華中科技大学

课 程 实 验 报 告

课程名称： 大数据分析

专业班级： **CS2202**

学　　号： **U202215399**

姓　　名： 吴渝东

指导教师： 崔金华

报告日期： **2024.06.19**

计算机科学与技术学院

目录

实验一 Map-reduce 算法及其实现

1.1 实验目的

1、理解 map-reduce 算法思想与流程；
2、应用 map-reduce 思想解决问题；
3、（可选）掌握并应用 combine 与 shuffle 过程。

1.2 实验内容

提供 9 个预处理过的文件夹（folder_1-9）模拟 9 个分布式节点中的数据，每个源文件夹中包含大约 6 千个文件，每个文件标题为维基百科条目标题，内容为对应的网页内容。提供 words.txt 文件作为待统计的词汇。

要求应用 map-reduce 思想，模拟 9 个 map 节点与 3 个 reduce 节点实现对维基百科条目词汇的词频的统计。

map 节点输出<((title1,key1),1),…,((titlem,keyn),1)>，其中 key 为文件 title.txt 中出现的且在 words.txt 中词。同时，要求最终的 reduce 节点输出出现次数最多的前 1000 个词汇，以及这些词汇的跳转关系。

输出对应的 map 文件和最终的 reduce 结果文件。要求使用多线程来模拟分布式节点。

学有余力的同学可以在 map-reduce 的基础上添加 combine 与 shuffle 过程，并可以计算线程运行时间来考察这些过程对算法整体的影响。

提示：实现 shuffle 过程时应保证每个 reduce 节点的工作量尽量相当，来减少整体运行时间。

1.3 实验过程

1.3.1 编程思路

将整个 MapReduce 算法分为 4 个部分来实现：map，combiner，shuffle，reduce。
map：共有 9 个目录，使用 9 个线程来处理每个目录中的文件，读取每个目录中的所有文件，并将每个出现的单词表示成<(title,key),1>的形式，将每个目录中文件的读取结果存储到对应的 map1~9.txt 文件中。

combiner: 从 map 操作生成的 9 个 map 文件分别中读取每一条记录，并将(title,key)相同的记录合并，然后存储到相应的 9 个 combiner 文件中。

shuffle：首先从 9 个 combiner 文件中读取记录，为了使每个 reduce 节点的工作量大致相当，采用哈希函数的方式来分配任务，哈希值的计算方式为对单词每个字母的 ascall 码值求和然后模 3，根据哈希值将每条记录存储到对应的 3 个 shuffle 文件中。

reduce：首先从 3 个 shuffle 文件中读取记录，合并单词相同的记录并以<key,frequence>的形式存储到 3 个相应的 reduce 文件中，然后合并 3 个 reduce 文件，选出其中出现次数最多的 1000 个在 words.txt 中出现的单词。

1.3.2 遇到的问题及解决方式

问题 1：map 运行时间很长，几乎运行不出结果。

解决方式：经过分析，在 map 操作时对每个单词都需要遍历依次 words.txt 中的单词，判断是否在其中出现过，导致运行时间太长，因此将这个判断放在最后对 3 个 reduce 文件的合并过程中，极大的减少了需要遍历的次数，也就减少了运行时间。

问题 2：最后生成的 reduce 结果与应有的结果不同

解决方式：通过分析生成的 reduce 结果，发现许多单词存在首字母大写，首字母是否大写应视作同一个单词，因此，在从文件读取单词的过程中，将所有字母同意转为小写形式。

1.3.3 实验测试与结果分析

最终 reduce 得到的结果如图 1-1 所示（内容过多，仅展示部分内容），程序能够正确实现所需功能。

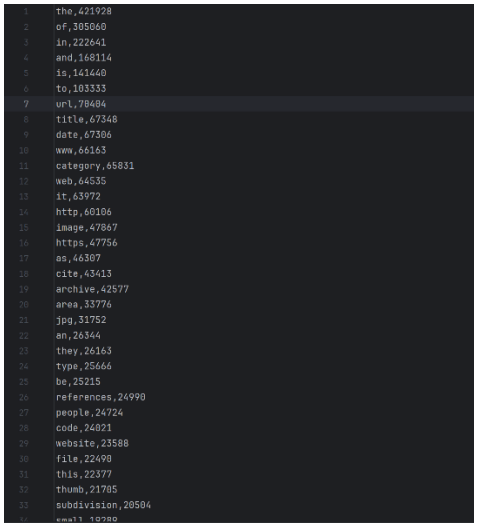

图 1-1 运行结果图

1.4 实验总结

通过这次实验，我对 MapReduce 算法有了更加深入的了解，体会了利用分布式节点的方式对规模庞大的数据进行处理的思想，在大数据分析的过程中，数据规模无疑都是巨大的，因此 MapReduce 算法是非常重要的。同时，这种将大量工作分散，从而可以使用多个节点同时进行处理以节省时间的思想值得我在以后的学习生活中不断体会和学习。