基于维基百科的 PageRank 算法实现

1800017821 李云济

October 27, 2020

Contents

1 引言

PageRank，是由 Google 的创始人拉里 • 佩奇和谢尔盖 • 布林于 1998 年在斯坦福大学发明的对网页进行排名的算法，使用网页间的链接计算网页的分数。

本文使用 python 实现了 PageRank 算法，并对维基百科页面 (enwiki-20180920-pages-articles-multistream.xml.bz2) 进行排名。从计算结果来看，PageRank 算法很好地计算出了网页的重要性。

2 数据处理

原始数据为 15.2G 的维基百科页面压缩包，首先使用 WikiExtractor，将数据提取成 1509 个 json 文件，储存在 16 个文件夹中，每个 json 文件包含若干 json 块，一个 json 块为一个网页的 id、url、titie 和 text 组成的字典，在 text 中包含 html 格式的链接，其余信息被舍弃。

第二步将所有网页包含的信息提取出来，构造 data_dicts，即网页字典的数组，损坏的数据均被舍弃，共获得 5719052 个 data_dict, 每个 data_dict 字典包含'id'、'title'、'outlinks'、'inlinks'、'outlinks_id' 和'inlinks_id' 六个 key，其中'inlinks' 和'inlinks_id' 在计算 PageRank 的过程中并不需要，但可以使数据分析更方便直观。links 均从 text 中提取，提取结束后 text 被舍弃，url 也被舍弃。

之后对数据进行抽样，由于数据中很多网页与其他网页的链接很少或没有，采用随机抽样可能获得许多独立网页，丢失许多重要网页，并且使得计算结果不直观。于是经试验设 threshold=24，丢弃 outlinks 和 inlinks 总数小于 threshold 的网页，经过滤得到 1046849 个 data_dict。下面将使用此数据实现 PageRank。

3 算法实现

本文采用改进后带 RWR 的 PageRank 算法，$\alpha = 0.9$, 公式为

$$r = \alpha P^T r + (1-\alpha)\frac{1}{n}e, \quad e = (1, 1, \cdots, 1)^T$$

r 初始化为 $(\frac{1}{n}, \frac{1}{n}, \cdots, \frac{1}{n})^T$, 由于本次实验中 n 巨大，为 1046849，于是为了计算方便及结果直观性，从公式中消去公因式 $\frac{1}{n}$, 并改写为

$$r = \alpha P^T r + (1-\alpha)e, \quad e = (1, 1, \cdots, 1)^T$$

r 初始化为 $(1, 1, \cdots, 1)^T$, 此改动并不影响最终的网页排名。

矩阵 P 从零矩阵开始构造，对每个 data_dicts[i]，对其 outlinks_id 中的每个 j，$P[i][j]+=\frac{1}{m},\ m=len(outlinks\_id)$.

显然 P 为稀疏矩阵，于是在实际计算中采用稀疏矩阵的储存方式及算法。迭代计算时设 $\epsilon=10^{-9}$, 当相邻两次所得 r 之差的二范数小于 ϵ 时，迭代停止，所得 r 即为最终 PageRank 计算结果，对结果进行排序并输出。

4 结果分析

首先显而易见，此结果难以避免存在误差，误差主要来源于对部分数据的丢弃和抽样，不同的超参数也可能产生不同的结果。但对于整体百万个网页的排名来看，影响不大。

在总共 1046849 个结果中，PageRank 值最大的网页的 title 是 United States (2618.97)，排名前 100 的网页大多都与国家或地区相关，范围扩大到 500 也是如此，猜测维基百科对于国家或地区的描述很丰富。

PageRank 值大于 100 的有 581 个网页，小于 1 的有 887785 个，PageRank 值最小为 0.1，可见大部分网页的 PageRank 值处于很小的区间范围内，只有极少部分的网页拥有很大的 PageRank 值。

PageRank 排名最高的人是 Barack Obama，以 336.3161 的 PageRank 值排在第 80 位，在他之后是第 121 位的 George W. Bush(263.7404) 和第 149 位的 Adolf Hitler(235.3352)。

在排名靠前的结果中也存在一些奇怪的 title，例如排名 73 位的 Köppen climate classification(柯本气候分类法)，拥有 356.8316 的 PageRank 值，高于 Cold War(337.1952) 和 Harvard University(293.6824)。查看数据发现有 4733 个链接指向 Köppen climate classification，绝大多数来源于与国家或地区相关的网页，猜测维基百科在叙述许多地区的气候时会引用到 Köppen climate classification，使其获得了很大的 PageRank 值。

下面列出了一些重要或有趣的网页以及他们的 PageRank 值。从中可看出网页

PageRank 值也部分反映了网页 title 在现实世界中的重要性和知名度等关系。例如微软谷歌等知名互联网公司拥有很高的 PageRank 值，星球大战比星际迷航观看人数略多一些 ···PageRank 自己拥有 4.2114 的 PageRank 值，排在第 33581 位。

排名	标题	PageRank 值
1	United States	2618.97
2	World War II	2059.48
3	United Kingdom	1362.12
5	New York City	1118.17
6	World War I	1063.66
9	Soviet Union	889.13
20	China	642.37
33	Marriage	496.00
100	Ukraine	294.02
184	Beijing	201.89
487	Shanghai	115.12
222	Microsoft	183.24
367	Google	136.46

排名	标题	PageRank 值
441	Facebook	124.11
464	Apple Inc.	118.71
909	Java	73.3633
2331	C++	36.7820
2552	Star Wars	34.0746
4188	Star Trek	23.4505
7793	Lionel Messi	14.3797
9463	Cristiano Ronaldo	12.2883
6307	Peking University	16.9712
10784	Tsinghua University	11.0503
33581	PageRank	4.2114
160005	Natural language processing	0.9946
329264	Computer vision	0.4630

5 结语

由本次实验及分析结果来看，PageRank 可以很好地体现网页的重要性，对于人们浏览搜索网页有很大的帮助，且其算法本身非常易于理解且便于计算，是难得一见的实用优秀的算法。不过仅就本文的实现来看，可改进的空间仍很多。

此外，本次实验中对数据进行提取和处理花费的时间要远大于实际计算 PageRank 花费的时间，可见数据挖掘和数据处理的重要性。