



**本科毕业论文**

**生活类爬虫系统**

**刘炜铭**

**201527010414**

|  |  |
| --- | --- |
| 指导教师 | **綦羽 副教授** |

|  |  |  |  |
| --- | --- | --- | --- |
| 学 院 名 称 | 软件学院 | 专 业 名 称 | 软件工程 |
| 论文提交日期 | 2018 年 4 月 15 日 | 论文答辩日期 | 2018 年 4 月 21 日 |

**摘 要**

网络爬虫其实是一个探测机器，它的基本行为就是模拟人去各个网站看看，然后带着我们所需的信息回来。我们使用的百度搜索引擎，它其实也是一个爬虫，每天放出无数个爬虫到各个网站，把信息带回来后，再提供给人们去检索。为了让某个网站更加容易被检索出来，我们可以通过爬虫提高它的点击率，这个过程就叫做 SEO（搜索引擎优化）。现在出现很多聚合电商，他们自己不生产任何产品，他们是搬运工，把天猫、淘宝、京东等商品聚合起来让用户筛选，成为电商中的百度，从而获益。可见网络爬虫与我们息息相关，它能极大的”优化”我们的生活。可以这样认为：每个爬虫都是你的“分身”，就像孙悟空拔了一撮汗毛，吹出一堆猴子一样。

本文将研究并开发一个生活类爬虫软件，它可以第一时间提示并且抢购附近低价美食、推送有关编程技术类型的文章、正在热播和准备上映的电影和机票价格等。为了方便使用，本文将用chrome插件作为客户端，让用户只要打开流浪器，就可以实时收到更新消息。从技术架构上，前端将使用react和material-ui框架开发。web服务后台使用python3的flask框架，为了提高系统稳定性，本文将介绍nginx加uwsgi等反向代理工具，开发一个多进程、单线程的web服务后台。爬虫系统使用crontab定时python3任务和xpath语法分析网页html结构，再使用正则表达式筛选正真所需的内容。使用mysql:5.70作为持久层，redis作为缓存，docker作为虚拟化工具，sqlalchemy作为orm框架，alembic作为数据库表版本控制。

关键词：网络爬虫 chrome插件 xpath crontab

**Implementation and Analysis web crawler application**

Liu Weiming

(College of Software Engineering, South China Agricultural University, Guangzhou 510642, China)

**Abstract:** In fact, web crawler is a detection machine. It`s base action is searching on the internet like common human do and back with some message, which is what we want. BaiDu Search Engines, a web crawler as well, BaiDu product a lot of web crawler every day, back and transfer to useful message, open for us searching. We can use web crawler to increase a website click through rate so that will be more easily retrieve, this skill is called SEO(Search Engine Optimization). Now is coming up so many polymeric e-commerce, they don`t need to produce own products, just showing commodity from TianMao, TaoBao, JinDong and so on, become the BaiDu of e-commerce. It easy to see, web crawler connected our life, it can make a hug convenient to our life. It can be considered that each web crawler is a divided yourself, like Sun Wukong pulling out a pinch of hair and blowing out a pile of monkeys.

This article is going to build a life-use web crawler application, the crawler will push and buy cheapest food around your city, showing the blog of programming, movie on show and upcoming, air ticket price and so on. For easier use, this article will build a chrome extensions for client, so user can receive message only by opening chrome browser. From the technical framework, using react and material-ui framework for client. Web server will use python3-flask framework, to more stable, this article will introduce nginx and uwsgi reverse proxy tools to build a multi process, single thread web server application. Web crawler system will use linux-crontab timing task python3 job and xpath syntax analysis website html structure, regex for select the information we want. Mysql:5.70 for database, redis for cache, docker for virtualization, sqlalchemy for orm framework, and alembic for database version control.

**Key words:** Web-crawler Chrome-extensions xpath crontab

目 录

[1 引言 5](#_Toc722695563)

[2 相关知识介绍 7](#_Toc107102891)

[3 需求分析 14](#_Toc1093373240)

[3 结论与展望 16](#_Toc138972382)

[参 考 文 献 18](#_Toc1777219774)

[致 谢 21](#_Toc1537623787)

# 引言

* 1. 选题的背景与意义

爬虫技术是一种最贴近人类生活的自动化技术。既可为AI自动提供有效的大数据，也可以方便人类生活。比如你的短期计划：租房、旅游、聚会。往往希望有个工具为了筛选心仪的对象。可是网上太多信息了，对于一个有其他重任的人来说，要精打细算浪费的时间（也就是金钱）是不划算的。那么很希望有这样一个软件，来帮你自动筛选，把你想要的消息推送给你。

计算机与人类息息相关，计算机的自动化特性更是取代了大多数疲惫又杂乱的工作，工人易犯错且累，计算机精准而不累。爬虫作为自动化技术的一员，对人类生活起着重要作用。

（1）搜索引擎

搜索引擎的建立需要大量的数据支撑，假如数据样本充足，人们采集信息将会极其对称且便利，人工采集显得有心无力，爬虫技术将可以代替人们源源不断采集数据。

（2）电商领域

如今电商领域，天猫、淘宝、京东、拼多多......信息五彩缤纷，价格也有高有底。如何快速搜索，快速对比成为一大难题。爬虫具有自动化的特征，可以夸张地说，爬虫可以模拟人类一切点击行为，于是聚合电商帮人类解决了这类问题。

（3）出行领域

堵车预报，天气预报，机票、动车价格，买票抢票等等，成为琐碎又不得不去关注的事情，人们每天都浪费很多时间在此。爬虫具有自动提醒功能，可以帮人们省去查看各种事儿的时间。

* 1. 国内外研究现状

自1993年初 Matthew Grays Wandered 在麻省理工学院开发出有史记载的第一个网络爬虫以来，爬虫技术已经发展了20多年，技术已日趋多样。为满足不同用户的不同需求，爬虫分为以下两大类:

（1）通用式爬虫

通用式爬虫指的是通过一系列预设的初始链接开始，根据规则提取当前网页的所需内容，然后提取当前网页里面所有的其他链接，再把这些新链接使用open-close表判重后，放去等待队列，直到到达系统预设指定条件为止的全网式爬虫技术。此类爬虫通常有两类不同的实现方式，可分为:

（1.1）深度优先搜索策略(Depth first search)，指的是从初始链接开始，选出第一个链接进行访问，处理出有用信息和新的链接集合，再从新的链接集合里面选出一个重复上诉过程，直到没有新的链接，就返回到上一层中选出一个链接进行访问。此策略具有极强的专攻性，可以爬取到最深入灵魂的隐蔽信息。

（1.2）广度优先搜索策略(Breadth first search)，又或者叫宽度优先搜索策略，指的是从初始链接开始，先把所有初始链接处理完毕，再处理从初试链接生成出来的其他链接，也就是说是一层一层的处理，这样就避免了深度优先搜索策略（1.1）会使得有些链接没机会处理的情况。

（2）倍增式爬虫(Incremental Web Crawler)，倍增式爬虫解决了僵尸链接问题，主要体现在：网站长久不更新，没有新的信息获取。这个问题对于通用式爬虫来说是一种诟病。倍增式爬虫根据网页更新的频率来自动调整网页权重，具有较高的权重的网页可以占用更高的硬件和网络资源，权重最低的只需分配剩下的计算资源即可。

* 1. 论文内容与结构安排

本文将研究并开发一个生活类爬虫软件，包括从生产到正式环境部署的过程，并且投入到实际生活使用。本文主要的工作和内容分为以下几个章节：

第一章主要介绍了网络爬虫这一领域的研究背景与意义，分析了这一领域的研究现状和应用方向，说明网络爬虫这种自动化技术已经长达20年的研究历史及人们生活对此的依赖，同时也说明爬虫与我们生活息息相关。

第二章主要介绍了整个爬虫软件所使用的相关技术知识。

第三章主要介绍了整个爬虫软件的架构与设计。

第四章主要对整个毕业设计作出总结，并分析可拓展之处。

# 相关知识介绍

* 1. React框架介绍

React 起源于 Facebook 的内部项目，因为该公司对市场上所有 JavaScript MVC 框架，都不满意，就决定自己写一套，用来架设 Instagram 的网站。做出来以后，发现这套东西很好用，就在2013年5月开源了。作为现在最火的前端框架react有着组件化，状态管理，虚拟DOM等优势，其扩展项目react-native目标更是宏大，希望用写 Web App 的方式去写 Native App。如果能够实现，整个互联网行业都会被颠覆，因为同一组人只需要写一次 UI ，就能同时运行在服务器、浏览器和手机。

React生命周期可以分为七个钩子函数，分别是：

1.componentWillMount

2.componentDidMount

3.componentWillReceiveProps

4.sholudComponentUpdate(这是一个性能优化点)

5.componentWillUpdate

6.componentDidUpdate

7.componentWillUnmount

每次钩子函数监听到变化，会重新调用render函数执行UI渲染，根据react-diff算法，不必刷新全部DOM，使得渲染效率大大提高。

React-diff算法的优势：

传统 diff 算法：传统 diff 算法通过循环递归对节点进行依次对比，效率低下，算法复杂度达到 O(n^3)，其中 n 是树中节点的总数。O(n^3) 到底有多可怕，这意味着如果要展示1000个节点，就要依次执行上十亿次的比较。这种指数型的性能消耗对于前端渲染场景来说代价太高了！

react-diff 策略，使得渲染达到O(n)复杂度。

1.Web UI 中 DOM 节点跨层级的移动操作特别少，可以忽略不计。

2.拥有相同类的两个组件将会生成相似的树形结构，拥有不同类的两个组件将会生成不同的树形结构。

3.对于同一层级的一组子节点，它们可以通过唯一 id 进行区分。

下图是React生命周期逻辑图。



* 1. Material-UI框架介绍

Material-UI是在react最流行的UI库，遵循material design，具有拟物化、质感风格、强烈的阴影和高光。

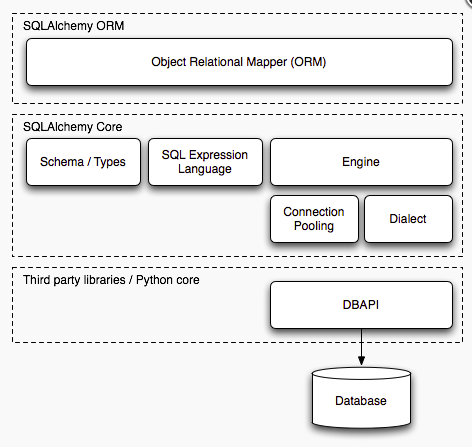
* 1. Python3-flask框架介绍

Flask是一个使用python编写的轻量级web应用框架，是基于werkzeug wsgi工具箱和jinjia2模板引擎的开源框架。由于它足够精简而且易上手，易配置扩展等特性，在github收到42731个start，相比与dingo，flask的可配置显得十分灵活。

* 1. Sqlalchemy框架介绍

Sqlalchemy是一个由python编写的orm库。提供了SQL工具包和对象关系映射工具，sqlalchemy把数据库表表达为python的list，每个list是一个truple，表示一行记录。Sqlalchemy采用的是数据映射模型，并非Active Record模型。加上python这类动态解析语言等特性，想到达到操作数据库的目的，往往只需要写一个公共使用的ORM文件即可，极其方便。SQLAlchemy首次发行于2006年2月，并迅速地在Python社区中最广泛使用的ORM工具之一。

Sqlalchemy采用两层架构，如图所示，最底层的操作数据库采用第三方库的形式完成。



* 1. Alembic框架介绍

Alembic是一个基于sqlalchemy的数据库版本迁移工具，在往后开发中，往往需要增表改表，alembic可以做到类似git一样的，保存数据库的各种历史版本，随着代码更新而部署不同的数据库结构。

* 1. MySQL:5.70数据库介绍

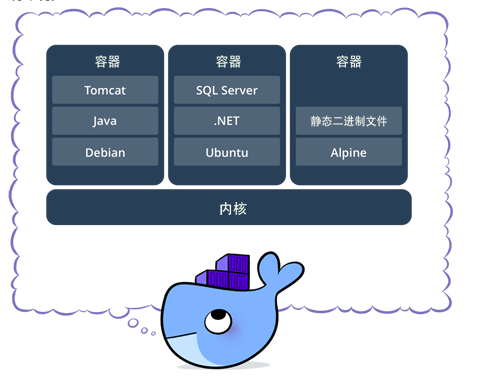
MySQL是一个关系形数据库管理系统，是最好用web-base应用数据库之一，支持事务、表级、行级锁等等。目前属于Oracle下的产品，是一个开源软件。其中，5.70是mysql一个比较稳定的版本，也是网上技术文章最多的版本，本文使用此版本作为主要存储工具。

* 1. Redis介绍

Redis是一个内存型的key-value数据库，支持hash，string, list, set，order\_set等多种数据结构。内存型数据库能火爆体现在：1、内存越来越廉价，网络速度越来越快。2、效率比落磁盘要高得多。3、redis也能极大可能保存用户存取的数据。Redis是一个开源项目，于2010年3月15日进行开发。

* 1. Docker虚拟化技术介绍

在Docker问世之前，部署项目代码视乎不是一件容易的事情。还有专门运维的团队，由于直接在服务器上安装依赖往往会弄坏系统变量，导致部署流程很繁琐。Docker提供一种虚拟化技术，支持把服务运行在docker-deam下的container中。一句话概括容器就是将软件打包成标准化单元，以用于开发、交付和部署。如下图所示，每一个container是独立的，因为每个container都有独立的PID，NET，IPC, USER, UTS, MNT，所以每一个container都是互不影响的，只有这样，我们才可以为每一个container安装不同的依赖而不影响宿主机。



相比于虚拟机，docker的优势是很明显的。如下图所示，传统的虚拟机技术首先需要虚拟出一套硬件，再在其上面运行一个完整的操作系统功能，在该系统上再运行所需的进程。而docker的容器内的应用进程则直接运行于宿主机内核中，容器没有自己的内核，而且没有进行硬件虚拟。因此容器要比传统虚拟机更为轻便。



容器与虚拟机对比总结图如下：



* 1. Linux-crontab介绍

Linux下的定时任务，crontab成为很多服务器定时任务的首选。在LINUX中，周期执行的任务一般由cron这个守护进程来处理。cron读取一个或多个配置文件，这些配置文件中包含了命令行及其调用时间。cron的配置文件称为“crontab”，是“cron table”的简写。

crontab是在后台维护一个守护进程，每一分钟调用一次，每一次调用，则去三个地方查找crontab文件，进行解析，运行对应语句，然后主程序休眠，如此反复。

1. /var/spool/cron/ 这个目录下存放的是每个用户包括root的crontab任务，每个任务以创建者的名字命名，比如vimi建的crontab任务对应的文件就是/var/spool/cron/vimi
2. /etc/crontab/ 这个文件负责安排由系统管理员制定的crontab
3. /etc/cron.d/ 这个目录用来存放任何要执行的crontab文件或脚本

* 1. xpath路径表达式语法介绍

xpath是一种选择XML文件中节点的方法，HTML文档也是XML文档的一种，利用xpath语法可以很方便地找出对应元素的对应DOM。

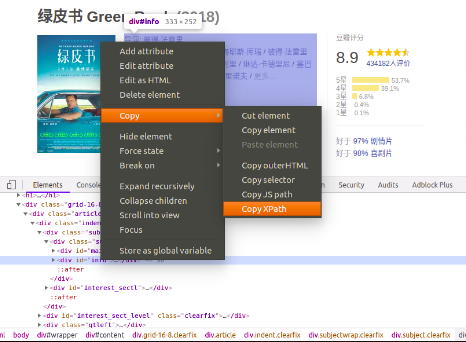
xpath语法介绍:

|  |  |
| --- | --- |
| nodename | 选取此节点的所有子节点。 |
| / | 从根节点选取。 |
| // | 从匹配选择的当前节点选择文档中的节点，而不考虑它们的位置。 |
| . | 选取当前节点。 |
| .. | 选取当前节点的父节点。 |
| @ | 选取属性。 |

谓语（Predicates）

|  |  |
| --- | --- |
| //div/dd[1] | 选取div元素的第一个dd子元素 |
| //div/dd[last()] | 选取div元素的最后一个dd子元素 |
| //title[@lang=’eng’] | 选取所有title元素，有值为end的lang元素 |
| //title[@lang] | 选取所有title元素，有lang元素 |
| \* | 匹配任何元素节点 |
| @\* | 匹配任何属性节点 |

更方便的是，对于chrome浏览器，打开调试模式可以很容易复制当前元素的xpth路径。



* 1. uwsgi介绍

WSGI(Web Server Gateway Interface)定义了web服务器(nginx, apache, iis等)和web应用(flask, django等)之间的接口规范，也就是说，只需要遵循wsgi协议，就可以和nginx等通信。

uwsgi基于c语言开发的一个web服务器，它实现了WSGI协议，它会启动若干个flask程序进程，每当有请求进来时，它会随机分配个一个空闲的进程处理，每个flask程序配置成单线程，就构造出一个多进程，单线程的web-server程序。因为web服务瓶颈在IO，不在CPU，所以设计成单线程可以减免线程和锁之间的切换花销。

* 1. nginx介绍

Nginx是一个网页服务器，它能方向代理http，https，smtp等等链接协议，以及是一个负载均衡器。

为什么需要使用nginx？作为web-server，很容易受到别人攻击，如果只单单使用uwsgi作为服务器的话，很容易被别人使用慢链接攻击，导致服务器所有flask程序进程处于busy状态，从而导致服务不可用。Nginx是一个使用c语言开发的高效http反向代理工具，处理慢链接是它的强项。Nginx会把整个http请求接受完整后，再反向代理转发到uwsgi中，uswgi再根据情况选择一个进程处理，然后再交给nginx返回给客户端。整个过程nginx处于第三者工作，抵挡外面先来的请求。保护后台服务进程。同时，nginx具有高效的转发静态资源能力，可以快速传输静态资源给客户端。

Slowloris慢速攻击是rsnake发明的一种http攻击技术，首先http协议规定，HTTP request以\r\n\r\n(0d0a0d0a)结尾表示客户端发送结束，服务端开始可以处理。那么如果永远都不发送\r\n\r\n，客户端以一定的时间间隔发送一个key-value格式的数据到服务端，就可以形成类似的攻击。

# 需求分析

* 1. 系统概述

生活类爬虫系统旨在“优化”人们生活，减少人类机械又低效的网页查询工作。系统采用chrome插件作为前端展示，结合nginx、uwsgi、flask、sqlalchemy框架作为web服务后台，xpath，regex，linux-crontab作为爬虫后台系统。实现了包括用户管理，信息模糊查询，热播、准备上映电影展示，附近美食展示，编程类技术博客展示等功能模块。

* 1. 功能需求

用户管理：用户注册，用户登录或者退出登录。

信息查询系统：把用户输入的信息在所有爬回来的消息中进行模糊查询。

电影展示模块：展示所有爬回来的电影信息，带有购票功能，查看电影详细信息，以及查看电影预告片等。

编程类技术博客展示模块：显示博客关键词信息，标题。跳转到博客页面等功能。

美食展示模块：爬取美团网信息，展示地址和图片信息。带有查看团购价和一键购买。

机票价格展示：查询来自北方航空，南方航空，春秋，九元航空等机票价格信息。

* 1. 系统可行性分析

1、从用户角度上，据调查大多数只坐在电脑前面的用户，都希望有一个自动提醒各种自

己心仪消息的软件。而自己只需要正常工作，就能看到最新的消息。其实和能自动提示邮件信息类似，每个人都想有一个软件自动提示收件箱的最新邮件，而不是自己担惊受怕的点开邮箱，生怕错过重要邮件。

2、从技术角度上，爬虫已经经过20余年的发展，网上成功例子或者开源的爬虫库也很多，可以参考。其实整个互联网世界充斥着爬虫，是有先人铺垫的。前端chrome插件，chrome作为最多人使用的浏览器，用户群大，背后的谷歌公司是世界上最大的IT公司。Web-server端采用nginx + uwsgi + python3-flask实现的这种多进程单线程服务是web开发主流，所以从技术角度上是完全可行的，都有前辈的经验指导。

3、从经济角度上，只需要一台电脑即可开发，一台服务器也可部署。

从以上三方面考虑，整个爬虫软件是很有开发前景的。

* 1. 系统运行环境

硬件：RAM: 512mb, Disk-size: 10GB, system: Ubuntu 16.04 x86\_64, CPU:QEMU Virtual CPU version (cpu64-rhel6)，

软件：docker、mysql:5.70、redis、linux-crontab

# 系统分析与设计

* 1. 系统解决思路

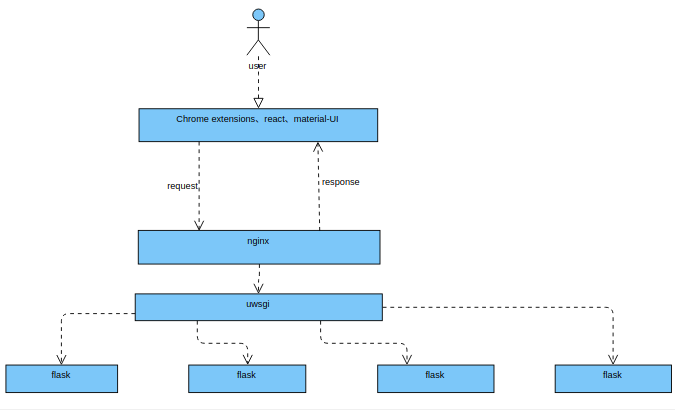
4.1.1 爬虫系统的实现：使用linux-crontab定时任务处理，不使用语言本身自带的定时器处理是因为不希望程序一直占用内存和数据库，redis的链接。每次唤醒爬虫服务，都会读取待爬的链接，然后一个一个去带着cookies请求，返回HTML后使用xpath语法获取所需元素的节点，regex判断内容是否合适，然后存去redis和mysql准备推送。

4.1.2 后台系统的实现：使用nginx + uwsgi + python3-flask构造一个多进程，单线程的web服务后台。对于每一个请求，先经过nginx反向代理到uwsgi守护进程，再由uwsgi拉起一个空闲的flask进程处理该请求。处理请求时，先判断cookies和所需参数是否合法。

4.1.3 前端系统的实现：使用react作为主要开发框架，material-ui作为主要UI框架，react-router作为前端路由控制，chrome插件notification作为主要提醒机制。前端首次加载会引起background.js的调用，设置每5分钟去轮询后台接口是否有更新数据。有则提示推送消息，没有则挂起。

* 1. 系统体系结构设计

本系统采用nginx反向代理web应用服务器以及前端使用react配合fetch进行前后端通信。流程图如下：



* 1. 系统包设计

4.3.1 爬虫系统没有使用任何框架，其中主要分包如下：

Local\_dev包：主要负责部署模块，包含makefile和dockerfile等部署文件。

Config包：爬虫系统的主要配置，包括redis主机IP和mysql主机IP等。

Database包：sqlalchemy库的DAO封装

Lib包：包括日志系统，还有错误处理逻辑。

Decorator包：一些装饰器包，包括python的单例实现等。

Testing包：单元测试用例。

Models包：model层操作。

Request\_client包：爬虫系统的主要部分。

4.3.2 后台系统使用flask框架，其中主要分包如下：

Local\_dev包：主要负责部署模块，包含makefile和dockerfile等部署文件。

Endpoint包：后台系统各种api。

Models包：model层操作。

Config包：后台系统的主要配置，包括redis主机IP和mysql主机IP等。

Database包：sqlalchemy库的DAO封装

Lib包：包括日志系统，还有错误处理逻辑。

Decorator包：一些装饰器包，包括python的单例实现等。

Testing包：单元测试用例。

4.3.3 前端展示使用flask框架，其中主要分包如下：

Api包：主要用来发起ajax请求，使用async和await封装fetch。

Lib包：主要负责前端日志系统。

Account包：前端账号系统包。

Page包：前端主要页面包。

* 1. 系统数据库设计

表一，用户表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 字段 | 数据类型 | 含义 | 是否主键 | 索引 |
| Id | Int | 用户id | YES | 主键索引 |
| Username | String | 用户名称 | NO | 唯一索引 |
| Password | String | 用户密码 | NO | NO |
| Create\_time | Timestamp | 用户账号创建时间 | NO | HASH索引 |
| Json | Json | 用户属性 | NO | NO |

表二，美团美食表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 字段 | 数据类型 | 含义 | 是否主键 | 索引 |
| Id | Int | 美食ID | YES | 主键索引 |
| poiId | Int | 美食poiId | NO | 唯一索引 |
| frontImg | Varchar | 美食图片url | NO | NO |
| Title | Varchar | 美食title | NO | HASH索引 |
| avgScore | Float | 美食评分 | NO | INDEX索引 |
| allCommentNum | Int | 美食评论数 | NO | NO |
| Address | Varchar | 美食地址 | NO | NO |
| avgPrice | Float | 美食平均价格 | NO | INDEX索引 |
| create\_time | Timestamp | 美食创建时间 | NO | HASH索引 |
| dealList | Json | 美食团购列表 | NO | NO |
| Json | Json | 美食其他属性 | NO | NO |

表三，豆瓣电影表。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 字段 | 数据类型 | 含义 | 是否主键 | 是否索引 |
| Id | Int | 电影表id | YES | 主键索引 |
| File\_id | Int | 电影豆瓣id | NO | INDEX索引 |
| Title | Varchar | 电影题目 | NO | NO |
| Score | Float | 电影平均评分 | NO | INDEX索引 |
| Region | Varchar | 电影源地 | NO | HASH索引 |
| Img | Varchar | 电影封面 | NO | NO |
| Url | Varchar | 电影链接 | NO | NO |
| buy\_ticket | Varchar | 购票链接 | NO | NO |
| create\_time | Timestamp | 电影创建时间 | NO | HASH索引 |
| onshow\_time | Varchar | 电影上映时间 | NO | NO |
| Json | Json | 电影其他属性 | NO | NO |