# 数据设计文档

**文档名称：单页应用搜索引擎优化系统数据设计文档**

**项目名称：单页应用搜索引擎优化系统**

**编写人：赖信涛**

1数据库选择

存储模块使用的数据库是MongoDB。MongoDB（来自于英文单词“Humongous”，中文含义为“庞大”）是可以应用于各种规模的企业、各个行业以及各类应用程序的开源数据库。存储在集合中的文档，被存储为键-值对的形式。键用于唯一标识一个文档，为字符串类型，而值则可以是各种复杂的文件类型。我们称这种存储形式为BSON（Binary JSON）[14]。

得益于NoSQL数据库不必定义数据表的结构的优点，在对一个页面进行持久化相关信息的时候，字段没有限制，可以随意添加。在将来，也可以任意扩展相关的字段，供扩展集群或调度爬虫使用。

因为系统使用的是NoSQL数据库，所以无需定义数据表的结构。

2 数据字段

因为没有数据表，所以字段可以灵活添加，一条标准的数据文档如下例。

{

"\_id" : ObjectId("58fedc329cd6844804c4c2f6"),

"path" : "/u/kxxoling",

"cookies" : {

},

"netloc" : "python-china.org",

"url" : "http://python-china.org/u/kxxoling",

"orig\_url" : "http://python-china.org/u/kxxoling",

"time" : 2.931,

"headers" : {

"X-XSS-Protection" : "1; mode=block",

"X-Content-Type-Options" : "nosniff",

"Set-Cookie" : "\_\_cfduid=d467504f8e0853b70629a22325cfaed261495469245; expires=Tue, 22-May-18 16:07:25 GMT; path=/; domain=.python-china.org; HttpOnly\nsession=eyJ0b2tlbiI6eyIgYiI6Ik0yUmhXRlF5V205SFExTlBOMlJWU3c9PSJ9fQ.DASePQ.kOmorQDLYeushG2BrePORbRZcCg; HttpOnly; Path=/",

"X-Sentry-ID" : "None",

"Server" : "cloudflare-nginx",

"Connection" : "close",

"Date" : "Mon, 22 May 2017 16:07:25 GMT",

"CF-RAY" : "36310741d4ef6bb0-SJC",

"Content-Type" : "text/html; charset=utf-8"

},

"status\_code" : 200,

"js\_script\_result" : null,

"md5" : "3e41da5ba07fcdb9b92384de07c90f86",

"status" : "success"

}

其中，各个字段的定义如下：

|  |  |  |
| --- | --- | --- |
| 字段名 | 类型 | 含义 |
| \_id | X16 | 唯一标志符 |
| Path | String | 路径 |
| Cookies | Dict | 使用的cookies |
| Netloc | String | 主机名 |
| url | String | url |
| Orig\_url | String | 源url |
| Time | Float | 处理所使用的时间 |
| Headers | Dict | 头信息 |
| Status\_code | Int | 网站返回码 |
| Js\_script\_result | Int | Js返回码 |
| Md5 | Hash | Md5文件名值 |
| Status | String | 处理结果状态 |
| Error | String | 错误信息 |