计算机科学与技术学院 大数据管理与分析 课程实验报告

|  |  |  |
| --- | --- | --- |
| 实验题目：文档倒排索引算法实现 | | 学号：201605130116 |
| 日期：2019.4.17 | 班级：2016级泰山学堂 | 姓名：杜洪超 |
| Email：[1503345074@qq.com](mailto:1503345074@qq.com) | | |
| 实验目的：  倒排索引（Inverted Index）被用来存储在全文搜索下某个单词在一个文档  或者一组文档中的存储位置的映射，是目前几乎所有支持全文索引的搜索引擎都  需要依赖的一个数据结构。通过对倒排索引的编程实现，熟练掌握 MapReduce 程序在集群上的提交与执行过程，加深对 MapReduce 编程框架的理解。 | | |
| 实验软件和硬件环境：  软件环境：  系统：Ubuntu16.04 LTS 64位，集群环境为 centos6.5  软件：openjdk-8-jre,openjdk-8-jdk,java1.8.0\_191  Hadoop 2.9.2，集群为2.9.0  Eclipse,ssh  硬件环境：  CPU: Intel® Core™ i5-6260U CPU @ 1.80GHz × 4  磁盘：121.8 GB  内存：7.7 GiB | | |
| 实验原理和方法：   1. 本地编写倒排索引算法的MapReduce程序，导出jar包后提交到集群执行； 2. 比较集群执行结果和标准答案，验证程序是否正确 | | |
| 实验步骤：（不要求罗列完整源代码）   1. 本地编写程序和调试   文档倒排索引算法的MapReduce程序主要有以下几个要点：   1. 设置合适的分词器；Java中的StringTokenizer类的分词功能不够灵活，可以使用正则表达式对文本的每一行进行模式匹配，实现更精细的分词 2. Map的输出；为了实现按文档id排序，Map输出的key为单词+文档名，这样在进入Reduce前同一单词的key-value对会自动按文档排序；value设置为1，用于计数； 3. 定制Partitioner；(2)中为了实现排序修改了key的定义，会导致reduce时相同单词被hash到不同节点，导致结果错误，因此使用定制Partitioner类来解决这个问题；Partitioner类处理的key仍然是单词+文档名个格式，但getPartition时只处理单词部分，从而保证了相同单词都映射到同一Reduce节点; 4. 为了把不同文件中的相同单词拼接在一起输出，借助Partitioner处理后的key-value对序列按单词和文档排序的特性，设置两个私有变量保存当前处理的单词以及目前次单词出现的文档和词频信息，如果下一个key-value对仍属于这个单词，就将改文档和词频信息添加到私有变量中，否则证明现在处理的单词以及处理完毕，需要输出，然后更新当前处理的单词和文档及词频信息；这样的作法会导致最后一个单词无法输出，在Reduce类的cleanup函数中输出当前单词和文档和词频信息即可； 5. 去停用词；为了去掉结果中无用的停用词，通过在Map中判断当前单词是否存在与停用词表中来实现筛选操作;把停用词表设置为缓存文件，在Map的setup函数中读取停用词表；   全部类和函数如下：     1. 在集群上提交作业并执行；     作业执行情况：    验证结果： | | |
| 结论分析与体会：  通过实现文档倒排索引算法，掌握了 MapReduce 程序在集群上的提交与执行过程，加深了对MaoReduce编程的理解；学习了自定义Partitioner的技巧,对Map和Reduce类的预处理和结束函数的作用有了更深的了解。 | | |
| 就实验过程中遇到和出现的问题，你是如何解决和处理的，自拟1－3道问答题：   1. 分词结果不理想；   通过正则表达式实现分词；   1. 无法设置自定义的Partitioner类；   Import类库错误；   1. 集群读取调用停用词失败；   将停用词文件设置成缓存文件，使各节点能够读取。 | | |