**基于Lucene的分析与应用**

**项目计划变更说明书**

小组成员：

刘宏宇

滕延林

顾泽鹏

杨帆

周晓懿

# 项目计划变成情况

|  |  |  |  |
| --- | --- | --- | --- |
| 项目内容 | 初始计划 | 改进计划 | 变更原因 |
| 工作内容 | ①了解工作流程； ②分析代码； ③开发与优化； ④测试； | 无 | - |
| 参加人员 | 刘宏宇 滕延林 顾泽鹏 杨帆 | 刘宏宇 滕延林 顾泽鹏 杨帆 周晓懿 | - |
| 项目期限 | 本学期第十七周 | 无 | - |
| 项目分工 | 刘宏宇：答辩；记录；索引机制部分代码研读 滕延林：资料收集；环境搭建；搜索机制部分代码研读 顾泽鹏：计划制定与记录；段合并机制部分代码研读 杨帆：答辩；记录；分词器代码研读 | 每名成员均承担一部分文档的撰写和项目开发的工作； | 充分发挥人员作用；分散每周任务压力； |
| 组织形式 | ①每周一次会议，讨论当前工作进展以及各自的问题； ②建微信群，保持实时的交流； ③资源的共享通过github进行实现； | 无 | - |
| 交付产品 | ①软件项目计划书 ②需求规格说明书 ③软件评审报告 ④软件产品改进与展示 ⑤软件测试分析报告 ⑥软件测试评审报告 ⑦软件进度分析报告 ⑧工作量分析报告 ⑨配置管理 | ①软件项目计划书 ②需求规格说明书 ③实现方案改进说明 ④测试需求规格说明书 ⑤软件测试报告 ⑥项目统计分析报告 ⑦配置管理总结报告 ⑧计划管理总结报告 | 根据各实验要求调整 |
| 软件功能 | 文本入库 | 无 | - |
| 文本检索 | 无 | - |
| 智能化检索 | 删除 | 缺少统一标准，概念比较模糊，测试难度大，比较耗时 |
| 非结构化数据管理 | 删除 | 过于复杂 |
| 数据挖掘 | 删除 | 数据量有限 |
| - | 增加：中文分词 | 在处理中文时，必须用到中文词汇作为索引表表头，而Lucene中没有中文分词功能 |
| - | 增加：图像检索 | 许多文档中含有图像，而这些图像往往包含重要信息。此外图像类型的数据十分丰富，而Lucene并不支持图像的检索 |
| - | 增加：哈希 | 为了提高系统的性能，降低检索时间 |

# 项目执行情况

|  |  |  |
| --- | --- | --- |
| 项目内容 | | 执行情况 |
| 工作内容 | 了解工作流程 | 按计划执行；已完成 |
| 分析代码 | 按计划执行；已完成 |
| 开发与优化 | 按计划执行；已完成 |
| 测试 | 按计划执行；已完成 |
| 参加人员 | | 按计划执行 |
| 项目期限 | | 按计划执行 |
| 项目分工 | | 按计划执行 |
| 组织形式 | | 按计划执行 |
| 交付产品 | | 按计划执行 |
| 软件功能 | 文本入库 | 按计划执行；已实现 |
| 文本检索 | 按计划执行；已实现 |
| 中文分词 | 按计划执行；已实现 |
| 图像检索 | 按计划执行；已实现 |
| 哈希 | 按计划执行；已实现 |

# 计划变更类型

项目最初的计划是对lucence进行改进，改进的方向包括缩小检索文件的大小、提高检索的效率、提高检索的准确率等。主要的工作重点在于改进。在于对lucence中的源代码进行阅读并对其进行修改。

变更后，我们的计划改为对lucence的功能进行扩展。原有的lucence只能进行英文的分词和检索，我们对其进行添加了中文的检索以及图像的检索。其中，图像并没有倒排索引的方式用来提高检索的效率，因此，我们计划采用hash的方式。项目变更后，我们的工作重点在于扩展。不需要对其源代码进行修改，但是需要对源代码进行调用。同时，利用已有的知识、已有的技术对lucence进行扩展。

项目类型由原来的改进，变更为现在的扩展。

# 变更原因

* 对lucence实现的原理和细节把握不够强。
* 有限的时间内，很难对lucence的性能进行提高。
* 了解一些图像检索的原理，能够对lucence进行图像检索的扩充。

# 计划与实际的偏差

计划对图像进行分批入库、建立索引。目的是为了提高检索的效率。比如当有100万张图片如果直接对其建立索引，可能在建立索引以及检索的时候，将花费大量的时间。因此我们将其分为5个block，分别建立索引，在hash的时候，分别对5分block进行hash，检索的时候，也分block进行检索，然后对block检索的内容进行合并。

但是，实际的操作中，我们很难得到大量的图片。我们的服务器磁盘空间也有限制。我们最终只有5万张测试图片，然而对5万张图片进行分批处理的时候，检索效率反而不如不分批处理，原因就是图像数量太少。同时，如果进行分批处理，也会影响到检索的效果。为了提高检索效率和效果，在我们当前的系统，仍然是不进行分批处理的。