渠道信息抓取功能

技术设计文档

一.涉及技术

网页抓取 httpclient

网页解析 jsoup

对象序列化 fastjson

数据库读取 sqlite

二.包结构

xiaobai.html 网络处理相关的类

SpiderHttpClient 页面抓取工具类

xiaobai.base 基础类

BasicStatistics 通用统计数据类

xiaobai.channel 渠道信息相关的类

Channel 渠道类

Article 文章类

Comment 评论类

User 渠道用户类

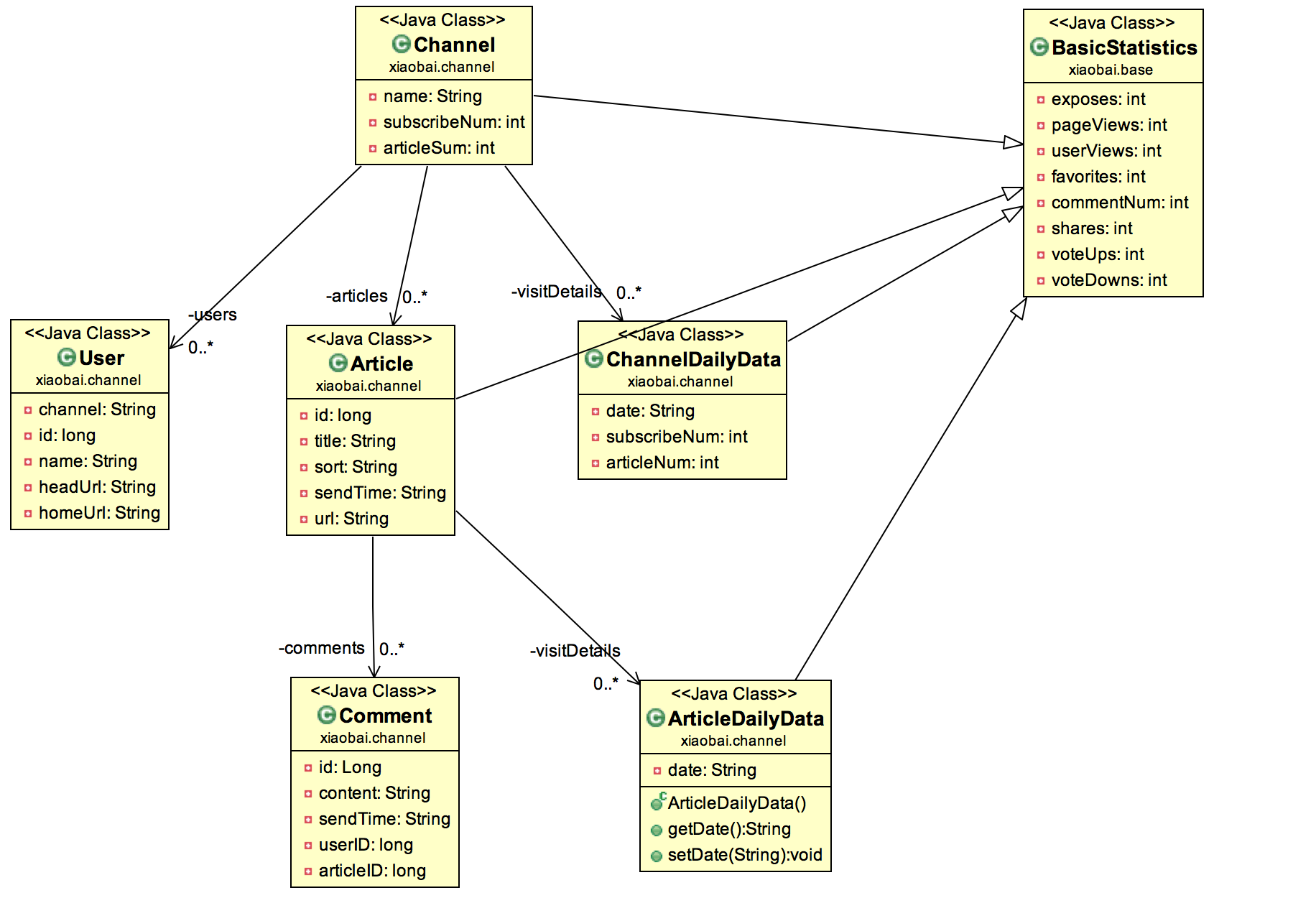
ArticleDailyData 文章每日详细数据

ChannelDailyData 渠道每日详细数据

xiaobai.analyze 渠道数据库分析的类

ToutiaoAnalyze 头条信息分析类

三.UML图

四.实现功能

1.渠道总访问量

2.渠道每日访问量

3.渠道所有文章的访问量

http://www.toutiao.com/group/3722003547

http://www.toutiao.com/preview\_article/?pgc\_id=3653125311



4.渠道所有文章的评论

http://www.toutiao.com/group/%article\_id%/comments/?count=%comment\_num%

该功能通过头条对外接口实现.



5.渠道所有对文章做评论的人

注意事项:

1.为了防止被当成攻击,需要控制页面抓取频率,现在控制为每秒抓取一次.

2.每次扫描将数据存储为json格式;方便做数据的二次处理.

3.为了使用fastjson,所有类的成员变量都需要有set-get函数;所有的类都需要有默认构造函数

后续改进

2.为了减少每次扫描的频度,并且保证每次数据抓取的时间;

程序实现为每天增量抓取;每周全量抓取一次.

3.加入log4j, 实现常规日志和调试日志分离;

4.

参考文献:

<http://www.lijiejie.com/python-http-request-with-chrome-cookies/>