

分布式系统 第二次作业

socket 编程 作业报告

1853444 崔鑫宇

一、 网络通信设计

本次作业我选择通过 Java 语言完成 socket 编程及后续任务，通过使用循环的方式，先后六次通过自行设定的 8888 号端口向不同的六个远程服务端进行查询。查询过程中首先由服务端创建 ServerSocket 并创建监听以监听用户连接，然后再由客户端发送连接请求至服务端，完成客户端和服务端的连接。

建立连接完成后，首先在用户端输入要查询的作者的姓名，并通过 `getOutputStream` 获取输出流，并将输出流中的数据发送给服务端，再由服务端的 `getInputStream` 获取输入流来得到客户端要查询的信息。服务端完成相应信息的查询后同样通过 `getOutputStream` 获取输出流，并将输出流中的数据返回给客户端，在将输出流数据发送给客户端后关闭当前服务端，客户端也同样通过 `getInputStream` 获取输入流并对获取到的数据进行相应处理，最后将处理后的得到的数据进行输出和累加。在获取到全部六个数据后关闭 socket 客户端并将查询结果和查询时间输出到屏幕，并将查询结果和所用时间写入到相应的 log 文件中。

二、 存储负载均衡设计

通过解压获取到完整的 DBLP 数据集后，在一台装有 Linux 系统的机器上先使用 `wc` 命令获取完整的 DBLP 数据集中的数据的总行数，之后再根据先前获取到的总行数通过使用 `split` 命令将 DBLP 数据集按行数平分分成六个分片，使得行数最多的分片和行数最少的分片的行数差不大于 1，获取到六个行数接近的分片后将每个分片分别存储到不同的虚拟机中，每个虚拟机分别存储两个分片，以达到存储负载均衡的目的。

三、 查询容错设计

每当服务端接收到客户端的查询请求和查询对象后，服务端会在本地先后执行两次查询操作，第一次查询的返回结果为其对应序号分片中查询对象的出现次数，第二次查询的返回结果为其对应序号的前一序号分片中查询对象的出现次数（服务端 1 的返回结果为 1 号和 6 号分片的查询结果）。将两个查询结果以空格相隔进行拼接将拼接后的字符串返回给客户端。

数据返回给客户端后，客户端对返回的结果进行分割并将分割结果存入一个字符串数组，之后客户端会根据自定义的标记判断前一个服务端是否存在异常，如果前一个服务端存在异常则将本次查询返回的前一序号分片的查询结果（最后循环的第七次不进行查询操作，其使用的是第一次查询时备份的第 6 号分片的查询结果）输出出来替补前一次异常查询的空缺。最后计算全部 6 次输出的查询结果之和并将其写入到 log 文件中。

四、性能优化

由于本次作业的查询对象仅涉及到 author 这一信息，而在 DBLP 数据集中每个 author 的信息仅存在于相应的 author 标签内，同时又基于我所采取的按行数平行平分的分片方式，故查询采用了通过使用 Linux 系统的 grep 命令进行查找与使用 wc 命令统计出现频次并行的方式来进行，在每个服务端使用上述方法进行查询并将查询结果按照相同格式返回给客户端，并且由客户端进行汇总，最后得到完整的查询结果。

五、功能实现样例

1) 查询功能演示

查询目标为 Paul Kocher 时：

```
1853444-hw2-q1.log x Client.java x
1  查询目标为:Paul Kocher
2  查询到的总次数:12
3  Difference in milliseconds:58854 ms
```

查询结果验证：

```
Inst1 : 2020-10-30T06:25:35.430959700Z
Server1:12次
查询到的总次数:12
Inst2 : 2020-10-30T06:25:38.832823900Z
Difference in milliseconds : 3401
```

查询目标为 Xiaohui Shen 时：

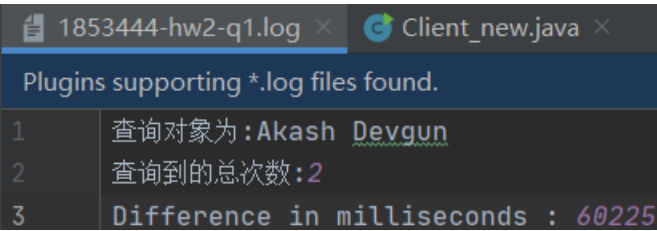
```
1853444-hw2-q1.log x Client.java x
1  查询目标为:Xiaohui Shen
2  查询到的总次数:142
3  Difference in milliseconds:68371 ms
```

查询结果验证：

```
Inst1 : 2020-10-30T06:07:33.974560400Z
Server1:142次
查询到的总次数:142
Inst2 : 2020-10-30T06:07:37.238553800Z
Difference in milliseconds:3263 ms
```

查询目标为 Akash Devgun 时:







```
Inst1 : 2020-10-31T03:49:18.509600800Z
Server1:0次
Server1 runtime in milliseconds : 1031
Server2:0次
Server2 runtime in milliseconds : 13029
Server3:0次
Server3 runtime in milliseconds : 13903
Server4:1次
Server4 runtime in milliseconds : 8650
Server5:0次
Server5 runtime in milliseconds : 9917
Server6:1次
Server6 runtime in milliseconds : 12596
server7 error:java.net.ConnectException
查询到的总次数:2
Inst2 : 2020-10-31T03:50:18.735352700Z
Difference in milliseconds : 60225
```



查询结果验证:

```
Inst1 : 2020-10-30T06:08:31.965175200Z
Server1:2次
查询到的总次数:2
Inst2 : 2020-10-30T06:08:34.952260500Z
Difference in milliseconds:2987 ms
```

2) 存储负载均衡机制实现:

 dblp_split_1.xml	2020/10/27 20:44	XML Source File	505,351 KB
 dblp_split_2.xml	2020/10/28 11:25	XML Source File	504,229 KB
 dblp_split_3.xml	2020/10/28 11:29	XML Source File	420,568 KB
 dblp_split_4.xml	2020/10/28 11:42	XML Source File	466,951 KB
 dblp_split_5.xml	2020/10/28 11:47	XML Source File	519,997 KB
 dblp_split_6.xml	2020/10/28 11:54	XML Source File	527,307 KB

3) 查询容错机制测试:

用例 1: 服务器 2、4、6 号异常时

```
1853444-hw2-q1.log x Client.java x
1  查询对象为:Paul Kocher
2  查询到的总次数:12
3  Difference in milliseconds : 87042

Inst1 : 2020-10-30T06:58:44.579981900Z
Server1:2次
server2 error:java.net.ConnectException: Connection timed out: connect
Server2:6次
Server3:0次
server4 error:java.net.ConnectException: Connection timed out: connect
Server4:1次
Server5:1次
server6 error:java.net.ConnectException: Connection timed out: connect
server7 error:java.net.ConnectException: Connection refused: connect
Server6:2次
查询到的总次数:12
Inst2 : 2020-10-30T07:00:11.622174700Z
Difference in milliseconds : 87042
```

用例 2: 服务器 1、3、5 号异常时

```
1853444-hw2-q1.log x Client_new.java x
Plugins supporting *.log files found.
1  查询对象为:Paul Kocher
2  查询到的总次数:12
3  Difference in milliseconds : 166204

Inst1 : 2020-10-31T11:43:33.442354Z
server1 error:java.net.ConnectException: Connection timed out: connect
Server1:2次
Server2:6次
Server2 runtime in milliseconds : 33634
server3 error:java.net.ConnectException: Connection timed out: connect
Server3:0次
Server4:1次
Server4 runtime in milliseconds : 25556
server5 error:java.net.ConnectException: Connection timed out: connect
Server5:1次
Server6:2次
Server6 runtime in milliseconds : 42953
server7 error:java.net.ConnectException: Connection refused: connect
查询到的总次数:12
Inst2 : 2020-10-31T11:46:19.647228600Z
Difference in milliseconds : 166204
```