实验 4 图的三角形计数

更新日期：2018-11-08

1. 实验要求

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。目前图的三角形计数问题已经成为了 Spark 系统中 GraphX 图计算库所提供的一个算法级 API。本次实验任务就是要在 Hadoop 系统上实现图的三角形计数任务。

实验任务

一个社交网络可以看做是一张图（离散数学中的图）。社交网络中的人对应于图的顶点；社交网络中的人际关系对应于图中的边。在本次实验任务中，我们只考虑一种关系——用户之间的关注关系。假设“王五”在 Twitter/微博中关注了“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 中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。。

本次的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，需要先进行有向边到无向边的转换，依据如下逻辑转换：

IF (A→B) **OR** (B→A) **THEN** A-B

“A→B”表示从顶点 A 到顶点 B 有一条有向边。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 1 的例子中，一共有 3 个三角形。

本次实验将提供一个 Twitter 局部关系图[1]作为输入数据（给出的图是有向图），请统计该图对应的无向图中的三角形个数。

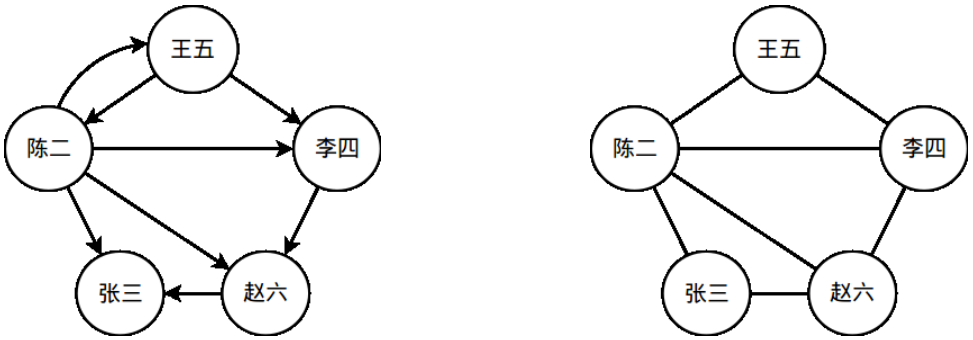

图 1 一个简单的社交网络示例。左侧的是一个社交网络图，右侧的图是将左侧图中的有向边转换为无向边后的无向图。

2. 输入输出

输入要求

输入数据仅一个文件。该文件由若干行组成，每一行由两个以空格分隔的整数组成：

A B

A，B 分别是两个顶点的 ID。这一行记录表示图中具有一条由 A 到 B 的有向边。整个图的结构由该文件唯一确定。

下面的框中是文件部分内容的示例：

```
87982906 17975898
17809581 35664799
524620711 270231980
247583674 230498574
348281617 255810948
159294262 230766095
14927205 5380672
.....
```

输出要求

请将统计出的三角形个数输出到一个 HDFS 的文件中。输出文件的路径可以自定义，但请在实验报告中说明你所采用的输出文件路径。

其他要求

本次实验需要多个 MapReduce Job 才能完成。请再编写一个 Driver 程序，将多个 MapReduce Job 组织在一个程序内自动执行。（如果不清楚 Driver 程序的作用，请参考课程参考书《深入理解大数据：大数据处理与编程实践》的第 8.6.3 节的第 3 部分，PageRankDriver 类的设计说明。）

3. 实验报告要求

请在报告中报告如下内容：

- 实验设计说明，包括主要设计思路、算法设计、程序和各个类的设计说明：
- Map 和 Reduce 的设计思路（含 Map、Reduce 阶段的 K、V 类型）。
- MapReduce 中 Map 和 Reduce 的伪代码（或者带注释的实际代码，如果使用实际代码，请做好排版）。
- 最终统计出的三角形个数：请在报告上附上填写好的表格 1。

表格 1 实验结果

数据集	三角形个数	Driver 程序在集群上的运行时间（秒）
Twitter		
Google+（选做）		

- 输出结果文件的截图，输出结果文件在 HDFS 上的路径（某些情况下助教会检查 HDFS 上的输出文件）。
- 程序运行性能的分析。
- 性能、扩展性等方面存在的不足和可能的改进之处。
- 源代码、可执行程序 JAR 包、JAR 包运行方式说明（助教可能会在集群上重新执行 JAR 包）。
- 请在报告中包含在集群上执行作业后，WebUI 执行报告的内容。请完整包括执行报告内容，否则影响分数。整个程序运行中的每个 MapReduce Job，都需要附上一个相应的 WebUI 执行报告。执行报告内容示例见第 3 次实验要求。

4. 选做内容

该部分内容不做要求，供学有余力的同学尝试练习。

选做 1：本次实验中，我们在做有向边到无向边的转换时，依赖如下的逻辑：

IF (A→B) **OR** (B→A) **THEN** A-B

现在请将该逻辑替换为：

IF (A→B) **AND** (B→A) **THEN** A-B

再次进行统计，看看统计出的三角形的个数是多少？请填写下述表格 2。

表格 2 选做 1 实验结果

数据集	三角形个数	Driver 程序在集群上的运行时间（秒）
Twitter		
Google+（选做）		

选做 2：挑战更大的数据集！使用 Google+的社交关系网[1]数据集作为输入数据集。请报告计算出的三角形个数和总用时。如果程序在集群上运行结果正确、性能较好，有补偿性的加分（但总分不会超过本次实验的满分）。输入文件的 HDFS 路径为：

/data/graphTriangleCount/gplus_combined.unique.txt

注意，该文件中可能有类似“x x”这样的自己指向自己的边，请注意处理。

5. 实验数据

本次实验提供 Twitter 局部关系图作为输入，输入文件在集群的 HDFS 上的路径为：

/data/graphTriangleCount/twitter_graph_v2.txt

该文件中可能有类似“x x”这样的自己指向自己的边，请注意处理。

调试用的小数据集，请自己构造。

数据集来源

[1]J. McAuley and J. Leskovec. Learning to Discover Social Circles in Ego Networks. NIPS, 2012.http://snap.stanford.edu/data/

友情提示

本次实验的程序运行会比较耗时，临近截止日期集群会比较繁忙影响性能，请尽早开始实验，以免最后非常仓促。