互联网数据挖掘第一次作业

——pagerank 计算

朱政烨 1700017760

数据处理

- 首先从 wikidump 下载 60+GB 的 xml 文件
- 解析 xml 的方法有很多，如 DOM、SAX、ElementTree 等，但是我这次采用了 ElementTree，它与 DOM 相比速度更快，与 SAX 相比同样可以用 iterparse 函数，不会一次性在内存中读入整个文档。这对于本次大文件读取是关键的。
- 具体来说，本次利用 python 的 lxml 库，使用 etree 的 iterparse 方法，逐个读取文件中的每个 element 并获取其标签。
- 经过研究数据结构，发现对于本次计算有用的标签有下列几个：
  <page>代表一个词条的整个页面
  <title>是词条的名称
  <id>是词条的序号（本次不采用，而是直接用 title 代表一个词条）
  <text>是词条正文，我们要求的外链就在此中，并且处于双中括号[[]]中
- 我们采用循环遍历每个 tag，每到一个 page 处，就记录下 title，并从 text 中用正则表达式提取出外链，最后都存储下来，存储格式为字典{“title1”:[“外链 1”,”外链 2”,…], “title2”:[“外链 1”,”外链 2”,…],…}，其实就是我们熟悉的邻接表结构。

算法描述

- pagerank 算法其实很简单，在我们上一步能够得出转移矩阵 P 后，直接用 P^T 与 pagerank 值向量 π 相乘，反复多次后就能收敛。（马尔科夫过程）

Transition Probability Matrix $P = \{p_{ij}\}$

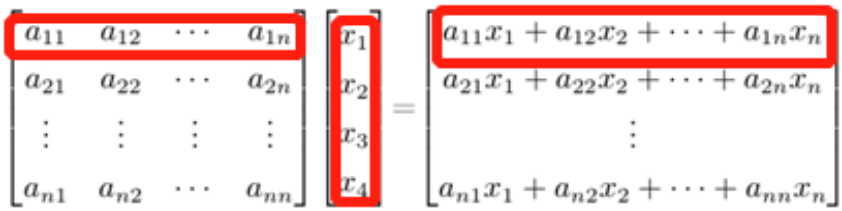
$$p_{ij} = \begin{cases} \dfrac{M(i,j)}{\sum_{v_k \in outlink[v_i]} M(i,k)}, & outlink[v_i] \neq 0 \\ M(i,j) = 0, & otherwise \end{cases}$$

$$P = \begin{bmatrix} 0 & 1/3 & 1/3 & 0 & 0 & 1/3 & 0 & 0 \\ 0 & 0 & 0 & 1/2 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1/2 & 0 & 1/2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1/4 & 0 & 0 & 1/4 & 0 & 0 & 1/4 & 1/4 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1/2 & 0 & 0 & 1/2 & 0 \end{bmatrix}$$

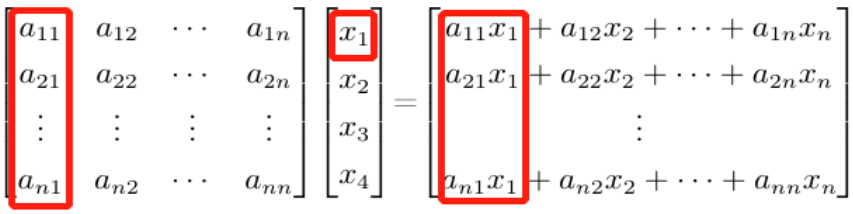
$$\pi(v_i) = \sum_{v_j \in inlink[v_i]} \frac{\pi(v_j)}{|outlink[v_j]|} \quad \Longrightarrow \quad \pi = P^T \pi$$

- 我们这里主要处理这个矩阵向量乘法。直接在 python 中乘是不现实的：首先这个矩阵是 $10^6 * 10^6$ 的，内存不够肯定不能直接运行；而且它也是稀疏的，之前已经用邻接表结构存储，完全没有必要用这个大矩阵。

- 于是我回想到之前学的并行与分布式计算，曾经提到了矩阵向量乘法不同的分解方法：（按行、按列、按块）



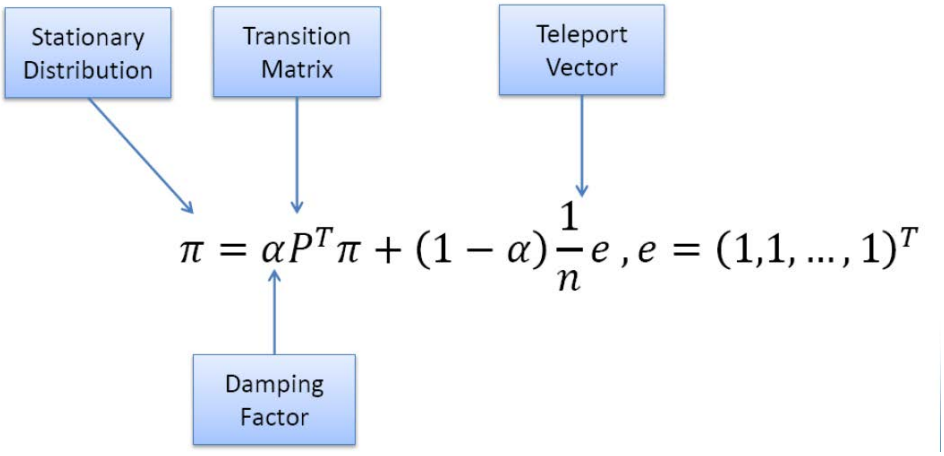
传统的按行分解



按列分解

那我们这里显然可以采用第二种方式。矩阵的第一列，转置前就是第一行，这一行我们很熟悉——它就是我们已经存在邻接表里的 x_1 网页相应的出链，把他与 x_1 相乘，就能得到 x_1 对所有页面的贡献，也就是乘后向量的第一列，然后 x_2、x_3…都这样依次相加，就得到了全部页面对全部页面的贡献，数学上也就是图中的按列相乘，列与列相加的最终结果。

- 实际操作后发现，我们采样的数据量很小，存在排序泄露和排序沉入问题，导致收敛速度很慢而且 pr 值不合理（迭代 20、50、100 次的结果相差很大，而且少数值过大，以及大量的 0 值）。于是我加入了 RWR 的改进，在随机游走过程中重新浏览新网页，也就是在原来的迭代中加入阻尼因数 α=0.85。事实证明这条改动是大有用处的。



结果分析

- 收敛情况

  分别设置迭代次数为 1、10、20、50，输出到文件中，观察数值变化及收敛情况。

  下图依次是 1、10、20、50 次迭代后第一页的词条及结果，可以发现，10 次后已经趋于稳定，而 50 次的结果已经和 20 次相差无几（pr 值小数点九位前都没有任何变化，100 万个词条的排名也完全一样），至此已经可以得出收敛的结论。

```
United States Census Bureau 0.0021208038152328614
Marriage 0.0007644432649006158
White (U.S. Census) 0.0007405483800640203
List of sovereign states 0.0006628857399271051
Hispanic (U.S. Census) 0.0006055275423950848
Latino (U.S. Census) 0.0006042817372393482
Geographic Names Information System 0.0005412708729629829
Native American (U.S. Census) 0.0005513971944494449
World War II 0.0004739688027510905
African American (U.S. Census) 0.0004723470807064254 5
Japan 0.0004676942130391854 7
Asian (U.S. Census) 0.0004545642755165889 5
United Kingdom 0.0004375972846559155 6
Germany 0.0004186649339423988
England 0.0004011319947273810 7
New York City 0.0003444631205689586 6
The New York Times 0.0003227239605476139
London 0.0002974570664889391
India 0.0002580606059830049
California 0.0002407683086071506
Italy 0.0002388307260957121 4
Brazil 0.0002325335809105412
Pennsylvania 0.0002262062889764946
Netherlands 0.0002124488159042642 5
Minnesota 0.0002035904466812979 3
Poland 0.0002015261507353246
China 0.0001906016096598456
Ontario 0.0001892441711695 71
Paris 0.0001863733180210424
Communes of France 0.0001813499560141294 3
Russia 0.0001718755568778735 8
Scotland 0.0001706286046462875
Soviet Union 0.0001679420034312966
Spain 0.0001674136262508849 6
Republican Party (United States) 0.0001610057978692001 3
List of Enochian angels 0.0001566066666666634
Texas 0.0001537295333198113
```

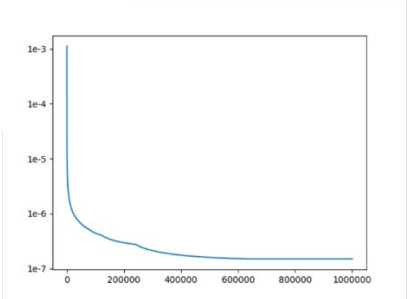
```
Race and ethnicity in the United States Census 0.0011158697870010288
United States Census Bureau 0.0008670301294969095
Marriage 0.0002719941051197725 5
White (U.S. Census) 0.0002626858749363128
List of sovereign states 0.0002187586853284309 4
Hispanic (U.S. Census) 0.0002171911377331179 8
Latino (U.S. Census) 0.0002166041647986756
Geographic Names Information System 0.0001912515740766663
Native American (U.S. Census) 0.0001897080368230776 8
African American (U.S. Census) 0.0001757525350966 4993
The New York Times 0.0001728752704514305 8
World War II 0.0001723018366981523 3
Asian (U.S. Census) 0.0001713133664466691 8
United Kingdom 0.0001455862818105785
Japan 0.0001402607547339286 6
New York City 0.0001274536718690047
Germany 0.0001210652931398691 2
England 0.0001065565157330754 3
London 9.488977226777743e-05
Pennsylvania 8.615025459483422e-05
California 8.578857306741958e-05
Republican Party (United States) 8.296767714164906e-05
India 7.991908361766908e-05
Italy 7.622330586146213e-05
Soviet Union 7.403000546328376e-05
Washington, D.C. 7.344804001070421e-05
Minnesota 7.292827564979789e-05
China 6.9815329056142e-05
Catholic Church 6.752408752495312e-05
Paris 6.629506310629072e-05
Latin 6.600937821711388e-05
Netherlands 6.517103737321558e-05
Russia 6.366781196163792e-05
Europe 6.259786986564783e-05
Poland 6.036805802982225e-05
Texas 6.011076991630416e-05
Pacific Islander (U.S. Census) 5.95677007426111e-05
```

```
Race and ethnicity in the United States Census 0.0011149293506275526
United States Census Bureau 0.0008665013205450971
Marriage 0.0002719047680699448
White (U.S. Census) 0.0002626098758633999
List of sovereign states 0.0002186628044935 4947
Hispanic (U.S. Census) 0.0002171197099755478
Latino (U.S. Census) 0.0002165330202066 6822
Geographic Names Information System 0.0001911836742009 5506
Native American (U.S. Census) 0.0001896407340065376
African American (U.S. Census) 0.0001756970716454 8566
The New York Times 0.0001723404624222104 6
World War II 0.0001719556733744 9638
Asian (U.S. Census) 0.0001712473228817 4933
United Kingdom 0.0001453254650851864
Japan 0.0001399868335212 7016
New York City 0.0001271698228138 4167
Germany 0.0001208687052439515 8
England 0.0001064089092958 6037
London 9.469129838356362e-05
Pennsylvania 8.804585779252049e-05
California 8.564571297844051e-05
Republican Party (United States) 8.276446331890432e-05
India 7.980057077773549e-05
Italy 7.605617183061 01e-05
Soviet Union 7.384758935997144e-05
Washington, D.C. 7.314791598234333e-05
Minnesota 7.287845351566879e-05
China 6.968454171417176e-05
Catholic Church 6.720450359533822e-05
Paris 6.615589960931748e-05
Latin 6.57910886054914e-05
Netherlands 6.503862851256966e-05
Russia 6.352549442788358e-05
Europe 6.247446129783996e-05
Poland 6.026344934429346e-05
Texas 6.002401192773314e-05
```

```
Race and ethnicity in the United States Census 0.0011149291388117756
United States Census Bureau 0.0008665012009306858
Marriage 0.0002719047503062613
White (U.S. Census) 0.0002626099583058004
List of sovereign states 0.0002186627755545263 6
Hispanic (U.S. Census) 0.0002171196934393144 3
Latino (U.S. Census) 0.0002165330037319481
Geographic Names Information System 0.0001911836584876276
African American (U.S. Census) 0.0001756070564052 0943
The New York Times 0.0001723403266510781
World War II 0.0001719555676142961 2
Asian (U.S. Census) 0.0001712473075490060 3
United Kingdom 0.0001453253865608537
Japan 0.0001399866555539801 16
New York City 0.0001271697504207217 4
Germany 0.0001208686434746546 29
England 0.0001064088633164641
London 9.469123767060 94e-05
Pennsylvania 8.804503243104607e-05
California 8.564567909182 23e-05
Republican Party (United States) 8.276441611715596e-05
India 7.980053651361685e-05
Italy 7.605611696832542e-05
Soviet Union 7.384753629156 56e-05
Washington, D.C. 7.314784421913962e-05
Minnesota 7.287844185074531e-05
China 6.968450448449325e-05
Catholic Church 6.720440279760835e-05
Paris 6.615565660420637e-05
Latin 6.5791018737684e-05
Netherlands 6.503858754096144e-05
Russia 6.352545198901926e-05
Europe 6.247442591019842e-05
Poland 6.026341717161062 6e-05
Texas 6.002399151508751 4e-05
Pacific Islander (U.S. Census) 5.953103630555582e-05
```

- 结果分布

  通过 matplotlib 绘图，得到数据的分布图。可以发现，pr 值虽然从 1e-7 到 1e-3 都有，但是主要还是集中在 1e-7~1e-6 之间，高于 1e-6 的仅有 23130 个词条，高于 1e-5 的仅有 589 条。



- 热门词条

  我选取了 100 万中排名前 100 个词条，翻译成中文，可以看出美国人口普查相关内容占很大的比重，而且其中人种中白人排第一。除此之外，英国是排名最高的国家，纽约是排名最高的城市。当然，我们只在 wiki 里选了 100 万个词条，得到的这些信息也只能代表前 100 万词条的趋势罢了。

```
美国人口普查中的种族和种族
美国人口调查局
结婚
白人（美国人口普查）
主权国家名单
西班牙裔（美国人口普查）
拉丁美洲人（美国人口普查）
地名信息系统
美洲原住民（美国人口普查）
非裔美国人（美国人口普查）
纽约时报
第二次世界大战
亚洲人（美国人口普查）
英国
日本
纽约城
德国
英格兰
伦敦
宾夕法尼亚州
加利福尼亚州
共和党（美国）
印度
意大利
苏联
华盛顿特区
明尼苏达州
中国
天主教堂
巴黎
```

总结反思

通过这次作业，我掌握了初步的数据挖掘知识，包括 xml 文件的解析、pagerank 算法的简单实现，编程能力也有了进一步的提升。这次作业还能有很多提升的地方，比如算法的优化、并行化，数据的优化等等，希望下次时间充裕可以更上一层楼。

2019.10.8