ИНФОРМАЦИОННЫЙ АНАЛИЗ

УДК [002 : 001.8] : 004

В.М. Московкин, Н.А. Голиков, О.В. Серкина

Конструирование метрик цитирования нового поколения*

Сделан учет импакт-факторов журналов, входящих в сеть цитирования автора, для нового поколения метрик цитирования scoring rules, представляющих собой некие суммирующие правила по подсчету всех статей автора их цитирований. Такой класс метрик назван как IF-scoring rules. Описана математическая модель этого класса метрик, а также особенности компьютерного алгоритма их расчетов, опирающегося на технологии Data Mining и Machine Learning.

Ключевые слова: *IF-scoring rules, импакт-фактор, публикационная активность, метрики цитирования, индекс Хирша, Data Mining, Machine Learning*

ВВЕДЕНИЕ

В научном обосновании механизмов повышения публикационной активности, большое внимание уделяется метрикам цитирования. Они могут агрегироваться с авторского на любой другой уровень (научный коллектив, университет, регион, страна). От этих метрик зависит рейтинг ученых, коллективов, институтов и т.д., а, следовательно, и их финансовая поддержка. Однако «ущербные» метрики сильно искажают состояние дел в науке. К таким метрикам сейчас относят индекс Хирша и все семейство хиршеподобных метрик. Оказалось, что семейство этих метрик не удовлетворяет трем простейшим постулатам сравнения. В связи с этим T. Marchant в 2009 г. вводит понятие scoring rule, а именно – метрик, основанных на подсчете всего спектра публикаций автора и их цитирований. Нами будет обоснована более совершенная метрика цитирования, которая, помимо всего спектра публикаций автора и их цитирований, использует и импакт-фактор всех журналов, входящих в сеть цитирования автора (журналы, в которых опубликовался автор, и журналы, из которых идут ссылки на его работы). Ясно, что какие-либо манипуляции с такой метрикой, в принципе, невозможны.

Проблема состоит не столько в математическом описании таких метрик, сколько в разработке компьютерного алгоритма и программы для автоматизированного их расчета. Эта проблема нами была решена с применением алгоритмов, близких к алгоритмам, используемым в некоторых областях Data Mining (Native Lanquage Processing). Программы были написаны на языке программирования Python. В настоящей работе будут представлены расчеты различных вариантов таких метрик по разработанному алгоритму и созданной программе для двух наиболее цитируемых ученых (физиков) Белгородского государственного университета (НИУ «БелГУ»). Максимальные расчетные значения этих метрик, для различных вариантов расчетов, лежали в разумном диапазоне, что не потребовало использования процедур нормирования.

Начиная с классической работы J. E. Hirsch [1], произошел бум по модификации индексов Хирша и созданию ему подобных индексов. Как показано в работе [2], в 2010 г. и 2011 г. почти каждая четвертая статья, опубликованная в журналах «Scientometrics» и «Journals of Informetrics», цитировала вышеупомянутую работу Хирша. В научной литературе мы можем обнаружить множество индексов (m, g, e, w, q^2, hg и др.), используемых для оценки научной продуктивности в терминах количества публикаций и цитирований. Как отмечено в [2], множество этих индексов основано на модификации индекса Хирша (h-index). Так, в работе [3] мы обнаружили перечень не менее чем 37 вариантов модификаций h-index.

Сущность всех этих исследований описал T. Marchant [4]: «Многие исследователи, анализируя предыдущие индексы, находят, что они обладают некоторыми недостатками и затем предлагают пути по их устранению или предлагают новые индексы, которые, как они предполагают, являются лучше старых. Но это не гарантирует, что эти модифицированные или совершенно новые индексы не обладают другими недостатками». Несовершенство h-index на фун-

* Работа выполнена при поддержке Госзадания на 2015 г., код проекта -516

даментальном уровне продемонстрировано в работе [2], в которой показано, что этот индекс не удовлетворяет трем принципиальным постулатам.

1. Если два ученых достигают одного и того же относительного улучшения научной результативности, то их ранжирование друг относительно друга должно оставаться неизменным.

2. Если два ученых достигают одного и того же абсолютного улучшения научной результативности, то их ранжирование друг относительно друга должно оставаться неизменным.

3. Если ученый X_1 имеет ранг выше, чем ученый Y_1, а ученый X_2 имеет ранг выше, чем ученый Y_2, то исследовательская группа, состоящая из ученых X_1 и X_2, должна иметь совокупный ранг выше, чем исследовательская группа, состоящая из ученых Y_1 и Y_2.

МЕТОДИКА ИССЛЕДОВАНИЯ

Все недостатки хиршеподобных индексов устраняются с помощью построения нового поколения метрик цитирования, основанных на так называемых scoring rules (summation-based rankings). Такие метрики были предложены в 2009 г. в работе T. Marchant [4].

В простейшем случае, чтобы вычислить scoring rule для множества публикаций, мы сначала должны вычислить score (рейтинг) для каждой отдельной публикации в этом множестве. Score отдельной публикации определяется количеством ее цитирований. После вычисления score для каждой публикации scoring rule вычисляется с помощью суммирования всех score для отдельных публикаций. Следовательно, для данного множества N, состоящего из n публикаций с C_1, C_2, .., C_n цитированиями, scoring rule равняется:

$$I(C_1, C_2, ..., C_n) = \sum_{i=1}^{n} f(C_i), \qquad (1)$$

где $f(C_i)$ – возрастающая функция, которая определяет score публикации на основе того, сколько раз эта публикаций была процитирована [2].

В качестве этой функции предлагается использовать слабовозрастающие выпуклые функции $f(C_i) = \sqrt{C_i}$ или $f(C_i) = \ln(C_{i+1})$ [5]. В работе [6], в формуле (1) используются функции $f(C_i) = \sqrt{C_i}$, $I(C_i) = \sqrt{\sum C_i}$.

В более строгом приближении T. Marchant записывает scoring rule в виде:

$$U(f) = \sum_{j \in J} \sum_{x \in N} \sum_{a \in N} f(i, x, a)\, u(j, x, a), \qquad (2)$$

где $f(i, x, a)$ – количество публикаций автора f в журнале j с цитированием x и a соавторами (количество авторов равно $a+1$);

$u(j, x, a)$ – значимость или score одной публикации в журнале j с цитированием x и a соавторами;

$J = \{j, k, l, \}... \subset N$ – множество журналов,

N – множество целых чисел.

Тройная сумма (2) представляет собой общий score автора. Как отмечает T.Marchant [4], многие популярные библиометрические ранжирования являются scoring rules. Например, если мы возьмем U равным положительной константе, то получим ранжирование, основанное на количестве публикаций. Если определим U выражением $u(j, x, a) = x$ для всех $j \in J$, $x, a \in N$, то получим ранжирование, основанное на количестве цитирований. Если определим u выражением $u = (j, x, a) = IF(j)$ для всех $j \in J$, $x, a \in N$, где $IF(j)$ является импакт-фактором журнала j, то получим ранжирование, основанное на сумме импакт-факторов, которое предложено в работе [7].

Для наших дальнейших исследований представляет интерес задание функции $u(j, x, a)$ в виде:

$$u(j, x, a) = x\, IF(j)/(a+1). \qquad (3)$$

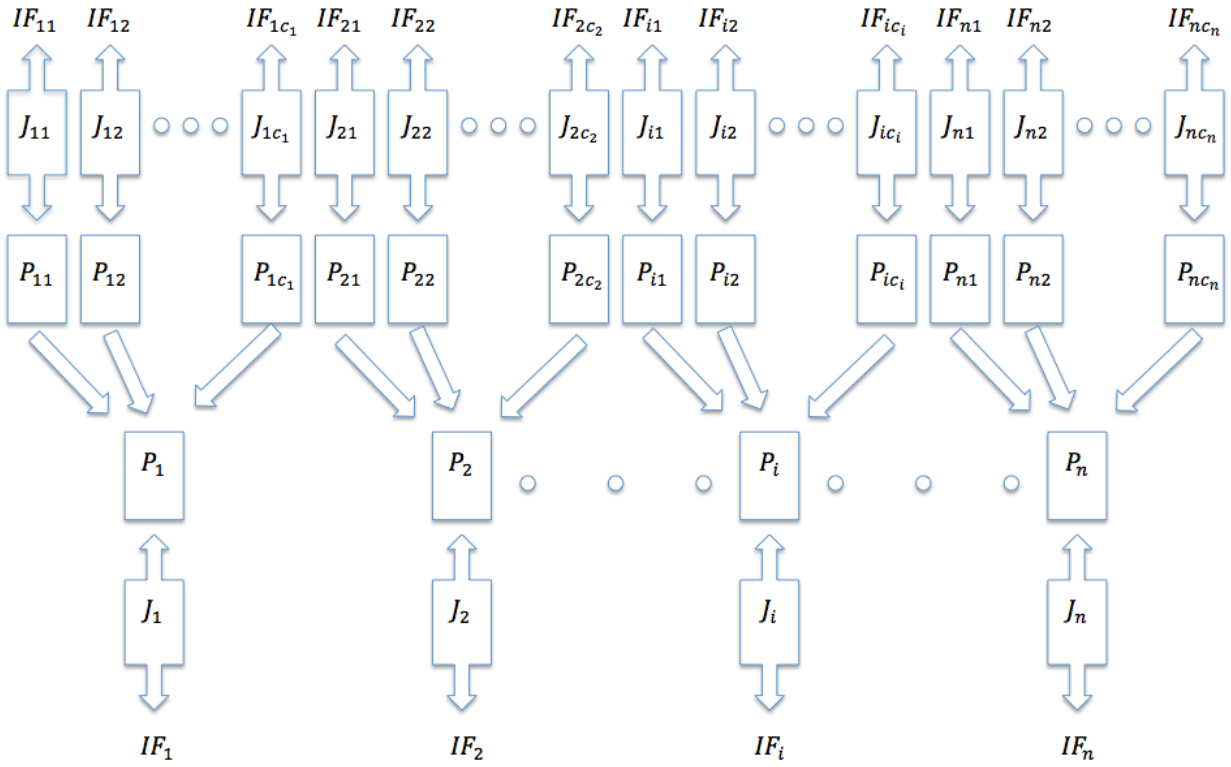
В этом случае мы получим простейший scoring rule для ранжируемых авторов в соответствии с числом цитирований, нормированных на количество авторов, и импакт-фактором [4]. В нашем исследовании мы будем абстрагироваться от количества авторов *a*, но будем дополнительно учитывать импакт-факторы журналов для тех статей, которые ссылаются на статьи рассматриваемого автора.

В работе [8] мы предложили идею построения IF-scoring rule. Здесь же рассмотрим его концепцию и математическую модель.

Блок-схема алгоритма предлагаемого нами IF-scoring rule показана на рисунке.

В простейшем случае, формулу расчета scoring rule, алгоритм которого показан на рисунке, запишем в виде:

$$\begin{aligned} U(P_1, P_2.., P_i, .., P_n) &= IF_1(IF_{11} + IF_{12} + ... + \\ &+ IF_{1j} + ... + IF_{1c_1}) + IF_2(IF_{21} + IF_{22} + ... + \\ &+ IF_{2j} + ... + IF_{2c_2}) + \\ &+ IF_i(IF_{i1} + IF_{i2} + ... + IF_{ij} + ... + IF_{ic_i}) + \\ &+ IF_n(IF_{n1} + IF_{n2} + ... + IF_{nj} + ... + IF_{nc_i}) = \\ &= \sum_{i=1}^{n} \sum_{j=1}^{C_i} IF_i IF_{ij} \end{aligned} \qquad (4)$$



Блок-схема алгоритма для вычисления IF-scoring rule:

$(P_1, P_2, .., P_i, .., P_n)$ – множество статей, опубликованных некоторым автором;

$(J_1, J_2, .., J_i, .., J_n)$ – множество журналов, где были опубликованы эти статьи;

$(IF_1, IF_2, .., IF_i, .., IF_n)$ – множество импакт-факторов, соответствующее этим журналам;

$(P_{i1}, P_{i2}, .., P_{ij}, .., P_{ici})$ – множество статей в количестве C_i, цитирующих P_i статью;

$(J_{i1}, J_{i2}.., J_{ij}, .., J_{ic_i})$ – множество журналов, соответствующее множеству статей $(P_{i1}, P_{i2}.., P_{ij}, .., P_{ic_i})$;

$(IF_{i1}, IF_{i2}.., IF_{ij}, .., IF_{ic_i})$ – множество импакт-факторов, соответствующих цитирующим журналам.

Отметим, что некоторые журналы из множеств $(J_1, J_2.., J_i, .., J_n)$ и $(J_{i1}, J_{i2}.., J_{ij}, .., J_{ic_i})$ могут совпадать.

Отсюда следует, что $U(P_1, P_2, .., P_i, .., P_n) = Q(IF_i, IF_{ij})$ является возрастающей квадратичной функцией многих переменных. Она обладает характерным свойством всех scoring rules: $U(IF_i + \Delta IF_i, IF_{ij} + \Delta IF_{ij}) > U(IF_i, IF_{ij})$, где $\Delta IF_i > 0$, $\Delta IF_{ij} > 0$ – малые приращения.

Рассмотрим частные случаи функции, приведенной в формуле (4):

1) If $IF_i = IF_{ij} = 1$, тогда $U = \sum_{i=1}^{n} C_i$,

2) If $IF_i \neq 1$, $IF_{ij} = 1$, тогда $U = \sum_{i=1}^{n} IF_i C_i$,

3) If $\sum_{i=1}^{C_i} IF_{ij} = 1$, тогда $U = \sum_{i=1}^{n} IF_i$.

Все эти случаи, следующие из формулы (4), получены так же в работе [4] из формулы (2).

Взяв $U = \overline{U} = \sum_{i=1}^{n} IF_i C_i$ за нормирующую функцию, можно показать, что при $\sum_{j=1}^{C_i} IF_{ij} > C_i$, когда суммарный импакт-фактор журналов цитирующих статей превышает количество цитирований, будет иметь место $q = \frac{U}{\overline{U}} > 1$, а в противном случае – $q < 1$.

В дополнение к функции, приведенной в формуле (4), введем еще пять функций многих переменных:

$$U = \sum_{i=1}^{n} IF_i^{\frac{1}{2}} \left(\sum_{j=1}^{C_i} IF_{ij} \right)^{\frac{1}{2}}, \qquad (5)$$

$$U = \sum_{i=1}^{n} IF_i^{\frac{1}{2}} \sum_{j=1}^{C_i} IF_{ij}^{\frac{1}{2}} = \sum_{i=1}^{n} \sum_{j=1}^{C_i} IF_i^{\frac{1}{2}} IF_{ij}^{\frac{1}{2}} \quad , \qquad (6)$$

$$U = \sum_{i=1}^{n} \sum_{j=1}^{C_i} (\delta + IF_i) IF_{ij} \quad , \qquad (7)$$

$$U = \sum_{i=1}^{n} (\delta + IF_i)^{\frac{1}{2}} \left(\sum_{j=1}^{C_i} IF_{ij} \right)^{\frac{1}{2}} \quad , \qquad (8)$$

$$U = \sum_{i=1}^{n} \left(\delta^{\frac{1}{2}} + IF_i^{\frac{1}{2}} \right) \sum_{j=1}^{C_i} IF_{ij}^{\frac{1}{2}} \quad , \qquad (9)$$

где δ является некоторым положительным параметром.

Для функций (5, 6, 8, 9) корень квадратный взят согласно работе [5], чтобы уменьшить рост функции U. Параметр δ в функциях (7-9) введен, чтобы придать значимость статьям, опубликованным в журналах с нулевым импакт-фактором $(IF_i = 0)$.

РЕЗУЛЬТАТЫ РАСЧЕТОВ И ДИСКУССИЯ

Для расчетов по формулам (4-9) нами были разработаны специальные алгоритмы и программа на языке Python, идентифицирующие названия «скопусовских» журналов, в которых статьи авторов были опубликованы при их поиске с помощью Google Scholar, и определяющие их импакт-факторы с помощью платформы SCIMAGO.

Приведем общее описание алгоритмов.

Задача вычисления scoring rule исследователя подразделяется на две подзадачи:

(а) – сбор информации, необходимой для расчёта, (б) – собственно расчёт.

В данном случае подзадача (а) является более трудоёмкой. Она, в свою очередь, подразделяется на следующие пункты.

1. Получение списка журналов с импакт-факторами.

2. Получение списка статей автора с идентификаторами (названиями) журналов, в которых статьи были опубликованы.

3. Получение для каждой публикации автора списка цитирующих её статей с идентификаторами (названиями) журналов, в которых статьи были опубликованы.

Стоит заметить, что, так как списки статей, упомянутые в п.п. 2 и 3, извлекались посредством скраппинга (scrapping) и последующего парсинга (parsing) поисковой выдачи Google Scholar, то это повлекло в ходе реализации этих пунктов возникновение наиболее трудоёмких пунктов всей программной части:

4. Преодоление защиты Google от поисковых роботов.

5. Идентификация и сопоставление названий журналов (либо их фрагментов), полученных из поисковой выдачи с названиями журналов из п. 1.

Отметим, что реализация п. 5 позволяет однозначно сопоставить название журнала из поисковой выдачи Google Scholar с названием журнала из списка SCIMAGO далеко не во всех случаях. Реализация, гарантирующая однозначное сопоставление в общем случае, по-видимому, не может быть выполнена, так как очень часто полученных фрагментов названий журналов (в общем случае Google Scholar выдаёт именно фрагменты, если название журнала слишком длинно, либо слишком длинен список авторов.) для однозначного восстановления названия недостаточно. В ходе парсинга и «собирания» названий журналов из фрагментов, могли возникнуть специфические трудности для данного конкретного названия.

Опишем более подробно пункты 1-5.

Пункт 1. «Получение списка журналов с их импакт-факторами» наиболее прост в реализации, так как платформа SCIMAGO предоставляет полный список научных журналов с их импакт-факторами. Физически этот список является простым html-документом, который легко парсится (используя lxml в качестве парсера) с извлечением необходимой информации путём обхода узлов html-дерева – html-тегов (tag) 'tr' (строки, содержащие информацию о журналах). Путь к узлам (node) задаётся в виде xpath-выражения. Далее осуществляется обращение к списку дочерних элементов каждого узла с извлечением текста, содержащегося в них.

Так как этот алгоритм: задание xpath-выражения, ведущего к соответствующим узлам html-дерева, получение списка узлов (в случае одного узла, соответствующего xpath-выражению, получаем список, состоящий из одного элемента), обращение к дочерним элементам (если необходимо), получение необходимой информации из атрибутов текста либо «хвоста» узла (текста, находящегося за «закрывающим тегом»), – является общим при парсинге всех html-документов, то в дальнейшем он будет опускаться.

Пункт 2. «Получение списка статей автора». Сразу же стоит оговорить, что действия, выполняемые в п.п. 2 и 3, велись посредством «прослойки» из библиотеки функций, реализующих п. 4, краткое описание алгоритмов работы которых дано далее. Список статей автора образовался посредством получения Google Scholar страницы профайла пользователя (автора), парсинга страницы профайла для выявления списка статей и ссылки на следующую страницу профайла (если таковая ссылка имелась). Пути к узлам html-дерева, содержащим как информацию о публикациях автора, так и ссылки на последующую страницу, задаются в виде xpath-выражений. И в том, и в другом случае данные узлы являются строками (тегами 'tr') некоторых таблиц. В цикле обрабатываются строки таблицы 'cit-table', содержащие в столбцах (путь к которым задаётся соответствующим xpath-выражением) информацию о публикациях автора, в частности url (интернет-адрес) запроса к Google Scholar, результатом которого является список статей, цитирующих данную публикацию. Обработка строк таблицы продолжается до тех пор, пока не встретится публикация автора, для которой нет

цитирующих статей. В запросе к профайлу автора указывается, что строки таблицы должны быть упорядочены по количеству цитирований – от наибольшего к наименьшему. Следовательно, если встретилась публикация, которая не цитируется, то у последующих публикаций, если таковые имеются, тоже не будет цитирований, либо не будет ссылки на следующую страницу профайла. Здесь же извлекается название журнала, в котором статья была опубликована, и количество цитирований этой статьи.

Пункт 3. «Получение для каждой публикации автора списка цитирующих её статей». Общий алгоритм аналогичен рассмотренному в описании п.2 и используемому для извлечения информации о публикациях автора из страниц Google Scholar профайла автора, т.е., осуществляется перебор в цикле узлов, содержащих информацию о цитирующей статье: если есть ссылка на следующую страницу, то получаем её и продолжаем цикл пока не исчерпаются страницы. Но этот алгоритм несколько сложнее предыдущего. Суть его состоит в следующем. Из узла, содержащего информацию о журнале (теге 'div', имеющем значение атрибута class, равное «gs_a»), необходимо извлечь само название журнала, которое может быть текстом, помеченным данным тегом, если нет гиперссылки на публикацию, либо может быть текстом гиперссылки (тега 'a' дочернего (child) по отношению к 'div'), либо быть «хвостом» (tail) тега 'a' (текстом находящимся сразу же за – закрывающей последовательностью тега), либо приемлема любая комбинация этих возможностей, и каждая из них должна быть охвачена одной из веток языковой конструкции if … elif … else. Из атрибута href тега 'a' необходимо извлечь и сам адрес ссылки, который может в дальнейшем помочь однозначно идентифицировать журнал.

Пункт 4. «Преодоление защиты Google от поисковых роботов». Google детектирует большое количество запросов к себе с одного и того же IP-адреса в течение небольшого промежутка времени, предлагая разгадать каптчу (CAPTCHA), подтвердив тем самым, что действует человек, а не средство автоматизированной добычи информации. CAPTCHA – это аббревиатура английских слов – «Completely Automatic Public Turing Test to Tell Computers and Humans Apart», в русском компьютерном языке используется русская трансляция английской аббревиатуры. Преодоление этой защиты состоит в попытке распознавания трудночитаемых надписей или в применении более радикальных мер к тому, что защитные механизмы Google считают средством автоматизированного сбора информации. В данном случае уточнить, что такое «большое количество запросов» и «небольшой промежуток времени» не представляется возможным, не зная алгоритмов защиты Google, либо не проведя специального исследования. Можно, почти наверняка, утверждать, что эти величины не являются фиксированными числами, но выяснение более точной информации о характере этих величин (выявление зависимости друг от друга, либо от других факторов) нам никак не поможет. Для преодоления защиты можно было бы попытаться научить робота разгадывать каптчу с той или иной степенью надёжности, но вероятно существование второго слоя защиты, который преодолеть было бы значительно труднее. Поэтому было принято решение пойти другим путём, устранив сами факторы которые позволяют защитным механизмам Google предположить, что он имеет дело с автоматизированным средством сбора информации.

Теперь перечислим методы устранения «вредных» факторов:

1) отправка запросов к Google Scholar с использованием различных IP-адресов, количество которых должно быть как можно больше. Смена IP-адресов в случайном порядке, по достижении некоторого предельного количества запросов с одного IP. Это самый важный и сложно реализуемый метод;

2) отправка запросов с одного и того же IP-адреса через случайные промежутки времени, выбираемые из некоторого диапазона длительности, ограниченного снизу не слишком маленькой величиной (порядка одной минуты). Иными словами, осуществляется примитивная имитация деятельности человека;

3) случайный выбор User-Agent – скраппера из некоторого списка User-Agent'ов, когда скраппер в случайном порядке «представляется» для Google Scholar в качестве различных версий браузеров Mozilla Firefox, Google Chrome, Internet Explorer и др.

Второй и третий методы реализуются достаточно просто: применяется функция random (или её вариаций) к соответствующим сущностям – длительности промежутка времени,через который отправлялись запросы, и номеру в списке User-Agent'ов.

Первый метод был реализован получением списка анонимных свободных HTTP и HTTPS прокси (proxy) с одного из Интернет-ресурсов, предоставляющих такие списки (hidemyass.com), и сменой прокси из списка по различным критериям. С данного прокси отправлялось несколько запросов подряд. Если при задержке с ответом на запрос возникала ошибка, либо исключение (Exeption), или на очередной запрос с данного прокси Google выдавал страницу с каптчей, то прокси удалялся из списка. Прокси был использован больше максимального числа раз (число выбиралось случайным образом из диапазона значений задаваемых пользователем) в течение некоторого промежутка времени, выбираемого случайным образом из диапазона длительностей временных промежутков.

Сложным в реализации оказался момент получения списка свободных прокси. Html, который выдаёт hidemyass.com, построен таким образом, чтобы человек воспринимал на экране компьютера IP-адрес в соответствующей ячейке таблицы списка как группу цифр, разделённых точками, как и положено IP-адресу. «Внутри» же html-документа в этой ячейке содержится смесь «мусорных» html-тегов и символов, окружающих «значимые» символы (цифры и точки), составляющие IP-адрес. Причём среди «мусорных» символов, в свою очередь, могли быть и точки. Потребовалось выявить закономерности, которым подчинялись «мусорные» теги и символы, и написать небольшую подпрограмму, ответственную за фильтрацию «не мусорных» символов. В результате задача преодоления защиты от роботов была решена полностью.

Пункт 5. «Идентификация и сопоставление названий журналов». Это наиболее сложный и громоздкий пункт. Ввиду громоздкости будут упомянуты только основные моменты реализации. Прежде всего, в списке журналов SCIMAGO к каждой записи списка (журнала) было добавлено название журнала, приведённое к единому регистру символов, используя только заглавные буквы, с исключёнными символами препинания (если таковые были) и составленная из названия аббревиатура. Так же были заменены все символы амперсанда (если таковые встречались) на «and». В описание журнала включалось «нормализованное» (приведённое к единому регистру с исключёнными знаками препинания и амперсандом, заменённым на and) название, разбитое на слова. В процессе получения названий (или их фрагментов) из поисковой выдачи Google Scholar в полученных фрагментах удалялись лишние пробелы, символы «-» в начале и конце, знаки препинания, косые черты, скобки. Далее, если фрагментов было несколько, то они «склеивались» и «нормализовались», а полученные строки разбивались на слова. Из поисковой выдачи извлекался Url (интернет-адрес) журнала, если он там был. Далее, название журнала, полученное из поисковой выдачи, сопоставлялось со всеми названиями журналов из списка SCIMAGO – нормализованное с нормализованным, если совпадения не было, то аббревиатура с аббревиатурой, – затем пословное с пословным. В последнем случае каждое слово из названия поисковой выдачи сопоставлялось со словами из представления названия в виде пословного списка SCIMAGO. Если для каждого слова было отмечено совпадение со словом из названия списка SCIMAGO и количество слов в обоих случаях совпадало, то названия считались совпавшими (в случае, если порядок слов в выдаче Google Scholar и в списке SCIMAGO различался). Если не было отмечено совпадений при применении всех трёх способов сопоставления, но у журнала был Url, то поэтому Url «вытягивался» расположенный там html-документ, и в нём проводился поиск (как в единой строке) нормализованного названия.

Если было отмечено только одно совпадение названия из выдачи Google Scholar при переборе всех названий из списка SCIMAGO, то журнал признавался однозначно идентифицированным, и Url (если он был) заносился в описание журнала в списке журналов. В дальнейшем при получении очередного названия из поисковой выдачи извлекался Url (если он был) и сопоставлялся со всеми Url, уже занесёнными в список журналов. Если было совпадение, то журнал считался опознанным однозначно. Высокоимпактные журналы идентифицировались с большей гарантией, ввиду более частого наличия у них Url адресов (собственных сайтов) и хорошо структурированных метаданных. Однозначно идентифицировать названия журналов удалось примерно для 35% цитирующих статей. Более изощрённые методы (введение «лингвистических» расстояний, различные статистические методы, более тонкая работа с сайтами журналов) не применялись ввиду того, что они изначально вероятностны, либо слишком сложны в реализации. Отметим, что доступ к платным наукометрическим базам данных делает пункты 4 и 5 (наиболее сложные в реализации) попросту не нужными.

Работа алгоритмов и программы были апробированы на основе Google Scholar-профилей двух наиболее цитируемых физиков НИУ “БелГУ” Рустама Кайбышева (Rustam Kaibyshev) и Андрея Белякова (Andrey Belyakov) (табл. 1). Автоматизированный сбор исходных данных по профилям с целью дальнейших вычислений по формулам (4-9) с помощью Google Scholar и SCIMAGO был осуществлен в августе 2013г. Так, для профиля Рустама Кайбышева количество публикаций, определяемых поисковой машиной Google Scholar, равняется 69, а количество их цитирований – 621. IF_i, IF_{ij} в формулах (4-9) были взяты с платформы SCIMAGO, как IF=Cites/Doc (2 years), что показано в табл. 1.

Таблица 1

Вычисление IF –scoring rule (авторской метрики цитирования) для двух наиболее цитируемых ученых НИУ “БелГУ”, август, 2013

Авторы	Формула номер / δ											
	7				8				9			
	0	0,01	0,1	1,0	0	0,01	0,1	1,0	0	0,01	0,1	1,0
Rustam Kaibyshev Cited articles: 69; Citing articles from identified journals: 621.	4014,9	4027,1	4137,3	5238,7	232,9	233,6	239,5	286,8	1389,9	1469,0	1639,9	2180,3
Andrey Belyakov Cited articles: 40; Citing articles from identified journals: 292.	1640,3	1646,5	1702,0	2257,0	173,2	173,6	177,5	210,6	592,6	630,2	711,5	968,7

В табл.1 расчеты по формуле (7) при δ=0 соответствуют расчетам по формуле (4); расчеты по формуле (8) при δ=0 – расчетам по формуле (5); расчеты по формуле (9) при δ=0 – расчетам по формуле (6). В табл.2 приведены значения $\Delta U(\delta) = \left(\frac{U_{(\delta=1)} - U_{(\delta=0)}}{U_{(\delta=0)}} \right) \times 100\%$, рассчитанные на основе табл. 1.

Из табл. 2 видим, что расчеты, проделанные по формуле (8), являются менее чувствительными к изменению параметра δ, причем эта формула дает на порядок меньшие абсолютные значения функции U по сравнению с расчетами по формулам (7) и (9). Отсюда следует, что для дальнейших расчетов следует рекомендовать формулу (5) как частный случай формулы (8).

Таблица 2

Значения ΔU(δ), рассчитанные на основе таблицы 1, %

	Номер формулы		
Author's Name	7	8	9
Rustam Kaibyshev	30,5	23,1	56,9
Andrey Belyakov	37,6	21,6	63,5

Итак, на основе scoring rule-подхода разработана метрика цитирования (citation metrics), учитывающая не только количество опубликованных статей автора и их цитирование, но и импакт-факторы журналов, в которых эти статьи были опубликованы, а также импакт-факторы журналов, в которых были опубликованы статьи, цитирующие статьи автора.

ЗАКЛЮЧЕНИЕ

В основе механизмов повышения публикационной активности лежат различные метрики цитирования – авторские, журнальные и другие. На волне жесткой критики хирше-подобных метрик возникло новое поколение метрик цитирования, учитывающих весь спектр публикаций автора и их цитирований. Для таких метрик T. Marchant в 2009 г. ввел понятие scoring rules. По сути, это некие суммирующие правила по подсчету всех статей автора и их цитирований. Помимо подсчета этих показателей нам удалось ввести в такую метрику импакт-факторы журналов, входящих в сеть цитирования автора, т. е. журналы с его публикациями и журналы, из которых идут ссылки на эти публикации.

Мы назвали такой класс метрик как IF-scoring rules от аббревиатуры термина «impact factor». Нами описана математическая модель этого класса метрик, а также особенности компьютерного алгоритма их расчетов, опирающегося на технологии Data Mining и Machine Learning.

Для расчета такой метрики было предложено шесть вариантов формул. Для автоматизированного расчета по этим формулам разработаны специальные алгоритмы и программа на языке Python, идентифицирующая названия Scopus-журналов и определяющая их импакт-факторы. Алгоритмы и программа были апробированы на Google Scholar-профилях двух наиболее цитируемых ученых (физиков) НИУ "БелГУ".

Результаты расчетов позволили выбрать наиболее приемлемую для дальнейшего использования формулу с точки зрения чувствительности этих формул к изменению предложенного нами неопределенного параметра. Кроме того, выбранная формула давала на порядок меньшие значения расчетных показателей. Расчеты показали, что исходные типы формул, различающиеся между собой разными видами агрегирования числа публикаций, их цитирований и значений импакт-фактора журналов, не требуют предварительной их нормировки, так как результаты расчетов варьировали в разумных пределах.

Как и любая другая метрика цитирования, IF-scoring rule может быть задействована в механизме повышения публикационной активности, но, в отличие от всех хирше-подобных метрик, она не подвержена манипулированию так же, как и более простые scoring rule.

СПИСОК ЛИТЕРАТУРЫ

1. Hirsch J. E. An index to quantify an individual's scientific research output // Proceedings of the National Academy of Sciences. – 2005. – Vol.102, № 46. – P. 16569–16572.
2. Waltman L., van Eck N.J. The inconsistency of the h-index // Journal of the American Society for Information Science and Technology. – 2012. – Vol. 63, № 2. – P. 406–415.
3. Bornmann L., Mutz R., Daniel H.-D. A multilevel meta-analysis of studies reporting correlations between the h index and 37 different h index variants // Journal of Informetrics. – 2011. – Vol. 5, № 3. – P. 346-359.
4. Marchant T. Score-based bibliometric rankings of authors // Journal of the American Society for Information Science and Technology. – 2011. – Vol.60, № 6. – P. 1132-1137.
5. Lundberg J. Lifting the crown—citation z-score // Journal of Informetrics. – 2007. – Vol. 1, № 2. – P. 145-154.
6. Levene M., Fenner T., Bar-Ilan J. A bibliometric index based on the complete list of cited publications // Cybermetrics. – 2012. – Vol.16, №1.
7. Fava G. A., Ottolini F. Impact factors versus actual citation // Psychotherapy and Psychosomatics. – 2000. – Vol.69. – P. 285 -286.
8. Moskovkin V.M., Golikov N.A. The new generation of citation metrics. Construction of IF-Scoring rules // In the proceedings of the 10-12 October 2013, Moscow International conference. – 2013. – P.92-93.

Материал поступил в редакцию 06.05.15.

Сведения об авторах

МОСКОВКИН Владимир Михайлович – доктор географических наук, директор Центра наукометрических исследований и развития университетской конкурентоспособности, профессор кафедры мировой экономики НИУ “БелГУ”, г. Белгород
e-mail: moskovkin@bsu.edu.ru

ГОЛИКОВ Николай Александрович – независимый разработчик, г. Харьков
e-mail: kolia_forme@mail.ru

СЕРКИНА Олеся Викторовна – кандидат педагогических наук, аналитик Центра наукометрических исследований и развития университетской конкурентоспособности, НИУ “БелГУ”, г. Белгород
e-mail: serkina@bsu.edu.ru