**May：IIIMMS-Crawler Work Log**

**HuYang 2014.05**

**第一部分：采集器站点数量的添加**

1. 为项目组全体成员进行讲座，培训内容包括：采集器的基本运行原理；采集器核心技术Jsoup的API讲解及使用方法；采集器grab预爬取阶段的参数配置方法；采集器fetch解析阶段的参数配置方法；页面解析模块分页方法编写。
2. 指导小组成员，戴丹，王秋森，俞浩亮配置BBS站点参数进入mysql数据库。
3. 本人添加豆瓣小组，搜狐圈子，百度百家等公共平台，新浪Blog的站点采集参数配置。

**第二部分：采集器压力测试**

1. 采集器在130服务器上经过单线程长时间压力测试和多线程短时间以及多线程长时间压力测试。
2. 测试结果，单线程：cpu占用率5%，速度：18页/min；多线程：cpu占用率：17%，速度：40页/min。
3. 测试过程中遇到采集器内核中的小Bug，及时进行修补，在此不再赘述。
4. 编写采集器即时中断所有线程的功能模块。

**ps**：方案简述，设置作用域为整个项目的全局变量threadEndFlag记录所有线程是否死亡的bool值，当线程开启的时候，只要主任务线程没有全部死亡，这个bool值都会为false（意义为：线程死亡的事实为假），手动终止线程的操作就是修改ThreadEndFlag的值为true，在采集器内核的每个细小阶段都会检查这个bool，发现值为true，则停止任务，线程随之结束。这样做结束线程不会造成线程垃圾，并且可以很好的记录当前运行的状态，方便断点续传。经过测试，此方法终止任务会有5s左右的延迟。

**第三部分：采集器内核功能增强**

1. AnalyzerTime和AnalyzerNum（见：ims.crawler.grab.util）类中时间处理和数字处理的算法增强，时间处理方法增加对字符串：“？分钟前”，“？小时前”，“今天？？”，“昨天？？”以及“年，月，日”等中文字符的解析，支持更多时间格式，支持更多数据格式。时间处理类增加如下功能：a.计算两段时间的时间间隔；b.将时间间隔由yyyy-MM-dd HH:mm:ss的时间格式转化为整数类型的秒数；c.b的求反。
2. mongodb操作基础接口的功能增强：增加的功能如下：a.获得数据库中所有集合（等同于sql数据库中的表）的名称；b.获得指定集合中记录的条数；c.获得指定集合中满足某条件的记录的条数；d.不设置分页的条件查询；e.设置分页的条件查询。

mongodb操作service接口对应基础操作接口功能进一步增强，衍生出更多外围操作service配合前台级联删除，查询工作。

1. 公共媒体的tool工具包（包括grab部分和fetch部分）的编写，hcBBS获取tool工具包（fetch部分）的编写，博客tool工具包（fetch部分）的编写，简单登录tool工具包（fetch部分）的编写。
2. Hc（httpclient）链接模式升级，加入模拟firefox获取页面html，解决basic连接模式获取html不完整的问题。
3. 在grab和fetch阶段中加入对特殊内容的检查和解析机制。

**ps**：机制共分为两个部分：检查是否在特殊内容判断列表中（一个以站点为单位的全局静态容器），根据站点区分解析出特殊内容，目前特殊内容支持，特殊解析方式的时间和特殊解析方式的数字。在原有的grab和fetch类中抽象出统一的时间和数字解析方法，采用这种设计模式，如果出现新的特殊情况，只需要在全局静态容器中加入对应的站点即可。

1. 添加通用分页实现类（只取每个帖子内部的第一页），废弃功能类似的分页实现类。

**第四部分：采集器内核性能优化**

1. MySql节点表（数据量大），之前设立了缓存池，方便增量爬取，在每次开启多线程任务之前，添加对缓存池的清空操作。
2. 由于出现对于C3P0连接池对数据库的访问量超过限制，需要对采集器内核中请求spring实例的方式进行改进，改进如下：设置全局静态链接池，只需请求一次，提供给采集器中所有的spring实例请求（单例模式在本项目中大量应用）。
3. 将总任务大线程的调度模式升级为线程池模式，使用java6新增加的线程池功能，选择固定数量的线程池，站点分任务的提交全部一次性提交给线程池，但是只有固定数量的线程被激活，其中的调度交给JVM实现。线程池的数量可以人为设定，默认设定为10个线程。完善在这种模式下的线程等待和同步机制，达到与之前一样的效果。

**第五部分：对后期任务的准备和规划**

1. 以上任务95%在本月23日前都已完成，后面一周查看资料，思考和准备后期NLP，机器学习的功能及任务规划，重点是采集文本的分类和不良信息的识别监测，对分类结果的统计，敏感语料的检索。分析对比多种开源机器学习工具包，为后面的选用作为参考，这一工作进展较慢，原因是英文文档偏多，系统性的教程缺乏，对相关算法的熟悉度欠缺。最迟在6.2之前提交相关分析规划文档。
2. 除此之外是配个其他团队成员完成采集器相关控制和逻辑功能。