Pagerank

本章介绍综述中提到的三种算法，我们将从如下几个方面介绍：

基本原理，算法输入、输出和算法应用。

1 pagerank

佩奇排名，又称网页排名，是一种由搜索引起根据网页之间的相互超链接计算的技术，以google公司的创始人拉里·佩奇的姓来命名。Google用它来体现网页的相关性和重要性。PageRank所采用的图是有向图，那些在图中被更多有向边指向为重点的点，将获得更高的权重。

1.1算法原理

方程式引入了*随机浏览*的概念，即有人上网无聊随机打开一些页面，点一些链接。一个页面的PageRank值也影响了它被随机浏览的概率。为了便于理解，这里假设上网者不断点网页上的链接，最终到了一个没有任何链出页面的网页，这时候上网者会随机到另外的网页开始浏览。

为了处理那些"没有向外链接的页面"（这些页面就像"黑洞"会吞噬掉用户继续向下浏览的概率）带来的问题，$\mathbf{d=0.85}$，这里的d被称为阻尼系数（damping factor），其意义是，在任意时刻，用户到达某页面后并继续向后浏览的概率，该数值是根据上网者使用浏览器书签的平均频率估算而得。$\mathbf{d=0.15}$（就是用户停止点击，随机跳到新URL的概率）的算法被用到了所有页面上。

所以，这个等式如下：

$$PageRank(p_i) = \frac{1-d}{N} + d \sum_{p_j \in M(p_i)} \frac{Pagerank(p_j)}{L(p_j)} \tag{1-1}$$

p_1，p_2,....p_n是被研究的页面，$M(p_i)$是链入p_i页面的集合，$L(p_j)$是p_j链出页面的数量，而N是所有页面的数量。

PageRank值是一个特殊矩阵中的特征向量。这个特征向量为：

$$\mathbf{R} = \begin{bmatrix} PageRank(p_1) \\ PageRank(p_2) \\ . \\ . \\ . \\ PageRank(p_N) \end{bmatrix} \tag{1-2}$$

R是等式的答案

$$\mathbf{R} = \begin{bmatrix} (1-d)/N \\ (1-d)/N \\ (1-d)/N \end{bmatrix} + d \begin{bmatrix} \iota(p_1,p_1) & \iota(p_1,p_2) & \dots & \iota(p_1,p_N) \\ \iota(p_2,p_1) & \dots & & \\ . & & & \\ . & & \iota(p_i,p_i) & \\ \iota(p_N,p_1) & \iota(p_N,p_2) & \dots & \iota(p_1,p_2) \end{bmatrix} \mathbf{R} \tag{1-3}$$

如果p_j不链向p_i，而且对每个j都成立，$\iota(p_i,p_j)$等于0

$$\sum_{i=1}^{N} \iota(p_i,p_j) = 1$$

通过若干次递归，最终对于任意的p_i，其上一次的递归值的差的绝对值小于δ，其中$\delta = 0.0001$。在spark Graphx中，pageRank的两个参数即为随机跳转概率d和递归停止容忍度δ。

1.2算法输入、输出

对于一般的pageRank算法，输入应当至少包含以下内容：

边的集合，包括每条边的src_id和dst_id;

点的集合，包括点的id;

默认条件下，随机跳转概率$d = 0.15$，递归停止容忍度$\delta = 0.0001$。

算法的输出包含：

点的id及其对应的pageRank值。

iFusion算法输入输出说明：

根据案例(图谱)的id获取案例中的vertex和其相关联的edge；

根据读出的案例(图谱),构建pageRank需要的graph对象;设置pageRank参数,调用pageRank算法；

将pageRank算法结果保存为一个新的案例(图谱),可在iFusion上查看。

1.3 算法应用

文学作品

最重要的文学作品是什么？PageRank算法可以帮助回答这个看似很主观的问题。Nebraska大学的一位文学教授开发了一款软件，使用了PageRank和其他的算法。他的研究对象是十九世纪的文学作家。经过分析了近3600部长篇小说，软件得出了结论：简•奥斯丁和沃尔特•司各特是那个年代最有影响力的作家。

体育运动

在特定的运动项目中，谁是历史上最好的球队或球员？这个问题如果交给球迷，那必将吵得不可开交，因为评判标准同样是主观的。有一则论文利用PageRank分析了1968年之后的所有职业网球比赛，它将相同的两个对手之间的比赛结果进行匹配，以“声望得分”为基础构建了一个网络。得出的结论是，在网球界，Jimmy Connors是史上最好的球员。就像Gleich说的，这些排名背后的基本思想是：假设一个粉丝会追随着一只球队或球员，直到他被打败，而后他会继续追随胜利的一方，直到结果出现。这类似于网上冲浪者在网站链接中做出选择。

神经科学

Gleich 在他的论文中写道：“人类的大脑是一个重要的网络，可我们对它的了解少的可怜”。PageRank当然也适用于此。最近，它被用来评估不同大脑区域之间的联结和重要性，以及随着年龄的变化结果会如何改变。

癌症研究

在一篇名为“Google Goes Cancer”的论文里，研究人员开发了一种基于Google算法的“新型计算方法”，该算法帮助他们确定了七个与遗传有关的肿瘤基因，这将帮助医生更好的指导癌症治疗的过程。

交通网络

PageRank的另一应用是用来预测城市里的交通流量和人流动向。有一项研究依赖于该算法的一个关键因素：一个叫做teleporting（传送）的概念，它模拟了人类的决定：开始或中止行程，或者在既定的街道停车。这有助于交通运输研究人员更好的创建测量模型，模拟道路的车流量和人流量。

2 参考资料

佩奇排名(PageRank)的原理

https://en.wikipedia.org/wiki/PageRank