实验四 图的三角计数

1. 实验要求

1.1 实验背景

图的三角形计数问题是一个基本的图计算问题，是很多复杂网络分析（比如社交网络分析）的基础。

1.2 实验任务

一个社交网络可以看做是一张图（理算数学中的图）。社交网络中的人对应于图的顶点，对应于图中的顶点；社交网络中的人际关系对应于图中的边。本次实验任务中，我们只考虑一种关系——用户之间的关注关系。假设“王五”在Twitter中关注了“李四”，则在社交网络图中，有一条对应的从“王五”指向“李四”的有向边，图1展示了一个简单的社交网络图，人之间的关注关系通过图中的有向边标识了出来。本次实验任务就是在给定的社交网络图中，统计图中所有三角形的数量。在统计前，需要进行有向边到无向边的转化，依据如下逻辑转换：

"A->B"表示从顶点A到顶点B的一条有向边。A-B表示顶点A和顶点B之间有一条无向边。一个示例见图1，图1右侧的图就是左侧的图去除边方向后对应的无向图。

请在无向图上统计三角形的个数。在图一的例子中一共有三个三角形。

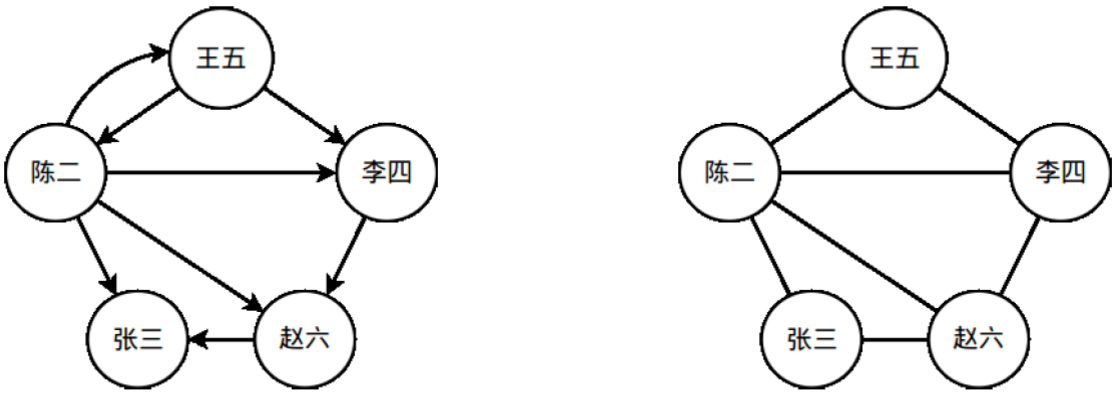

图 1 一个简单的社交网络示例。左侧的是一个社交网络图，右侧的图是将左侧图中的有向边转换为无向边后的无向图。

2.实验设计说明

2.1 主要设计思路

考虑到三角形计数，将与选定点相连的边设为“+”，与选定点相连的两个点之间若有边设为“-”，最后计算“-”的个数即可。

2.2 算法设计

使用MapReduce来实现这个算法， 可以通过使用三次Job来实现。

第一次， Map读入数据，将数据整理一下，例如：读入A，B，保证A<B，输出键值对为（A+B， +），然后Reduce去重。

第二次，Map读入第一次Reduce后的数据，将数据键值对变为（A，B），第二次Reduce将与指定点相连的边的值设为"+",例如存在AB，AC两条边，则存入（A+B，+）和（A+C，+）两个键值对，然后检查是否有BC边，若有记为（B+C，-)并保证B <C.

第三次，这次Map什么都不做，Reduce需要计算三角形个数，对于一条边，若有"+"说明存在着条边，这时可以计算”-“的个数，即为包含这条边的三角形个数，将所有的count加在一起即可以计算出三角形个数。

Class	Input Key	Input Value	Output Key	Output Value
Map_1	Object	Text	Text	Text
Reduce_1	Text	Text	Text	Text
Map_2	LongWritable	Text	Text	Text
Reduce_2	Text	Text	Text	Text
Map_3	LongWritable	Text	Text	Text
Reduce_3	Text	Text	Text	Text

2.3 代码实现

2.3.1 Map_1

```
public class Map_1 extends Mapper<Object, Text, Text, Text> {
    public void map(Object key, Text value, Context context) throws IOException,
InterruptedException {
        String[] line = value.toString().split(" ");
        String a = new String(line[0]);
        String b = new String(line[1]);
        if(a.compareTo(b) > 0) {//保证A<B
            context.write(new Text(b + "+" + a), new Text("+"));
        }else if(a.compareTo(b) < 0) {
            context.write(new Text(a + "+" + b), new Text("+"));
        }else {
            return ;
        }
    }
}
```

2.3.2 Reduce_1

```
public class Reduce_1 extends Reducer<Text, Text, Text, Text> {
    public void reduce(Text key, Iterable<Text> values, Context context) throws
IOException, InterruptedException {
        context.write(key, new Text("+"));//去重
    }
}
```

2.3.3 Map_2

```
public class Map_2 extends Mapper<LongWritable ,Text ,Text, Text> {
    public void map(LongWritable key, Text value, Context context) throws IOException,
InterruptedException {
        StringTokenizer st = new StringTokenizer(value.toString());
        String[] line = st.nextToken().toString().split("\\+");
        context.write(new Text(line[0]), new Text(line[1]));
    }
}
```

2.3.4 Reduce_2

```
public class Reduce_2 extends Reducer<Text, Text, Text, Text> {
    public void reduce(Text key, Iterable<Text> values, Context context) throws
IOException, InterruptedException {
        ArrayList<String> array = new ArrayList<String>();
        for(Text value : values) {
            array.add(value.toString());
            context.write(new Text((key.toString() + "+" + value.toString())), new
Text("+"));
        }
        for(int i=0; i<array.size(); i++) {
            for(int j=i+1; j<array.size(); j++) {
                String a = array.get(i);
                String b = array.get(j);
                if(a.compareTo(b) < 0) {
                    context.write(new Text(a + "+" + b), new Text("-"));//"-"表示邻边关系
                }
                else {

                    context.write(new Text(b + "+" + a), new Text("-"));
                }
            }
        }
    }
}
```

2.3.5 Map_3

```
public class Map_3 extends Mapper<LongWritable, Text, Text, Text> {
    public void map(LongWritable key, Text values, Context context) throws IOException,
InterruptedException {
        StringTokenizer st=new StringTokenizer(values.toString());
        context.write(new Text(st.nextToken()), new Text(st.nextToken()));//什么都不做
    }
}
```

2.3.6 Reduce_3

```
public class Reduce_3 extends Reducer<Text, Text,Text, Text> {
    private static int result = 0;
    public void cleanup(Context context) throws IOException, InterruptedException {
        context.write(new Text("Result: "), new Text("" + result));//输出结果
    }
    public void reduce(Text key, Iterable<Text> values, Context context) throws
IOException, InterruptedException {
        boolean flag = false;
        int count = 0;
        for(Text value: values) {
            if(value.toString().equalsIgnoreCase("+")){//判断是否有这条边
                flag = true;
            }else if(value.toString().equalsIgnoreCase("-")) {//计数
                count ++;
            }
        }
        if(flag) {
            result += count;
        }
    }
}
```

3.结果

3.1 输出结果

数据集	三角形个数	**Driver**程序在集群上的运行时间（秒）
Twitter	13082506	19min40s(1180s)

- Twitter运行截图（/user/2018st18/exp4/result3

```
[2018st18@master01 ~]$ hdfs dfs -cat /user/2018st18/exp4/result3/*
18/11/19 18:40:06 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Result:         13082506
```

- Google+运行截图

3.2 作业截图

3.2.1Twitter作业截图

application_1540647657689_1258	2018st18	Job2	MAPREDUCE	root.2018st18	Mon Nov 19 18:20:04 +0800 2018	Mon Nov 19 18:22:58 +0800 2018	FINISHED	SUCCEEDED		History
application_1540647657689_1257	2018st18	Job1	MAPREDUCE	root.2018st18	Mon Nov 19 18:19:10 +0800 2018	Mon Nov 19 18:20:01 +0800 2018	FINISHED	SUCCEEDED		History

application_1540647657689_1260	2018st18	Job3	MAPREDUCE	root.2018st18	Mon Nov 19 18:23:00 +0800	N/A	RUNNING	UNDEFINED		Application

Job Overview	
Job Name:	Job1
User Name:	2018st18
Queue:	root.2018st18
State:	SUCCEEDED
Uberized:	false
Submitted:	Mon Nov 19 18:19:10 CST 2018
Started:	Mon Nov 19 14:10:38 CST 2018
Finished:	Mon Nov 19 14:11:20 CST 2018
Elapsed:	42sec
Diagnostics:	
Average Map Time	16sec
Average Shuffle Time	2hrs, 57mins, 47sec
Average Merge Time	6sec
Average Reduce Time	-2hrs, -57mins, -39sec

ApplicationMaster

Attempt Number	Start Time	Node	Logs
1	Mon Nov 19 14:10:35 CST 2018	slave015:8042	logs

Task Type	Total	Complete
Map	1	1
Reduce	1	1

Attempt Type	Failed	Killed	Successful
Maps	0	0	1
Reduces	0	0	1

Logged in as: dr.who

Counters for job_1540647657689_1257

- Application
- Job
  - Overview
  - Counters
  - Configuration
  - Map tasks
  - Reduce tasks
- Tools

Counter Group	Counters			
File System Counters	Name	Map	Reduce	Total
	FILE: Number of bytes read	0	8,549,072	8,549,072
	FILE: Number of bytes written	8,665,028	8,664,981	17,330,009
	FILE: Number of large read operations	0	0	0
	FILE: Number of read operations	0	0	0
	FILE: Number of write operations	0	0	0
	HDFS: Number of bytes read	32,467,364	0	32,467,364
	HDFS: Number of bytes written	0	27,328,241	27,328,241
	HDFS: Number of large read operations	0	0	0
	HDFS: Number of read operations	3	3	6
	HDFS: Number of write operations	0	2	2
Job Counters	Name	Map	Reduce	Total
	Launched map tasks	0	0	1
	Launched reduce tasks	0	0	1
	Rack-local map tasks	0	0	1
	Total megabyte-seconds taken by all map tasks	0	0	138,346,496
	Total megabyte-seconds taken by all reduce tasks	0	0	125,083,648
	Total time spent by all map tasks (ms)	0	0	16,888
	Total time spent by all maps in occupied slots (ms)	0	0	33,776
	Total time spent by all reduce tasks (ms)	0	0	15,269
	Total time spent by all reduces in occupied slots (ms)	0	0	30,538
	Total vcore-seconds taken by all map tasks	0	0	16,888
	Total vcore-seconds taken by all reduce tasks	0	0	15,269
Map-Reduce Framework	Name	Map	Reduce	Total
	Combine input records	0	0	0
	Combine output records	0	0	0
	CPU time spent (ms)	15,200	14,590	29,790
	Failed Shuffles	0	0	0
	GC time elapsed (ms)	135	66	201
	Input split bytes	130	0	130
	Map input records	1,768,135	0	1,768,135
	Map output bytes	36,003,504	0	36,003,504
	Map output materialized bytes	8,549,064	0	8,549,064
	Map output records	1,768,135	0	1,768,135
	Merged Map outputs	0	1	1
	Physical memory (bytes) snapshot	1,085,218,816	328,196,096	1,413,414,912
	Reduce input groups	0	1,342,296	1,342,296
	Reduce input records	0	1,768,135	1,768,135
	Reduce output records	0	1,342,296	1,342,296
	Reduce shuffle bytes	0	8,549,064	8,549,064
	Shuffled Maps	0	1	1
	Spilled Records	1,768,135	1,768,135	3,536,270
	Total committed heap usage (bytes)	1,354,235,904	758,120,448	2,112,356,352
	Virtual memory (bytes) snapshot	4,767,363,072	8,114,343,936	12,881,707,008
Shuffle Errors	Name	Map	Reduce	Total
	BAD_ID	0	0	0
	CONNECTION	0	0	0
	IO_ERROR	0	0	0
	WRONG_LENGTH	0	0	0
	WRONG_MAP	0	0	0
	WRONG_REDUCE	0	0	0
File Input Format Counters	Name	Map	Reduce	Total
	Bytes Read	32,467,234	0	32,467,234
File Output Format Counters	Name	Map	Reduce	Total
	Bytes Written	0	27,328,241	27,328,241

Job Overview	
Job Name:	Job2
User Name:	2018st18
Queue:	root.2018st18
State:	SUCCEEDED
Uberized:	false
Submitted:	Mon Nov 19 18:20:04 CST 2018
Started:	Mon Nov 19 17:09:20 CST 2018
Finished:	Mon Nov 19 17:12:03 CST 2018
Elapsed:	2mins, 42sec
Diagnostics:	
Average Map Time	11sec
Average Shuffle Time	-1hrs, -24mins, -57sec
Average Merge Time	5sec
Average Reduce Time	1hrs, 27mins, 5sec

ApplicationMaster

Attempt Number	Start Time	Node	Logs
1	Mon Nov 19 17:09:17 CST 2018	slave017:8042	logs

Task Type	Total	Complete
Map	1	1
Reduce	1	1

Attempt Type	Failed	Killed	Successful
Maps	0	0	1
Reduces	0	0	1

Logged in as: dr.who

Counters for job_1540647657689_1258

- Application
- Job
  - Overview
  - Counters
  - Configuration
  - Map tasks
  - Reduce tasks
- Tools

Counter Group	Counters			
	Name	Map	Reduce	Total
File System Counters	FILE: Number of bytes read	0	7,847,355	7,847,355
	FILE: Number of bytes written	7,963,293	7,963,246	15,926,539
	FILE: Number of large read operations	0	0	0
	FILE: Number of read operations	0	0	0
	FILE: Number of write operations	0	0	0
	HDFS: Number of bytes read	27,328,366	0	27,328,366
	HDFS: Number of bytes written	0	1,676,044,590	1,676,044,590
	HDFS: Number of large read operations	0	0	0
	HDFS: Number of read operations	3	3	6
	HDFS: Number of write operations	0	2	2
Job Counters	Name	Map	Reduce	Total
	Data-local map tasks	0	0	1
	Launched map tasks	0	0	1
	Launched reduce tasks	0	0	1
	Total megabyte-seconds taken by all map tasks	0	0	94,920,704
	Total megabyte-seconds taken by all reduce tasks	0	0	1,088,438,272
	Total time spent by all map tasks (ms)	0	0	11,587
	Total time spent by all maps in occupied slots (ms)	0	0	23,174
	Total time spent by all reduce tasks (ms)	0	0	132,866
	Total time spent by all reduces in occupied slots (ms)	0	0	265,732
	Total vcore-seconds taken by all map tasks	0	0	11,587
	Total vcore-seconds taken by all reduce tasks	0	0	132,866
Map-Reduce Framework	Name	Map	Reduce	Total
	Combine input records	0	0	0
	Combine output records	0	0	0
	CPU time spent (ms)	10,870	89,460	100,330
	Failed Shuffles	0	0	0
	GC time elapsed (ms)	168	322	490
	Input split bytes	125	0	125
	Map input records	1,342,296	0	1,342,296
	Map output bytes	24,643,649	0	24,643,649
	Map output materialized bytes	7,847,347	0	7,847,347
	Map output records	1,342,296	0	1,342,296
	Merged Map outputs	0	1	1
	Physical memory (bytes) snapshot	714,006,528	1,687,547,904	2,401,554,432
	Reduce input groups	0	70,840	70,840
	Reduce input records	0	1,342,296	1,342,296
	Reduce output records	0	82,505,643	82,505,643
	Reduce shuffle bytes	0	7,847,347	7,847,347
	Shuffled Maps	0	1	1
	Spilled Records	1,342,296	1,342,296	2,684,592
	Total committed heap usage (bytes)	1,337,458,688	993,001,472	2,330,460,160
	Virtual memory (bytes) snapshot	4,767,363,072	8,114,524,160	12,881,887,232
Shuffle Errors	Name	Map	Reduce	Total
	BAD_ID	0	0	0
	CONNECTION	0	0	0
	IO_ERROR	0	0	0
	WRONG_LENGTH	0	0	0
	WRONG_MAP	0	0	0
	WRONG_REDUCE	0	0	0
File Input Format Counters	Name	Map	Reduce	Total
	Bytes Read	27,328,241	0	27,328,241
File Output Format Counters	Name	Map	Reduce	Total
	Bytes Written	0	1,676,044,590	1,676,044,590

Job Overview	
Job Name:	Job3
User Name:	2018st18
Queue:	root.2018st18
State:	SUCCEEDED
Uberized:	false
Submitted:	Mon Nov 19 18:23:00 CST 2018
Started:	Mon Nov 19 14:14:29 CST 2018
Finished:	Mon Nov 19 14:30:45 CST 2018
Elapsed:	16mins, 16sec
Diagnostics:	
Average Map Time	1mins, 15sec
Average Shuffle Time	1hrs, 40mins, 41sec
Average Merge Time	1mins, 59sec
Average Reduce Time	–1hrs, –29mins, –17sec

ApplicationMaster

Attempt Number	Start Time	Node	Logs
1	Mon Nov 19 14:14:26 CST 2018	slave015:8042	logs

Task Type	Total	Complete
Map	13	13
Reduce	1	1

Attempt Type	Failed	Killed	Successful
Maps	0	0	13
Reduces	0	0	1

Logged in as: dr.who

Counters for job_1540647657689_1260

- Application
- Job
  - Overview
  - Counters
  - Configuration
  - Map tasks
  - Reduce tasks
- Tools

Counter Group	Counters			
File System Counters	Name	Map	Reduce	Total
	FILE: Number of bytes read	381,555,498	268,367,103	649,922,601
	FILE: Number of bytes written	736,724,341	268,482,995	1,005,207,336
	FILE: Number of large read operations	0	0	0
	FILE: Number of read operations	0	0	0
	FILE: Number of write operations	0	0	0
	HDFS: Number of bytes read	1,676,095,367	0	1,676,095,367
	HDFS: Number of bytes written	0	18	18
	HDFS: Number of large read operations	0	0	0
	HDFS: Number of read operations	39	3	42
	HDFS: Number of write operations	0	2	2
Job Counters	Name	Map	Reduce	Total
	Data-local map tasks	0	0	9
	Launched map tasks	0	0	13
	Launched reduce tasks	0	0	1
	Rack-local map tasks	0	0	4
	Total megabyte-seconds taken by all map tasks	0	0	7,997,210,624
	Total megabyte-seconds taken by all reduce tasks	0	0	6,576,128,000
	Total time spent by all map tasks (ms)	0	0	976,222
	Total time spent by all maps in occupied slots (ms)	0	0	1,952,444
	Total time spent by all reduce tasks (ms)	0	0	802,750
	Total time spent by all reduces in occupied slots (ms)	0	0	1,605,500
	Total vcore-seconds taken by all map tasks	0	0	976,222
	Total vcore-seconds taken by all reduce tasks	0	0	802,750
Map-Reduce Framework	Name	Map	Reduce	Total
	Combine input records	0	0	0
	Combine output records	0	0	0
	CPU time spent (ms)	1,001,350	486,550	1,487,900
	Failed Shuffles	0	0	0
	GC time elapsed (ms)	2,301	2,232	4,533
	Input split bytes	1,625	0	1,625
	Map input records	82,505,643	0	82,505,643
	Map output bytes	1,676,044,590	0	1,676,044,590
	Map output materialized bytes	353,661,529	0	353,661,529
	Map output records	82,505,643	0	82,505,643
	Merged Map outputs	0	13	13
	Physical memory (bytes) snapshot	17,454,018,560	1,816,150,016	19,270,168,576
	Reduce input groups	0	39,031,397	39,031,397
	Reduce input records	0	82,505,643	82,505,643
	Reduce output records	0	1	1
	Reduce shuffle bytes	0	353,661,529	353,661,529
	Shuffled Maps	0	13	13
	Spilled Records	165,011,286	82,505,643	247,516,929
	Total committed heap usage (bytes)	21,552,431,104	1,681,915,904	23,234,347,008
	Virtual memory (bytes) snapshot	61,976,018,944	8,113,152,000	70,089,170,944
Shuffle Errors	Name	Map	Reduce	Total
	BAD_ID	0	0	0
	CONNECTION	0	0	0
	IO_ERROR	0	0	0
	WRONG_LENGTH	0	0	0
	WRONG_MAP	0	0	0
	WRONG_REDUCE	0	0	0
File Input Format Counters	Name	Map	Reduce	Total
	Bytes Read	1,676,093,742	0	1,676,093,742
File Output Format Counters	Name	Map	Reduce	Total
	Bytes Written	0	18	18

4小结

本次实验使用了三个job来实现，第三次job的map阶段什么都没有做，浪费了很多时间，如果深入的话，可以减少一次job，还有很多MapReduce的技巧没有用上，比如combiner和Partioner，这些可以更快的提高效率。

同时发现，集群由于太多人在跑，所以花费了很多时间，通过实验室同学在本地docker环境里运行我们这个程序，一个Twitter的三角形计数只需要三分多钟就可以实现。

附录：JAR包的运行方式

```
hadoop jar TwitterGraph.jar Assignment //完成Twitter
```