PageRank实验报告

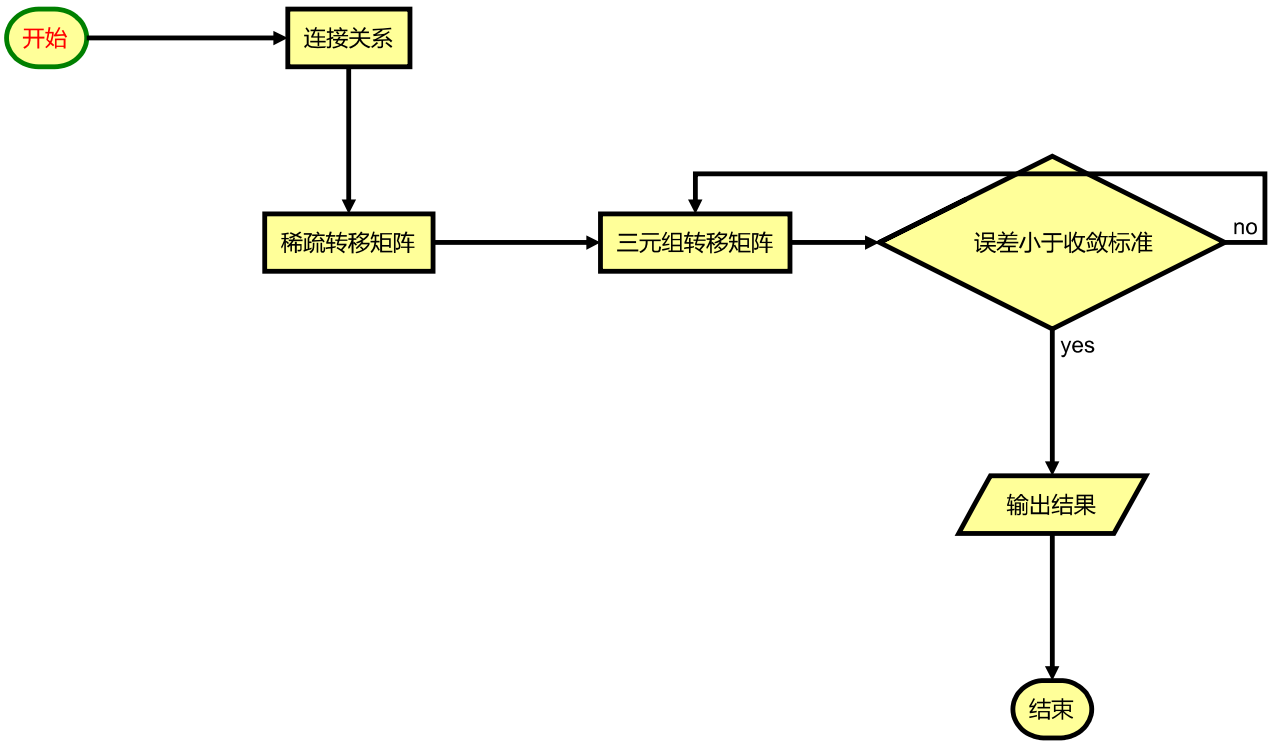
PageRank实验报告，内容包括PageRank算法的设计思路、算法流程图、关键代码描述、实验结果和实验分析。

1.算法设计思路

PageRank,即网页排名。是Google创始人拉里·佩奇和谢尔盖·布林于1997年构建早期的搜索系统原型时提出的链接分析算法。算法的基本思想是：若网页A存在一个指向网页B的链接，则表明A的所有者认为B比较重要，从而把一部分的重要性得分赋予B，而B的PageRank值为一系列类似于A的页面重要性得分值的累加。所有网页的PageRank值通过转移矩阵的迭代得到。
根据实现要求，算法的具体设计思路如下：

- 因为不能使用Python标准库，所以对矩阵乘法进行了封装。算法实现可以直接使用封装函数来进行矩阵计算。
- 在6-15中随机取一个整数，作为对应网页的出度，然后选取出度数量的网页编号作为对应结点指向的网页编号组，生成N行格式为(src,degree,[dest1,...])的三元组。
- 将三元组形式的迁移矩阵进行分块，必须至少分为两块。
- 遍历分块的三元组迁移矩阵，对相应的rank值进行累加。为防止spider trap，rank值需算上随机转移概率。
- 进行迭代，直到误差小于设置的收敛标准。误差计算即为求新旧rank向量的的差的模。
- 输出算法总时间、迭代次数以及PageRank值最大的10个网页编号与对应的PageRank值。

2.算法流程图



3.关键代码描述

matrix.py中的代码封装了矩阵乘法，包括矩阵相乘和数乘矩阵。
矩阵相乘函数：首先判断行列相等情况，然后初始化C的行数和列数分别等于A的行数和B的列数，接着进行A、B矩阵乘法，最后返回C。

```
    def __mul__(self, B):
        """
        重载乘法运算符，用于矩阵乘矩阵
        """
        if self.col != B.row:
            return Matrix(1, 1)
        C = Matrix(self.row, B.col)
        for i in range(C.row):
```

```
            for j in range(C.col):
                for p in range(B.row):
                    C[i][j] += self.A[i][p] * B[p][j]
        return C
```

数乘矩阵函数:对矩阵的每个元素进行数乘运算。

```
    def __rmul__(self, B):
        """
        反向重载乘法运算符，用于数乘矩阵
        """
        for i in range(self.row):
            self.A[i] = list(map(lambda x: x*B, self.A[i]))
        return self
```

generator.py中的代码生成原始的稀疏迁移矩阵和三元组分块表示的稀疏矩阵。
matrix为N行的原始三元组稀疏迁移矩阵，下面的代码进行分块：part_size为每个分块的大小，part_num为分块数，m[2][i]表示matrix第m个网页所指向的第i个网页编号，part为分块表示的稀疏迁移矩阵。

```
    for m in matrix:
        left = 0
        old_seq = m[2][0] // part_size
        for i in range(len(m[2])):
            new_seq = m[2][i] // part_size
            if old_seq != new_seq:
                new_m = (m[0], m[1], m[2][left:i])
                part[old_seq].append(new_m)
                if new_seq == part_num - 1:
                    new_m = (m[0], m[1], m[2][i:])
                    part[new_seq].append(new_m)
                    break
                else:
                    left = i
                    old_seq = new_seq
```

计算误差函数：采用rank向量差的模进行表示。

```
def error(rank: list, last: list) -> float:
    """计算r_new与r_old差的模，即停止条件"""
    mysum = sum(list(map(lambda x: (x[0]-x[1])**2, zip(rank, last))))
    mysum = math.sqrt(mysum)
    return mysum
```

下面的代码进行PageRank值的计算：M是迁移矩阵，step是分块大小，为防止spider trap，还需要算上随机转移概率。

```
    while True:
        rnew = [0 for i in range(N)]
        pointer = 0
        for m in M:  # 遍历每个分块
            for line in m:  # 查询该分块的每一行
                for j in line[2]:  # 遍历每行dest字段包含的节点
                    rnew[j] += rank[line[0]]/line[1]  # 对r_new进行累加
            pointer += step
        # r_new乘beta再加(1-beta)/N，防止spider trap
        rnew = list(map(lambda x: x*beta + (1 - beta)/N, rnew))
        iterations += 1
        if error(rnew, rank) < epsilon:  # 判断收敛条件
            rank = rnew
            break
        rank = rnew  # 更新rank值
```

4.实验结果

设置三个收敛标准，取N=1000,10000,100000，分块数为10的实验结果：

```
PS E:\Code\DMML\PageRank> python main.py
矩阵大小N=1000，生成用时：0.008886000000000002秒。
迭代完成，共迭代1次，epsilon=0.01。
PageRank算法用时：0.009128999999999986秒。
rank    number  PageRank value
1       993     0.0018504939504939506
2       975     0.0018203552003552006
3       286     0.0017752558552558554
4       86      0.001740961260961261
5       646     0.0017228171828171828
6       977     0.0017184504384504386
7       43      0.0017065689865689865
8       209     0.0016723609723609727
9       479     0.0016621334221334223
10      576     0.0016298501498501498
PS E:\Code\DMML\PageRank> python main.py
矩阵大小N=10000，生成用时：0.0915329秒。
迭代完成，共迭代1次，epsilon=0.01。
PageRank算法用时：0.021135299999999996秒。
rank    number  PageRank value
1       3065    0.00022484537684537686
2       1858    0.00022327072927072933
3       9774    0.00019414985014985022
4       9798    0.00019203063603061
5       9442    0.0001918670218670219
6       9910    0.00019136885336885534
7       9641    0.00018990853590853594
8       4589    0.0001869590409590041
9       7646    0.00018522632922632928
10      2115    0.00018454057054057056
PS E:\Code\DMML\PageRank> python main.py
矩阵大小N=100000，生成用时：1.2478942秒。
迭代完成，共迭代1次，epsilon=0.01。
PageRank算法用时：0.2532235999999999秒。
rank    number  PageRank value
1       95542   2.2967277167277165e-05
2       42210   2.283807303807304e-05
3       23775   2.282959262959263e-05
4       49193   2.2272816072816072e-05
5       25166   2.154791874791875e-05
6       90977   2.153610833610834e-05
7       54252   2.1482828282828286e-05
8       31666   2.1241647241647243e-05
9       25654   2.1212787212787218e-05
10      35799   2.121221001221002e-05
PS E:\Code\DMML\PageRank>
```

```
PS E:\Code\DMML\PageRank> python main.py
矩阵大小N=1000，生成用时：0.009258399999999998秒。
迭代完成，共迭代19次，epsilon=1e-05。
PageRank算法用时：0.033828399999999995秒。
rank    number  PageRank value
1       137     0.0015898345433128433
2       232     0.001556117877787802
3       648     0.0014779131524464697
4       432     0.0013716444128839605
6       946     0.0013556845726052165
7       252     0.00134106712189557
8       173     0.0013359778647535625
9       278     0.0013340827328876131
10      996     0.0013288746053060803
PS E:\Code\DMML\PageRank> python main.py
矩阵大小N=10000，生成用时：0.0920468秒。
迭代完成，共迭代15次，epsilon=1e-05。
PageRank算法用时：0.26546810000000004秒。
rank    number  PageRank value
1       238     0.0001774779793452018
2       433     0.00017662269351238544
3       918     0.0001741135123978059
4       3542    0.00017157245312358915
5       2589    0.00016543916810671995
6       5391    0.00016364097736695714
7       4546    0.0001626746935132571
8       5060    0.0001621457869346803
9       5337    0.00015968736012988058
10      889     0.00015955513361134367
PS E:\Code\DMML\PageRank> python main.py
矩阵大小N=100000，生成用时：1.2668873999999999秒。
迭代完成，共迭代11次，epsilon=1e-05。
PageRank算法用时：2.9420288999999995秒。
rank    number  PageRank value
1       19700   1.966967889899187e-05
2       4736    1.9519425121483806e-05
3       31375   1.8746081969720928e-05
4       45581   1.855323146960658e-05
5       39231   1.8307688045572068e-05
6       54315   1.8266198945142537e-05
7       91689   1.82458935160335e-05
8       29245   1.8115426899860587e-05
9       1260    1.8006048119578014e-05
10      22925   1.787804616515725e-05
PS E:\Code\DMML\PageRank>
```

```
PS E:\Code\DMML\PageRank> python main.py
矩阵大小N=1000，生成用时：0.008801299999999998秒。
迭代完成，共迭代44次，epsilon=1e-08。
PageRank算法用时：0.07569119999999999秒。
rank    number  PageRank value
1       396     0.001543504221241367
3       231     0.001500410858300496 4
4       12      0.0014959953498306904
5       952     0.00148589241029322
6       510     0.0014223369790390434
7       354     0.0013912129279434438
8       418     0.00137744586036482
9       149     0.0013696277931698323
10      496     0.0013621106733311458
PS E:\Code\DMML\PageRank> python main.py
矩阵大小N=10000，生成用时：0.0918524秒。
迭代完成，共迭代39次，epsilon=1e-08。
PageRank算法用时：0.7123969秒。
rank    number  PageRank value
1       9839    0.0001728547425938880 6
2       2931    0.000168981936363354 2
3       2364    0.0001688157611169824 8
4       5971    0.0001684942880243464 7
5       5609    0.0001666272622537000 1
6       5802    0.0001626638503688272 6
7       6653    0.0001614038555865450 7
8       2522    0.0001613843438697713 2
9       2220    0.0001597079913390251 2
10      9964    0.0001590360900787779 2
PS E:\Code\DMML\PageRank> python main.py
矩阵大小N=100000，生成用时：1.2747492秒。
迭代完成，共迭代35次，epsilon=1e-08。
PageRank算法用时：9.493896300000001秒。
rank    number  PageRank value
1       40820   1.957220366091735e-05
2       96290   1.9559966621111103e-05
3       51950   1.8483685715079335e-05
4       52548   1.8405023702142324e-05
5       25812   1.8319286951441238e-05
6       5920    1.8213857631367152e-05
7       65136   1.8110952367762887e-05
8       1877    1.799881025072281e-05
9       1522    1.7840753460442053e-05
10      1429    1.7833689335452994e-05
```

```
PS E:\Code\DMML\PageRank> python main.py
矩阵大小N=1000，生成用时：0.5945568秒。
迭代完成，共迭代1次，epsilon=0.01。
PageRank算法用时：0.013613699999999995秒。
排名    编号    PageRank值
1       567     0.002064018204018204
2       460     0.001934332334332334
3       412     0.0018290975690975692
4       544     0.0018179775779775783
5       697     0.0017403840603840604
6       575     0.0017313597513597516
7       256     0.0016244333444333443
9       902     0.0015966189366189367
10      783     0.001594831834831835
PS E:\Code\DMML\PageRank> python main.py
矩阵大小N=10000，生成用时：0.7363771秒。
迭代完成，共迭代1次，epsilon=0.01。
PageRank算法用时：0.1016994秒。
排名    编号    PageRank值
1       3377    0.00021271928071928074
2       4138    0.00020179775779775787
3       502     0.0002007607947607948
4       2744    0.00019601021201021208
5       6827    0.0001957131757131757
6       9558    0.0001955564435564436
7       9275    0.00019500987900987905
8       408     0.0001940497280497281
9       5989    0.00019402575202575207
10      2340    0.00019351004551004556
PS E:\Code\DMML\PageRank> python main.py
矩阵大小N=100000，生成用时：2.4170259秒。
迭代完成，共迭代1次，epsilon=0.01。
PageRank算法用时：1.0890228秒。
排名    编号    PageRank值
1       686     2.348629148629149e-05
2       39802   2.2145343545343546e-05
3       90686   2.2111199911199918e-05
4       44534   2.2104162504162508e-05
5       54644   2.1643956043956047e-05
6       15235   2.1575912975912985e-05
7       61099   2.1568631368631375e-05
8       21523   2.1504650904650907e-05
9       43116   2.13955377955378e-05
10      1229    2.114958374958375e-05
PS E:\Code\DMML\PageRank>
```

```
PS E:\Code\DMML\PageRank> python main.py
矩阵大小N=1000，生成用时：0.6219174秒。
迭代完成，共迭代19次，epsilon=1e-05。
PageRank算法用时：0.22943479999999994秒。
排名    编号    PageRank值
1       709     0.0017351757638943315
2       295     0.0016357182903538655
3       468     0.0015417759161911282
4       959     0.0015219275289961576
6       337     0.001508383138650763
7       945     0.001499995678204426
8       74      0.00149509261253005 2
9       121     0.0014721952042373746
10      45      0.0014490204429499188
PS E:\Code\DMML\PageRank> python main.py
矩阵大小N=10000，生成用时：0.7131193秒。
迭代完成，共迭代15次，epsilon=1e-05。
PageRank算法用时：1.5471678秒。
排名    编号    PageRank值
1       5986    0.00017880531088150167
2       9148    0.00017183029444329658
3       1612    0.00016046615518289375
4       5847    0.00015916641343851476
5       5924    0.00015874132324060044
6       1027    0.00015863280622504915
7       3906    0.00015811661879309245
8       7751    0.0001576750737789123
9       5072    0.00015730424455490992
10      9261    0.00015726737469883785
PS E:\Code\DMML\PageRank> python main.py
矩阵大小N=100000，生成用时：2.3968867秒。
迭代完成，共迭代11次，epsilon=1e-05。
PageRank算法用时：11.7635103秒。
排名    编号    PageRank值
1       93972   2.1067301450011695e-05
2       14129   2.0033002239933732e-05
3       29071   1.9820979314419228e-05
4       97856   1.9562916559533593e-05
5       97133   1.9477786820460175e-05
6       5306    1.947294571091248e-05
7       26752   1.945279809832342e-05
8       8878    1.9229371485825047e-05
9       5946    1.9156326743772193e-05
10      64734   1.9020638297778013e-05
PS E:\Code\DMML\PageRank>
```

```
PS E:\Code\DMML\PageRank> python main.py
矩阵大小N=1000，生成用时：0.50507秒。
迭代完成，共迭代43次，epsilon=1e-08。
PageRank算法用时：0.5186748秒。
排名    编号    PageRank值
1       235     0.0015062446375119316
2       351     0.0014932361130220381
3       353     0.001483479358976693
4       630     0.0014565335737148288
6       420     0.0014091024269299696
7       166     0.0014056921621842012
8       373     0.001394780328369904
9       391     0.0013926942724002526
10      443     0.0013888595985558224
PS E:\Code\DMML\PageRank> python main.py
矩阵大小N=10000，生成用时：0.8016835999999999秒。
迭代完成，共迭代39次，epsilon=1e-08。
PageRank算法用时：3.9175445000000004秒。
排名    编号    PageRank值
1       455     0.00018663676356391601
2       942     0.0001769508078992021
3       9177    0.0001757482622829857
4       1058    0.00017263099232443438
5       4515    0.00017062963298441374
6       2853    0.0001684655401224594
7       6641    0.00016820695452213414
8       1535    0.00016578241119458615
9       386     0.00016505740770316488
10      148     0.00016457593612528986
PS E:\Code\DMML\PageRank> python main.py
矩阵大小N=100000，生成用时：2.330263秒。
迭代完成，共迭代35次，epsilon=1e-08。
PageRank算法用时：36.7687733秒。
排名    编号    PageRank值
1       38137   2.0524078238259196e-05
2       47991   2.012030113766361e-05
3       96496   1.957652085591175e-05
4       50876   1.896383204985931e-05
5       98689   1.8893663810957967e-05
6       99147   1.884685487891727e-05
7       20763   1.8518987068909066e-05
8       46223   1.8508003063958327e-05
9       41681   1.8501281542490803e-05
10      92273   1.8431973714793596e-05
PS E:\Code\DMML\PageRank>
```

5.实验分析

从实验结果图可以看出，当收敛标准为10e-2时，只需要一次迭代，生成矩阵的时间约为迭代用时的四到五倍，算法整体耗时主要在生成矩阵上面；当收敛标准变为10e-5时，迭代次数在10到20次之间，迭代的用时就超过生成矩阵的用时；当收敛标准达到10e-8时，迭代次数达到30次以上，迭代占了算法用时大部分的时间。因此，当收敛标准较严时，主要需改进迭代算法；当收敛标准较松时，主要需改进生成矩阵算法。在同一收敛标准下，迭代次数随N增加而减少。时间复杂度为O(c·n^2),其中c为迭代次数，n为节点个数。