# 搜索引擎实现的技术方案

## 小组成员

| 组员姓名 | 学号 | 教学班 |
| --- | --- | --- |
| 刘博文 | 1120201883 | 07112004 |
| 张开元 | 1120202079 | 07112004 |
| 许康哲 | 1120201702 | 07112003 |
| 刘勇奇 | 1120201334 | 07152001 |

## 功能描述

本搜索引擎可以将用户输入的内容切分成各个关键词，根据关键词进行搜索。然后将搜索到的相关内容，根据相关度进行排序，并展示给用户。

结构分为三层：原始数据，倒排索引，二级索引。

原始数据：保存链接地址、标题以及网站内容，对于本实验为数据集。

倒排索引：倒排索引不再是按文件记录，而是按单个词记录。它记录每个词在该文件中存在的位置，以及对于该文件的相关度。

二级索引：二级索引记录每一个词，在倒排索引中的位置。

## 技术选型

编写一个搜索引擎，需要对网络上的大量信息进行处理，由于其庞大的数据量，所以采用Hadoop组件中的HDFS和MapReduce组件。

HDFS是整个Hadoop体系的基础，负责数据的存储与管理。HDFS有着高容错性的特点，并且是一个分布式的文件系统，可以用来部署在多台设备上。而且它提供高吞吐量来访问应用程序的数据，适合那些有着超大数据集的应用程序。

MapReduce是一种基于磁盘的分布式并行批处理计算模型，用于处理大数据量的计算。其中Map对应数据集上的独立元素进行指定的操作，生成键-值对作为中间结果，Reduce则对中间结果中相同的键的所有值进行规约，以得到最终结果。

本项目使用Hadoop的HDFS存储所有文件的内容，并使用爬虫，定期更新网络数据。之后使用MapReduce对得到的网络数据进行处理，得到倒排索引，但是由于倒排索引文件过大，甚至可能会大于原始数据文件，所以再将它处理成二级索引文件，这个文件较小，可以直接读到内存中，可以实现快速搜索。

## 功能实现

首先是获得原始数据，对于本实验来说，将数据集按每十万个单词切分成一个文件即可获得原始数据。而对于一个真正的搜索引擎来说，应该使用爬虫技术，爬取搜索范围内的所有网站，存储其链接地址、标题、以及网站内容。

获得原始数据后将原始数据的每个文件切成一个个单词，然后使用TFIDF分析工具，计算每一个单词在它出现过的网页中的重要程度。最后输出的内容为：第一个单词，它所出现的第一个文件、对于该文件的重要程度，它所出现的第二个文件……这些内容就构成了倒排索引文件。

而倒排索引文件太大，不利于快速搜索。所有需要根据倒排索引文件文件生成一个二级索引文件，它的内容只有单词和对应单词出现在倒排索引文件中的位置。

至此，对数据的处理工作全部完成。

当用户输入搜索内容后，搜索引擎会将搜索内容切分成一个个单词，然后在二级索引文件中找到它在倒排索引文件文件中的位置。再将它们提取出来，将出现在同一个文件中的单词的重要程度之求和后排序，按照从大到小的顺序将这些文件输出给用户。

搜索引擎的工作完成。

**关于倒排索引与二级索引的具体实现请参照InvertedIndex.java, SecondaryIndex.java与实验报告.doc三个文件**

## 工作计划

8.27—9.5 上课

9.6—9.11 搜集资料并学习Hadoop的使用

9.12 完成工作安排及成员分工

9.13--9.20 完成相关工作

## 组织结构

技术文档：许康哲、刘勇奇

代码及环境搭建：张开元、刘博文

## 软件质量保证、非功能性保证

### 软件功能：

#### 建立搜索引擎库。

以hadoop为底层的存储和访问工具，利用网络爬虫，使用深度、广度优先，pagerank，OPIC等技术爬取用户查询相关网页，查看其中页面内容读取和保存相关信息。对于保存在hdfs中的网页的文本信息使用运用MapReduce算法计算,构建一个倒排索引, 将倒排索引存储在HBase中，从而构建快捷的索引数据库。

#### 在索引数据库中检索关键内容。

当用户输入搜索内容，单击搜索按钮后，系统检索程序开始根据分词结果，利用二级索引，抽取索引库中的有关分词结果的文档，利用倒排索引，从网页索引数据库中，找到包含用户搜索内容的所有相关网页，同时对搜索结果进行排序处理。将处理好的文档呈现到客户端。

### 软件质量保证

#### 软件开发过程实施了控制。

通过软件过程的分工，每一个软件过程经过了测试和运行，保证软件正常功能的实现以及故障的及时排除

#### 过程文档进行了记录和报告

软件开发过程中记录了工作的结果，并写入到报告之中，发布给相关的人员。生成的文档共享，保证个阶段开发交流畅通。

#### 及时发现和解决错误。

对于软件开发的每个步骤进行了及时的测试和评估，对于倒排索引索造成的数据量增大的问题，使用二级索引进行解决，该过程以及最终的测试结果记录在软件的实现文档中。

### 软件的非功能性保证

搜索引擎满足快速性和便捷性的要求，对于用户的打开和检索能够及时响应，这对检索算法和读取速度有很高的要求。测试过程中的检索能基本满足要求。之后通过对hadoop集群的性能调优，以及及时更新索引，可以进一步有效提升搜索引擎的响应速度。

同时搜索引擎的web端应用也应该满足简洁和美观的需求，对于web页面的设计在保证响应速度的同时增加美观度，提升用户体验。