**数据汇报与分析**

**1.PageRank值与页面入链数分布情况**

[备注:程序分析了约170万个页面的信息，但是result.txt只给出了排名前100万页面的信息]

关于PageRank值：

PageRank值之和为0.998257890567 [因为计算中的舍入，所以约等于1.0]

最大值：0.0457785

最小值：9.50692e-08

平均值：5.78062514552e-07

标准差：3.8215007776e-05

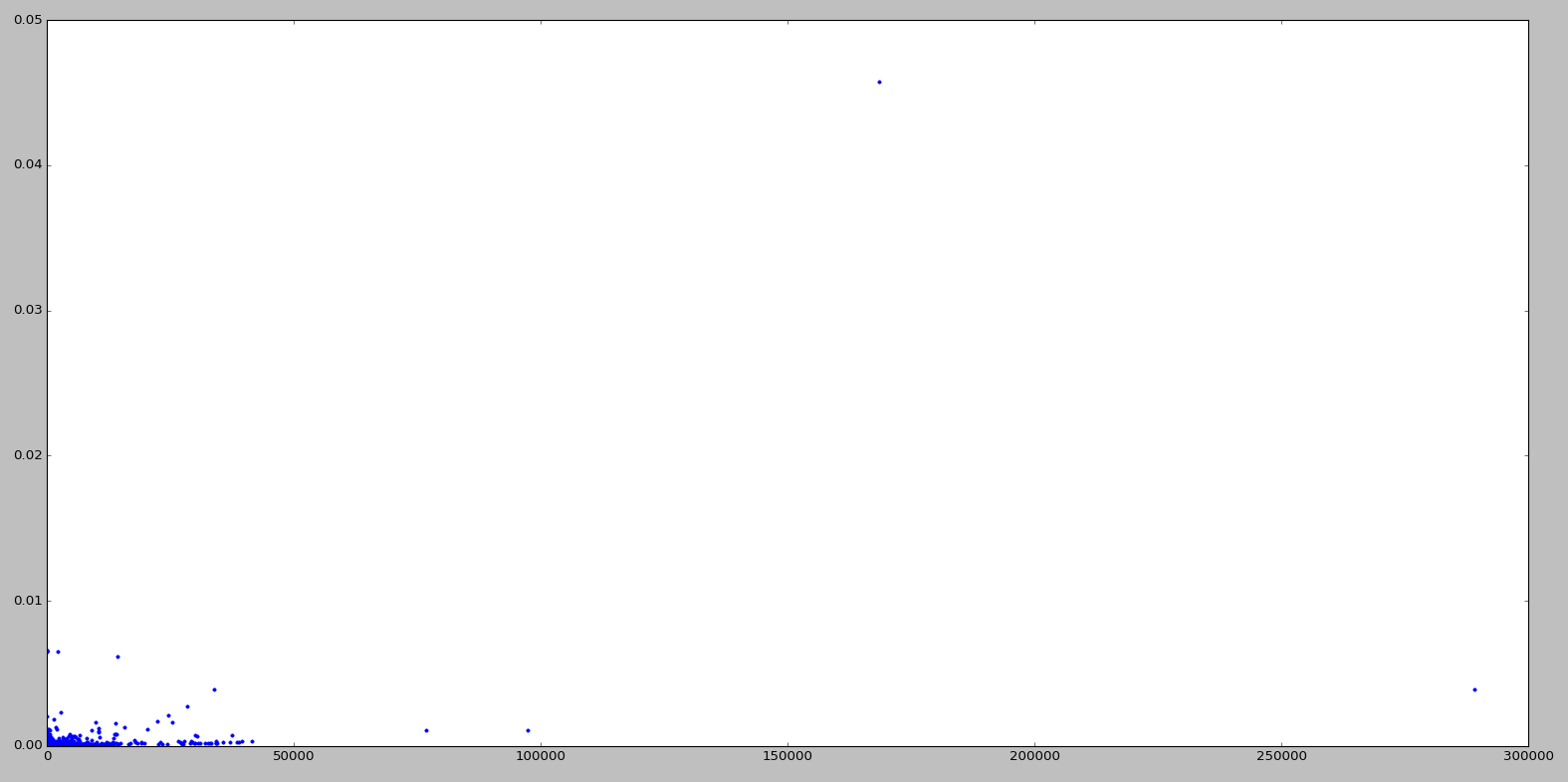
关于链接：

链接总数：36909885

一个页面的入链数最大值为：289186；最小值为：0

平均每个页面被其他的21不同的页面链接

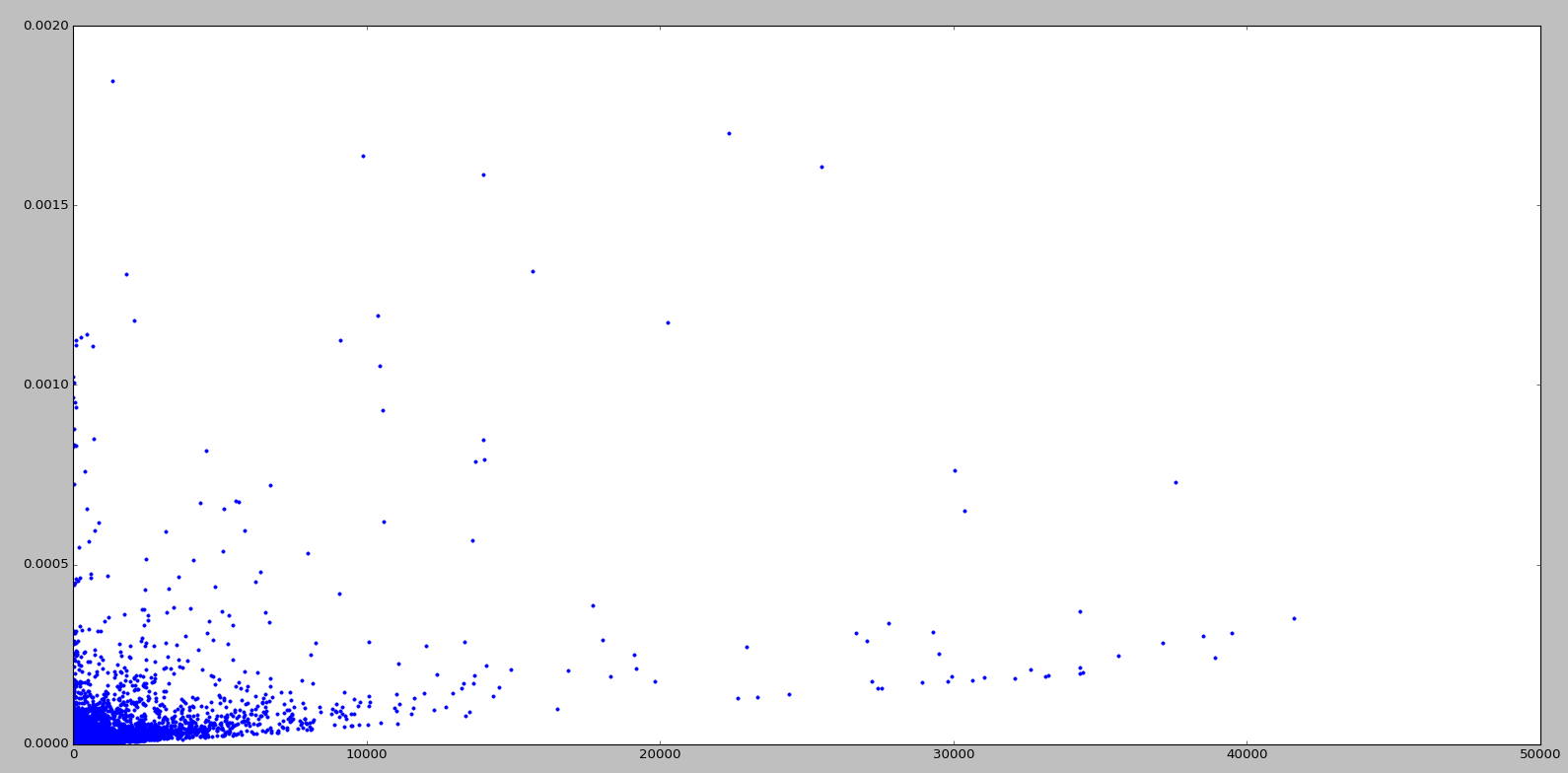
以页面入链数为横坐标，以PageRank值为纵坐标可以画出如下图：



从该图可以看出PageRank值分布并不均匀，只有少数几个页面拥有较高的pageRank值，绝大多数的页面PageRank值集中在0.0000到0.005之间。另外，入链情况也分布不均匀，只有少数几个页面被众多页面链接，绝大多数的页面入链数集中在0到40000之间。

从图中可以看出，入链数最多的页面（假设为页面A）并不是PageRank值最高的页面（假设为页面B），这可能是因为链接到页面A的页面的PageRank值没有链接到页面B的页面大。

仔细观察入链数[0, 50000]和PageRank值[0, 0.002]的那部分图：



可以看出，虽然总体上入链数和PageRank值不是严格的正相关，但是大体上PageRank值还是随着入链数增加而增加。

特别的是，图中部分页面，其入链数非常少，但是PageRank值很大。根据在维基百科上的观察，有的页面是一些词条的共同缩写，维基百科会为其单独创建一个页面，该页面会与缩写是它的词条页面相互链接。以上两种页面分别组成的集合很可能具有很高的重合度，从而造成图中所示的特殊分布情况。

**2.对排名前15名的数据进行分析：**

排名前十五的网页分别是：<页面标题, PageRank值>

AWB 0.0457785

Average White Band 0.00659519

Adaptive Multi-Rate Wideband 0.00655875

AWB (album) 0.00655864

Wikipedia:Autowikibrowser 0.0065377

Afrikaner Weerstandsbeweging 0.00649978

AWB Limited 0.00648602

Wikipedia:Deletion review 0.00613741

← 0.00389242

Thumb 0.00388286

New York city 0.00271605

BOT 0.00230541

Canada 0.00209531

BZip2 0.00201783

2003 Invasion of Iraq 0.00184774

经过在Wikipedia的实际搜索，会发现第一个AWB是一个多义词，其含有多个义项，而其在Wikipedia中的可能引用对象包括排名第二到第七的六个词条。经过查阅，AWB是它们的共同缩写，因此AWB另外六个词条之间任意一个可以形成一个圈，从而在PageRank迭代中相互传递PageRank值；另一种情况就是，别的页面想应用某个词条时，有可能会引用到该词条的缩写页面，这让类似AWB这样的多义词一定程度上能够综合来自各个词条的PageRank值。另一方面，第二到第七这六个词条又有一定的热度：Average White Band是一支活跃了46年的乐队，Adaptive Multi-Rate Wideband是一种音乐文件格式，AWB (album)是Average White Band的一个唱片（曾经是Billboard's Hot 100 第一名）。这三个词条之间的意思相近，很容易和其他同类词条组成一个更大的连通图，在PageRank算法中能得到更高的PageRank值。另外，Afrikaner Weerstandsbeweging是欧洲血统南非人抵抗运动，这是从1973年活跃至今的极端政党，拥有一定的影响力。AWB Limited是澳洲小麦局，也有一定的热度。Wikipedia:Autowikibrowser是一个半自动的MediaWiki编辑器，应用较广泛。所以，即使PageRank算法已经用damping value来应对这种节点相互连通的现象，但是又因为五个词条有一定的热度（即被较多其他词条引用），且都有指向AWB的出链，而AWB又指向它们，这在PageRank算法中更容易获得较高的PageRank值。综上，他们占据了排名的前五名。

接下来，排名第8的Wikipedia:Deletion review和排名第5的Wikipedia:Autowikibrowser都是应用广泛的工具，很容易被其他领域相关的页面引用，PageRank较高有其道理。排名第9的←和排名第12的BOT都是多义词，其PageRank值较高的原因和AWB类似。

排名第13的Canada，排名第11的New York city，排名第10的Thumb都是日常生活中较常使用的名词。排名第15的2003 invasion of Iraq是伊拉克战争，是一个热点话题。排名第14的bzip2是一个基于Burrows-Wheeler 变换的无损压缩软件，是一种较出名的压缩软件。这些页面较高的PageRank值来源于知名度。