

**课 程 实 验 报 告**

**课程名称： 大数据分析**

**专业班级： CS1804（交换）**

**学 号： X2020I1007**

**姓 名： 刘日星**

**指导教师： 杨 驰**

**报告日期： 2021年12月10日**

**计算机科学与技术学院**

**目录**

[实验二 PageRank算法及其实现 1](#_Toc90027732)

[**2.1实验目的** 1](#_Toc90027733)

[**2.2 实验内容** 1](#_Toc90027734)

[**2.3 实验过程** 1](#_Toc90027735)

[2.3.1 编程思路 1](#_Toc90027736)

[2.3.2 遇到的问题及解决方式 1](#_Toc90027737)

[2.3.3 实验测试与结果分析 1](#_Toc90027738)

[**2.4 实验总结** 1](#_Toc90027739)

# 实验二 PageRank算法及其实现

## **2.1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。

## **2.2 实验内容**

提供的数据集包含邮件内容（Emails.csv），人名与id映射（Persons.csv），别名信息（Aliases.csv），Emails文件中只考虑MetadataTo和MetadataFrom两列，分别表示收件人和寄件人的名称，但这些名称包含许多别名，因此需要对邮件中的名称进行统一并映射到唯一id。（提供预处理代码preprocess.py参考）。

完成这些后，由寄件人和收件人为节点构造有向图，不考虑重复边，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。

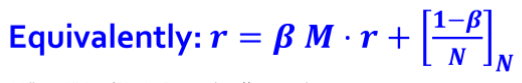
输出人名id及其对应的pagerank值。

加分项：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。

## **2.3 实验过程**

### 2.3.1 编程思路

将预处理preprocess后所生成的sent\_receive.csv文件内的内容读取进程序里，并且按格式分割好数据。通过循环遍历后得到节点总数和边长度分别是N和L，构建一个初始数值都为0的N\*N的转移矩阵matrix，使矩阵的每条边对应的值为1。然后使用matrix矩阵重新构造，计算出每列数据之和col\_sum并且使用每列上不为0的值除以col\_sum（/=），最终得出martix矩阵上的每列不为0的值为1/col\_sum。

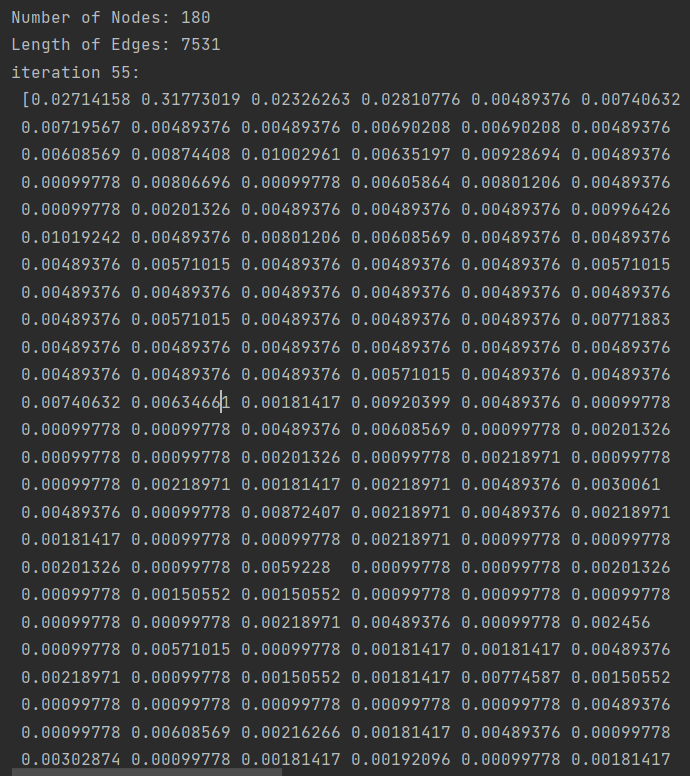
建立一个名为r的矩阵为pagerank初始贡献值矩阵，每一个值都为1/N。然后定义一个误差初始化大小为300000，一个初始迭代次数为0和一个阻尼系数βBeta值为0.85（范围：0.8~0.9取平均值）。用一个while循环，循环套用迭代公式 “next\_r = np.dot(matrix, r) \* beta + (1-beta) / N \* np.ones(N)” 和公式“einit = next\_r – r”，意思是从einit的所有矩阵误差值中选择一个最大的值赋值给einit，每次循环都判断误差是否大于0.00000001同时每循环一次计算一次迭代次数，也就是iteration += 1并且指向下一个r矩阵，也就是r = next\_r。当循环完毕，则会得到最后的迭代结果和迭代次数。

### 2.3.2 遇到的问题及解决方式

在代码写完debug调试程序的时候发现迭代次数有些不太正确，发现在读取文件数据提取节点的时候将sent\_receive.csv的第一行内容“sent\_id”和“receive\_id”这两个单词也读取进去了，导致id的数据在矩阵里迭代计算不正确，后来修改了一下提取内容的起始位置后，就没有再出现相同的错误了。

### 2.3.3 实验测试与结果分析

实验测试结果如下图显示：



结果分析：总的迭代次数为55次，以一个数组的形式输出所有pagerank值，每个pagerank值对应一个id。其与上一个pagerank值误差小于0.00000001。

## **2.4 实验总结**

通过本次实验，对pagerank算法加深了印象，了解了阻尼系数βBeta在pagerank里的作用以及pagerank里的随机传送公式在python里的实现方式，熟悉了python内置库numpy对矩阵的操作和处理。在网上看了一些关于pagerank算法的应用，发现更多是应用在网络网页层面上，好比如说一个网页被很多其他网页链接到，说明该网页比较重要同时它的pagerank值也会相对较高，或者是一个pagerank值很高的网页链接到其他网页，那么被链接的网页的pagerank值也会相应提高很多，代表该网页也很重要。