

Pseudocode for calculating EigenfactorTM Score and Article InfluenceTM Score using data from Thomson-Reuters Journal Citations Reports

Jevin West and Carl T. Bergstrom*

November 25, 2008

1 Overview

There are seven steps for calculating journal-level EigenfactorTM Metrics using data from the Thomson-Reuters Journal Citation Reports (JCR) database:

1. Data Input
2. Creating an Adjacency Matrix
3. Modifying the Adjacency Matrix
4. Identifying the Dangling Nodes
5. Calculating the Stationary Vector
6. Calculating EigenfactorTM Score and Article InfluenceTM Score

*Both authors are at the Department of Biology, University of Washington, Seattle WA 98115. Questions: email Jevin West at jevinw@u.washington.edu.

7. Outputting the Results

Like Thomson’s Impact Factor metric, the Eigenfactor Metrics measure the number of times that articles published during a *census period* provide citations to papers published during an earlier *target window*. The Impact Factor as reported by Thomson Scientific has a one year census period and uses the two previous years for the target window. In its current form, the Eigenfactor Metrics use a one year census period and uses the five previous years for the target window.

1.1 Data Input

Four inputs — two files and two constants — are needed:

- Journal File: using the CD-ROM version of the Thomson-Reuters (CR) database, create a list of unique journals included in the Science and Social Science JCR.¹ This list should contain journals from the Sciences and the Social Sciences. For the Eigenfactor Metrics we combine these two lists instead of treating them as two separate lists. Then list how often each journal cites each other journal, where we count citations that are given during census period (e.g. 2006) to papers published during the target window (e.g. 2001–2005).
- Article File: this is the file that contains the number of articles that each journal produces in the five previous years. On each JCR CD-ROM, Thomson lists article counts for the two previous years; this means that we have two sources of article information for each year:

¹For 2006, there were 7611 unique journals for the combined Science and Social Science combined list in the JCR. We refer to these JCR-listed journals as “ISI journals”.

one from the following year's JCR and one from the JCR two years later. These numbers do not exactly agree because Thomson tends to refine their counts over time. The rule for which to use is simple: take article counts from the most current CD on which those counts are available. For example, to compute a 5 year article count for 2007, Take the 2006 and 2005 article counts from 2007 CD, the 2004 counts from the 2006 CD, the 2003 counts from the 2005 CD, and the 2002 counts from the 2004 CD.

- Alpha constant ($\alpha = 0.85$)
- Epsilon constant ($\epsilon = 0.00001$)

1.2 Creating an Adjacency Matrix

The journal citation network can be conveniently represented as an adjacency matrix \mathbf{Z} , where the \mathbf{Z}_{ij} -th entry indicates the number of times that articles published in journal j during the census period cite articles in journal i published during the target window. The dimension of this square matrix is $n \times n$ where n is the number of unique ISI journals. For example, suppose there are journals A , B , and C .

	A	B	C
A	2	0	3
B	4	1	1
C	0	2	7

In the adjacency matrix above, journal A cites itself 2 times, it cites journal

B 4 times, and it doesn't cite journal C at all. Journal B receives 4 citations from journal A , 1 citation from itself, and 1 citation from journal C .

1.3 Modifying the Adjacency Matrix

There are some modifications that need to be done to \mathbf{Z} before the eigenvectors can be calculated.

- First, we set the diagonal of \mathbf{Z} to zero (i.e., we set all of the entries $Z_{ii} = 0$). This is done so that journals do not receive credit for self-citations.
- Second, we normalize the columns of the matrix \mathbf{Z} (i.e., divide each entry in a column by the sum of that column). To do this, compute the column sums for each column j as $Z_j = \sum_i \mathbf{Z}_{ij}$. Then divide the entries from each column by the corresponding column sum to get the entries of the \mathbf{H} matrix: $\mathbf{H}_{ij} = \mathbf{Z}_{ij}/Z_j$. There may be columns that sum up to zero (i.e., journals that cite no other journals). These are dangling nodes, and we will deal with them in the next section.

In the example below, we take an adjacency matrix through these two modifications. The matrix you get after these two modifications is \mathbf{H} . This example matrix will be used throughout the pseudocode as an example of how to calculate the Eigenfactor Score of a journal. The numbers in parentheses next to each journal letter represent the number of papers that each journal has published.

Example raw adjacency matrix (\mathbf{Z})

	<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>	<i>E</i>	<i>F</i>
<i>A</i> (3)	1	0	2	0	4	3
<i>B</i> (2)	3	0	1	1	0	0
<i>C</i> (5)	2	0	4	0	1	0
<i>D</i> (1)	0	0	1	0	0	1
<i>E</i> (2)	8	0	3	0	5	2
<i>F</i> (1)	0	0	0	0	0	0

1. Set the diagonal to zero

↓

	<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>	<i>E</i>	<i>F</i>
<i>A</i> (3)	0	0	2	0	4	3
<i>B</i> (2)	3	0	1	1	0	0
<i>C</i> (5)	2	0	0	0	1	0
<i>D</i> (1)	0	0	1	0	0	1
<i>E</i> (2)	8	0	3	0	0	2
<i>F</i> (1)	0	0	0	0	0	0

2. Normalize the columns. This matrix is H.

↓

	<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>	<i>E</i>	<i>F</i>
<i>A</i> (3)	0	0	2/7	0	4/5	3/6
<i>B</i> (2)	3/13	0	1/7	1	0	0
<i>C</i> (5)	2/13	0	0	0	1/5	0
<i>D</i> (1)	0	0	1/7	0	0	1/6
<i>E</i> (2)	8/13	0	3/7	0	0	2/6
<i>F</i> (1)	0	0	0	0	0	0

1.4 Identifying the Dangling Nodes

As mentioned in the previous section, there will be journals that don't cite any other journals. These journals are called dangling nodes and can be identified by looking for columns that contain all zeros. These columns need to be identified with a row vector of 1's and 0's. Call this vector d . The 1's indicate that a journal is a dangling node; the 0's indicate a non-dangling node. For the example above, d would be the following row vector:

	<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>	<i>E</i>	<i>F</i>
d_i	0	1	0	0	0	0

1.5 Calculating the Influence Vector

The next step is to construct a transition matrix \mathbf{P} and compute its leading eigenvector. This eigenvector, normalized so that its components sum to 1, will be called the influence vector π^* . This vector gives us the journal weights that we will use in assigning eigenfactor scores.

To calculate the influence vector π^* , we need six inputs: the matrix \mathbf{H} that we just created, an initial start vector $\pi^{(0)}$, the constants α and ϵ , the dangling node vector d and the article vector a .

Article Vector. Let A_{tot} be the total number of articles published by all of the journals. The article vector a is a column vector of the number of articles published in each journal over the (five-year) target window, normalized so that its entries sum to 1. (To do this normalization, divide the number of articles that each journal publishes by A_{tot}). Using the example from above, $A_{\text{tot}} = 3 + 5 + 2 + 1 + 2 + 1 = 14$ and the article vector would be

Article Vector

	a_i
A	3/14
B	2/14
C	5/14
D	1/14
E	2/14
F	1/14

Initial start vector $\pi^{(0)}$. This vector is used in iterating the influence vector. Set each entry of this column vector to $1/n$. For our example, this vector would look like

	$\pi_{\mathbf{i}}^{(0)}$
A	1/6
B	1/6
C	1/6
D	1/6
E	1/6
F	1/6

Calculating the influence vector π^* . The influence vector π^* is the leading eigenvector (normalized so that its terms sum to one) of the matrix \mathbf{P} , defined as follows:²

$$\mathbf{P} = \alpha \mathbf{H}' + (1 - \alpha) a.e^T,$$

Here e^T is a row vector of all 1's and $a.e^T$ is thus a matrix with identical columns a . The matrix \mathbf{H}' is the matrix \mathbf{H} , with all columns corresponding to dangling nodes replaced with the article vector a . In the example, \mathbf{H}' would be the following matrix (notice the replacement of the B column):

²This matrix describes a stochastic process in which a random walker moves through the scientific literature; it is analogous to the “google matrix” that Google uses to compute the PageRank scores of websites. The stochastic process can be interpreted as follows: a fraction α of the time the random walker follows citations and a fraction $1 - \alpha$ of the time the random walker “teleports” to a random journal chosen at a frequency proportional to the number of articles published.

	<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>	<i>E</i>	<i>F</i>
<i>A</i> (3)	0	3/14	2/7	0	4/5	3/6
<i>B</i> (2)	3/13	2/14	1/7	1	0	0
<i>C</i> (5)	2/13	5/14	0	0	1/5	0
<i>D</i> (1)	0	1/14	1/7	0	0	1/6
<i>E</i> (2)	8/13	2/14	3/7	0	0	2/6
<i>F</i> (1)	0	1/14	0	0	0	0

Because \mathbf{P} will be an irreducible aperiodic Markov chain by construction, it will have a unique leading eigenvector by the Perron-Frobenius theorem. We could compute the normalized leading eigenvector of the matrix P directly using the power method, but this involves repeated matrix multiplication operations on the dense matrix \mathbf{P} and thus is computationally intensive. Instead, we can use an alternative approach that involves only operations on the sparse matrix \mathbf{H} and thus is far faster³. To compute the influence vector rapidly, we will iterate the following equation

$$\pi^{(k+1)} = \alpha \mathbf{H} \pi^{(k)} + [\alpha d \cdot \pi^{(k)} + (1 - \alpha)]a$$

This iteration will converge uniquely to the leading eigenvector of \mathbf{P} , normalized so that its terms sum to 1. To find this eigenvector, iterate repeatedly. After each iteration, check to see if the residual ($\tau = \pi^{(k+1)} - \pi^{(k)}$) is less than ϵ . If it is, then $\pi^* \approx \pi^{(k+1)}$ is the influence vector. Typically, this does not take more than 100 iterations with $\epsilon = 0.00001$. Using the raw adjacency matrix example above and the corresponding article vector, the

³Notice that the equation below uses the matrix \mathbf{H} , without the dangling node columns replaced, not the matrix \mathbf{H}' . In fact, one does not need to ever construct the matrix \mathbf{H}' in the process of doing these calculations.

stationary vector converges after 16 iterations to the following vector with $\alpha = 0.85$ and $\epsilon = 0.00001$:

	π_i^*
A	0.3040
B	0.1636
C	0.1898
D	0.0466
E	0.2753
F	0.0206

1.6 Calculating Eigenfactor Score (EF_i) and Article Influence Score (AI_i)

The vector of eigenfactor values for each journal is given by the dot product of the H matrix and the influence vector π^* , normalized to sum to 1 and then multiplied by 100 to convert the values from fractions to percentages:

$$EF = 100 \frac{\mathbf{H} \cdot \pi^*}{\sum_i [\mathbf{H} \cdot \pi^*]_i}$$

The Eigenfactor Scores for our example are thus

	EF_i
A	34.0510
B	17.2037
C	12.1755
D	3.6532
E	32.9166
F	0.0000

The Article Influence Score \mathbf{AI}_i for each journal (i) is calculated using the following equation:

$$\mathbf{AI}_i = 0.01 \frac{\mathbf{EF}_i}{a_i}$$

where \mathbf{EF}_i is the Eigenfactor Score for journal i and a_i is the normalized article vector. In words, the Article Influence Score is essentially the Eigenfactor Score/100, divided by the fraction of all articles that each journal has published. The Article Influence Scores for our example are

	AI_i
A	1.5890
B	1.2043
C	0.3409
D	0.5114
E	2.3042
F	0.0000

1.7 Calculating (\mathbf{EF}_i) and (\mathbf{AI}_i) for non-ISI journals

Because the JCR lists citations from listed journals to many non-listed journals (and other reference items such as the *New York Times*), Eigenfactor Scores can be calculated for these non-ISI journals. Article Influence Scores can also be calculated for non-ISI journals if article information is available. Article information for these journals are not found in the JCR database, so this information would have to come from other sources.

To calculate non-ISI Eigenfactor Scores, first retrieve the matrix \mathbf{Z} . Zero the diagonals and then find the sum of each column. Second, construct a matrix \mathbf{N} that contains the number of citations from the ISI journals. The matrix below illustrates what it would look like when these two matrices are sewed together. The journal R , S and T are non-ISI journals of the matrix \mathbf{N} . As you can see, the non-ISI journals receive citations from ISI journals, but since they are not listed in the JCR, we do not have a tally of the citations that they give to ISI journals $A-F$.

	A	B	C	D	E	F
$A(3)$	0	0	2	0	4	3
$B(2)$	3	0	1	1	0	0
$C(5)$	2	0	0	0	1	0
$D(1)$	0	0	1	0	0	1
$E(2)$	8	0	3	0	0	2
$F(1)$	0	0	0	0	0	0
$R(\text{n/a})$	3	0	0	0	0	2
$S(2)$	0	0	1	0	0	0
$T(\text{n/a})$	0	0	1	0	1	0

In the example above, the ISI journal, A , cites the non-ISI journal, R , 3 times. The ISI journal, E , cites the non-ISI journal, T , 1 time. The numbers in parentheses again indicate the number of articles that each non-ISI journal produced in the five year target window. In this example, we have data only for journal S ; we do not know how many articles were published by R or T .

Now, divide each number in \mathbf{N} by the corresponding column sum in the \mathbf{Z} matrix.⁴ . This new matrix \mathbf{N}' would look like

	A	B	C	D	E	F
$R(n/a)$	3/13	0	0	0	0	2/6
$S(2)$	0	0	1/7	0	0	0
$T(n/a)$	0	0	1/7	0	1/5	0

The Eigenfactor Score for each non-ISI journal in this \mathbf{N}' matrix is the product of that row vector times the influence vector π^* for the ISI journals times 100. In vector notation, the vector of Eigenfactor Scores is simply $100 \mathbf{N}' \cdot \pi^*$. For example, the row vector for journal R is $\{3/13, 0, 0, 0, 0, 2/6\}$ and the influence vector is the column vector that we calculated before, $\pi^* = \{0.3040, 0.1636, 0.1898, 0.0466, 0.02753, 0.0206\}$. Thus the extended eigenfactor for journal B is the product of these vectors: $EF_{(R)} = 100 \left(\frac{3}{13} \times 0.3040 + \frac{2}{6} \times 0.0206 \right)$

Thus calculated, the Eigenfactor Scores for the non-ISI journals are

⁴Recall that the j -th column sum of this \mathbf{Z} matrix indicates how many citations are given out by that journal to all ISI-listed journals excluding itself. We use this — rather than the column sum of the extended matrix formed by appending \mathbf{N} to \mathbf{Z} — because this was the denominator we used in computing Eigenfactor Scores for ISI-listed journals in Section 1.5. We want to make the Eigenfactor Scores of for the non-ISI journals directly comparable, so we use the same denominator here.

	EF_i
R	7.7041
S	2.7114
T	8.2176

Article Influence Scores can be calculated for the non-ISI journals so long as we have article counts. If we don't have the article count for a non-ISI journal, its Article Influence Score is listed as *NA*. To calculate Article Influence for non-ISI journals, use the same equation used for the ISI journals

$$\mathbf{AI}_i = 0.01 \frac{\mathbf{EF}_i}{a_i}$$

Here a_i represents the entries in an extended version of the article vector computed in step 1.5. The denominator for the a_i 's should be the total number of articles A_{tot} published by ISI-listed journals, not the total number of articles published by all journals, ISI-listed or otherwise.⁵ Thus in our example the a_i value for journal S should be 2/14, not 2/16. Thus calculated, the Article Influence Scores for the non-ISI journals are

	AI_i
R	NA
S	0.1898
T	NA

⁵Again, we want our Article Influence Scores for non-ISI journals to be directly comparable to those for ISI journals, so we have to use the same denominator in our calculations.

1.8 Outputting the Results

To get the journal rankings, just sort in descending order the **EF** and **AI** vectors. Output the results in whatever format is easiest to compare rankings. Right now, we are using Excel. The following is what we include in our data output:

- Year
- Short Name
- Long Name
- Group (Science or Social Science)
- Field (e.g., Physics)
- Eigenfactor Score
- Article Influence Score
- Impact Factor
- Total Articles (5 yrs)
- Total Citations Received (5 yrs)