## 关于爬取百度知道数据库设计说明

**主表(bdzd\_master)字段设计说明**

|  |  |  |
| --- | --- | --- |
| **字 段** | **类 型** | **说 明** |
| Id | int | 数据记录索引，可设置为自增 |
| QuestionID | Varchar(25) | 百度知道每个问题ID |
| AskTitle | Varchar(50) | 问题标题 |
| QContent | Varchar(255) | 问题内容 |
| AskDate | Varchar(15) | 提问时间 |
| Answer | Varchar(1000) | 回答内容(超过500字，一般为复制粘贴，可以保存到文件) |
| AnswerDate | Date | 回答时间 |
| Good | Int | 赞同该回答的数量(动态变化) |
| Bad | Int | 反对该回答的数量(动态变化) |
| AdoptRate | Varchar(3) | 回答者的采纳率 |
| AnswerType | Varchar(1) | 分3类：  ‘’没有回答  ‘0’代表最佳回答  ‘1’代表匿名回答 |
| CrawlDate | Date | 首次抓取时间（可选） |
| UpdateDate | Date | 最后一次更新的时间（可选） |
| OthAnsCount | ShortInt | 其它回答的记录数 |
| Cur\_key | Varchar(20) | 当前关键字(学校名称或简称) |
| Cur\_alias | Varchar(10) | 学校英文缩写 |

**副表(bdzd\_minor)字段设计说明（数据的增长主要在该表）**

|  |  |  |
| --- | --- | --- |
| **字 段** | **类 型** | **说 明** |
| Id | int | 数据记录索引，可设置为自增，也可以主表的ID值作为该值 |
| QuestionID | Varchar(25) | 百度知道每个问题ID，通过该字段与主表进行联接查询 |
| OthersNo | ShorInt | 该条记录的序号 |
| Answer | Varchar(500) | 回答内容(超过250字，一般为复制粘贴，可以保存到文件) |
| Good | Int | 赞同该回答的数量(动态变化) |
| Bad | Int | 反对该回答的数量(动态变化) |
| AnswerDate | Date | 回答时间 |
| UpdateDate | Date | 最后一次更新的时间（可选） |

说明：

1、百度知道每个问题都有一个具体的问题ID，可根据该ID拼接完整的URL

2、对于文字量过大的评论一般是从百度百科里复制过来的，可以考虑存在文件里