РАНЖИРОВАНИЕ КОЛЛЕКЦИИ ПЕРИОДИЧЕСКИХ ИЗДАНИЙ БАЗЫ ДАННЫХ *RePEc* НА ОСНОВЕ МЕТРИК *Eigenfactors*

С. В. Бредихин, В. М. Ляпунов, Н. Г. Щербакова

Институт вычислительной математики и математической геофизики СО РАН,
630090, Новосибирск, Россия

УДК 001.12+303.2

Представлен обзор метрик *Eigenfactor* и *Article Influence.* Основным содержанием статьи являются описание процедур извлечения данных из БД *RePEc*, формирование матрицы цитирования и ее модификация; обоснование методики вычисления метрики *Eigenfactor*, определяющей степень влияния одного элемента коллекции на другие; ранжирование коллекции периодических изданий БД *RePEc* на основе метрики *Article Influence.*

Ключевые слова: матрица цитирования, матричные вычисления, ранжирование, метрика импакт-фактор, метрика *Eigenfactor,* метрика *Article Influence,* алгоритм *PageRank.*

The review of metrics *Eigenfactor* and *Article Influence* is submitted. The main contents of the article is the description of procedures of data extraction from the database *RePEc*, the cross citation matrix creation and modification; justification of the technique of calculation of *Eigenfactor* metric, that provides extent of the total influence of one element of the collection on others; *RePEc* journal ranking based on *Article Influence* metric.

Keywords: citation matrix, array computation, impact factor metric, eigenfactor metric, article influence metric, algorithm PageRank.

Введение. Задача оценки научных журналов, авторов статей, институтов, высших учебных заведений и научных школ на основе подсчета цитирований была сформулирована в работе [1]. Тридцать лет спустя появилась и приобрела популярность метрика импакт-фактор (Impact Factor, далее — *IF*), которая измеряет среднее количество цитирований, полученных статьями журнала за некоторый период времени после публикации [2]. Достоинством этой метрики является простота определения и расчета. Однако *IF* не учитывает престиж цитирующего журнала, т. е. цитирование, полученное от реферативной публикации, будет цениться так же, как цитирование, полученное от исследовательской статьи, содержащей ссылки только на те работы, которые соотносятся с содержанием исходной работы [3]. Также *IF* не учитывает стиль цитирования, присущий различным научным дисциплинам [4]. Например, цитирования в области математики, где, как правило, библиографии коротки и цитирования недавно опубликованных работ редки, оцениваются не выше, чем цитирования в области иммунологии, где приняты длинные библиографии, и цитирование недавно вышедших работ является обычным явлением. Метрика *Eigenfactor* (далее — *EF*) [5] была разработана в дополнение к *IF* и другим мерам, основанным на арифметике “сырых” счетчиков цитирования, с целью устранения указанных недостатков.

Акцентируем внимание на двух фактах. Во-первых, метрики *IF* и *EF* зависят от временного интервала сбора цитирований, т. е. учитывают цитирования, полученные журналами,

вышедшими в течение предыдущих двух (пяти) лет, что отражает особенности цитирования во многих исследовательских областях [6]. Во-вторых, метрика *EF* измеряет влияние журнала, представляя его в виде "сборника статей", т. е. оценивает, насколько часто "этот сборник" используется исследователями. Иными словами, *EF* предоставляет агрегированную оценку, образованную вкладами всех статей, опубликованных в журнале за пятилетний период. Отсюда следует, что число статей в журнале влияет на значение *EF*, поскольку журналы с большим числом статей имеют больше цитирований, следовательно, получают больший "вес", что приводит к увеличению значений *EF*.

Метрика *Article Influence* (далее — *AI*) отражает "престижность" журнала [5], ее значение пропорционально значению *EF*, деленному на количество публикаций в журнале. Данная метрика нормализуется так, что среднее значение по статьям в исследуемой базе данных составляет единицу [3]. Метрика *AI* может рассматриваться как среднее влияние индивидуальных статей, опубликованных в одном и том же журнале. Значение метрики $AI_i = m$ для журнала i означает, что средняя статья этого журнала в m раз влиятельнее, чем средняя статья в базе. Метрики *EF* и *AI* включены в отчеты по цитированию журналов *JCR* (*Journal Citation Reports*) и доступны с 2007 г.

Настоящая работа посвящена описанию процедуры вычисления значения метрик *EF* и *AI* (*Eigenfactors metrics*) для коллекции периодических изданий БД *RePEc* [7] (*Research Papers in Economics* — "Исследовательские статьи по экономике"), которая представляет собой интернет-проект, посвященный систематизации исследовательских работ в области экономики, финансов, менеджмента и маркетинга. Основой проекта является децентрализованная библиографическая база данных (далее — ДББД) по рабочим документам, статьям, книгам и главам из книг. *RePEc* не содержит полных текстов статей журнала, а предоставляет ссылки на полнотекстовый материал. Информацию поставляют различные организации и волонтеры.

Базовая модель. При разработке метрики *EF* в основу было положено наблюдение, что сами цитирования не являются изолированными событиями, они образуют сеть взаимосвязанных объектов [8]. Введение *EF* направлено на изучение этой сети с целью определения важности каждого ее узла, включая все объекты цитирования и все источники цитирования. Значения метрики *EF* можно рассматривать как результат свободного блуждания по научной литературе. Предположим, что гипотетический читатель случайным образом извлекает из некоторого архива журнал за год Y и публикации в нем. После прочтения исследователь случайным образом выбирает одну из ссылок и извлекает соответствующий журнал за год Y, вновь выбирает случайную статью и ссылку в ней. И так сколько угодно раз. Частота, с которой исследователь выбирает каждый журнал, определяет меру важности журнала в сети цитирования. Эта частота, выраженная в процентах, является значением метрики *EF* для этого журнала. На практике ожидаемая частота выбора журнала вычисляется на основании матрицы цитирования, являющейся матрицей смежности исследуемой сети, рассматриваемой как граф.

Предпосылки. Алгоритм нахождения метрик *EF* и *AI* опирается на понятие "центральности собственного вектора", определяющего важность вершин графа на основании информации о структуре сети [9]. В этой работе сеть с n вершинами представляется в виде матрицы смежности A размера n: a_{ij} — количество ребер из вершины i в вершину j. Сделаем начальное предположение о центральности x_i для вершины i. Например, установим $x_i = 1$ для любого i. Используем это равенство для определения нового значения x'_i, которое определим как сумму центральностей соседей вершины i:

$$x_i' = \sum_j a_{ij} x_j.$$

В матричном виде это можно записать как $\mathbf{x}' = A\mathbf{x}$, где $\mathbf{x}$ — вектор элементов x_i. Процедура вычисления вектора, задающего степень центральности каждой вершины графа, начинается с некоторого начального значения $\mathbf{x}(0)$. После t шагов имеем вектор центральности вершин $\mathbf{x}(t) = A^t\mathbf{x}(0)$. Предел вектора центральности пропорционален главному собственному вектору матрицы смежности. Таким образом, искомый вектор $\mathbf{x}$ удовлетворяет условиям $A\mathbf{x} = \lambda_1\mathbf{x}$, где λ_1 — максимальное собственное значение, $\mathbf{x}$ — главный собственный вектор, определяющий ранжирование (упорядочивание).

В работе [10] обобщен подход к оценке статуса объектов на основании собственных векторов и собственных значений матрицы, представляющей отношения между объектами, названный спектральным ранжированием (*spectral ranking*). Этот подход неоднократно применялся в различных формах, взаимосвязь между которыми была обнаружена благодаря изучению популярного алгоритма ссылочного ранжирования *PageRank* [11], применяемого для ранжирования веб-страниц. Суть алгоритма состоит в формировании матрицы, элемент которой a_{ij} равен единице, если со страницы i есть указатель на страницу j, и нулю в противном случае. "Важность" определяется следующим образом:

$$x_i = \alpha \sum_j a_{ij} \frac{x_j}{k_j^{out}} + \beta,$$

где α и β — константы, k_j^{out} — количество дуг, исходящих из вершины j. В терминологии [10] это называется спектральным ранжированием с затуханием (*damping*, α) и граничным условием (*boundary condition*, β). Если у вершины j нет исходящих дуг, то определим $k_j^{out} = 1$, чтобы избежать деления на ноль. Уравнение для вектора центральности можно представить в виде

$$\mathbf{x} = \alpha A D^{-1}\mathbf{x} + \beta\mathbf{I},$$

где D — диагональная матрица с элементами $d_{ii} = \max(k_i^{out}, 1)$; $\mathbf{I}$ — вектор размерности n, состоящий из единиц.

В определении *PageRank* "влияние" распространяется в равной степени по всем ссылкам со страницы, поскольку ссылки не взвешены (отсутствуют оценки их значимости). В сетях цитирования периодических изданий в узлах расположены журналы, представляющие единый контент (т. е. все статьи всех выпусков за рассматриваемый год), а связи отражают цитирования журналов друг другом. Связи взвешены и направлены, бóльшие веса соответствуют большему количеству цитирований.

Метод. Ниже представлены пошаговая процедура преобразования исходных данных и алгоритм вычисления метрик *EF* и *AI* [12].

Шаг 1. *Извлечение исходных данных.* Зафиксируем год Y, для которого вычисляется метрика *EF*. На основе данных, извлеченных из некоторой БД, построим файл журналов, содержащий информацию для каждого журнала о том, насколько часто этот журнал цитировал в рассматриваемом году другие журналы, вышедшие за предыдущие пять лет и проиндексированные в этой базе. Затем построим файл статей, содержащий информацию о количестве статей, которые опубликовал каждый журнал за предыдущие пять лет.

Шаг 2. *Построение матрицы цитирования.* Пусть $C_{ij}(Y_1, Y_2)$ обозначает, сколько раз статьи в журнале j, опубликованном в году Y_1, цитировали статьи журнала i, опубликованные в году Y_2. Построим квадратную матрицу Z размерности n, где n — количество

рассматриваемых журналов. Элемент z_{ij} равен количеству цитирований от журнала j в рассчитываемом году Y на публикации в журнале i, относящиеся к пяти предыдущим годам:

$$z_{ij} = \sum_{k=1}^{5} C_{ij}\,(Y, Y-k).$$

Получаем матрицу смежности графа цитирования.

Шаг 3. *Удаление самоцитирований* (*self-citations*). По определению метрика *EF* не учитывает цитирования между статьями одного журнала, поэтому все диагональные элементы матрицы цитирования обнуляются: $z_{ii} = 0$ для всех i.

Шаг 4. Нормализация матрицы цитирования по столбцам. Для каждого столбца вычисляется сумма элементов, затем каждый элемент столбца делится на эту сумму. Пусть $z_j = \sum\limits_i z_{ij}$. Строим новую матрицу H, элементы которой $h_{ij} = z_{ij}/z_j$. Если сумма элементов столбца равна нулю, т. е. $z_j = 0$, то столбец не модифицируется. Такой столбец соответствует журналу, который никого (возможно, кроме себя) не цитирует, а узлы сети называются "отвисшими" (*dangling*).

Шаг 5. *Построение вектора публикаций.* Пусть J_i — количество публикаций в журнале i за весь период окна цитирования, в нашем случае — за пять предыдущих лет.

Пусть $A_{tot} = \sum\limits_i J_i$. Построим вектор-столбец $\mathbf{a}$, такой что $a_i = J_i/A_{tot}$.

Шаг 6. *Удаление отвисших узлов.* В матрице H все столбцы, соответствующие отвисшим узлам, заменяем на вектор публикаций $\mathbf{a}$ и получаем стохастическую матрицу H'.

Шаг 7. *Построение матрицы "читателя".* Построим матрицу P, описывающую марковский процесс свободного блуждания "читателя", действия которого могут быть интерпретированы следующим образом. Часть времени читатель действует в соответствии с базовой моделью, перемещаясь от журнала к журналу по цитированиям. Остальную часть времени читатель действует стихийно: случайным образом выбирает журнал для рассмотрения (без учета цитирования), при этом на его выбор влияет количество публикаций в журнале. Итак, с вероятностью α он действует в соответствии с первым случаем, а с вероятностью $(1-\alpha)$ — в соответствии со вторым случаем. По аналогии с подходом, предложенным в алгоритме *PageRank* [11], определяем новую стохастическую матрицу P:

$$P = \alpha H' + (1-\alpha)A,$$

где матрица A — квадратная матрица, все столбцы которой совпадают с вектором публикаций $\mathbf{a}$ ($A = \mathbf{a}\mathbf{I}^T$). В соответствии с *PageRank*, $\alpha = 0,85$.

Шаг 8. Вычисление вектора "влияния" $\boldsymbol{\pi}$. Определим $\boldsymbol{\pi}$ [12] как главный собственный вектор матрицы P [13]. Переход от матрицы H' к матрице P позволяет воспользоваться рядом утверждений теории матриц и цепей Маркова. По построению стохастическая матрица P является неразложимой и апериодической матрицей. Согласно [14], неотрицательная неразложимая и апериодическая матрица является примитивной, а для примитивной матрицы верна теорема Перрона [13], которая устанавливает свойства спектра положительных матриц. Таким образом, матрица P имеет единственное наибольшее положительное вещественное собственное значение λ_0 (любое другое собственное значение строго меньше λ_0 по абсолютной величине), а соответствующий ей собственный вектор имеет положительные компоненты.

Простым методом вычисления собственного вектора в случае наличия строго большего по абсолютной величине собственного значения является "степенной метод", который впервые был изложен в работе [15]. Пошаговое описание этого метода можно найти, например, в монографии [16]. Главный цикл "степенного метода" выглядит так. Пусть имеется матрица A, для которой существует единственное наибольшее положительное вещественное собственное значение. Пусть $\mathbf{x}_0 = \mathbf{y}_0$ — начальный вектор. Шаг а: $\mathbf{x}_k = A\mathbf{y}_{k-1}$; шаг б: $\mathbf{y}_k = \dfrac{\mathbf{x}_k}{\|\mathbf{x}_k\|}$. Повторять шаги (а) и (б) до тех пор, пока $||\mathbf{x}_k|| - ||\mathbf{x}_{k-1}|| > \varepsilon$, где ε — фактор остановки (здесь $\varepsilon = 0{,}00001$), а $||\mathbf{x}||$ — норма вектора $\mathbf{x}$. Доказательство сходимости процесса к собственному вектору, соответствующему доминирующему собственному значению, можно найти в работе [17].

В рамках стохастических процессов главный собственный вектор $\boldsymbol{\pi}$ исследуемой матрицы P соответствует вектору устойчивых вероятностей [17], а каждый элемент вектора соответствует стабильной части времени, потраченного на каждый журнал, представленный в P.

Шаг 9. Вычисление метрик EF и AI. Для вычисления указанных метрик используется вектор $\boldsymbol{\pi}$, задающий веса (влияние) журналов. Вектор EF получается путем скалярного умножения матрицы H на вектор $\boldsymbol{\pi}$, нормирования до суммы 1 и умножения на 100, чтобы получить значение в процентах:

$$EF = 100\frac{H\boldsymbol{\pi}}{\sum\limits_i [H\boldsymbol{\pi}]_i}.$$

Таким образом, элемент EF_i задает количество очков, полученных журналом i. Теперь вычислим значения AI для журнала i по следующей формуле:

$$AI_i = 0{,}01EF_i \;/\; a_i.$$

Результат. Основным результатом работы является вычисление метрик EF и AI, представленных всеми коллекциями ДББД *RePEc*. На дату извлечения данных (ноябрь 2013 г.) в ДББД *RePEc* содержалась информация о 551 990 документах, 13 750 948 ссылках и 5 260 242 цитированиях (определения понятий "ссылка" и "цитирование" см. в работе [18]). На сайте [19] приводится информация для журналов, имеющих значение импакт-фактора больше нуля, содержащих более 5 цитирований, опубликовавших более 20 статей в год вычисления метрик, и процент самоцитирований в которых меньше или равен 50 %. Элементы этой коллекции на сайте [19] упорядочены по метрике *IF*. Далее в таблице приведены результаты ранжирования для первых двадцати журналов согласно метрике AI за 2012 г.

Заключение. В работе представлены метрики, основанные на подсчете цитирований и учитывающие "важность" журналов, от которых они получены. Сравнение метрик AI и *IF* не проводилось, поскольку в работе [12] в ответ на [20] показано, что нужно с осторожностью делать заключения, основываясь на величине коэффициента корреляции, особенно вне формальных определений тестирующих гипотез.

Таблица

Метрика *AI* журналов ДББД *RePEc*

Ранг	Название журнала и издательство	*AI*	*EF*
1	Journal of Economic Literature / American Economic Association	27,290	0,61639
2	The Quarterly Journal of Economics / Oxford University Press	26,773	1,35281
3	Econometrica / Econometric Society	22,828	1,49900
4	Journal of Economic Perspectives / American Economic Association	14,772	0,84271
5	Journal of Political Economy / University of Chicago Press	14,672	0,60127
6	Review of Economic Studies / Oxford University Press	13,348	0,77392
7	Journal of Finance / American Finance Association	12,774	1,13330
8	IMF Economic Review / Palgrave Macmillan	12,142	0,40429
9	American Economic Review / American Economic Association	11,886	3,15502
10	American Economic Journal: Macroeconomics / American Economic Association	10,778	0,22588
11	Journal of International Economics / Elsevier	10,605	0,95075
12	Review of Economic Dynamics / Elsevier for the Society for Economic Dynamic	10,596	0,49838
13	Journal of Labor Economics / University of Chicago Press	10,558	0,28026
14	Economic Journal / Royal Economic Society	10,310	1,12835
15	Journal of Development Economics / Elsevier	10,278	0,94058
16	The Review of Economics and Statistics / MIT Press	9,874	0,82314
17	American Economic Journal: Economic Policy / American Economic Association	9,316	0,17570
18	Journal of Monetary Economics / Elsevier	8,974	0,97164
19	American Economic Journal: Applied Economics / American Economic Association	8,509	0,22191
20	Journal of the European Economic Association / European Economic Association	8,483	0,59261

Авторы благодарят Г. И. Забиняко за ценные практические советы в области матричных вычислений.

Список литературы

1. Gross P. L. K., Gross E. M. College libraries and chemical education // Science. 1927. V. 66, iss. 1713. P. 385–389.

2. Garfield E. Citation indexes to science: a new dimension in documentation through association of ideas // Science. 1955. V. 122, iss. 3159. P. 108–111.

3. Bergstrom C. T. Eigenfactor: Measuring the value and prestige of scholarly journals // College and Research Libraries News. 2007. V. 68. N 5.

4. Althouse B. M., West J. D., Bergstrom T. C., Bergstrom C. T. Differences in impact factor across fields and over time // [Electron. resource]. http://escholarship.org/uc/item/76h442pg.

5. Bergstrom C. T., West J. D. Assessing Citations with the Eigenfactor Metrics // Neurology. 2008. V. 71. P. 1850–1851.

6. Bergstrom C. T. Eigenfactor Project. Ranking and mapping scientific knowledge // [Electron. resource]. http://eigenfactor.org.

7. RePEc. General principles. [Electron. resource]. http://repec.org/.

8. Price D. Networks of scientific papers // Science. 1965. V. 149, N 3683. P. 510–515.

9. Bonacich P. Factoring and weighting approaches to clique identification // J. Math. Sociology. 1972. V. 2. P. 113–120.

10. Vigna S. Spectral Ranking. [Electron. resource]. http://vigna.dsi.unimi.it/papers.php.

11. Page L., Brin S., Motwani R., Winograd T. The pagerank citation ranking: Bringing order to the web // Techn. rep., Stanford Digital Library Technologies Project. 1998. [Electon. Resource]. http://ilpubs.stanford.edu:8090/422/.

12. West J. D. Eigenfactor: ranking and mapping scientific knowledge. [Electron. resource]. http://octavia.zoology.washington.edu/people/jevin/Documents/Dissertation_JevinWest.pdf.

13. Гантмахер Ф. Р. Теория матриц. М.: Наука, 1988.

14. Seneta E. Non-negative matrices and Markov chains. N. Y.: Springer, 2006.

15. Von Mises R., Pollaczek-Geiringer H. Praktische Verfahren der Gleichungsauflösung // ZAMM – Zeitschriftfür Angewandte Mathematik und Mechanik. 1929. N 9. P. 152–164.

16. Вержбицкий В. М. Численные методы. Линейная алгебра и нелинейные уравнения. М.: Высш. шк., 2000.

17. Poole D. Linear algebra. A modern introduction. Canada: Thomson, 2006.

18. Бредихин С. В., Кузнецов А. Ю., Щербакова Н. Г. Анализ цитирования в библиометрии. Новосибирск: ИВМиМГ СО РАН, НЭИКОН, 2013.

19. [Electron. resource]. http://citec.repec.org/search.html#journals.

20. Davis P. M. Eigenfactor: Does the principle of repeated improvement result in better estimates than raw citation counts? // J. Amer. Soc. Inform. Sci. Tech. 2008. V. 59, iss. 13. P. 2186–2188.

Бредихин Сергей Всеволодович — канд. техн. наук, зав. лабораторией Института вычислительной математики и математической геофизики СО РАН; e-mail: bred@nsc.ru;
Ляпунов Виктор Михайлович — ведущий инженер Института вычислительной математики и математической геофизики СО РАН; e-mail: vic@nsc.ru;
Щербакова Наталья Григорьевна — ст. научн. сотр. Института вычислительной математики и математической геофизики СО РАН; e-mail: nata@nsc.ru

Дата поступления — 29.11.2013