

**密级：机密**

**大数据品类分析与企业口碑爬虫数据结构设计规范**

**版本 1.2**

**修订历史记录**

|  |  |  |  |
| --- | --- | --- | --- |
| **日期** | **版本** | **说明** | **作者** |
| 2017.3.10 | 1.0 |  | 冯石佳，许烨 |
| 2017.3.14 | 1.1 |  | 冯石佳，许烨 |
| 2017.3.15 | 1.2 | 整理表字段，添加字段详细说明 | 冯石佳，许烨 |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

目录

1. **说明**
2. **命名规范**
3. **数据表设计**

3.1.概念结构设计--------------------------------------------------------------------3

3.2. 逻辑结构设计--------------------------------------------------------------------3

3.3．数据结构--------------------------------------------------------------------------4

1. **数据表字段详细说明**

4.1．searchresult数据表字段说明-----------------------------------------------5

4.2．comment数据表字段说明--------------------------------------------------7

4.3．userinfo 数据表字段说明---------------------------------------------------8

**1.说明**

本设计规约的目的是统一各个网站数据存储的结构，使不同的网站数据在存储数据的时候有统一的结构，便于之后对数据的操作。

**2. 命名规范**

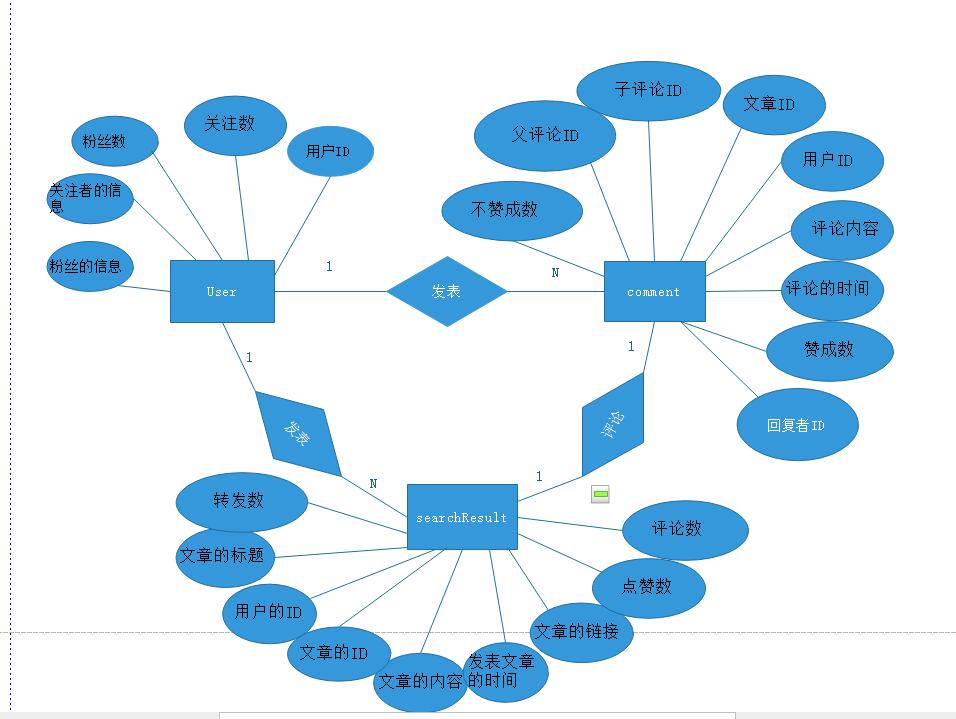
小组成员的命名规范均以自己负责的网站名+下划线作为前缀，后面跟上统一后数据表名，最后再加上下划线+before（未清洗的数据）或者是下划线+after（表示已经经过清洗的数据）

字段中的所有命名均采用小写字母

例如：weibo\_searchresult\_before表示的是微博未经过清洗的搜索结果数据

**3. 数据表设计**

**3.1. 概念结构设计**

****

**3.2. 逻辑结构设计**

searchresult 表（文章ID,文章标题，文章作者ID,文章内容，文章链接，文章发布时间，点赞数，评论数，转发数）

comment表（文章ID,父评论ID,评论ID,用户ID，目标ID，评论内容，评论时间，赞成数，不赞成数）

userinfo表（用户ID，用户基本信息（四个字段），关注数，粉丝数，关注者个人主页链接，粉丝个人主页链接）

**3.3. 数据结构**

**表一：searchresult表**

|  |  |  |
| --- | --- | --- |
| **字段名** | **类型** | **描述** |
| acticleid | Varchar(50) | 文章ID,主键,详见4.1.1 |
| authored | Varchar(50) | 文章作者ID,详见4.1.2 |
| title | Text | 文章标题，文章标题，详见4.1.3 |
| content | Text | 文章内容，包含文章的详细内容，详见4.1.4 |
| url | Varchar(200) | 文章链接，打开此链接可以直接到文章页面，详见4.1.5 |
| time | Data | 文章发布时间，详见4.1.6 |
| supportnum | Int(10) | 文章的点赞数，详见4.1.7 |
| commentnum | Int(10) | 文章的评论数，详见4.1.8 |
| transpondnum | Int(10) | 文章的转发数，详见4.1.9 |

**表二：commen表**

|  |  |  |
| --- | --- | --- |
| **字段名** | **类型** | **描述** |
| articleid | Varchar(50) | 文章的ID,主键，详见4.2.2 |
| commentid | Varchar(50) | 评论ID,主键,本条评论评论唯一标识符，详见4.2.2 |
| fathercommentid | Varchar(50) | 父评论ID,主键，本条评论的目标评论ID，详见4.2.3 |
| userid | Varchar(50) | 发布评论的用户ID,详见4.2.4 |
| atid | Varchar(50) | 评论中涉及其他用户的ID,详见4.2.5 |
| content | Text | 评论的具体内容，详见4.2.6 |
| time | Data | 评论发布的时间，详见4.2.7 |
| supportnum | Int(10) | 评论的点赞数，详见4.2.8 |
| unsupportnum | Int(10) | 评论的不赞成数，详见4.2.9 |

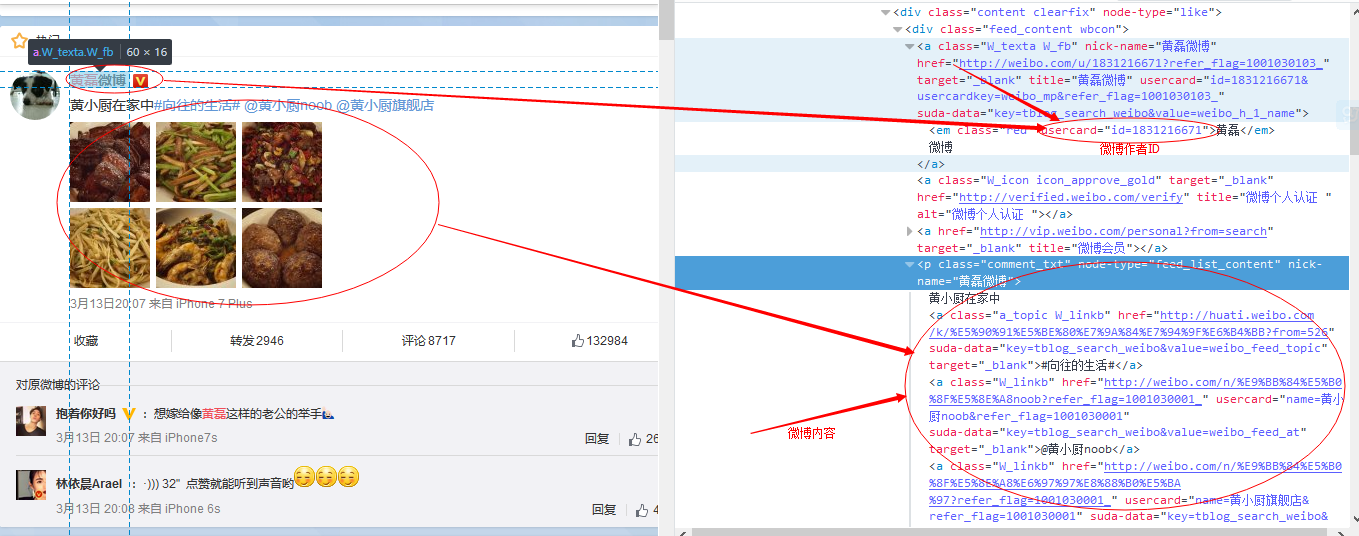
**表三：userinfo表**

|  |  |  |
| --- | --- | --- |
| **字段名** | **类型** | **描述** |
| userid | Varchar(50) | 用户的ID,主键，详见4.3.1 |
| info\_one | Varchar(50) | 用户信息，详见4.3.2 |
| info\_two | Varchar(50) | 同上 |
| info\_three | Varchar(50) | 同上 |
| info\_four | Varchar(50) | 同上 |
| attentionnum | Int(10) | 用户关注的人数，详见4.3.3 |
| fannum | Int(10) | 用户的粉丝数，详见4.3.3 |
| attention | Text | 用户关注人的个人主页链接，每一个关注之间用逗号 分隔，详见4.3.4 |
| fan | Text | 用户粉丝的个人主页链接，每一个粉丝主页之间用逗号分割，详见4.3.4 |

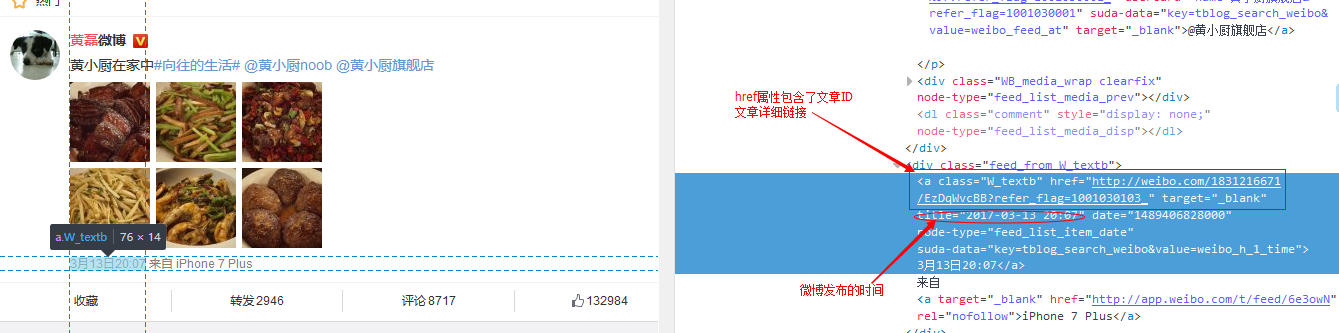
**4.数据表字段详细说明**

**说明：**下面的演示将以微博网站作为样板，截取的内容涉及的用户均只是为了说明数据字段作用，并无其它意图

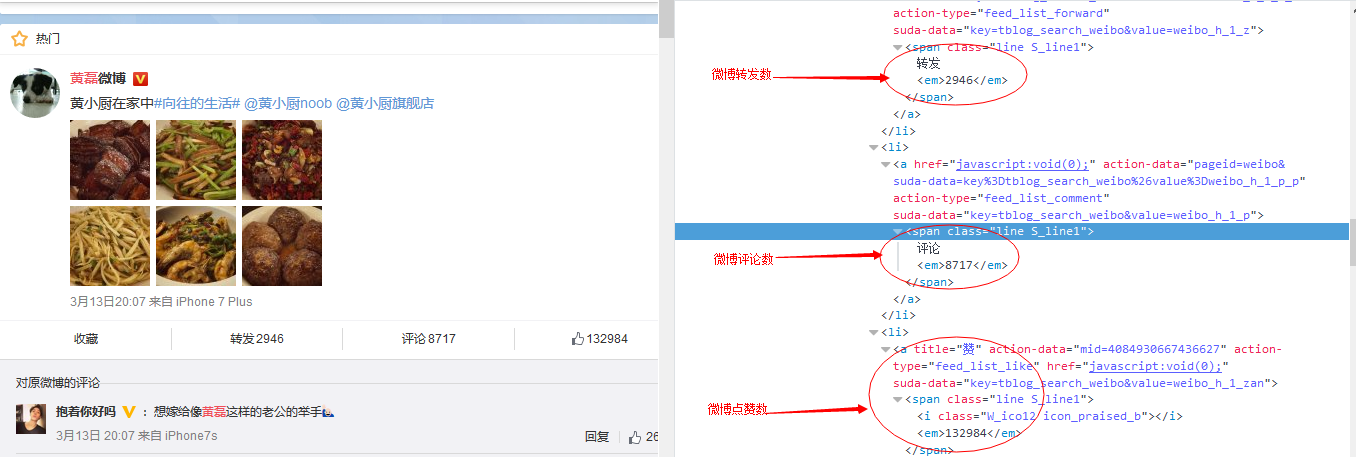
**4.1 , searchresult数据表字段说明**



**图4-1-1**



**图4-1-2**



**图4-1-3**

4.1.1：表示的是文章的ID属性，在图4-1-2中指出的href属性的值，其构成是<http://weibo.com/> +用户名+文章专属ID，文章ID从这个链接中截取出来

4.1.2：表示的是文章作者的ID编号，在图4-1-1中指出了微博作者的ID,该ID会出现在源码中的很多位置，如获取文章ID时也可得到作者ID,可以自由选取位置得到信息

4.1.3：表示文章的标题，示例的微博中没有标题，如果有标题的话可以从微博的内容中提取出来

4.1.4：表示文章的内容，图4-1-1中指出了微博的内容在页面中的显示以及在源码中的显示。

4.1.5：表示文章的详细链接，文章的详细链接在图4-1-2中指出，在页面中是在微博发布的时间位置有一个链接，在图右侧显示了其在源码中的显示

4.1.6：文章的发布时间，发布时间在图4-1-2中指出，跟随在文章的详情链接后面

4.1.7：文章的点赞数，见图4-1-3

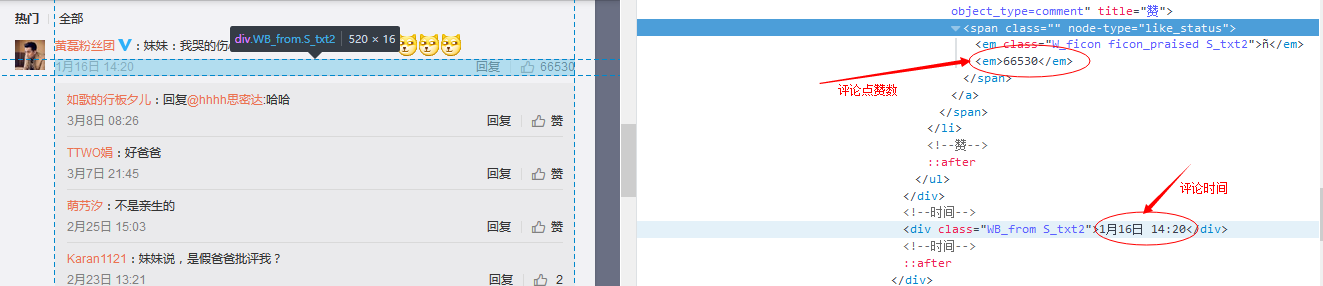
4.1.8：文章的评论数，见图4-1-3

4.1.9：文章的转发数，见图4-1-3

**4.2,comment数据表字段说明**



**图4-2-1**



**图4-2-2**



**图4-2-3**

4.2.1：文章ID来源于上一张表中的文章ID字段，需要有文章的ID才能得

到对于该文章的评论，也就是说这个字段应从上表传过来

4.2.2：评论的ID编号，评论的唯一表示符，在图4-2-1中指出在源码中评论ID的位置

4.2.3：父评论id号，在图4-2-3中存在有一个用户回复的情况，第一层的

用户是针对文章的评论，第二层的评论是对第一层评论的评论，那么第

一层评论的父评论Id为文章的ID,第二层评论的父评论是第一层评论的ID，

第三层评论的父评论ID是第二层评论的ID

4.2.4：表示的是发布评论的用户的ID,图4-2-1标出了评论用户ID在源码中的位置

4.2.5：atid表示的是用户评论中@到的人，表示的是想要其他人注意到这个微博或者是注意到这条评论，在图4-2-3中用矩形框表示的部分，在矩形框之前的@符号只是表示一种回复关系，这个@的用户已经注意到这个微博或是评论

4.2.6：本字段表示的是评论的内容，图4-2-1中有指出在微博网站中评论内

容的展示

4.2.7：，本字段表示的是评论发布的时间，图4-2-2中指出了评论发布的时

间信息

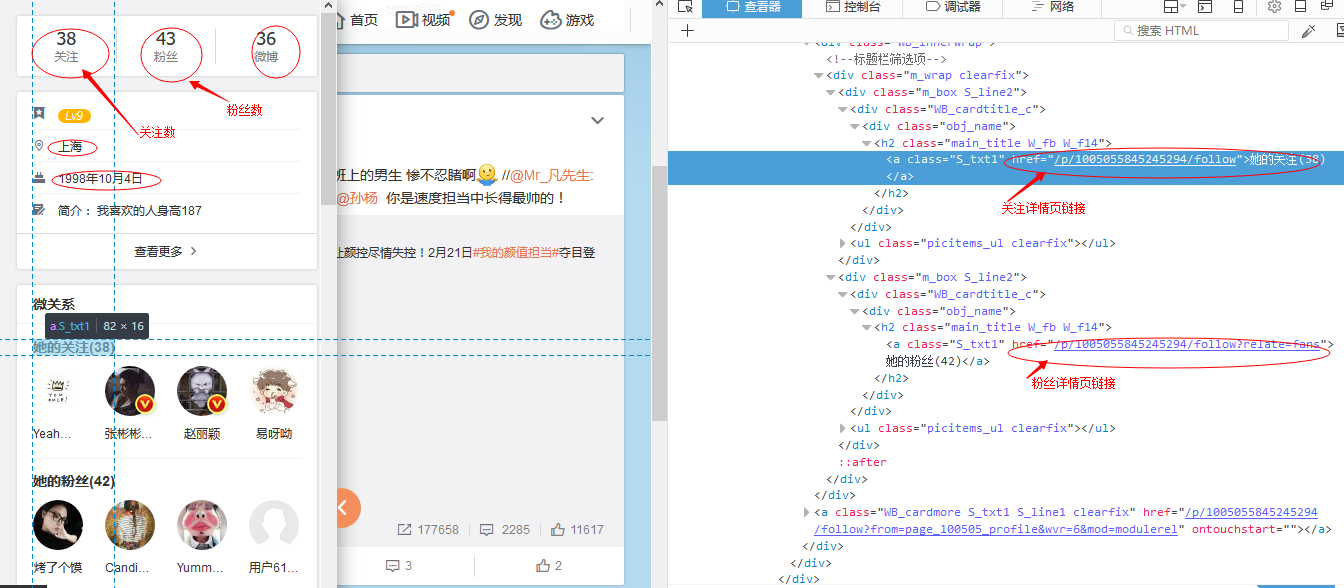
4.2.8：表示对评论内容的赞成数，如果没有该数据，本字段可以为空，在

微博中表现如图4-2-2所示

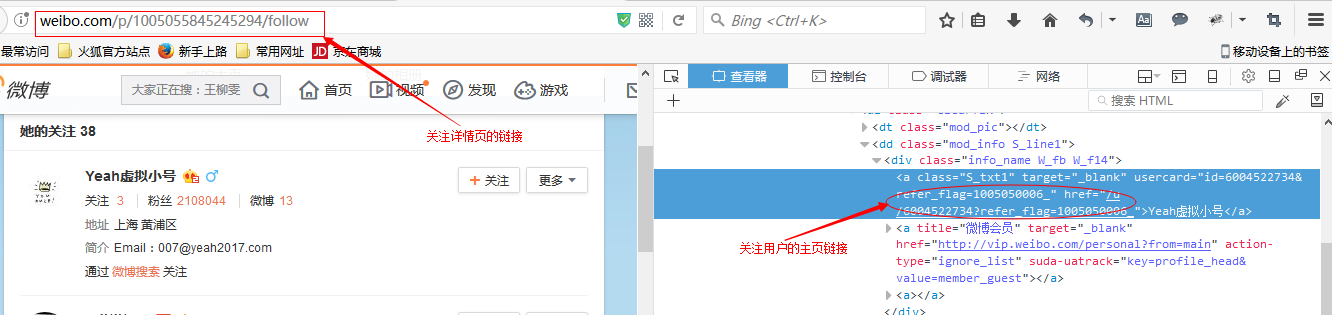
4.2.9：评论的不赞成数，微博网站没有该数据，所以在存储的时候该字段

为空

**4.3, userinfo 数据表字段说明**



**图4-3-1**



**图4-3-2**

**4.3.1：**用户的ID由链接中得到，因为要通过用户个人主页的链接才能

进入个人主页获取信息，可以从哪个链接中得到用户的ID

**4.3.2：**如图4-3-1所示：个人信息的四个字段在微博网站中会存储用户发布的微博数,用户的注册地,用户的生日等信息，这四个字段可以根据网站的情况填写，如果没有足够的信息，字段可以为空

**4.3.3：**在图4-3-1中指示出关注数和粉丝数，只需将这两个字段存入即可

**4.3.4：**由图4-3-1中的关注详情页链接可以得到类似与图4-3-2的网页，在 这样的页面中显示的是用户关注人的情况，选择保留用户的主页链接，这

样就可以通过这个链接获取其关注或是粉丝的详细情况，保存的内容如图

4-3-2中圈出的href属性，在微博中不需要问号后的参数也能打开网页，所以将以<http://weibo.com/u/6004522734>，的格式存储一个关注用户，链接后面的逗号不可省略，其作用是用来分割每一个主页，如有多个关注，按上面格式直接写入即可