谷歌搜索问题

Q. 创建谷歌时，拉里佩奇和谢尔盖布林对互联网的基本认识是什么？

A. 在互联网越来越普及的时代行情下，为了去访问某个网站，用户不可能将每一个网站的网址都记录下来以便访问。因此，搜索会变得越来越重要。

在当年的市场上，YAHOO作为龙头具有75%的占有率。

Q. 谷歌的创始人之一拉里佩奇凭借网页排名的数学模型**Page Rank**当选美国工程院院士。在该模型的构建中，有向图的模型可以看成是网页之间的相互投票。按民主的原则，每一票之间就是没有差别的。这种思想是否正确？

A. 不正确。网页之间的投票其实并不是民主的。网页得到的票数越多，则说明该网页的重要性越大。而重要性大的网页投出的票也必然占的分量会很大。因此，在该模型中我们会对网页的票数做出加权来平衡这个问题。

Q.请基于网页排序模型的核心思想查找资料，进一步改进该模型。

Page Rank模型回顾

$$x_i = \sum_{j:j\to i} \frac{x_j}{cj}$$

$$\sum x_i = 1$$

参数说明：

- x_i表示网页i得到的加权票数
- c_i 表示网页投出的加权票数

在改进模型之前，我们需要知道模型的用途。**Page Rank** 算法是**Google** 用来衡量网络重要性的。更直接地说，他其实跟我们日常使用Google做的搜索引擎并没有很大的关系。

对于某个互联网网页A来说，该网页Page Rank的计算基于以下两个基本假设：

- 数量假设：在Web图模型中，如果一个页面节点接收到的其他网页指向的入链数量越多，那么这个页面越重要。

- 质量假设：指向页面A的入链质量不同，质量高的页面会通过链接向其他页面传递更多的权重。所以越是质量高的页面指向页面A，则页面A越重要。

参数改进

在这两种假设的情况下，我们考虑互联网上的这些网站：(不同域名之间的链接)

1. 没有任何入链的网站(如某小学生自己搭建的博客有很多出链，但是没有入链)
2. 没有任何出链的网站(如某西交软院学生写的全是BUG和ERROR的html文件，被老师放在了服务器上)

显然，这样两种网站无法通过Page Rank算法获得一个合适的排位，并且还会导致他们的出链权重失衡。

因此，我们需要对每个网站的重要性增加一个基数，便于对一些小型网站和新网站做出评估。

我们可以基于访问量或者规模等参数给予网站一个连接系数α_i。通过这个系数我们可以获得用户在经过某个指向该网页的节点时访问到该节点的概率，即网站**i**的吸引力。因此，我们将上面的模型可以修正为：

$$x_i = \alpha_i \sum_{j:j\to i} \frac{x_j}{cj}$$

$$\sum x_i = 1$$

计算改进

很明显，除了像谷歌主页这样的网页之外，其他网页的入链和出链涵盖的范围都很小。因此，这个图矩阵的大多数元素都是0，是一个稀疏矩阵。稀疏矩阵可以方便做矩阵运算，我们可以通过PCA算法将重要性相似的网站筛选出来，并且给予他们合适的分级。

PCA算法简介：

PCA（Principal Components Analysis）即主成分分析，是经常用到的降维方法，例如，我们在处理有关数字图像处理方面的问题时，比如经常用的图像的查询问题，在一个几万或者几百万甚至更大的数据库中查询一幅相近的图像。这时，我们通常的方法是对图像库中的图片提取响应的特征，如颜色，纹理，sift，surf，vlad等等特征，然后将其保存，建立响应的数据索引，然后对要查询的图像提取相应的特征，与数据库中的图像特征对比，找出与之最近的图片。这里，如果我们为了提高查询的准确率，通常会提取一些较为复杂的特征，如sift，surf等，一幅图像有很多个这种特征点，每个特征点又有一个相应的描述该特征点的128维的向量，设想如果一幅图像有300个这种特征点，那么该幅图像就有300*vector（128维）个，如果我们数据库中有一百万张图片，这个存储量是相当大的，建立索引也很耗时，如果我们对每个向量进行PCA处理，将其降维为64维，就会加速计算并且减小存储空间。