搜索引擎是如何判断网页的重要性的

[马少平，王东]

搜索引擎提供的搜索结果，最基础的要求是与查询词相关，另外一个要求网页足够重要。那么，搜索引擎是如何判断一个网页的重要性的呢？这要归结为一个称为 PageRank 的排序算法。

我们知道一个网页中往往会有一些超链接，用户点击这些链接会指向其他网页。仔细思考一下，网页设计者为什么会设计这些链接呢？这是因为被链接的网页与当前网页相关，或者有某种联系，或者对了解该网页有所帮助……，总之链接本身提供了一种和网页内容本身无关的信息，这种信息更重要的价值是评价网页彼此之间的相关性。可以认为，如果一个网页被很多网页链接，那么有理由说明该网页比较重要。

如果我们将网页之间的链接关系表示成一张有向图，其中节点表示网页，节点间的边表示两个网页之间存在链接，边的方向代表链接指向，如图 1 所示。依据前述讨论，如果某个节点被很多边链接，则该节点是一个重要节点，对应的网页比较重要。由此，我们得到了一个判断网页重要性的方法。

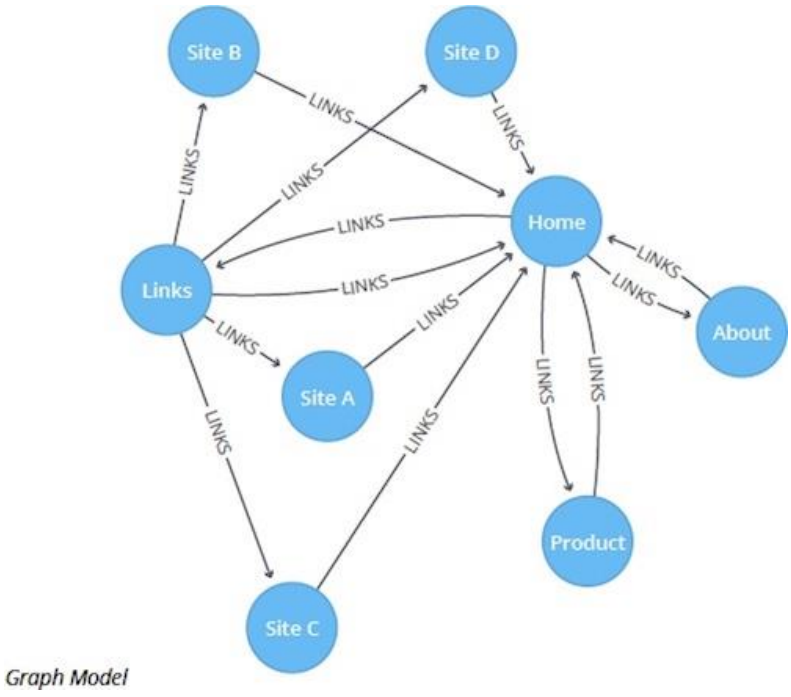

图 1：描述网页间的链接关系的有向图

然而，实际情况要复杂一点点，这是因为不同的链接本身的重要性是不一样的，**来自重要网页的链接比来自普通网页的链接应更受重视**，这是因为如果当前网页被某个重要网页链接，说明这个网页自身也是重要的。基于这一思路，在计算网页重要性时可以将~~你~~链接来源的重要性[illegible]链接 B->A 的源网页 B 的重要性，回过头来，计算 B 的重要性时，也有可能用到源网页 A 的重要性，这样就形成了一个循环，需要一个迭代算法来求解。谷歌的 PageRank 算法[1]正是这样一种迭代求解算法。

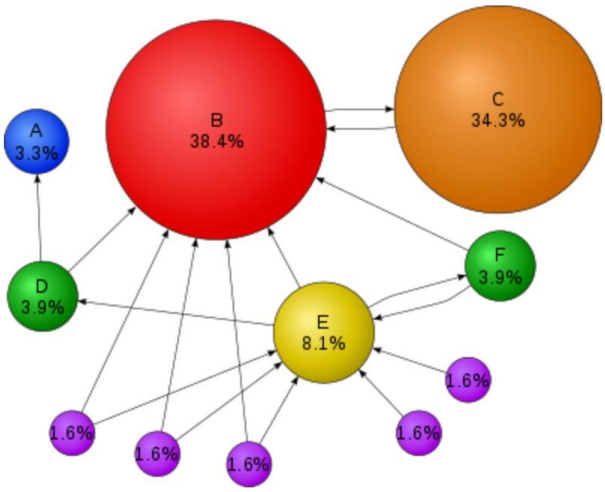

图 2：基于 PageRank 算法对网页计算重要性 [2]

在 PageRank 算法中，一个网页的 PageRank 值反应了该网页被访问的概率，概率越大说明网页的重要性越大。假定网页 P 链接到了网页 A，并且 P 链接到包括 A 在内的 K 个链接，PageRank(P)为访问到 P 的概率，再假定用户访问到 PA 后，按照均匀分布访问 PA 链接到的 K 个网页，则从 P 访问 A 的概率为 PageRank(P)/K。如果有 n 个网页链接到了 A，分别为 P_i(i=1,2,⋯,n)，而网页 P_i 分别链接到 K_i 个网页，则通过这些网页访问 A 的总概率为通过这些网页访问 A 的概率之和，即：

$$\mathrm{PageRank(A)} = \sum_{i=1}^{n} \frac{PageRank(P_i)}{K_i} \qquad (1)$$

除了通过链接访问网页 A 之外，该网页还有可能被用户随机访问。假设共有 N 个网页，且被随机访问的概率是均匀分布，则网页 A 被随机访问的概率为 1/N。这样的话，网页 A 被访问的总概率应为随机访问概率和链接访问概率的加权和，即：

$$PageRank(\mathrm{A}) = \alpha\frac{1}{N} + (1-\alpha)\sum_{i=1}^{n} \frac{PageRank(P_i)}{K_i} \qquad (2)$$

其中$0 \le \alpha \le 1$为调节系数。

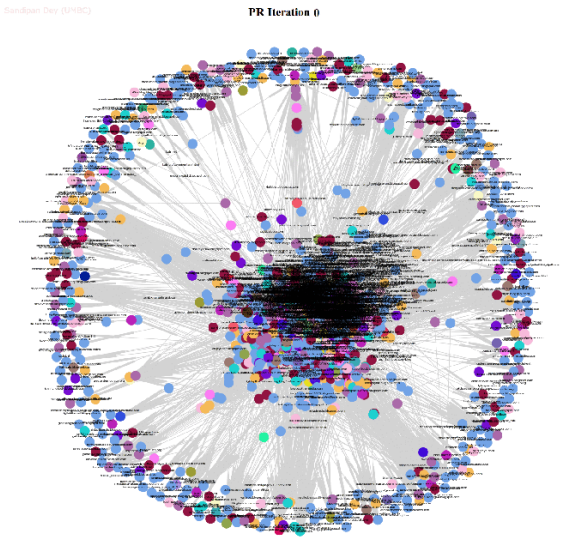

图 3：PageRank 迭代过程【动图】[3]

需要注意的是，公式（1）是一个迭代计算公式，因为 PageRank(A)的计算需要用到 PageRank(P_i)，因此 PageRank 是一个迭代算法，每次迭代会对所有网页的 PageRank 值进行更新，更新后的值用到下一轮迭代中，这一迭代过程持续进行，直到 PageRank 的计算值收敛为止。实际搜索引擎中，网页的 PageRank 值是线下计算好的，处理搜索请求的时候直接应用就可以了。由于互联网上的网页是动态变化的，每隔一段时间，PageRank 就需要重新计算一次，以满足网页动态变化的需求。

[1]Page L, Brin S, Motwani R, et al. The PageRank citation ranking: Bringing order to the web[R]. Stanford InfoLab, 1999.

[2] How to get web traffic from Google, https://www.ethanhein.com/wp/2010/how-to-get-web-traffic-from-google/

[3] https://sandipanweb.files.wordpress.com/2017/09/anipr.gif?w=676