

华中科技大学

课程实验报告

课程名称： 大数据分析

专业班级： CS2202
学 号： U202215399
姓 名： 吴渝东
指导教师： 崔金华
报告日期： 2024.06.19

计算机科学与技术学院

目录

实验二 PageRank 算法及其实现.....	1
1.1 实验目的	1
1.2 实验内容	1
1.3 实验过程	1
1.3.1 编程思路.....	1
1.3.2 遇到的问题及解决方式.....	1
1.3.3 实验测试与结果分析.....	2
1.4 实验总结	3

实验二 PageRank 算法及其实现

1.1 实验目的

- 1、学习 pagerank 算法并熟悉其推导过程；
- 2、实现 pagerank 算法，理解阻尼系数的作用；
- 3、将 pagerank 算法运用于实际，并对结果进行分析。

1.2 实验内容

利用实验一得到的出现次数最多前 1000 个的 title 之间的引用关系 $\langle \text{title}, \langle \text{title}_1, \dots, \text{title}_k \rangle \rangle$ ，由 title 为节点构造有向图，编写 pagerank 算法的代码，根据每个节点的入度计算其 pagerank 值，迭代直到误差小于 10^{-8} 。

实验进阶版考虑加入 teleport β ，用以对概率转移矩阵进行修正，解决 dead ends 和 spider trap 的问题。

输出 title 及其对应的 pagerank 值。

1.3 实验过程

1.3.1 编程思路

首先对实验一得到的引用关系进行处理，将其构造成（节点，边）的有向图形式，同时记录每个节点的出度。

设置两个矩阵 $\text{nodes}, \text{next_nodes}$ ， nodes 表示原始的 pagerank 值， next_nodes 初始值为 0，然后每次迭代时遍历所有的边，对每条边 e_{ij} ，通过 nodes 中节点 i 的 pagerank 值 r_i ，更新 next_nodes 中节点 j 的 pagerank 值 r_{jn} ， $r_{jn} += r_i * \beta / d_i$ （ β 表示概率矩阵的修正因子， d_i 表示节点 i 的出度）。在边遍历结束后，更新 $r_{jn} += (1 - \beta) / n$ （ n 表示总节点数），表示 pagerank 中随机跳转的部分，然后计算该次迭代过程中泄露的 pagerank 值，进行归一化处理，最后将计算得到的 pagerank 写入文件。

1.3.2 遇到的问题及解决方式

问题：迭代结束后计算 pagerank 值得总和远小于 1

问题解决：经分析，有向图中存在只进不出得死角节点，因此 pagerank 值会泄露，需要在迭代结束后进行归一化处理。

1.3.3 实验测试与结果分析

最后计算得到的 pagerank 值如图 1-1 所示（内容过多，仅展示部分内容），迭代后 pagerank 总和如图 1-2 所示，程序能够正确实现所需功能。

```
pagerank.py [D:\Desktop\大数据分析\实验\gitrepository\实验1_实验2\MapReduce] - pagerank
pagerank x
17 people,0.004037100302720404
18 they,0.00243820870898442
19 can,0.005822488677226919
20 mean,0.0054752168960200155
21 person,0.0024281079457242802
22 go,0.0009949018094203652
23 word,0.004745649695147396
24 day,0.0008674355246965946
25 now,0.0007056739850653201
26 utc,0.0004205373390151765
27 fruit,0.0004366726560080527
28 of,0.016709073214767244
29 wikt,0.0017402299777283506
30 be,0.003836337704907992
31 out,0.0010193196454193025
32 world,0.009881277702364622
33 wide,0.008820031275127477
34 web,0.010281808959353331
35 as,0.004089997135692361
36 wiktionary,0.004254675119450002
37 insect,0.0009260427640526517
38 common,0.002639035544691743
39 short,0.001175117391740127
40 feet,0.0008200295004598101
41 material,0.0011516739985283206
42 similar,0.0010780999722216837
43 thumb,0.002984601523603099
44 picture,0.0013264528157518678
45 citation,0.0014211982184957278
46 http,0.0013833972653755696
47 www,0.0015495351838077846
48 archive,0.0012288665726411716
49 url,0.009277994183754674
50 an,0.004281614344282996
```

图 1-1 运行结果图

```
python pagerank x
D:\Anaconda\python.exe D:\Desktop\大数据分析
total pagerank:
0.99999999999999988

Process finished with exit code 0
```

图 1-2 运行结果图

1.4 实验总结

在本次实验中，实现了将引用关系转换为有向图，同时体会了 pagerank 算法的运行流程，体会到了蜘蛛陷阱和死角问题给 pagerank 计算带来的问题，了解了如何通过随机跳转和归一化对这些问题进行处理。