

**《基于平台的编程》实践报告**

|  |  |
| --- | --- |
| 姓 名： | 刘本嵩 |
| 学 院： | 计算机科学与技术 |
| 专 业： | 计算机科学与技术 |
| 班 级： | CS1601 |
| 学 号： | U201614531 |
| 指导老师： | 刘海坤 |

2019年10月28日

**目 录**

[实验任务一 PageRank算法 3](#_Toc1890358325)

[1.1 实验环境 3](#_Toc969089260)

[1.2 实验目的 3](#_Toc2118508663)

[1.3 设计思路 3](#_Toc141539692)

[算法 3](#_Toc1299008921)

[实现 4](#_Toc1375789300)

[1.4 源代码 5](#_Toc1598063942)

[1.5 实验结果 6](#_Toc1721052851)

# 实验任务一 PageRank算法

## 实验环境

Amazon Web Service，大太平洋区域，东京数据中心提供计算资源。

Node Name: login.hpc-lan.recolic.org

Internet address: login.hpc.recolic.org

Intel Xeon E5-2676 v3 @ 2x 2.4GHz, RAM 4GiB

Ubuntu Linux 18.04, Linux kernel 4.15.0-1052-aws

Hadoop HDFS 3.1.3, Spark 2.4.4 with Hadoop 2.7, Python 3.6.8

Node Name: 1.hpc-lan.recolic.org

Intel Xeon E5-2676 v3 @ 2x 2.4GHz, RAM 4GiB

Ubuntu Linux 18.04, Linux kernel 4.15.0-1052-aws

Hadoop HDFS 3.1.3, Spark 2.4.4 with Hadoop 2.7, Python 3.6.8

Node Name: 2.hpc-lan.recolic.org

Intel Xeon E5-2676 v3 @ 2x 2.4GHz, RAM 4GiB

Ubuntu Linux 18.04, Linux kernel 4.15.0-1052-aws

Hadoop HDFS 3.1.3, Spark 2.4.4 with Hadoop 2.7, Python 3.6.8

## 实验目的

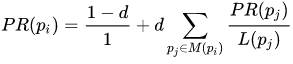
本实验要求采用MapReduce开源系统Hadoop实现PageRank算法，来对课堂上的知识进行巩固，达到此次教学的目的。

## 设计思路

### 算法

PageRank算法有很多版本的变种，出于一定考量，选用最原始论文的版本。

数学语言定义：



编程语言风格的严格定义：

rank[node[i], time[t]] = (1-d) + d \* sum([rank[inbound\_link.src\_node, time[t-1]]/inbound\_link.src\_node.outbound\_links.count() for inbound\_link in inbound\_links(node[i])])

### 实现

Spark从HDFS读入文本，首先转换为RDD[(k, v), ...]的格式，也就是links，cache起来备用。

然后我们对着维基百科上的公式实现这个简单的算法就好了。

ranks是RDD[(k, rank), ...]的格式，初始化为1,然后迭代。每一次迭代中，对每一个节点A的所有邻居B，计算A对B的贡献，放回一个RDD[(B, contrib), ...]，然后reduce掉相同的key，拿到contrib，更新就好了。

最后，print出rank最高的30个结果。

## 源代码

值得注意的是，源代码和配置文件在https://git.recolic.org/recolic-hust/hust-bigdata-platformprog可供下载。

# spark-submit --master spark://login.hpc-lan.recolic.org:7077 pagerank.py hdfs://login.hpc-lan.recolic.org:54310/first-soc-Epinions1.txt 10

import sys

from operator import add

from pyspark.sql import SparkSession

def computeContribs(urls, rank):

num\_urls = len(urls)

for url in urls:

yield (url, rank / num\_urls)

if \_\_name\_\_ == "\_\_main\_\_":

if len(sys.argv) != 3:

print("Usage: pagerank <file> <iterations>", file=sys.stderr)

sys.exit(-1)

# Arguments

spark = SparkSession.builder.appName("PythonPageRank").getOrCreate()

fileUrl, iterations = sys.argv[1], int(sys.argv[2])

# Prepare

lines = spark.read.text(fileUrl).rdd.map(lambda r: r[0])

links = lines.map(lambda urls: tuple(urls.split())).distinct().groupByKey().cache()

ranks = links.map(lambda url\_neighbors: (url\_neighbors[0], 1.0))

# Do the iteration

for i in range(iterations):

contribs = links.join(ranks).flatMap(lambda ele: computeContribs(ele[1][0], ele[1][1]))

ranks = contribs.reduceByKey(add).mapValues(lambda rank: rank \* 0.85 + 0.15)

ranks.first() if i % 50 == 49 else 0

# Collect result

[print('%s has rank: %s.'%r) for r in ranks.sortBy(lambda ele: -ele[1]).take(30)]

spark.stop()

## 实验结果

依次进行10/20/30/40/70/100次迭代，可以发现结果在30-40次迭代后快速收敛了。这里给出前30名的结果。

Iteration=10

18 has rank: 162.4428081148736.

737 has rank: 114.0162002362675.

118 has rank: 78.24350219557793.

1719 has rank: 77.54210439018783.

143 has rank: 73.68780953192217.

136 has rank: 72.36789591800222.

790 has rank: 70.94083184966586.

40 has rank: 70.46300211332742.

725 has rank: 56.80342393432776.

401 has rank: 55.687565755826114.

1619 has rank: 55.68304316737551.

849 has rank: 54.55457198377165.

27 has rank: 53.816178479010496.

1179 has rank: 53.59825137665851.

77 has rank: 52.83786516515877.

135 has rank: 52.051283976453746.

128 has rank: 51.777312483158575.

1191 has rank: 51.40062002536622.

1401 has rank: 51.319418584370915.

550 has rank: 49.987737678464406.

1164 has rank: 49.74903956183953.

34 has rank: 49.408978917836976.

918 has rank: 49.34320761700201.

726 has rank: 47.375636216458226.

1247 has rank: 47.194195739503044.

28 has rank: 46.47485337643704.

301 has rank: 45.73734492017987.

59 has rank: 45.30930631495248.

1909 has rank: 44.41205564240727.

31 has rank: 43.74246625907862.

Iteration=20

18 has rank: 144.1714919790289.

737 has rank: 97.60380957811051.

1719 has rank: 68.46467690038762.

118 has rank: 67.76505011999369.

143 has rank: 65.4811428351346.

136 has rank: 63.59431171813178.

790 has rank: 63.185106237089535.

40 has rank: 59.85300935542242.

725 has rank: 48.438017084770365.

1619 has rank: 48.303049032303065.

401 has rank: 47.02660786529866.

1179 has rank: 46.97073339435029.

849 has rank: 46.290511228162615.

27 has rank: 46.16665535290123.

77 has rank: 45.253544310501645.

128 has rank: 44.82042903826988.

135 has rank: 44.273335776275054.

1191 has rank: 44.01629030712916.

1401 has rank: 43.83869424702122.

1164 has rank: 42.45089782183669.

550 has rank: 42.21058019914778.

918 has rank: 42.016530131810605.

34 has rank: 42.00406768017598.

726 has rank: 40.42480942061105.

1247 has rank: 39.922891679355565.

28 has rank: 39.85475302503735.

301 has rank: 38.9835334722502.

59 has rank: 38.95168204769983.

1909 has rank: 38.22467700900163.

1621 has rank: 37.48868048310641.

Iteration=30

18 has rank: 141.1780450766436.

737 has rank: 94.80141375710701.

1719 has rank: 67.07283922696725.

118 has rank: 66.03677822420502.

143 has rank: 64.12731657831326.

136 has rank: 62.196296344072536.

790 has rank: 62.02052692653999.

40 has rank: 58.00436331347218.

1619 has rank: 47.140209572422386.

725 has rank: 46.9050381313319.

1179 has rank: 45.831212465745516.

401 has rank: 45.447366833790056.

849 has rank: 44.80971530903063.

27 has rank: 44.75970975500974.

77 has rank: 43.94857127887025.

128 has rank: 43.69035180902982.

135 has rank: 42.91962041935763.

1191 has rank: 42.77527256647698.

1401 has rank: 42.5338267270746.

1164 has rank: 41.23993194879988.

550 has rank: 40.77918648661756.

918 has rank: 40.76356576875844.

34 has rank: 40.650570099872056.

726 has rank: 39.20830412336428.

28 has rank: 38.71643473479704.

1247 has rank: 38.62076156031959.

59 has rank: 37.902524010652265.

301 has rank: 37.87036683958358.

1909 has rank: 37.2446291115351.

1621 has rank: 36.83043128174186.

Iteration=40

18 has rank: 140.68042241635737.

737 has rank: 94.33414249942437.

1719 has rank: 66.84311736993082.

118 has rank: 65.74907195646438.

143 has rank: 63.902090143069316.

136 has rank: 61.96427070994222.

790 has rank: 61.82838329560182.

40 has rank: 57.69526181667362.

1619 has rank: 46.947568594018065.

725 has rank: 46.64772707232791.

1179 has rank: 45.640767135899914.

401 has rank: 45.18231061649808.

849 has rank: 44.56174484504087.

27 has rank: 44.523453348839155.

77 has rank: 43.73054565835891.

128 has rank: 43.50249118759436.

135 has rank: 42.693406923725945.

1191 has rank: 42.568344620836974.

1401 has rank: 42.315707055399656.

1164 has rank: 41.038276812522646.

918 has rank: 40.55455033196816.

550 has rank: 40.5387343506575.

34 has rank: 40.423328394985596.

726 has rank: 39.00474242218114.

28 has rank: 38.52619671211394.

1247 has rank: 38.40258904012402.

59 has rank: 37.727876929774276.

301 has rank: 37.68508002441284.

1909 has rank: 37.082206865950795.

1621 has rank: 36.72265423794547.

Iteration=70

18 has rank: 140.58169854163813.

737 has rank: 94.2413890639969.

1719 has rank: 66.79758807152793.

118 has rank: 65.69196689225804.

143 has rank: 63.857411577495114.

136 has rank: 61.91824753123123.

790 has rank: 61.79029436116879.

40 has rank: 57.63387524555418.

1619 has rank: 46.909359600973076.

725 has rank: 46.59660215229408.

1179 has rank: 45.602948362067835.

401 has rank: 45.12964331802549.

849 has rank: 44.51249192168089.

27 has rank: 44.476509221730055.

77 has rank: 43.687250498143996.

128 has rank: 43.46521184631009.

135 has rank: 42.64848865623191.

1191 has rank: 42.52726149921069.

1401 has rank: 42.27239464527909.

1164 has rank: 40.998250123769324.

918 has rank: 40.51305692961377.

550 has rank: 40.490950048141826.

34 has rank: 40.37817278995167.

726 has rank: 38.9643069778806.

28 has rank: 38.48841463265166.

1247 has rank: 38.35925018756414.

59 has rank: 37.69320920640966.

301 has rank: 37.648304064908196.

1909 has rank: 37.049991217243864.

1621 has rank: 36.70132904080081.

Iteration=100

18 has rank: 140.5812459266121.

737 has rank: 94.24096374647925.

1719 has rank: 66.79737937707239.

118 has rank: 65.69170503679193.

143 has rank: 63.85720680822491.

136 has rank: 61.91803662699611.

790 has rank: 61.79011975473052.

40 has rank: 57.633593724272366.

1619 has rank: 46.90918442910913.

725 has rank: 46.596367668311395.

1179 has rank: 45.60277493171017.

401 has rank: 45.129401753167315.

849 has rank: 44.5122660374037.

27 has rank: 44.476293914059134.

77 has rank: 43.68705195050176.

128 has rank: 43.46504091469184.

135 has rank: 42.648282678396775.

1191 has rank: 42.527073099196954.

1401 has rank: 42.27219602510829.

1164 has rank: 40.998066584425686.

918 has rank: 40.51286666131764.

550 has rank: 40.4907308729098.

34 has rank: 40.377965675617105.

726 has rank: 38.96412152376558.

28 has rank: 38.48824135700618.

1247 has rank: 38.35905142441312.

59 has rank: 37.69305023491136.

301 has rank: 37.64813543241229.

1909 has rank: 37.04984351778924.

1621 has rank: 36.70123138497974.