ISSN: 2037-6367

MatematicaMente

Pubblicazione mensile della sezione veronese della MATHESIS – Società Italiana di Scienze Matematiche e Fisiche – Fondata nel 1895 – Autorizzazione del Tribunale di Verona n. 1360 del 15 – 03 – 1999 – I diritti d'autore sono riservati. Direttore: Luciano Corso – Redazione: Alberto Burato, Fabrizio Giugni, Michele Picotti, Sisto Baldo, Andrea Sellaroli, Bruno Stecca – Via IV Novembre, 11/b – 37126 Verona – tel. 338 6416432 – e-mail: lcorso@iol.it – info@mathesisverona.it – Stampa in proprio – Numero 241 – Pubblicato il 02 - 10 - 2018

How Google Ranks Webpages

Len Bos [*]

Search Engines need to rank webpages by their "importance" in order to know which ones to display first. Google does this by a very specific algorithm based on some basic Linear Algebra. In order to see what's involved let's consider a very simple example. Suppose that we have an "Internet'' consisting of exactly three pages, P_1, P_2 and P_3. Each of these will (in principle) have links to the other pages (and even possibly to itself). We can count these outgoing links and calculate their percentages. For example, suppose that P_1 has 0 links to itself, 5 links to P_2 and 5 links to P_3, i.e., 0% to itself, 50% to P_2 and 50% to P_3. These can be collected into a (column) vector decribing the outgoing links for the first page:

$$\begin{bmatrix}0.00\\0.50\\0.50\end{bmatrix}.$$

If we do the same for the other two pages we form the so-called Link matrix. Continuing with our little example, we could end up with a matrix such as

$$L=\begin{pmatrix}0.00 & 0.25 & 0.50\\ 0.50 & 0.00 & 0.50\\ 0.50 & 0.75 & 0.00\end{pmatrix}. \qquad (1)$$

Here the jth column gives the percentages of outgoing links from page P_j to the other pages P_i. In general, $L_{i,j}$ is the percentage of all outgoing links from page P_j that refer to page P_i. Further, row i tells us what percentages of the links from all the other pages go to page P_i. Notice that, by definition, the sum of each column must be exactly 1! This will turn out to be an important property of the Link matrix.

Now, what do we do with this Link matrix to get a ranking? The Google idea is as follows. Suppose that x_i denotes the "importance'' of page P_i. First look at the first page. It receives 25% of the outgoing links from P_2 and 50% of those from P_3. Hence its "importance'' should be 25% of P_2's importance plus 50% of P_3's importance, i.e., $x_1=0.25\times x_2+0.50\times x_3$.
Similarly, $x_2=0.50\times x_1+0.50\times x_3$ and $x_3=0.50\times x_1+0.75\times x_3$. In terms of matrices this is exactly the same as

$$Lx=x$$

where

$$x=\begin{bmatrix}x_1\\x_2\\x_3\end{bmatrix}$$

is the "importance'' vector.
Equivalently, the importance vector is given by the equation $(L-I)x=0$ (the zero vector). With a little bit of old-fashioned Gaussian elimination, we can solve this system to find that

$$x=t\begin{bmatrix}5\\6\\7\end{bmatrix},\ \text{for any } t\in\mathbb{R}.$$

There are several things to notice. First of all, there is a solution! This is not automatic as the system $(L-I)x=0$ is homogeneous and so for there to be a non-trivial solution it is necessary that $\det(L-I)=0$. This is indeed the case (as is easily verified from (1)), and, as we will see, also true in general. Secondly, if we normalize so that, say, $x_1+x_2+x_3=1$, then the solution *is* unique. Doing this we get

$$x=\begin{bmatrix}5/18\\6/18\\7/18\end{bmatrix}$$

and, in particular, each importance $x_i\geq 0$. We will see that this also holds more generally. Now, to complete the ranking of these three pages we just compare their importances. Obviously, P_3 is more "important'' than P_2 which is more "important'' than P_1.

I hope the idea is clear! Let's look now at the general case. We have say n webpages (which could be a very large number!) P_i, $1\leq i\leq n$. The Link matrix $L\in\mathbb{R}^{n\times n}$ has entries $L_{i,j}$ = the percentage of outgoing links from page P_j that go to page P_i. The column sums of L, $\sum_{i=1}^{n}L_{i,j}=1$, $1\leq j\leq n$, i.e., the columns all sum to 1. Alternatively, we may express this condition as

$$\mathbb{1}^tL=\mathbb{1}^t \qquad (2)$$

where $\mathbb{1}\in\mathbb{R}^n$ is the (column) vector of all ones.
We let $x\in\mathbb{R}^n$ be the vector of the importances of the *n* pages. The ranking equation is

$$Lx=x \iff (L-I)x=0. \qquad (3)$$

In other words, the importance vector $x\in\mathbb{R}^n$ is a *fixed point* of the Link matrix (it is also an eigenvector for eigenvalue $\lambda=1$, but we won't really make use of this fact).

We should first remark that there are situations where it is not possible to give a coherent ranking. For example, if the internet broke down into two polarized "cliques" A and B where the members of A referred only to other members of A and *never* to any of B, and vice versa, then we would have two groups that don't interact and so we could only hope to rank the members of A only relative to each other and similarly for B – there is no way to compare across groups. This is only common sense! Mathematically, it means that (after re-indexing the pages) we would have a Link matrix of the form

$$L=\begin{pmatrix}A & 0\\ 0 & B\end{pmatrix}$$

where A and B are two square matrices and the 0's are rectangular matrices of all zeros. If u and v were two vectors such that $Au=u$ and $Bv=v$ then $x=\begin{bmatrix}au\\bv\end{bmatrix}$ for *any* scalars *a* and *b* would be such that $Lx=x$. In particular, the solution would *not* be unique (even after normalization), nor would it have to be positive! We have to avoid this break-down possibility! The easiest (although certainly not the only) thing to do is to ask that each entry $L_{i,j}>0$ which can be accomplished by artificially changing each zero element into an arbitrary small positive value. Under this assumption we will be able to prove that our little example above actually does illustrate what happens in general! Of course, as Linear Algebra is involved we will have to use some of its basics. Indeed it is possible just to appeal to some standard theorems available in the literature, but it is also possible to show what we need to show in a completely elementary manner, which is what I will attempt to now do.

1. The ranking equation (3) always has some non-zero solution $x \in \mathbb{R}^n$.

To see this we make use of a very basic fact about matrices. Indeed, for every matrix $A \in \mathbb{R}^{n\times n}$ there are two important subspaces: (a) its *kernel*, i.e.,

$$\ker(A) = \{x \in \mathbb{R}^n : Ax = 0\}$$

and (b) its *image*, i.e.,

$$\mathrm{Im}(A) = \{y \in \mathbb{R}^n : y = Ax, \text{for an } x \in \mathbb{R}^n\}.$$

Then

$$\dim(\ker(A)) + \dim\big(\mathrm{Im}(A)\big) = n. \tag{4}$$

(This is actually not hard to prove! Just start with a basis for ker(A) and extend it to a basis for all of $\mathbb{R}^n$. It is easy to check that the extension is a basis for Im(A). Details can be found in any text on Linear Algebra.) Now assuming (4), property (2) of a Link matrix means that for any $x \in \mathbb{R}^n$, we have

$$\mathbb{1}^t(Lx) = (\mathbb{1}^t L)x = \mathbb{1}^t x$$

and so

$$\mathbb{1}^t(L-I)x = \mathbb{1}^t x - \mathbb{1}^t x = 0,$$

i.e., for $y = (L-I)x \in \mathrm{Im}(L-I)$, $\mathbb{1}^t y = 0$. Hence, as $\mathbb{1} \in \mathbb{R}^n$, *cannot* be the case that $\mathrm{Im}(L-I) = \mathbb{R}^n$, i.e., $\dim\big(\mathrm{Im}(L-I)\big) \le n-1$ and consequently, by (4), $\dim\big(\ker(L-I)\big) \ge 1$. In other words, the kernel of $L-I$ has to be more than just $\{0\}$; there has to be a non-zero $x \in \ker(L-I)$, i.e., a non-zero $x \in \mathbb{R}^n$ such that $(L-I)x = 0 \iff Lx = x$, as we claimed.

2. If $L_{i,j} > 0$ for all $1 \le i,j \le n$, and $x \in \mathbb{R}^n$ is a non-zero importance vector, i.e., $Lx = x$, then either $x_i > 0, 1 \le i \le n$ or else $x_i < 0, 1 \le i \le n$. (Consequently, if x is normalized so that $\sum_{i=1}^n x_i = 1$ then it will have *strictly positive* entries!)

To see this let $P := \{i : x_i > 0\}$ the set of indices where the entry in the importance vector is strictly positive and similarly let $N := \{i : x_i < 0\}$ be the strictly negative indices. We claim that either $P = \{1,2,\dots,n\}$ or else $N = \{1,2,\dots,n\}$. Indeed, if P is *neither* the empty set *nor* all of $\{1,2,\dots,n\}$ then from $Lx = x$ we have

$$x_i = \sum_{j=1}^n L_{i,j}\, x_j \le \sum_{j\in P} L_{i,j}\, x_j, 1 \le i \le n$$

as taking the sum only over P leaves out some non-positive contributions to the sum. Hence,

$$\sum_{i\in P} x_i \le \sum_{i\in P}\left\{\sum_{j\in P} L_{i,j}x_j\right\} = \sum_{j\in P} x_j\left\{\sum_{i\in P} L_{i,j}\right\}.$$

But, as the columns of L all sum to 1 (by (2)) and P is not the set of all indices (by assumption), we must have $\sum_{i\in P} L_{i,j} < 1$ so that $\sum_{i\in P} x_i < \sum_{j\in P} x_j$, a contradiction. In other words, if P is *not* the empty set it must be all of $\{1,2,\dots,n\}$, i.e., all the components $x_i > 0$.

By a similar argument we can show that if N is not the empty set it must be all of $\{1,2,\dots,n\}$, i.e., all the components $x_i < 0$.

3. If $L_{i,j} > 0$ for all $1 \le i,j \le n$, then, up to normalization $\sum_{i=1}^n x_i = 1$, the importance vector x is unique.

To see this, suppose that there are *two* vectors $x, y \in \mathbb{R}^n$ such that $Lx = x$, $Ly = y$ with $\sum_{i=1}^n x_i = 1$ and $\sum_{i=1}^n y_i = 1$. We will show that in fact, $x = y$. Indeed, consider $z := x - y$. Then

$$Lz = L(x-y) = Lx - Ly = x - y = z,$$

i.e., z is also an importance vector, but with

$$\sum_{i=1}^n z_i = \sum_{i=1}^n (x_i - y_i) = 1 - 1 = 0.$$

Now, by the preceeding, either $z_i > 0$, $1 \le i \le n$, or else $z_i < 0$, $1 \le i \le n$. In either case it is not possible that $\sum_{i=1}^n z_i = 0$, a contradiction.

We have shown that the importance vector always exists, is unique (after normalization) and then has strictly positive entries. But how can we calculate it? The number of pages on the Internet is currently well over 1.5 billion and doing Gaussian elimination on such an excessively large matrix is prohibitive. But there is a faster way – we can do what's called a Fixed Point Iteration. This consists of starting with some initial approximation $x^{(0)} \in \mathbb{R}^n$, normalized so that $\sum_{i=1}^n x_i^{(0)} = 1$, and then iterating

$$x^{(k+1)} = Lx^{(k)}, k = 0,1,2,\dots.$$

We will show that this will always converge to the importance vector (and in practice, quite quickly provided you start with a reasonably good approximation!).

First note that the iterates remain normalized, i.e.,

$$\mathbb{1}^t x^{(k+1)} = \mathbb{1}^t L x^{(k)} = (\mathbb{1}^t L)x^{(k)} = \mathbb{1}^t x^{(k)}$$

so that the sum of the components doesn't change. Next, looking at the error vectors $e^{(k)} := x^{(k)} - x$, we may calculate $e^{(k+1)} = x^{(k+1)} - x = Lx^{(k)} - x = L\big(x^{(k)} - x\big) = Le^{(k)}$, so the error at stage $k+1$ is just L times the preceeding error. We also note that

$$\mathbb{1}^t e^{(k)} = \mathbb{1}^t\big(x^{(k)} - x\big) = 1 - 1 = 0$$

which means that the error vectors are always in the subspace of vectors *orthogonal* to $\mathbb{1} \in \mathbb{R}^n$. Or, in other words, the sum of their components is always 0.

4. The Link matrix L is a *contraction*, i.e., it reduces the size of the error (measured using the right norm).

Indeed, the appropriate vector norm is the so-called 1-norm, defined as

$$||z||_1 := \sum_{i=1}^n |z_i|, \ \ z \in \mathbb{R}^n.$$

Then,

$$||Lz||_1 = \textstyle\sum_{i=1}^n |(Lz)_i| = \sum_{i=1}^n |\sum_{j=1}^n L_{i,j}z_j\,|$$

$$\le \textstyle\sum_{i=1}^n \sum_{j=1}^n L_{i,j}|z_j\,| = \sum_{j=1}^n |z_j\,|\{\sum_{i=1}^n L_{i,j}\}$$

$$= \textstyle\sum_{j=1}^n |z_j| = ||z||_1 \tag{5}$$

as the column sums of L are all equal to 1. In particular

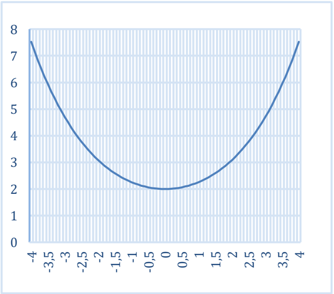
$$\left\|e^{(k+1)}\right\|_1 = \left\|Le^{(k)}\right\|_1 \le \left\|e^{(k)}\right\|_1, \tag{6}$$

i.e., the error at least doesn't increase! But more is true. Notice that in the calculations we did in (5), we can only have equality if the components of the vector z are either all positive or all negative. But, as remarked above, the sum of the components of the error vector is always 0 and hence in (6) we must have *strict* inequality. Actually, with a little bit more work we could show that there is some factor $R < 1$ such that

$$\left\|e^{(k+1)}\right\|_1 \le R \times \left\|e^{(k)}\right\|_1 \text{ so that } \left\|e^{(k)}\right\|_1 \le R^k \times \left\|e^{(0)}\right\|_1$$

which tends to 0! It converges!

[*] Professore Ordinario di Analisi Numerica, Università degli Studi di Verona. E-mail: leonardpeter.bos@univr.it



Catenaria

Il grafico è stato realizzato con Excel

$$y = a \cdot cosh\left(\frac{x}{a}\right)$$

$$a = 2$$

$$-4 \le x \le +4\,, x \in \mathbb{R}$$

$$\Delta x = 0.1$$