实验二：PageRank 算法实现

梁业升 2019010547（计 03）

2022 年 4 月 4 日

1 实验结果

评测详情

#	用户	题目	语言	状态	时间
36044	2019010547	搜索引擎技术基础作业二-PageRank	gcc	Accepted	2022-04-04 02:51:28 PM

图 1: OJ 测试结果

2 实验报告

2.1 请分析在迭代过程中，为什么 PageRank 值的和始终为 1

第 $k(k \geq 1)$ 次迭代中，出度为 0 的节点的 PageRank 值之和为

$$S^{(k)} = \sum_{\{j|j \Rightarrow i\}=\emptyset} \mathtt{PR}[i]$$

节点 i 的 PageRank 值为

$$\mathtt{PR}[i]^{(k)} = \frac{\alpha}{N} + (1-\alpha)\Big[\sum_{j \Rightarrow i} \frac{\mathtt{PR}[j]^{(k-1)}}{\mathtt{out\_degree}[j]} + \frac{S^{(k)}}{N}\Big]$$

PageRank 的和为

$$\begin{aligned}\sum_{i\in V(G)} \texttt{PR}[i]^{(k)} &= N\cdot\frac{\alpha}{N} + (1-\alpha)\Big[\sum_{(j,i)\in E(G)} \frac{\texttt{PR}[j]^{(k-1)}}{\texttt{out\_degree}[j]} + \sum_{\{i|j\Rightarrow i\}=\emptyset} \texttt{PR}[j]^{(k-1)}\Big] \\ &= \alpha + (1-\alpha)\sum_{\{i|j\Rightarrow i\}\neq\emptyset} \texttt{PR}[j]^{(k-1)} + (1-\alpha)\sum_{\{i|j\Rightarrow i\}=\emptyset} \texttt{PR}[j]^{(k-1)} \\ &= \alpha + (1-\alpha)\sum_{i\in V(G)} \texttt{PR}[i]^{(k-1)}\end{aligned}$$

$k=0$ 时,

$$\sum_{i\in V(G)} \texttt{PR}[i]^{(0)} = 1$$

由数学归纳法

$$\sum_{i\in V(G)} \texttt{PR}[i]^{(k)} = \alpha + (1-\alpha) = 1$$

2.2 语料入链接数和出链接数分布情况分布

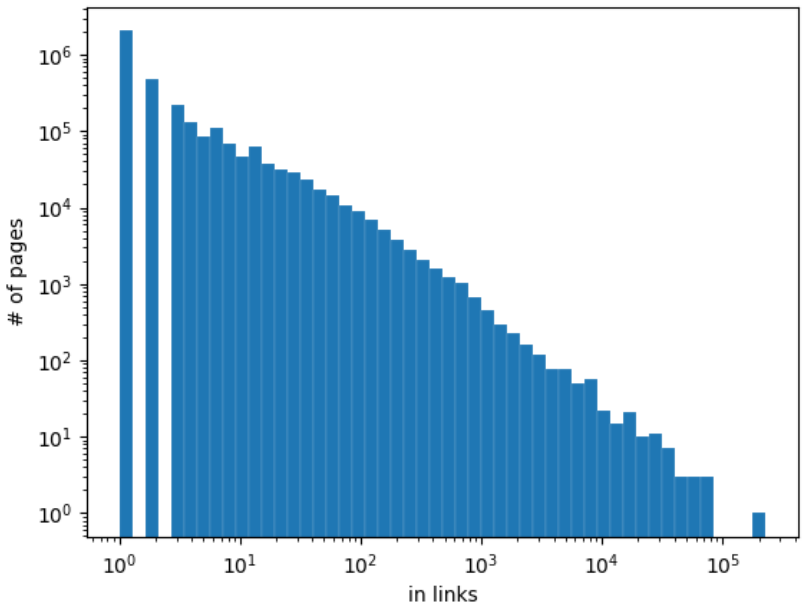

图 2: 入链接数分布

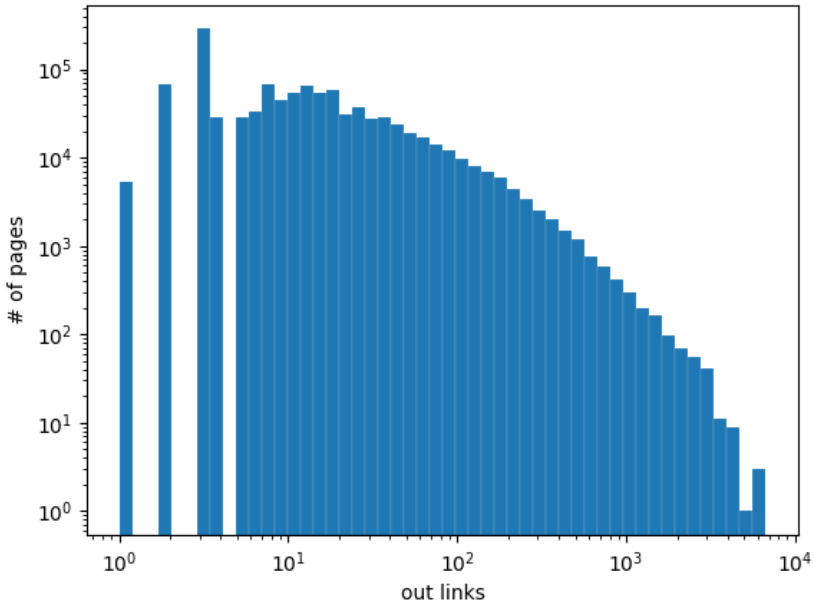

图 3: 出链接数分布

大多数的网页入链接和出链接数均较少，具有较高入链接或出链接的网页较少。相较而言，入链接数较少的网页的比例比出链接数较少的网页的比例多，而最大出链接数量少于最大入链接数量。

2.3 PageRank 结果分布

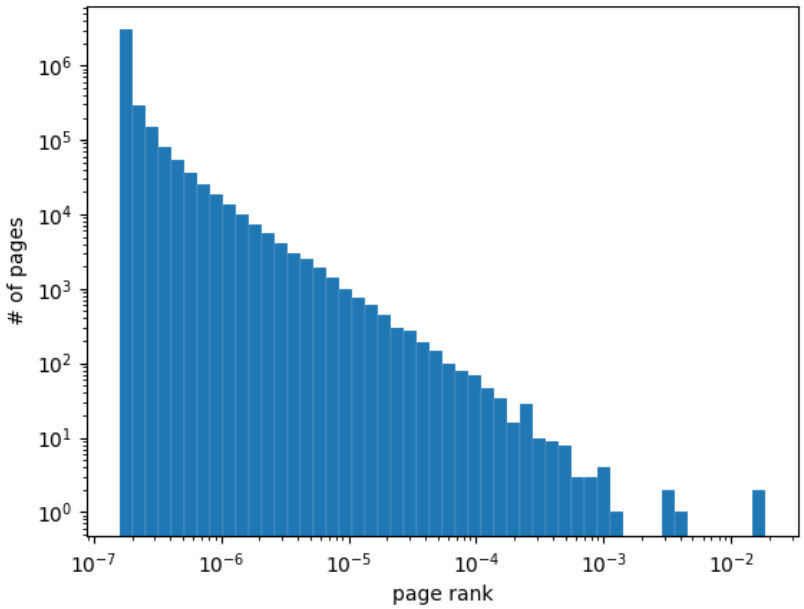

图 4: PageRank 结果分布

大多数的 PageRank 均较低，具有较大 PageRank 的网页较少。

2.4 PageRank 得分与入链接的关联分析

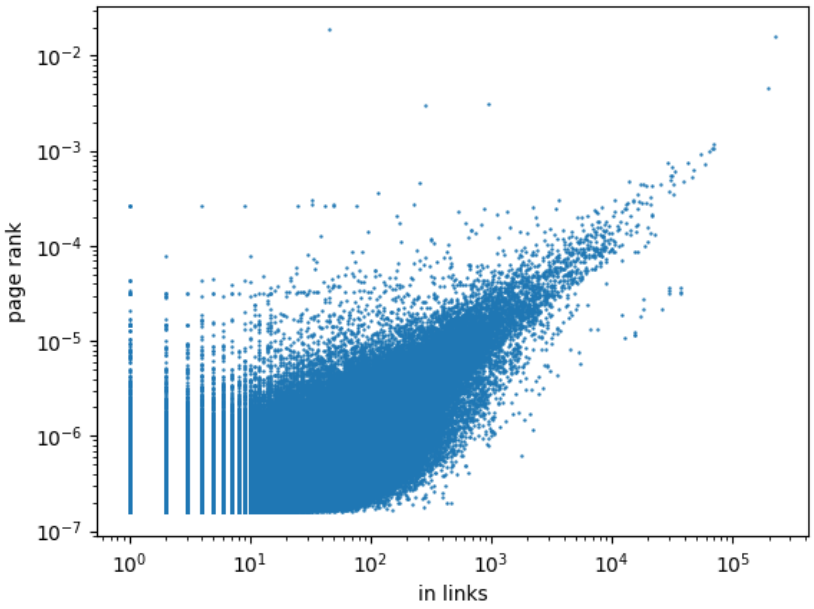

图 5: PageRank 得分与入链接的关系

由上图可见，PageRank 得分与入链接数量大致成正相关，且入链接越多，相关性越强。

2.5 PageRank 得分与相应条目语义内容分析

PageRank 排名前 3 的条目为：

1. 箭头：1.87×10^{-2}
2. ←：1.59×10^{-2}
3. 维基数据：3.1×10^{-3}

PageRank 排名后 2 的条目为：

1. 李东🄵：1.586×10^{-7}
2. 金炯：1.586×10^{-7}

可见，PageRank 较高的条目为搜索或引用的高频词，较低的条目大多为如

部分姓名的冷门词。