On Page Rank

C. Hoede

Faculty of Electrical Engineering, Mathematics and Computer Science
University of Twente
P.O.Box 217
7500 AE Enschede, The Netherlands

Abstract

In this paper the concept of page rank for the world wide web is discussed.
The possibility of describing the distribution of page rank by an exponential law is considered.
It is shown that the concept is essentially equal to that of status score, a centrality measure discussed already in 1953 by Katz. A structural classification of users in the web is given in terms of graph theoretical concepts.

Key words: page rank, status score, graph

AMS Subject Classifications 2000: 05C99, 68M10, 94C15

1 Introduction

The concept of page rank was introduced by Brin and Page in 1998 [1]. The world wide web, to be called just web from now on, can be described by a directed graph. For graph theoretical terminology we refer to one of the many text books. As there is no uniformity of terminology we have to state that a directed graph $\vec{G}(V, A)$ consists of two sets, the set V of *vertices* and the set A of *arcs*. A is a subset of $V \times V$, the Cartesian product of V with V, where V is a set of labeled elements. The arcs may be labeled as well. Labels of vertices describe the sites in the web. If site v can connect to site w then arc (v,w) is present in $\vec{G}$. The label l of (v,w) can be chosen to express the extent to which this is done. In principle $\vec{G}$ is a complete directed graph. The arcs going out from a vertex v, including the *loop* (v,v), to the N vertices of $\vec{G}$ may carry labels l_i, chosen such that

$$\sum_{i=1}^{N} l_i(v, v_i) = C_v, \tag{1}$$

where i refers to vertex v_i in some numbering of V and C_v is a constant specific for vertex v.

Based on the pattern of choices made by the sites, the problem is to determine a measure for the centrality of a site. The measure should incorporate the fact that some sites are more often visited than others. This is what the concept of *page rank* is about. We will come back to the concept more explicitly in Section 3.

2 Description of a measure distribution

Presumably a minority will have a high rank and the number a(r) of sites with rank r, still to be defined, will diminish from the minimum value R_m for r to the maximum value R_M for r, see Figure 1.

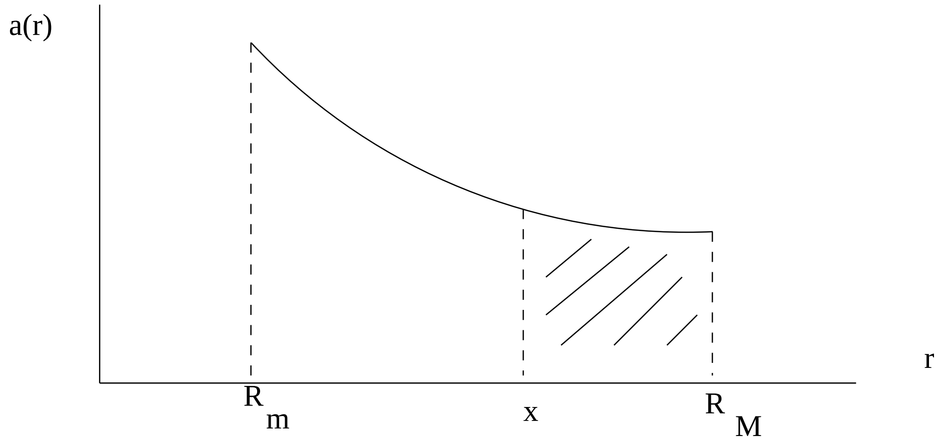

Figure 1: A measure distribution

We focus on the arced area, giving the total number T(x) of sites greater than or equal to x, where, as N is very large, we have gone over from a discrete description to a continuous one.

$$T(x) = \int_x^{R_M} a(r)dr.$$

In the literature for T(x) assumptions are made, in particular that T(x) behaves as

$$T(x) = C_1 x^{-\alpha} + C_2, \tag{2}$$

with $\alpha > 1$.

We want to see which consequences such an assumption has, given α, R_m and R_M, for the average rank $\bar{R}$.
From $T(R_M) = 0$ follows

$$C_2 = -C_1.R_M^{-\alpha}$$

and from $T(R_m) = N$ follows

$$C_1.R_m^{-\alpha} + C_2 = N,$$

leading to

$$C_1 = N.\frac{1}{[R_m^{-\alpha} - R_M^{-\alpha}]}$$

and

$$C_2 = -N.\frac{R_M^{-\alpha}}{[R_m^{-\alpha} - R_M^{-\alpha}]}.$$

So

$$T(x) = N.\frac{[x^{-\alpha} - R_M^{-\alpha}]}{[R_m^{-\alpha} - R_M^{-\alpha}]}. \tag{3}$$

Differentiating Equation (1) to x on both sides gives

$$\frac{dT(x)}{dx} = -a(x),$$

from which we derive

$$a(r) = N.\frac{\alpha}{[R_m^{-\alpha} - R_M^{-\alpha}]}.r^{-\alpha-1}.$$

Now $\bar{R}$ follows as

$$\bar{R} = \frac{1}{N}\int_{R_m}^{R_M} a(r).rdr = \frac{\alpha}{[R_m^{-\alpha} - R_M^{-\alpha}]}\int_{R_m}^{R_M} r^{-\alpha}dr = ... = \frac{\alpha}{\alpha-1}.R_m.\frac{1-(\frac{R_m}{R_M})^{\alpha-1}}{1-(\frac{R_m}{R_M})^{\alpha}}. \tag{4}$$

If $R_m = f.R_M$ we have

$$\bar{R} = \frac{\alpha}{\alpha - 1}.R_m.\frac{1 - f^{\alpha-1}}{1 - f^{\alpha}}.$$

If, for example, $\alpha = 2$ and $f = \frac{1}{2}$ we find $\bar{R} = \frac{4}{3}.R_m$ and

$$\lim_{f \to 1} \bar{R} = \frac{\alpha}{\alpha - 1}.R_m.\frac{\alpha - 1}{\alpha} = R_m,$$

Equation (4) is an important consequence of the assumptions we made, in particular that on the shape of the function T(x) in Equation (2).

3 Page rank and centrality

In a directed graph $\vec{G}$ each vertex j has an outdegree od(j) and an indegree id(j), being the number of outgoing arcs and incoming arcs respectively. For the complete graph both are N. We will discuss this case first as, in principle, it covers all other cases.

The labels on the arcs are assumed to satisfy Equation (1). We now interpret C_v as the intensity with which from site v sites are visited. "Intensity" may be seen as a frequency, say the number of jumps to other, possibly also the same, sites per unit of time. The labels l_i, i = 1,2,...,N, can then be seen as differentiating frequencies of jumps to these sites.

As formulated by Volkovich, Litvak and Donato [7], *Page Rank* PR is a method for evaluating " popularity of nodes in information networks". It is therefore a kind of centrality measure. The basic modeling given by Brin and Page [1] is by an equation for the page rank values PR(i), i = 1,2,...,N. This equation is to capture the idea that the rank for a site depends on the ranks of the sites that choose the site to jump to and on the frequency with which this is done. This basic modeling is expressed by Equation (5)

$$PR(i) = \sum_{j=1}^{N} l_j(v_j, v_i).PR(j). \tag{5}$$

The important next step is, of course, the choice of the labels l_j.
First we can choose $C_v = 1$ for all vertices v of $\vec{G}$. The labels can then be interpreted as probabilities, the setting chosen in [7]. Second, we can, for each vertex v, determine a set O_v of outgoing arcs of which one is chosen to jump along. This with probability c, $0 \leq c \leq 1$, whereas a jump following one of the other outgoing arcs happens with probability 1-c. For the labels this means that they are $\frac{c}{|O_v|}$ in the first case and $\frac{1-c}{N-|O_v|}$ in the second case. O_v might be called a set of preferred outgoing arcs. If $O_v = \emptyset$, then the vertex is called a "dangling" vertex in [7]. In that paper the jump is either along a preferred arc with total probability c or to any other vertex with total probability 1-c, meaning that also preferred

arcs are to be taken into consideration. If $\mathcal{D}$ is the set of dangling vertices this leads to Equation (6)

$$PR(i) = c\sum_{j\rightarrow i} \frac{1}{od(j)}.PR(j) + \frac{c}{N}\sum_{j\in\mathcal{D}} PR(j) + \frac{1-c}{N}, i = 1, ..., N, \tag{6}$$

given in [7]. The notation $j \rightarrow i$ here indicates that (j,i) is a preferred outgoing arc of j. This modeling slightly deviates from the one given above. Other choices of the labels are possible. A variety of page rank equations is therefore possible.

For our purpose we will stick to the general Equation (5).

The main point of this section is that the concept of page rank is not new. Centrality of actors in social networks has been considered in sociology as far back as 1953 by Katz [5]. The measure searched for was that for influence or *status* of an actor. Each actor is supposed to influence all other actors, and Katz's approach was to calculate all weighted paths going out from a vertex. That is, the social network was modeled as a complete directed graph $\vec{G}$, with labels on the arcs, possibly zero, and the weight of a path from vertex v was defined as the product of the weights on the arcs of the path.
Now these weights are just labels and the sum of the weights of arcs going out from v may be C_v. When we now remark that the paths from vertex v can all be seen as consisting of one first arc, to a neighbour, followed by a path from that neighbour, then defining S, status, as the sum of weights of all outgoing paths we obtain

$$S(i) = \sum_{j=1}^{N} l_i(v_i, v_j).S(j). \tag{7}$$

But now we see that Equation (5) is precisely Equation (7), when all arcs in the graph $\vec{G}$ for the web are reversed, (v,w) becomes (w,v), while the labels are kept the same, i.e. putting

$$l_i(v_i, v_j) = l_j(v_j, v_i).$$

The conclusion is that the page rank equation is nothing but the status equation for the "inverted web". This means that all the literature on that centrality measure is applicable.

As a final remark we point out that both equations determine page rank respectively status up to a constant factor.

4 Page rank and in-degree

To discuss the relationship between page rank and indegree in the webgraph, we first discuss a status score introduced by Hoede [3] in an internal report, but contained in the GRADAP manual for software used in the social sciences [2].

Let W be the adjacency matrix A of $\vec{G}$ but with entries 1 for $A_{i,j}$ replaced by entries $W_{i,j}$. As is well-known $(A^l)_{i,j}$ gives the number of walks of length l from vertex i to vertex j. $(W^l)_{i.j}$ gives the sum of weights of walks of length l from vertex i to vertex j. Status seen as sum of the weights of all paths starting in vertex i gives

$$\underline{S}(i) = \sum_{l=1}^{\infty} \sum_{j=1}^{N} (W^l)_{i,j} \tag{8}$$

As

$$(W^l)_{i,j} = \sum_{k=1}^{N} (W^{l-1})_{i,k} (W)_{k,j}$$

for $l > 1$, we have

$$\underline{S}(i) = \sum_{j=1}^{N} (W)_{i,j} + \sum_{k=1}^{N} (W)_{i,k} . \underline{S}(k).$$

We now put $\underline{d}(i) = \sum_{j=1}^{N} (W)_{i,j}$, the i-th component of a vector $\underline{d}$ called the *direct influence* vector. $\underline{S}(i)$ is the i-th component of the *status score* vector. This vector satisfies the equation

$$\underline{S} = \underline{d} + W.\underline{S}. \tag{9}$$

However, for this the series of elements of powers of W should converge. This was one of the main problems in the theory. A solution can be obtained by using the idea of *renormalization* as common in theoretical physics. Instead of attaching attenuation factors to weights of arcs to control the convergence, as was e.g. done by Hubbell [4], we can take Equation (9) as **definition** of $\underline{S}$ to obtain

$$\underline{S} = [I - W]^{-1}.\underline{d}$$

for the status score. Only in case the matrix I - W is not invertible, which can be corrected by a minute change in W, this formula cannot be applied. Thus convergence problems are avoided.

We now recall that page rank is essentially status score for the inverted web. Development of $[I - W]^{-1}$ in Equation (10) shows that

$$\underline{S} = I.\underline{d} + W.\underline{d} + W^2.\underline{d} + ... \tag{10}$$

The first term in this development is $I.\underline{d}$ so, in that approximation, $\underline{S}(i) = \underline{d}(i)$, the direct influence of vertex i. But that is essentially the weighted outdegree in the inverted web. So the weighted indegree in the web in first approximation determines the page rank in the web.

A second way to arrive at this result is to consider Equation (5). Without any further information, a first approximate solution for the page rank is obtained by putting PR(j) equal to the average page rank $\overline{PR}$, leading to

$$PR(i) = \sum_{j=1}^{N} l_j(v_j, v_i).\overline{PR}.$$

But, apart from the factor $\overline{PR}$, this is precisely the weighted indegree of vertex i.

We refer to [7] for a more refined treatment of the relationship between page rank and indegree in the web, that turn out to have the same distribution, as might be expected.

5 Structural classification of sites

The centrality of a web site, its "popularity" if one wants, is an obvious concept to consider. However, there are other interesting questions. Is there a way to distinguish between sites on the basis of structural aspects?
In this case we consider a, not necessarily complete, directed graph $\vec{G}$ without any labels on the arcs. Generalization to a complete graph with labels can come later.

A first distinction between vertices of $\vec{G}(V, A)$ can be made on the basis of outdegree od(v) and indegree id(v), $v \in V$.
A vertex with id(v)=0 is a *source*, whereas a vertex v with od(v)=0 is a *sink*. Clearly in the web the sites that are sources are extremely impopular, whereas the sites that are sinks are the ones to which jumps are made, but as an end-destination for a path of jumps from site to site by some user of the web. A first natural feature to consider is the quotient

$$Q(v) = \frac{id(v)}{od(v)}$$

for vertex v. The range of Q is from zero to infinity. Q partitions the vertex set V into classes of vertices with the same value. Within a class the actual values of id(v) or od(v) may distinguish between the class members. Consider Figure 2.

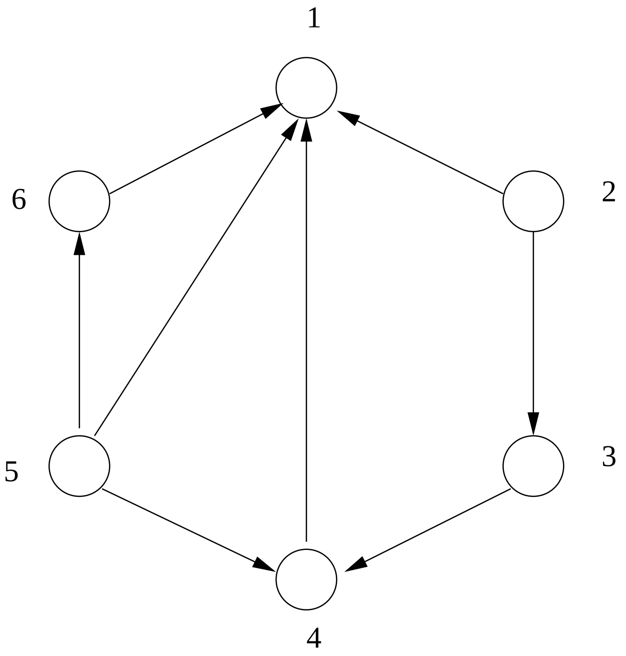

Figure 2: A graph $\vec{G}$

For Q we find Q(1) = ∞, Q(2) = 0, Q(3) = 1, Q(4) = 2. Q(5) = 0, Q(6) = 1.

In a large web different sites v may have identical values for id(v) and od(v). Yet there may be a feature that distinguishes them. Consider the out-neighbours of the vertices v_1 and v_2 with $od(v_1) = od(v_2)$. There may be significant difference between the ways these neighbouring sets are connected. In Figure 3 we consider two vertices with the same out-degree 4.

The "out-neighbourhood" of v_1 contains only one arc, whereas that of v_2 contains seven arcs. Various measures to distinguish between these two situations are possible.
One, simple, way to compare the "complexity" of two out-neighbourhoods is to consider the "density" of the graph induced by the neighbouring vertices. In our example, we may compare the numbers 1 and 7 with the maximum number of arcs possible. Including loops there are $4^2 = 16$ possible arcs, which would lead to complexities $\frac{1}{16}$ and $\frac{7}{16}$ for the out-neighbourhoods of v_1 and v_2.
Another way, more akin to the concept of complexity in graph theory, would be to consider all outgoing spanning trees on the vertex set consisting of a vertex v and all its out-neighbours, with root in v. For vertex v_1 there would be two such trees, for vertex v_2 only one arc is obligatory and the loop can be left out of consideration. The problem is then to count the outgoing spanning trees in the graph of Figure 4, where the obligatory arc is drawn bold. The number may increase rapidly for larger neighbourhoods. For the undirected complete labeled graph on n vertices there are n^{n-2} different spanning trees.

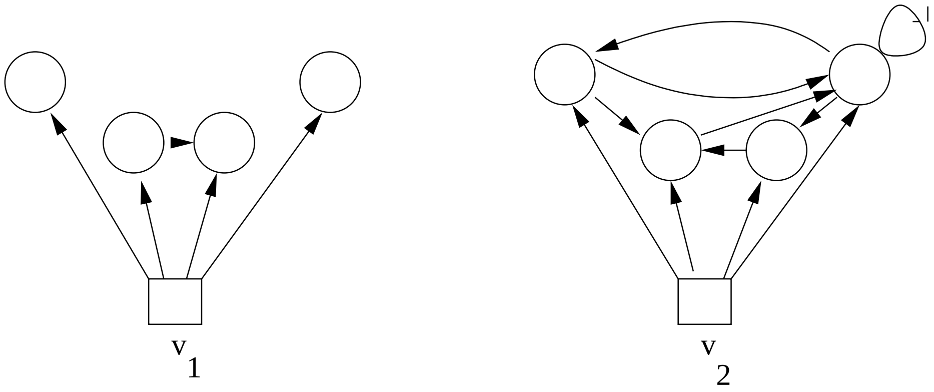

Figure 3: Two out-neighbourhoods

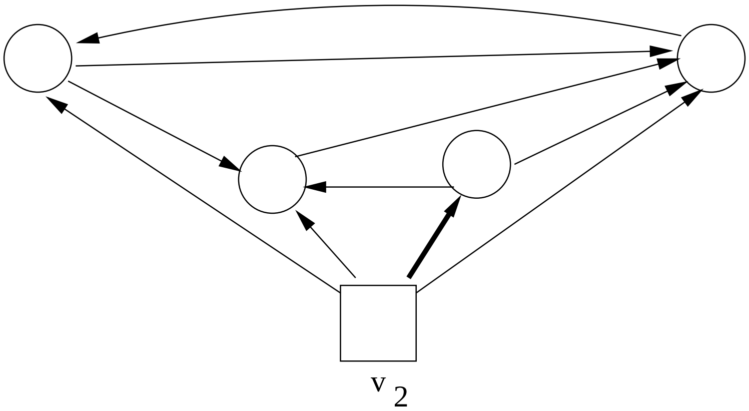

Figure 4: Obligatory arc

For the "in-neighbourhood" an analogous calculation can be made. Let $C_{in}(v)$ and $C_{out}(v)$ denote these complexities. Then, next to id(v), od(v) and Q(v), we can distinguish vertices on the basis of $C_{in}(v)$ and $C_{out}(v)$. Clearly a vertex for which $C_{in}(v) = C_{out}(v) = 0$ has a position in the web, different from one for a vertex w for which $C_{in}(w)$ and $C_{out}(w)$ are both high. Note that these numbers may be normed by dividing by the maximum value that they can assume, like we did for the simple measure.

The generalization to labeled directed graphs is immediate by replacing indegree and outdegree by weighted indegree and outdegree, possibly also allowing that $\vec{G}$ is complete now. For the complexities one might, for example, not count the number of arcs in an in-neighbourhood or an out-neighbourhood, but the sum of the weights of such arcs. In all cases the interpretation should make sense.

The quotient Q may be used, also in the weighted version, in the following way. Vertices of $\vec{G_0}$, the whole web graph, may be partitioned into two sets. A set for which $Q \geq 1$ and a set of vertices for which $Q < 1$. The choice of the constant 1 is arbitrary here, one might also consider the set with minimum value of Q and the rest of the vertices. The vertices with the higher values of Q may be said to get more *attention* from other vertices than they give attention to other vertices.
Now we delete the vertices of the second set and their adjacent arcs and focus on the graph $\vec{G_1}$ induced by these vertices. Note that the degrees respectively the weighted values of indegrees and outdegrees now may have changed. A vertex that received much attention from vertices of the second set, in $\vec{G_0}$, may receive little attention from the other vertices in $\vec{G_1}$. Again Q's can be considered, now in $\vec{G_1}$, and a second partitioning takes place. The "surviving" vertices, for which $Q \geq 1$, then induce a graph $\vec{G_2}$. The procedure can be repeated to any desired extent.

For the graph in Figure 2, $\vec{G_0}$, the first partition yields {1,3,4,6} for the first set and {2,5} for the second set. Figure 5 gives the graph $\vec{G_1}$.

For $\vec{G_1}$ we find Q(1) = ∞, Q(3) = 0, Q(4) = 1 and Q(6) = 0. Consequently the partitioning gives {1,4} and {3,6}. $\vec{G_2}$ is induced by {1,4} and consists of the single arc (4,1). In $\vec{G_2}$ Q(1) = ∞ and Q(4) = 0, partitioning $V(\vec{G_2})$ into {1} and {4}. The outcome of a simple procedure like this is a ranking:

$$\{2,5\} \succ \{3,6\} \succ \{4\} \succ \{1\}.$$

As getting to the top of such a ranking requires incoming arcs from vertices that rank higher and higher themselves, the basic idea of the page rank Equation (5), as well as the status score Equation(7) is captured. The procedure does not give a value for the ranking, although the values for Q calculated in the procedure may be used to distinguish between different vertices in a class for which $Q < 1$ and is therefore not considered anymore.

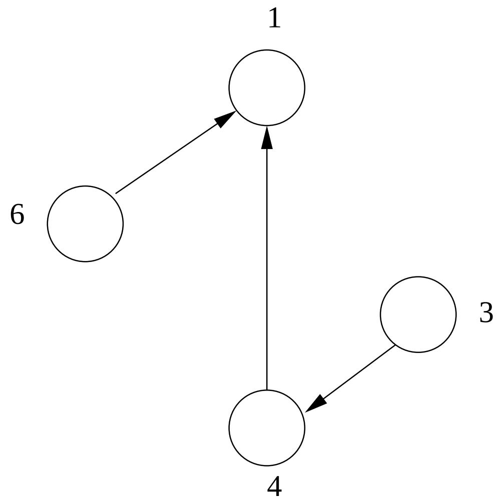

Figure 5: $\vec{G_1}$

References

[1] Brin, S. and L. Page, The Anatomy of a Large-Scale Hypertextual Web Search Engine, *Computer Networks and ISDN Systems* **33**, 107-117, (1998).

[2] GRADAP (GRAph Definition and Analysis Package), Iec ProGAMMA, Groningen, ISBN 90-5144-004-9, (1989).

[3] Hoede, C., *A new status score for actors in a network*, Memorandum nr.243, Department of Applied Mathematics, University of Twente, ISSN 0169-2690, (1979).

[4] Hubbell, C.H., An Input-Output approach to clique identification, *Sociometry*, Vol.28, no.4, 377-399, (1965).

[5] Katz, L., A new status score derived from sociometric analysis, *Psychometrika*, Vol.18, no.1, 39-43, (1953).

[6] Litvak, N., W.R.W. Scheinhardt and Y. Volkovich, *In-degree and Page Rank of Web pages: Why do they follow similar power laws*, Memorandum nr.1807, Department of Applied Mathematics, University of Twente, ISSN 1874-4850, (2006).

[7] Volkovich, Y., N. Litvak and D. Donato, *Determining factors behind the Page Rank log-log plot*, Memorandum nr.1823, Department of Applied Mathematics, University of Twente, ISSN 1874-4850, (2007).