本文档是小组成员superhy对后期项目功能以及实现方式的一些设想。

——superhy 2014.5.5

**第一部分：采集器功能总结和原始情报库的使用**

1. 采集器所用功能完善后，应当达到如下成果：实现多线程多站并行爬取，爬取内容直接存入NoSql数据库，NoSql数数据库存储情报的内容，来源节点对应关系型数据库的指纹信息，情报的作者，子情报（如帖子的回复）的内容，子情报的作者等信息；采集器日志系统能够较为及时真实地反映采集器的运行情况；采集器才去增量爬取，不重复爬去已有内容；采集器实现基本的阶段任务记忆机制，支持暂停，终止，恢复，取消等操作。**采集器最终压力测试，功能逻辑检查，性能调优**。
2. 采集器工作的成果是存储在mysql数据库当中的主题列表（tbl\_theme），节点列表（tbl\_post）和存储在mongodb中的节点内容文件集合（corpus），其中最主要的是存储在mongodb.corpus中的节点内容信息，需要编写mongodb查询类实现对内容的查询，并能够关联查询出所属的节点，节点所属的站点，以及发布内容作者等等。这里总结为工作模块：**编写mongodb查询功能，内容属性关联查询功能**。

**第二部分：不良信息种类定义，不良信息种类训练集合生成，不良信息自动分类功能的实现**

1. 定义不良信息的类别，通过常识定义出不良信息的几种大类，比如：淫秽信息，暴力信息，反动言论，虚假恐慌煽动信息等。
2. 原始情报库文档关键词提取，**启动分词器（中科院分词器，mmseg4j分词器<java实现，轻量级，分词效果优秀，可以整合Lucene>），过滤停用词，tf.idf计算词的权重（Lucene检索计算词频），提取文本关键词，对文本关键词进行情感识别，提取不良信息关键词**。
3. \*原始情报库聚类服务，将原始情报库中部分文档的内容聚类为不良信息的各种分类，清除健康文档，保存不良信息文档及其所属类别信息，**不良信息语料库聚类功能实现（备选算法：层次聚类法，K均值，DBScan算法）**，聚类结果成为后期不良信息分类的训练集。Ps：不良信息种类训练集的生成也可以使用其他方法。
4. 其余文档及新增加文档不良信息的自动分类功能实现，结合文档关键词提取模块对待分类文档进行特征提取操作，**相似度计算模块编写（LCS），特征选择模块编写（CHI特征提取方法）**，**文本自动分类模块编写（备选算法：K近邻算法<KNN>，中心向量法<Rocchio>，支持向量机算法<SVM>）**。Ps：SVM算法实用设计简单规划：①第一次总体分类：SVM\_All学习：文档k∈不良信息 or 文档k！∈不良信息，分出不属于不良信息的文档，即时清除。②第二次具体分类：SVM\_i学习：不良文档k ∈ 不良信息种类I or ！∈不良信息种类i。③完成分类。④拉格朗日乘子法解决SVM线性约束优化问题⑤\*按照①，②两步的思路操作SVM学习多级分类。

**第三部分：不良信息各项指数统计（何佳师兄指导）**

1. 文档不良信息各项指数统计，各站点不良信息各项指数统计，各类不良信息出险情况统计等等，**统计算法模块编写，统计指数图表制作**。
2. \*统计结果反馈指导采集器侧重参数，体现在tbl\_grabUserParame参数表，**统计结果分析模块实现，侧重权值算法设计，自动修改参数表模块实现**。

**第四部分：不良信息文档简报生成，各类语料库（不良信息库，原始情报库）全文检索功能实现**

1. **不良信息文档自动摘要功能实现（Lucene-highlighter动态摘要），**不良信息统计情况显示，**不良信息统计结果查询聚合功能实现**，不良程度评级，**不良信息评级定义，不良信息权值求和自动评级功能实现，不良信息程度排序功能实现**。
2. **语料库全文检索功能实现（Lucene），\*按不良信息评级程度排序功能实现（Lucene索引权值管理）**。

**第五部分：不良信息自动识别，其他功能**

1. 不良信息自动识别，手动输入文字段落或输入URL链接，采集链接中的文本内容（调用采集器接口），形成文字段落，调用分类算法，判别不良信息情况，判断是否存入不良信息库，**采集接口调用整合，分类算法调用整合，手动添加不良信息（经过不良情况判断，已知部分自动添加，未知部分手动添加）功能实现。**
2. **系统管理等其他功能按照新系统调整**。

【注1】：除标题以外加黑部分为工作量预测。

【注2】：其他开源工具：开源关键词提取开源工具：KEA（<http://www.nzdl.org/Kea>）；数据挖掘与机器学习工具weka（<http://www.cs.waikato.ac.nz/ml/weka/>）；索引管理工具Lucene-luke（<http://code.google.com/p/luke/>）；信息抽取框架GATE（<http://gate.ac.uk/>）；文本自动分类开源工具classifier4j<整合中文分词器可实现自动摘要>（<http://sourceforge.net/projects/classifier4j/>）