Searching the web with eigenvectors

Herbert S. Wilf
University of Pennsylvania, Philadelphia, PA 19104-6395

April 13, 2001

How might we measure the importance of a web site? Well, it's important if other important web sites link to it (if that sounds circular, just hold on for a moment). Suppose $x_1, x_2, \ldots, x_n$ are the importances of all n of the web sites in the world. We want *the importance of each web site to be proportional to the sum of the importances of all of the sites that link to it.* That's a system of equations that might look like this-

$$\begin{aligned} x_1 &= K(x_{14} + x_{97} + x_{541}) \\ x_2 &= K(x_{1003} + x_{3224} + x_{9773} + x_{10029}) \\ \ldots &= \ldots, \end{aligned}$$

in which K is the constant of proportionality, and on the right side of each equation is the sum of the importances of all of the sites that link to the one on the left.

Now imagine a gigantic $n \times n$ matrix A whose (i, j) entry is 1 if web site j links to web site i, and is otherwise 0. Then we can rewrite the above equations as

$$x_i = K \sum_{j=1}^{n} a_{i,j} x_j \qquad (i = 1, \ldots, n) \tag{1}$$

which, remarkably enough, is an eigenvalue and eigenvector problem! It says that the vector of importances that we are looking for is an eigenvector of the gigantic matrix A. To find it, we might, for instance, find all of the eigenvectors of A, and use the one that has all of it components positive. Once you have the eigenvector, the most important web site is the one with the largest entry in that eigenvector, the next most important has the second largest entry, and so forth.

The web search engine Google (`www.google.com`) uses a variant of this idea to find the importances of a large number of web sites (the inventors of Google ave described their methods in considerable detail [2]). Then when you request a search, you get your list of hits in decreasing order of importance, which might save you a lot of time in finding the one you want. The idea of using the eigenvector to do ranking is due to Kendall and Wei [3, 4], in the 1950's, and the method has acquired considerable currency today because of web applications.

Think about all of the college football teams in the USA. How could we find "the top ten"? Well, if x_i is the strength of the ith football team, let's say that x_i is proportional to the sum of the strengths of all of the teams that the ith team defeated. Then we have the equations (1) again, where now the matrix entry $a_{i,j}$ is 1 if team i defeated team j and is 0 otherwise. So you would find its eigenvectors and look for the one in which all of the entries have the same sign. The largest entry would tell you the team of first rank, the next largest entry would belong to the team of second rank, etc.

The idea has lots of applications to ranking things or people in order of importance based on some measure of the influence that they have over each other. Notice that we *might* have decided that the importance of a web site is proportional to the *number* of web sites that link to it, which would have saved us the big eigenvector computation. But this isn't nearly as smart a thing to do because some web sites are linked to by only a few other sites, but important ones, while other web sites are linked to by many other sites, but unimportant ones. The eigenvector method has a higher IQ.

Here's a small numerical example. Let

$$A = \begin{pmatrix} 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 1 & 0 \end{pmatrix}.$$

This represents six football teams, where team 1 defeated teams 2 and 5, team 2 defeated teams and 5, etc. The eigenvector of this matrix whose entries are all positive is $(.31, .31, .22, .57, .50, .43)$. The best team, according to this ranking system, is team 4, followed by teams 5,6,1,2,3. Notice that teams 4 and 5 each defeated four other teams, but the method prefers team 4 to team 5, because it considers that team 4 defeated better teams than team 5 did.

References

[1] Claude Berge, The theory of graphs and its applications, Wiley, New York, 1962.

[2] Sergey Brin and Lawrence Page, The anatomy of a large-scale hypertextual web search engine, Computer Networks and ISDN Systems, **30** (1998), 107-117. [This paper is available at numerous sites on the web. Use Google to search for it!- HW]

[3] M.G. Kendall, Further contributions to the theory of paired comparisons, Biometrics **11** (1955), p. 43.

[4] T.H. Wei, The algebraic foundations of ranking theory, Cambridge University Press, London, 1952.