现代信息检索项目组会议记录

团队成员：陈新荃(组长) 林裕杰 高妍 肖卡飞

目录

[2014-9-23 2](#_Toc402732636)

[2014-9-28 3](#_Toc402732637)

[2014-10-18 4](#_Toc402732638)

# 2014-9-23

项目进度

* 小组成员确定为4名。关于本小组：不仅仅是项目组，也是学习小组，可相互交流探讨学业问题，互帮互助，取长补短。
* 一致决定定期进行一次小组会议
  + 时间地点：**周日晚八点半，地点待定。**
  + 会议内容为：
    - 分别汇报项目各个模块的进度
    - 对遇到的问题进行交流
    - 安排本周工作重点
    - 知识输出，即谁有接触到比较impressive的知识可以介绍给大家（optional）
* 确定了项目题目为：搜索类第一题，**体育新闻检索系统**
* **系统要实现的功能（按照流程顺序）**：
  + 信息数据抽取
    - 设计网页爬虫工具：
      * 可参考现有开源工具，需对抽取算法进行修改
      * 注意提高抽取算法鲁棒性，而不是只针对固定几种网页模板
      * 实现定期更新（e.g.网站添加了新的url或者原有url网页内容被修改过）
  + 建立索引
    - 可先利用lucene自带方法建立倒排索引，后期优化尝试多种索引建立的实现方法，e.g. MapReduce等
  + 查询语句处理
    - 可利用好的开源中文分词器
      * 需对现有中文分词器进行比较，选出优秀的一个
      * 将其嵌入lucene框架
  + 检索结果处理
    - 对相似新闻进行聚类
      * 聚类算法
      * UI上如何显示
    - 排序：
      * 默认按照相关度排序
      * User可以自己定义排序标准，e.g.相关度(location)、热度(评论数)、时间等
  + Web UI设计与交互
    - 搜索页：
      * 可配置项，高级搜索等
      * 查询自动补齐
    - 结果页：参考百度
      * Snippet生成
      * 结果预览（快照）
      * 相关搜索推荐
  + 系统测试
    - 召回率
    - 精准度
    - 响应速度

本周工作安排

* Lucene上手
  + 上网查阅相关资料
  + 尝试在自己电脑配置Eclipse，按照网上教程运行一个简单的lucene检索实例，有感性认识。
* 讨论检索系统开发基础内容：
  + 团队名称
  + 检索系统名称
  + 统一开发环境
    - 检索框架lucene
    - 结合网页显示的系统框架
    - 使用代码控制工具：github、svn等
    - 系统模块划分以及最初版本的功能设计

Tips

* 在项目分工之后，大家根据自己的模块搜索相应论文来看，同时借以完成本课程的阅读作业。

# 2014-9-28

项目进度

* 确定了项目的各部分命名：
  + 团队名称：GeeK
  + web搜索引擎名称：Geeking
    - project以及类命名（以geek为开头命名），示例：
      * 爬虫project名称：geek-spider …
      * 索引部分project名称：
        + geek-analyzer
        + geek-parser
        + geek-searcher
        + …
  + 确定以github为代码协同工具，维护代码以及开发日志等。
  + 确定开发环境：
    - Eclipse + tomcat7.0 + JDK1.7 + Jersey
  + 确定了小组成员分工：
    - 高妍：web 框架，UI设计与交互等
    - 卡飞：爬虫，关键信息抽取等
    - 新荃&裕杰：索引建立，查询，排序等

本周工作安排

* 高妍：Jersey框架等🡪almost done
* 裕杰：学习java，mysql，文本检索等
* 卡飞：学习java，爬虫等
* 新荃：建立github项目，学习lucene等

Tips

* 关于学习java：个人建议直接看现有的代码，有疑问可群上讨论或自行谷歌。
* 关于分工：各个成员虽然侧重点不一样，但是分工的目的是充分提高团队的效率，因此不建议个人封闭式冥思苦想，有想法大家讨论，集思广益！make it better and better！
* 以及，建议各成员自己维护一份项目开发日志，记录每周的所想所得，事后回顾会发现这个过程学到了很多。
* 预祝大家国庆快乐！！~~

# 2014-10-18

项目进展情况

* Github环境基本搭建完成：
  + Github团队名称：UCAS-GeeK
  + 项目名称：UCAS-GeeK/Geeking
* 学习IBM developerWorks中介绍的搜索引擎实现
  + 深入理解程序运作流程
  + 对爬虫、建立索引以及查询过程有更深入理解
* 卡飞：基本实现体育新闻网页抓取
  + 遇到的问题：
    - 网页url混乱，格式不一致
    - 网页源文件存储目录太深
  + 下一步工作：
    - 化简存储目录，以url为名称存储网页文件
    - 考虑网页内容抓取的完整性
      * E.g.热度所需关键字是否可抓取（评论数由java script计算）
    - 考虑程序稳定性
    - 考虑周期更新
* 高妍：提供非常值得学习的搜索引擎相关资料
  + 下一步工作
    - 有爬虫及编程经历，协助卡飞完成爬虫内容
    - 研究搜索词补全功能等
* 裕杰：搭建mysql数据库环境
  + 下一步工作
    - 研究网页预处理
      * 设计数据库要存储的属性列表
      * 通过jdbc调用mysql实现基本的数据库操作
* 新荃：搭建github团队协同工作项目
  + 下一步工作
    - 协同裕杰研究网页预处理及索引构建的设计
      * 基于位置信息的倒排索引
      * 考虑分层索引及文档权重排序

Tips

* 为了避免冲突，请在GeekSearch工程下新建包之后截图到群上公示，并尽快pull request

# 2014-11-02

项目进展情况

* 卡飞：实现了四类门户网站体育新闻网页抓取，优化了网页存储目录结构
  + 遇到的问题：
    - 存储了部分非html文件
    - 工程main函数太多
  + 下一步工作：
    - 在存储前过滤掉url后缀名不是html/htm/shtml…等网页
    - 将多个门户网站的抓取main函数合并成一个
    - 将工程编码改为UTF-8；
    - 规范类、方法以及变量的命名
    - 考虑网页内容抓取的完整性
      * E.g.热度所需关键字是否可抓取（评论数由java script计算）
* 新荃：基本实现倒排索引的设计与实现
  + 遇到的问题：
    - 网页日期未获取
    - 未过滤停用词
    - 网页编码
  + 下一步工作
    - 实现日期获取以及停用词过滤
    - 考虑基于“标题/描述”构建一级索引，以及tf-idf计算，实现搜索结果排序
    - 协同高妍实现query部分的结果排序
* 高妍：已完成界面部分的设计与编写
  + 下一步工作
    - 基本实现查询功能，以及查询结果的显示等
    - 考虑词组搜索的位置信息。
    - 考虑搜索词补全，动态摘要，快照，搜索建议等功能
* 裕杰：完成各类索引在mysql的设计与建立，以及数据库调用的封装
  + 下一步工作
    - 将程序中的路径用户名等配置信息迁移到properties配置文件中
    - 考虑使用redis数据库
    - 研究搜索引擎，主要为本项目工程的测试方法，设计测试方案

Tips

* 项目进展顺利，各位加油！