幕信息  新闻网站用户信息 |
| 3 | 采集新闻手机客户端信息资源 | 新闻手机客户端主体信息，  新闻手机客户端评论信息 |
| 4 | 采集微博信息资源 | 微博主体信息  微博用户信息 |
| 5 | 采集微信公众号信息资源 | 微信公众号主体信息  微信公众号账号信息 |
| 6 | 采集论坛贴吧信息资源 | 论坛贴吧消息  论坛贴吧回帖消息，  论坛贴吧用户消息，  论坛贴吧讨论消息 |
| 7 | 采集博客信息资源 | 博客主体信息  博客用户消息 |
| 8 | 采集搜索引擎信息资源 | 搜索引擎主体信息 |
| 9 | 采集短视频手机客户端信息资源 | 短视频手机客户端主体信息  短视频手机客户端评论信息 |
| 10 | 采集移动应用商店信息资源 | 移动应用主体信息  移动应用评论信息  移动应用商店信息  移动应用分类信息  移动应用开发商信息 |
| 11 | 采集境外新闻网站信息资源 | 境外新闻网站主体信息  境外新闻网站评论信息 |
| 12 | 采集境外Facebook信息资源 | 境外Facebook主体信息  境外Facebook评论信息  境外Facebook账号信息 |
| 13 | 采集境外Twitter信息资源 | 境外Twitter主体信息  境外Twitter评论信息  境外Twitter用户信息 |

采集门户网站信息资源

采集门户网站信息资源主要针对网上采集到门户网站信息，由门户网站主体信息，门户网站评论信息构成。

实体目录

采集门户网站信息资源目录设计详见下表。

**表 1.1‑3 采集门户网站信息资源目录**

| **序号** | **数据表** | **表描述** | **备注** |
| --- | --- | --- | --- |
| 1 | 门户网站主体信息 | 门户网站的主体信息 |  |
| 2 | 门户网站评论信息 | 门户网站的评论信息 |  |

实体

门户网站主体信息

**表 1.1‑4门户网站主体信息表**

| **序号** | **主要数据项** | **数据类型** | **数据长度** | **备注** |
| --- | --- | --- | --- | --- |
| 1 | 新闻消息唯一主键 | 字符型 | 32字节 |  |
| 2 | 发布时间 | 时间 |  |  |
| 3 | 入库日期 | 时间 |  |  |
| 4 | 获取时间 | 时间 |  |  |
| 5 | 更新时间 | 时间 |  |  |
| 6 | 消息地址 | 长文本型 |  |  |
| 7 | 标题 | 长文本型 |  |  |
| 8 | 摘要 | 长文本型 |  |  |
| 9 | 内容 |  |  |  |
| 10 | 语种 | 字符型 | 32字节 |  |
| 11 | 首页地址 | 长文本型 |  |  |
| 12 | 网站备案号 | 字符型 | 128字节 |  |
| 13 | 网站名称 | 字符型 | 32字节 |  |
| 14 | 用户唯一标识 | 字符型 | 32字节 |  |
| 15 | 用户名 | 字符型 | 32字节 |  |
| 16 | 评论数 | 整数 |  |  |
| 17 | 阅读数 | 整数 |  |  |
| 18 | 图片数 | 整数 |  |  |
| 19 | 图片地址集 | 长文本型 |  |  |
| 20 | 视频数 | 整数 |  |  |
| 21 | 视频地址集 | 长文本型 |  |  |
| 22 | 境内外标识 | 字符型 | 128字节 |  |
| 23 | 消息状态 | 字符型 | 32字节 |  |
| 24 | 处理时间 | 时间 |  |  |
| 25 | 访问方式 | 整数 |  |  |
| 26 | 跟帖数 | 整数 |  |  |
| 27 | 转发数 | 整数 |  |  |
| 28 | 点赞数 | 整数 |  |  |
| 29 | 被踩数 | 整数 |  |  |
| 30 | 图片文件列表 | 长文本型 |  |  |
| 31 | 视频文件列表 | 长文本型 |  |  |
| 32 | 消息分类 | 字符型 | 128字节 |  |
| 33 | 关键词列表 | 字符型 | 128字节 |  |
| 34 | 消息发布年份 | 字符型 | 32字节 |  |
| 35 | 消息发布月份 | 字符型 | 32字节 |  |
| 36 | 消息发布日期 | 字符型 | 32字节 |  |
| 37 | 消息发布小时 | 字符型 | 32字节 |  |
| 38 | 正文语种 | 字符型 | 8字节 |  |
| 29 | 正文关键词 | 字符型 | 128字节 |  |
| 40 | 倾向性 | 字符型 | 8字节 |  |
| 41 | 倾向性关键词 | 字符型 | 128字节 |  |
| 42 | 文本判别类别 | 字符型 | 8字节 |  |
| 43 | 文本判别关键词 | 字符型 | 128字节 |  |
| 44 | 文本判别时间 | 时间 |  |  |
| 45 | 图片ocr判别关键词 | 字符型 | 128字节 |  |
| 46 | 图片ocr判别类型 | 字符型 | 8字节 |  |
| 47 | 图片ocr有害内容 | 字符型 | 128字节 |  |
| 48 | 图片ocr图片地址 | 字符型 | 128字节 |  |
| 49 | 图片ocr判别时间 | 时间 |  |  |
| 50 | 图片ocr结果 | 字符型 | 1024字节 |  |
| 51 | 图片判别结果 | 字符型 | 8字节 |  |
| 52 | 图片模型判别地址 | 字符型 | 128字节 |  |
| 53 | 图片模型判别时间 | 时间 |  |  |
| 54 | 音频模型判别结果 | 字符型 | 8字节 |  |
| 55 | 音频模型判别地址 | 字符型 | 128字节 |  |
| 56 | 音频模型判别时间 | 时间 |  |  |
| 57 | 视频模型判别结果 | 字符型 | 8字节 |  |
| 58 | 视频模型判别地址 | 字符型 | 128字节 |  |
| 59 | 视频模型判别时间 | 时间 |  |  |
| 60 | 有害台标判别结果 | 字符型 | 8字节 |  |
| 61 | 有害台标判别地址 | 字符型 | 128字节 |  |
| 62 | 有害台标判别时间 | 时间 |  |  |
| 63 | 地域-国家 | 字符型 | 128字节 |  |
| 64 | 地域-省 | 字符型 | 128字节 |  |
| 65 | 地域-市 | 字符型 | 128字节 |  |
| 66 | 地域-区县 | 字符型 | 128字节 |  |
| 67 | 地域名字串 | 字符型 | 256字节 |  |

门户网站评论信息

**表 1.1‑5门户网站评论信息表**

| **序号** | **主要数据项** | **数据类型** | **数据长度** | **备注** |
| --- | --- | --- | --- | --- |
| 1 | 评论唯一标识符 | 字符型 | 32字节 |  |
| 2 | 被评论信息标识符 | 字符型 | 32字节 |  |
| 3 | 发布时间 | 时间 |  |  |
| 4 | 获取时间 | 时间 |  |  |
| 5 | 入库时间 | 时间 |  |  |
| 6 | 更新时间 | 时间 |  |  |
| 7 | 标题 | 长文本型 |  |  |
| 8 | 内容 | 长文本型 |  |  |
| 9 | 语种 | 字符型 | 32字节 |  |
| 10 | 网站备案号 | 字符型 | 128字节 |  |
| 11 | 网站名称 | 字符型 | 32字节 |  |
| 12 | 用户唯一标识 | 字符型 | 32字节 |  |
| 13 | 用户名 | 字符型 | 32字节 |  |
| 14 | 评论数 | 整型 |  |  |
| 15 | 点赞数 | 整型 |  |  |
| 16 | 原始页地址 | 字符型 | 128字节 |  |
| 17 | 评论属性 | 字符型 | 128字节 |  |
| 18 | 评论状态 | 字符型 | 128字节 |  |
| 19 | 处理时间 | 时间 |  |  |
| 20 | 被评论新闻用户 | 字符型 | 128字节 |  |
| 21 | 访问方式 | 整型 |  |  |
| 22 | 点赞用户数 | 整型 |  |  |
| 23 | 被踩数 | 整型 |  |  |
| 24 | 跟帖推荐书 | 整型 |  |  |
| 25 | 市县编码 | 整型 |  |  |
| 26 | 省份编码 | 整型 |  |  |
| 27 | 关键词列表 | 长文本型 |  |  |
| 28 | 标题相似度 | 字符型 | 128字节 |  |
| 29 | 评论内容长度 | 整型 |  |  |
| 30 | 内容相似度 | 字符型 | 128字节 |  |
| 31 | 网站归属地编号 | 整型 |  |  |
| 32 | 发布年份 | 整型 |  |  |
| 33 | 发布月份 | 整型 |  |  |
| 34 | 发布日期 | 整型 |  |  |
| 35 | 发布小时 | 整型 |  |  |
| 36 | 正文语种 | 字符型 | 8字节 |  |
| 37 | 正文关键词 | 字符型 | 128字节 |  |
| 38 | 倾向性 | 字符型 | 8字节 |  |
| 39 | 倾向性关键词 | 字符型 | 128字节 |  |
| 40 | 文本判别类别 | 字符型 | 8字节 |  |
| 41 | 文本判别关键词 | 字符型 | 128字节 |  |
| 42 | 文本判别时间 | 时间 |  |  |
| 43 | 图片ocr判别关键词 | 字符型 | 128字节 |  |
| 44 | 图片ocr判别类型 | 字符型 | 8字节 |  |
| 45 | 图片ocr有害内容 | 字符型 | 128字节 |  |
| 46 | 图片ocr图片地址 | 字符型 | 128字节 |  |
| 47 | 图片ocr判别时间 | 时间 |  |  |
| 48 | 图片ocr结果 | 字符型 | 1024字节 |  |
| 49 | 图片判别结果 | 字符型 | 8字节 |  |
| 50 | 图片模型判别地址 | 字符型 | 128字节 |  |
| 51 | 图片模型判别时间 | 时间 |  |  |
| 52 | 音频模型判别结果 | 字符型 | 8字节 |  |
| 53 | 音频模型判别地址 | 字符型 | 128字节 |  |
| 54 | 音频模型判别时间 | 时间 |  |  |
| 55 | 视频模型判别结果 | 字符型 | 8字节 |  |
| 56 | 视频模型判别地址 | 字符型 | 128字节 |  |
| 57 | 视频模型判别时间 | 时间 |  |  |
| 58 | 有害台标判别结果 | 字符型 | 8字节 |  |
| 59 | 有害台标判别地址 | 字符型 | 128字节 |  |
| 60 | 有害台标判别时间 | 时间 |  |  |
| 61 | 地域-国家 | 字符型 | 128字节 |  |
| 62 | 地域-省 | 字符型 | 128字节 |  |
| 63 | 地域-市 | 字符型 | 128字节 |  |
| 64 | 地域-区县 | 字符型 | 128字节 |  |
| 65 | 地域名字串 | 字符型 | 256字节 |  |

E-R 图



**图 2.1‑1采集门户网站信息资源-数据设计**

采集新闻网站信息资源

###### 中央接入数据分类

###### 属地新闻单位接入数据分类

###### 属地互联网单位接入数据分类

###### CNCERT分中心接入数据分类

##### 特征数据分类

###### 基础数据分类

###### 模式识别特征数据分类

###### 多源索引特征数据分类

###### 有害识别数据分类

###### 主题分析数据分类

##### 领域知识信息资源分类

###### 领域业务知识数据分类

###### 领域样本知识数据分类

###### 通用样本知识数据分类

##### 业务应用数据分类

###### 应急发现业务数据分类

###### 专网舆情研判数据分类

###### 属地应急管控业务数据分类

###### 网上舆论引导业务数据分类

网评员协同数据分类

机器网评数据分类

网评案例数据分类

境外账号数据分类

虚拟账号仿真数据分类

###### 重点目标监管业务数据分类

###### 网上传播监管业务数据分类

从业服务数据分类

内容巡查数据分类

传播分析数据分类

##### 支撑数据分类

###### 业务应用支撑数据分类

电子取证数据分类

授权管理数据分类

多语言翻译数据分类

知识管理业务数据分类

### 数据库设计

数据库设计结合信息资源规划和总体结构设计内容，按照数据的种类不同分别建设共36类数据资源库，包含负责存储和管理舆情监测系统和探针系统采集得到原始数据（包括新闻、论坛、博客等）的原始数据库；大数据平台清洗、过滤、去重和汇聚后的过程数据库；以及应用系统使用的样本数据库及产生的结果数据库等。主要数据库类型包含如下内容及部分内容。

#### 原始数据库

##### 网上采集数据

###### 采集网站数据

采集门户网站信息资源主要针对网上采集到门户网站信息，由门户网站主体信息，门户网站评论信息构成。

数据库目录

采集门户网站信息资源目录设计详见下表。

**表 11 采集门户网站信息资源目录**

| **序号** | **数据表** | **表描述** | **备注** |
| --- | --- | --- | --- |
| 1 | 新闻网站主体信息表 | 门户网站的主体信息 |  |
| 2 | 新闻网站评论信息表 | 门户网站的评论信息 |  |

数据库表结构

新闻网站主题信息表

**表 12 采集门户网站信息资源**

| **序号** | **主要数据项** | **数据类型** | **数据长度** | **备注** |
| --- | --- | --- | --- | --- |
| 1 | 新闻消息唯一主键 | 字符型 | 32字节 |  |
| 2 | 发布时间 | 时间 |  |  |
| 3 | 入库日期 | 时间 |  |  |
| 4 | 获取时间 | 时间 |  |  |
| 5 | 更新时间 | 时间 |  |  |
| 6 | 消息地址 | 长文本型 |  |  |
| 7 | 标题 | 长文本型 |  |  |
| 8 | 摘要 | 长文本型 |  |  |
| 9 | 内容 |  |  |  |
| 10 | 语种 | 字符型 | 32字节 |  |
| 11 | 首页地址 | 长文本型 |  |  |
| 12 | 网站备案号 | 字符型 | 128字节 |  |
| 13 | 网站名称 | 字符型 | 32字节 |  |
| 14 | 用户唯一标识 | 字符型 | 32字节 |  |
| 15 | 用户名 | 字符型 | 32字节 |  |
| 16 | 评论数 | 整数 |  |  |
| 17 | 阅读数 | 整数 |  |  |
| 18 | 图片数 | 整数 |  |  |
| 19 | 图片地址集 | 长文本型 |  |  |
| 20 | 视频数 | 整数 |  |  |
| 21 | 视频地址集 | 长文本型 |  |  |
| 22 | 境内外标识 | 字符型 | 128字节 |  |
| 23 | 消息状态 | 字符型 | 32字节 |  |
| 24 | 处理时间 | 时间 |  |  |
| 25 | 访问方式 | 整数 |  |  |
| 26 | 跟帖数 | 整数 |  |  |
| 27 | 转发数 | 整数 |  |  |
| 28 | 点赞数 | 整数 |  |  |
| 29 | 被踩数 | 整数 |  |  |
| 30 | 图片文件列表 | 长文本型 |  |  |
| 31 | 视频文件列表 | 长文本型 |  |  |
| 32 | 消息分类 | 字符型 | 128字节 |  |
| 33 | 关键词列表 | 字符型 | 128字节 |  |
| 34 | 消息发布年份 | 字符型 | 32字节 |  |
| 35 | 消息发布月份 | 字符型 | 32字节 |  |
| 36 | 消息发布日期 | 字符型 | 32字节 |  |
| 37 | 消息发布小时 | 字符型 | 32字节 |  |
| 38 | 正文语种 | 字符型 | 8字节 |  |
| 29 | 正文关键词 | 字符型 | 128字节 |  |
| 40 | 倾向性 | 字符型 | 8字节 |  |
| 41 | 倾向性关键词 | 字符型 | 128字节 |  |
| 42 | 文本判别类别 | 字符型 | 8字节 |  |
| 43 | 文本判别关键词 | 字符型 | 128字节 |  |
| 44 | 文本判别时间 | 时间 |  |  |
| 45 | 图片ocr判别关键词 | 字符型 | 128字节 |  |
| 46 | 图片ocr判别类型 | 字符型 | 8字节 |  |
| 47 | 图片ocr有害内容 | 字符型 | 128字节 |  |
| 48 | 图片ocr图片地址 | 字符型 | 128字节 |  |
| 49 | 图片ocr判别时间 | 时间 |  |  |
| 50 | 图片ocr结果 | 字符型 | 1024字节 |  |
| 51 | 图片判别结果 | 字符型 | 8字节 |  |
| 52 | 图片模型判别地址 | 字符型 | 128字节 |  |
| 53 | 图片模型判别时间 | 时间 |  |  |
| 54 | 音频模型判别结果 | 字符型 | 8字节 |  |
| 55 | 音频模型判别地址 | 字符型 | 128字节 |  |
| 56 | 音频模型判别时间 | 时间 |  |  |
| 57 | 视频模型判别结果 | 字符型 | 8字节 |  |
| 58 | 视频模型判别地址 | 字符型 | 128字节 |  |
| 59 | 视频模型判别时间 | 时间 |  |  |
| 60 | 有害台标判别结果 | 字符型 | 8字节 |  |
| 61 | 有害台标判别地址 | 字符型 | 128字节 |  |
| 62 | 有害台标判别时间 | 时间 |  |  |
| 63 | 地域-国家 | 字符型 | 128字节 |  |
| 64 | 地域-省 | 字符型 | 128字节 |  |
| 65 | 地域-市 | 字符型 | 128字节 |  |
| 66 | 地域-区县 | 字符型 | 128字节 |  |
| 67 | 地域名字串 | 字符型 | 256字节 |  |

新闻网站评论信息表

**表 13 采集门户网站信息资源评论信息**

| **序号** | **主要数据项** | **数据类型** | **数据长度** | **备注** |
| --- | --- | --- | --- | --- |
| 1 | 评论唯一标识符 | 字符型 | 32字节 |  |
| 2 | 被评论信息标识符 | 字符型 | 32字节 |  |
| 3 | 发布时间 | 时间 |  |  |
| 4 | 获取时间 | 时间 |  |  |
| 5 | 入库时间 | 时间 |  |  |
| 6 | 更新时间 | 时间 |  |  |
| 7 | 标题 | 长文本型 |  |  |
| 8 | 内容 | 长文本型 |  |  |
| 9 | 语种 | 字符型 | 32字节 |  |
| 10 | 网站备案号 | 字符型 | 128字节 |  |
| 11 | 网站名称 | 字符型 | 32字节 |  |
| 12 | 用户唯一标识 | 字符型 | 32字节 |  |
| 13 | 用户名 | 字符型 | 32字节 |  |
| 14 | 评论数 | 整型 |  |  |
| 15 | 点赞数 | 整型 |  |  |
| 16 | 原始页地址 | 字符型 | 128字节 |  |
| 17 | 评论属性 | 字符型 | 128字节 |  |
| 18 | 评论状态 | 字符型 | 128字节 |  |
| 19 | 处理时间 | 时间 |  |  |
| 20 | 被评论新闻用户 | 字符型 | 128字节 |  |
| 21 | 访问方式 | 整型 |  |  |
| 22 | 点赞用户数 | 整型 |  |  |
| 23 | 被踩数 | 整型 |  |  |
| 24 | 跟帖推荐书 | 整型 |  |  |
| 25 | 市县编码 | 整型 |  |  |
| 26 | 省份编码 | 整型 |  |  |
| 27 | 关键词列表 | 长文本型 |  |  |
| 28 | 标题相似度 | 字符型 | 128字节 |  |
| 29 | 评论内容长度 | 整型 |  |  |
| 30 | 内容相似度 | 字符型 | 128字节 |  |
| 31 | 网站归属地编号 | 整型 |  |  |
| 32 | 发布年份 | 整型 |  |  |
| 33 | 发布月份 | 整型 |  |  |
| 34 | 发布日期 | 整型 |  |  |
| 35 | 发布小时 | 整型 |  |  |
| 36 | 正文语种 | 字符型 | 8字节 |  |
| 37 | 正文关键词 | 字符型 | 128字节 |  |
| 38 | 倾向性 | 字符型 | 8字节 |  |
| 39 | 倾向性关键词 | 字符型 | 128字节 |  |
| 40 | 文本判别类别 | 字符型 | 8字节 |  |
| 41 | 文本判别关键词 | 字符型 | 128字节 |  |
| 42 | 文本判别时间 | 时间 |  |  |
| 43 | 图片ocr判别关键词 | 字符型 | 128字节 |  |
| 44 | 图片ocr判别类型 | 字符型 | 8字节 |  |
| 45 | 图片ocr有害内容 | 字符型 | 128字节 |  |
| 46 | 图片ocr图片地址 | 字符型 | 128字节 |  |
| 47 | 图片ocr判别时间 | 时间 |  |  |
| 48 | 图片ocr结果 | 字符型 | 1024字节 |  |
| 49 | 图片判别结果 | 字符型 | 8字节 |  |
| 50 | 图片模型判别地址 | 字符型 | 128字节 |  |
| 51 | 图片模型判别时间 | 时间 |  |  |
| 52 | 音频模型判别结果 | 字符型 | 8字节 |  |
| 53 | 音频模型判别地址 | 字符型 | 128字节 |  |
| 54 | 音频模型判别时间 | 时间 |  |  |
| 55 | 视频模型判别结果 | 字符型 | 8字节 |  |
| 56 | 视频模型判别地址 | 字符型 | 128字节 |  |
| 57 | 视频模型判别时间 | 时间 |  |  |
| 58 | 有害台标判别结果 | 字符型 | 8字节 |  |
| 59 | 有害台标判别地址 | 字符型 | 128字节 |  |
| 60 | 有害台标判别时间 | 时间 |  |  |
| 61 | 地域-国家 | 字符型 | 128字节 |  |
| 62 | 地域-省 | 字符型 | 128字节 |  |
| 63 | 地域-市 | 字符型 | 128字节 |  |
| 64 | 地域-区县 | 字符型 | 128字节 |  |
| 65 | 地域名字串 | 字符型 | 256字节 |  |

###### 采集短视频手机客户端信息资源

采集短视频手机客户端信息资源主要针网上采集短视频手机客户端信息数据，由短视频手机客户端主体信息，短视频手机客户端评论信息组成。

数据库表目录

采集短视频手机客户端信息资源目录设计详见下表。

**表 2.1-38 采集短视频手机客户端信息资源目录**

| **序号** | **数据表** | **表描述** | **备注** |
| --- | --- | --- | --- |
| 1 | 短视频手机客户端主体信息 | 短视频手机客户端的主体信息 |  |
| 2 | 短视频手机客户端评论信息 | 短视频手机客户端的评论信息 |  |

数据实体表

**表 2.1-39 采集短视频手机客户端信息资源短视频手机客户端主体信息**

| **序号** | **主要数据项** | **数据类型** | **数据长度** | **备注** |
| --- | --- | --- | --- | --- |
| 1 | 消息的唯一标识符 | 字符型 | 128字节 |  |
| 2 | 发布时间 | 时间 |  |  |
| 3 | 获取时间 | 时间 |  |  |
| 4 | 入库时间 | 时间 |  |  |
| 5 | 更新时间 | 时间 |  |  |
| 6 | 消息ID | 字符型 | 128字节 |  |
| 7 | 消息地址 | 字符型 | 128字节 |  |
| 8 | 数据类型 | 整型 |  |  |
| 9 | 消息内容 | 长文本型 |  |  |
| 10 | 语种 | 字符型 | 128字节 |  |
| 11 | 消息状态 | 字符型 | 32字节 |  |
| 12 | 用户标识符 | 字符型 | 128字节 |  |
| 13 | 用户名 | 字符型 | 128字节 |  |
| 14 | 访问方式 | 整型 |  |  |
| 15 | 关注数 | 整型 |  |  |
| 16 | 粉丝数 | 整型 |  |  |
| 17 | 消息评论数 | 整型 |  |  |
| 18 | 消息转发数 | 整型 |  |  |
| 19 | 根节点标识符 | 字符型 | 128字节 |  |
| 20 | 根节点ID | 字符型 | 128字节 |  |
| 21 | 根节点用户标识符 | 字符型 | 128字节 |  |
| 22 | 根节点评论数 | 整型 |  |  |
| 23 | 根节点转发数据 | 整型 |  |  |
| 24 | 根节点点赞数 | 整型 |  |  |
| 25 | 图片数 | 整型 |  |  |
| 26 | 图片地址集 | 长文本型 |  |  |
| 27 | 图片文件地址 | 长文本型 |  |  |
| 28 | 视频数 | 整型 |  |  |
| 29 | 视频地址集 | 长文本型 |  |  |
| 30 | 视频文件地址集 | 长文本型 |  |  |
| 31 | 关键词列表 | 长文本型 |  |  |
| 32 | 内容相似度 | 字符型 | 128字节 |  |
| 33 | 市县编码 | 整型 |  |  |
| 34 | 省份编码 | 整型 |  |  |
| 35 | 话题 | 字符型 | 128字节 |  |
| 36 | 发布年份 | 整型 |  |  |
| 37 | 发布月份 | 整型 |  |  |
| 38 | 发布日期 | 整型 |  |  |
| 39 | 发布小时 | 整型 |  |  |
| 40 | 正文语种 | 字符型 | 8字节 |  |
| 41 | 正文关键词 | 字符型 | 128字节 |  |
| 42 | 倾向性 | 字符型 | 8字节 |  |
| 43 | 倾向性关键词 | 字符型 | 128字节 |  |
| 44 | 文本判别类别 | 字符型 | 8字节 |  |
| 45 | 文本判别关键词 | 字符型 | 128字节 |  |
| 46 | 文本判别时间 | 时间 |  |  |
| 47 | 图片ocr判别关键词 | 字符型 | 128字节 |  |
| 48 | 图片ocr判别类型 | 字符型 | 8字节 |  |
| 49 | 图片ocr有害内容 | 字符型 | 128字节 |  |
| 50 | 图片ocr图片地址 | 字符型 | 128字节 |  |
| 51 | 图片ocr判别时间 | 时间 |  |  |
| 52 | 图片ocr结果 | 字符型 | 1024字节 |  |
| 53 | 图片判别结果 | 字符型 | 8字节 |  |
| 54 | 图片模型判别地址 | 字符型 | 128字节 |  |
| 55 | 图片模型判别时间 | 时间 |  |  |
| 56 | 音频模型判别结果 | 字符型 | 8字节 |  |
| 57 | 音频模型判别地址 | 字符型 | 128字节 |  |
| 58 | 音频模型判别时间 | 时间 |  |  |
| 59 | 视频模型判别结果 | 字符型 | 8字节 |  |
| 60 | 视频模型判别地址 | 字符型 | 128字节 |  |
| 61 | 视频模型判别时间 | 时间 |  |  |
| 62 | 有害台标判别结果 | 字符型 | 8字节 |  |
| 63 | 有害台标判别地址 | 字符型 | 128字节 |  |
| 64 | 有害台标判别时间 | 时间 |  |  |
| 65 | 地域-国家 | 字符型 | 128字节 |  |
| 66 | 地域-省 | 字符型 | 128字节 |  |
| 67 | 地域-市 | 字符型 | 128字节 |  |
| 68 | 地域-区县 | 字符型 | 128字节 |  |
| 69 | 地域名字串 | 字符型 | 256字节 |  |

**表 2.1-40 采集短视频手机客户端信息资源短视频手机客户端评论表**

| **序号** | **主要数据项** | | | **数据类型** | | **数据长度** | | **备注** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | | 评论唯一标识符 | 字符型 | | 32字节 | |  | |
| 2 | | 被评论信息标识符 | 字符型 | | 32字节 | |  | |
| 3 | | 发布时间 | 时间 | |  | |  | |
| 4 | | 获取时间 | 时间 | |  | |  | |
| 5 | | 入库时间 | 时间 | |  | |  | |
| 6 | | 更新时间 | 时间 | |  | |  | |
| 7 | | 标题 | 长文本型 | |  | |  | |
| 8 | | 内容 | 长文本型 | |  | |  | |
| 9 | | 语种 | 字符型 | | 32字节 | |  | |
| 10 | | 用户唯一标识 | 字符型 | | 32字节 | |  | |
| 11 | | 用户名 | 字符型 | | 32字节 | |  | |
| 12 | | 评论数 | 整型 | |  | |  | |
| 13 | | 点赞数 | 整型 | |  | |  | |
| 14 | | 评论属性 | 字符型 | | 128字节 | |  | |
| 15 | | 评论状态 | 字符型 | | 128字节 | |  | |
| 16 | | 处理时间 | 时间 | |  | |  | |
| 17 | | 被评论新闻用户 | 字符型 | | 128字节 | |  | |
| 18 | | 访问方式 | 整型 | |  | |  | |
| 19 | | 点赞用户数 | 整型 | |  | |  | |
| 20 | | 被踩数 | 整型 | |  | |  | |
| 21 | | 市县编码 | 整型 | |  | |  | |
| 22 | | 省份编码 | 整型 | |  | |  | |
| 23 | | 关键词列表 | 长文本型 | |  | |  | |
| 24 | | 标题相似度 | 字符型 | | 128字节 | |  | |
| 25 | | 评论内容长度 | 整型 | |  | |  | |
| 26 | | 内容相似度 | 字符型 | | 128字节 | |  | |
| 27 | | 网站归属地编号 | 整型 | |  | |  | |
| 28 | | 发布年份 | 整型 | |  | |  | |
| 29 | | 发布月份 | 整型 | |  | |  | |
| 30 | | 发布日期 | 整型 | |  | |  | |
| 31 | | 发布小时 | 整型 | |  | |  | |
| 32 | | 正文语种 | 字符型 | | 8字节 | |  | |
| 33 | | 正文关键词 | 字符型 | | 128字节 | |  | |
| 34 | | 倾向性 | 字符型 | | 8字节 | |  | |
| 35 | | 倾向性关键词 | 字符型 | | 128字节 | |  | |
| 36 | | 文本判别类别 | 字符型 | | 8字节 | |  | |
| 37 | | 文本判别关键词 | 字符型 | | 128字节 | |  | |
| 38 | | 文本判别时间 | 时间 | |  | |  | |
| 39 | | 图片ocr判别关键词 | 字符型 | | 128字节 | |  | |
| 40 | | 图片ocr判别类型 | 字符型 | | 8字节 | |  | |
| 41 | | 图片ocr有害内容 | 字符型 | | 128字节 | |  | |
| 42 | | 图片ocr图片地址 | 字符型 | | 128字节 | |  | |
| 43 | | 图片ocr判别时间 | 时间 | |  | |  | |
| 44 | | 图片ocr结果 | 字符型 | | 1024字节 | |  | |
| 45 | | 图片判别结果 | 字符型 | | 8字节 | |  | |
| 46 | | 图片模型判别地址 | 字符型 | | 128字节 | |  | |
| 47 | | 图片模型判别时间 | 时间 | |  | |  | |
| 48 | | 音频模型判别结果 | 字符型 | | 8字节 | |  | |
| 49 | | 音频模型判别地址 | 字符型 | | 128字节 | |  | |
| 50 | | 音频模型判别时间 | 时间 | |  | |  | |
| 51 | | 视频模型判别结果 | 字符型 | | 8字节 | |  | |
| 52 | | 视频模型判别地址 | 字符型 | | 128字节 | |  | |
| 53 | | 视频模型判别时间 | 时间 | |  | |  | |
| 54 | | 有害台标判别结果 | 字符型 | | 8字节 | |  | |
| 55 | | 有害台标判别地址 | 字符型 | | 128字节 | |  | |
| 56 | | 有害台标判别时间 | 时间 | |  | |  | |
| 57 | | 地域-国家 | 字符型 | | 128字节 | |  | |
| 58 | | 地域-省 | 字符型 | | 128字节 | |  | |
| 59 | | 地域-市 | 字符型 | | 128字节 | |  | |
| 60 | | 地域-区县 | 字符型 | | 128字节 | |  | |
| 61 | | 地域名字串 | 字符型 | | 256字节 | |  | |

###### 采集境外新闻网站信息资源

采集境外新闻网站信息资源主要针对网上采集到境外新闻网站信息，由境外新闻网站主体信息，境外新闻网站评论信息构成。

数据库目录

采集境外新闻网站信息资源目录设计详见下表。

**表 2.1-47 采集境外新闻网站信息资源目录**

| **序号** | **数据表** | **表描述** | **备注** |
| --- | --- | --- | --- |
| 1 | 境外新闻网站主体信息 | 境外新闻网站的主体信息 |  |
| 2 | 境外新闻网站评论信息 | 境外新闻网站的评论信息 |  |

数据库实体表

**表 2.1-48 采集境外新闻网站信息资源境外新闻网站主体消息**

| **序号** | **主要数据项** | **数据类型** | **数据长度** | **备注** |
| --- | --- | --- | --- | --- |
| 1 | 新闻消息唯一主键 | 字符型 | 32字节 |  |
| 2 | 发布时间 | 时间 |  |  |
| 3 | 入库日期 | 时间 |  |  |
| 4 | 获取时间 | 时间 |  |  |
| 5 | 更新时间 | 时间 |  |  |
| 6 | 消息地址 | 长文本型 |  |  |
| 7 | 标题 | 长文本型 |  |  |
| 8 | 摘要 | 长文本型 |  |  |
| 9 | 内容 |  |  |  |
| 10 | 语种 | 字符型 | 32字节 |  |
| 11 | 首页地址 | 长文本型 |  |  |
| 12 | 网站备案号 | 字符型 | 128字节 |  |
| 13 | 网站名称 | 字符型 | 32字节 |  |
| 14 | 用户唯一标识 | 字符型 | 32字节 |  |
| 15 | 用户名 | 字符型 | 32字节 |  |
| 16 | 评论数 | 整数 |  |  |
| 17 | 阅读数 | 整数 |  |  |
| 18 | 图片数 | 整数 |  |  |
| 19 | 图片地址集 | 长文本型 |  |  |
| 20 | 视频数 | 整数 |  |  |
| 21 | 视频地址集 | 长文本型 |  |  |
| 22 | 境内外标识 | 字符型 | 128字节 |  |
| 23 | 消息状态 | 字符型 | 32字节 |  |
| 24 | 处理时间 | 时间 |  |  |
| 25 | 访问方式 | 整数 |  |  |
| 26 | 跟帖数 | 整数 |  |  |
| 27 | 转发数 | 整数 |  |  |
| 28 | 点赞数 | 整数 |  |  |
| 29 | 被踩数 | 整数 |  |  |
| 30 | 图片文件列表 | 长文本型 |  |  |
| 31 | 视频文件列表 | 长文本型 |  |  |
| 32 | 消息分类 | 字符型 | 128字节 |  |
| 33 | 关键词列表 | 字符型 | 128字节 |  |
| 34 | 消息发布年份 | 字符型 | 32字节 |  |
| 35 | 消息发布月份 | 字符型 | 32字节 |  |
| 36 | 消息发布日期 | 字符型 | 32字节 |  |
| 37 | 消息发布小时 | 字符型 | 32字节 |  |
| 38 | 正文语种 | 字符型 | 8字节 |  |
| 29 | 正文关键词 | 字符型 | 128字节 |  |
| 40 | 倾向性 | 字符型 | 8字节 |  |
| 41 | 倾向性关键词 | 字符型 | 128字节 |  |
| 42 | 文本判别类别 | 字符型 | 8字节 |  |
| 43 | 文本判别关键词 | 字符型 | 128字节 |  |
| 44 | 文本判别时间 | 时间 |  |  |
| 45 | 图片ocr判别关键词 | 字符型 | 128字节 |  |
| 46 | 图片ocr判别类型 | 字符型 | 8字节 |  |
| 47 | 图片ocr有害内容 | 字符型 | 128字节 |  |
| 48 | 图片ocr图片地址 | 字符型 | 128字节 |  |
| 49 | 图片ocr判别时间 | 时间 |  |  |
| 50 | 图片ocr结果 | 字符型 | 1024字节 |  |
| 51 | 图片判别结果 | 字符型 | 8字节 |  |
| 52 | 图片模型判别地址 | 字符型 | 128字节 |  |
| 53 | 图片模型判别时间 | 时间 |  |  |
| 54 | 音频模型判别结果 | 字符型 | 8字节 |  |
| 55 | 音频模型判别地址 | 字符型 | 128字节 |  |
| 56 | 音频模型判别时间 | 时间 |  |  |
| 57 | 视频模型判别结果 | 字符型 | 8字节 |  |
| 58 | 视频模型判别地址 | 字符型 | 128字节 |  |
| 59 | 视频模型判别时间 | 时间 |  |  |
| 60 | 有害台标判别结果 | 字符型 | 8字节 |  |
| 61 | 有害台标判别地址 | 字符型 | 128字节 |  |
| 62 | 有害台标判别时间 | 时间 |  |  |
| 63 | 地域-国家 | 字符型 | 128字节 |  |
| 64 | 地域-省 | 字符型 | 128字节 |  |
| 65 | 地域-市 | 字符型 | 128字节 |  |
| 66 | 地域-区县 | 字符型 | 128字节 |  |
| 67 | 地域名字串 | 字符型 | 256字节 |  |

**表 2.1-49 采集境外新闻网站信息资源评论信息**

| **序号** | **主要数据项** | **数据类型** | **数据长度** | **备注** |
| --- | --- | --- | --- | --- |
| 1 | 评论唯一标识符 | 字符型 | 32字节 |  |
| 2 | 被评论信息标识符 | 字符型 | 32字节 |  |
| 3 | 发布时间 | 时间 |  |  |
| 4 | 获取时间 | 时间 |  |  |
| 5 | 入库时间 | 时间 |  |  |
| 6 | 更新时间 | 时间 |  |  |
| 7 | 标题 | 长文本型 |  |  |
| 8 | 内容 | 长文本型 |  |  |
| 9 | 语种 | 字符型 | 32字节 |  |
| 10 | 网站备案号 | 字符型 | 128字节 |  |
| 11 | 网站名称 | 字符型 | 32字节 |  |
| 12 | 用户唯一标识 | 字符型 | 32字节 |  |
| 13 | 用户名 | 字符型 | 32字节 |  |
| 14 | 评论数 | 整型 |  |  |
| 15 | 点赞数 | 整型 |  |  |
| 16 | 原始页地址 | 字符型 | 128字节 |  |
| 17 | 评论属性 | 字符型 | 128字节 |  |
| 18 | 评论状态 | 字符型 | 128字节 |  |
| 19 | 处理时间 | 时间 |  |  |
| 20 | 被评论新闻用户 | 字符型 | 128字节 |  |
| 21 | 访问方式 | 整型 |  |  |
| 22 | 点赞用户数 | 整型 |  |  |
| 23 | 被踩数 | 整型 |  |  |
| 24 | 跟帖推荐书 | 整型 |  |  |
| 25 | 市县编码 | 整型 |  |  |
| 26 | 省份编码 | 整型 |  |  |
| 27 | 关键词列表 | 长文本型 |  |  |
| 28 | 标题相似度 | 字符型 | 128字节 |  |
| 29 | 评论内容长度 | 整型 |  |  |
| 30 | 内容相似度 | 字符型 | 128字节 |  |
| 31 | 网站归属地编号 | 整型 |  |  |
| 32 | 发布年份 | 整型 |  |  |
| 33 | 发布月份 | 整型 |  |  |
| 34 | 发布日期 | 整型 |  |  |
| 35 | 发布小时 | 整型 |  |  |
| 36 | 正文语种 | 字符型 | 8字节 |  |
| 37 | 正文关键词 | 字符型 | 128字节 |  |
| 38 | 倾向性 | 字符型 | 8字节 |  |
| 39 | 倾向性关键词 | 字符型 | 128字节 |  |
| 40 | 文本判别类别 | 字符型 | 8字节 |  |
| 41 | 文本判别关键词 | 字符型 | 128字节 |  |
| 42 | 文本判别时间 | 时间 |  |  |
| 43 | 图片ocr判别关键词 | 字符型 | 128字节 |  |
| 44 | 图片ocr判别类型 | 字符型 | 8字节 |  |
| 45 | 图片ocr有害内容 | 字符型 | 128字节 |  |
| 46 | 图片ocr图片地址 | 字符型 | 128字节 |  |
| 47 | 图片ocr判别时间 | 时间 |  |  |
| 48 | 图片ocr结果 | 字符型 | 1024字节 |  |
| 49 | 图片判别结果 | 字符型 | 8字节 |  |
| 50 | 图片模型判别地址 | 字符型 | 128字节 |  |
| 51 | 图片模型判别时间 | 时间 |  |  |
| 52 | 音频模型判别结果 | 字符型 | 8字节 |  |
| 53 | 音频模型判别地址 | 字符型 | 128字节 |  |
| 54 | 音频模型判别时间 | 时间 |  |  |
| 55 | 视频模型判别结果 | 字符型 | 8字节 |  |
| 56 | 视频模型判别地址 | 字符型 | 128字节 |  |
| 57 | 视频模型判别时间 | 时间 |  |  |
| 58 | 有害台标判别结果 | 字符型 | 8字节 |  |
| 59 | 有害台标判别地址 | 字符型 | 128字节 |  |
| 60 | 有害台标判别时间 | 时间 |  |  |
| 61 | 地域-国家 | 字符型 | 128字节 |  |
| 62 | 地域-省 | 字符型 | 128字节 |  |
| 63 | 地域-市 | 字符型 | 128字节 |  |
| 64 | 地域-区县 | 字符型 | 128字节 |  |
| 65 | 地域名字串 | 字符型 | 256字节 |  |

# 标题模板

## 二级

### 三级

#### 四级

##### 五级

###### 六级

七级

八级

九级