搜索引擎第三次作业-链接结构分析

2012011307 黄必胜

1 实验内容

基于维基百科网站中各网页之间的链接关系图，计算该网站中各网页对应的 PageRank 值。PageRank 的核心公式为

$$\mathrm{PR}^{(k)} = \alpha \cdot \frac{1}{N} \cdot I + (1-\alpha) \cdot \sum_{P_i \to n} \frac{PR^{(k-1)}}{Outdegree(P_i)}$$

其中$\mathrm{PR}^{(k)}$为网页结点 n 的第 k 次迭代的 PageRank 值，α 为随机浏览网页的概率。

2 算法实现

利用 python 实现 pagerank 的求值，利用 matlab 对结果进行统计、作图以及分析。

输入文件： node.map.utf8, wiki.graph

PageRank： pagerank.py

实验参数： 随机概率 $\alpha = 0.15$， 迭代次数 $\mathrm{TN} = 30$

实验结果分析： analyse.m

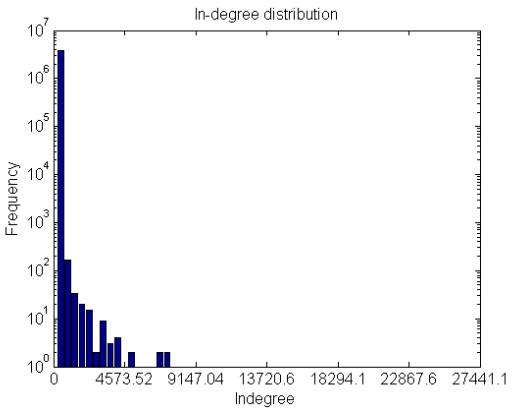
3 实验结果分析

3.1 入链接数/出链接数分布情况

入链接数分布情况

Min	0
Mean	6.98
Median	1
Max	228676

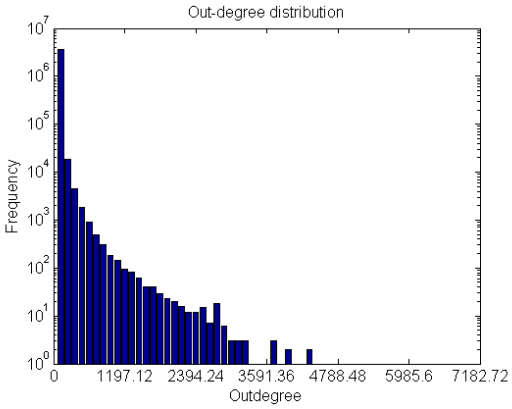
结合右图的入链接数频率分布及入链接数数据统计得知，网页频率在入链接数较小位置有所聚集，绝大多数多数的网页入链接数都较小。



出链接数分布情况

Min	0
Mean	55.7
Median	1
Max	59856

同入链接数的情况类似，随着出链接数的增大，频率逐渐下降，但是下降的趋势更为平缓，且平均值较入链接数大。

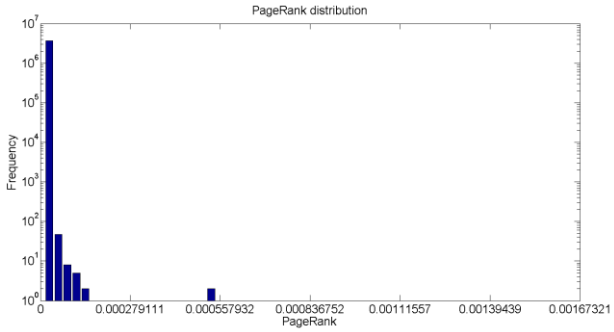


3.2 PageRank 算法结果分布情况

PageRank 分布情况

Min	1.4529e-7
Mean	2.6807e-7
Median	1.5242e-7
Max	0.0139

页面的 pagerank 主要集中在10^{-7}的量级，这是由于 N 的数目大致为$3 * 10^{6}$，每个页面平均分配到 pagerank 大概就是这个量级。



3.3 PageRank 与入链接数的关联关系

右图为所有页面在 PageRank 和入链接数的二维平面上的作图结果。当入链接数和 PageRank 较小的时候，没有明显的趋势，在图像中表现为在左下角一大片矩形区域的聚集。但当入链接数较大时，可以看到 PageRank 与之有正相关的趋势。

对入链接数和 pagerank 求相关系数，得

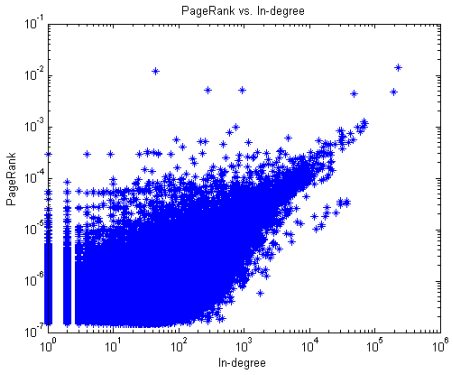
$$r = 0.6159$$

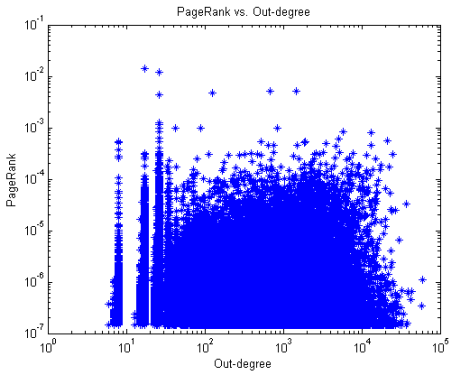
所以可以认为两者之间有较好的相关。

对比右图中 PageRank 与出链接数的关系，观察图像并不能得出较为明显的规律，对出链接数和 pagerank 求相关系数，得

$$r = 0.0247$$

所以基本可以认为二者的相关性不大。





3.4 PageRank 得分与相应条目语义内容的分析

将 pagerank 得分的最大和最小区域对应的条目列出如下：

Top 10	PageRank	In-degree	Out-degree
←	0.013941	228676	17
箭头	0.011853	45	26
Unicode	0.005166	945	1487
符号	0.005085	283	688
维基数据	0.004778	196429	125
台湾	0.004363	47474	26
中国	0.001258	68798	26
美国	0.001133	70222	26
学名	0.000987	68384	26
县级行政区	0.000985	741	845

Last 10	PageRank	In-degree	Out-degree
陈相镇	1.45287e-07	0	17
国际互联网络	1.45287e-07	0	26
三维计算机图像	1.45287e-07	0	163
台湾位置图	1.45287e-07	1	0
通行语言	1.45287e-07	1	0
台湾原住民族诸语	1.45287e-07	1	0
岛屿面积列表	1.45287e-07	1	0
延续迄今	1.45287e-07	1	0
5 大都会区	1.45287e-07	1	0
网域缩写	1.45287e-07	1	0

从图表中即可推测，导致 PageRank 较高的原因可能有：

- 由于庞大的入链接数，其它的页面贡献值累计起来较大，而导致 pagerank 非常高。这些页面多为符号←，维基数据，或是国家如中国，美国，或是一些普遍的概念如学名等。
- 由于被 pagerank 很高的页面指向，比如箭头，虽然其入链接数和出链接数都不大，但是由于被 pagerank 最大的”←”所指向，所以大大提高了其 pagerank 值。

导致 PageRank 较低的原因可能是：

- 词汇较为生僻，或是入链接数太小，从而被访问的概率较小。

四 实验总结

以上实验结果基本与课程内容中的解释相符合。总之，这是一次很有意义的实验。通过本次实验，我们更好地理解了 pagerank 的原理和求解过程，对于一些网站的 pagerank 作弊行为也有了理解，例如通过给其他页面提供大量的出链接来提高其 pagerank 值等。