ИНФОРМАЦИОННЫЙ АНАЛИЗ

УДК 001:002:311.311WOS

Р.С. Гиляревский

О некорректности использования индексов цитирования для вычислений по сопоставлению разделов науки*

Обсуждается заведомая неточность вычислений вклада ученых, организаций и стран, а также периода полужизни публикаций, количества статей и их цитирований по разделам науки при использовании данных платформ индексации и цитирования (на примере Web of Science). Некорректность объясняется тем, что на этих платформах по тематическим категориям распределяются журналы, тогда как объекты вычислений требуют классификации по статьям. Распределение статей конкретной тематики по журналам неравномерно и подчиняется закону рассеивания Брэдфорда, который в этом случае не учитывается. Выясняется невозможность исправления результатов подобных вычислений. Выход видится в использовании для них реферативных журналов, в которых по рубрикаторам индексируются статьи.

Ключевые слова: *неточность вычислений, тематические категории, платформы индексации и цитирования, Web of Science, закон Брэдфорда, распределение журналов*

DOI: 10.36535/0548-0027-2022-02-3

ВВЕДЕНИЕ

Вычислениями сопоставлений по разделам науки с использованием данных *Web of Science* (*WoS*) мне пришлось заниматься с 2019 г. в связи с присоединением к возглавляемому мною Отделению теоретических и прикладных проблем информатики ВИНИТИ РАН подразделения, для которого это было государственным заданием. Об Указателе библиографических ссылок по естественным наукам (*Science Citation Inde*x) мне стало известно с момента его появления в 1962 г., он описан в монографии А.И. Михайлова «Основы научной информации» в 1965 г. [1], с его создателем Ю. Гарфилдом мы были знакомы лично, с 1991 г. я неоднократно посещал *Institute for Scientific Information* (*ISI*) в Филадельфии на Маркет-стрит и его вычислительный центр в пригороде. На протяжении прошедших лет мы в ВИНИТИ РАН следили за всеми этапами развития платформы *WoS* и отражали его на страницах нашего сборника «Научно-техническая информация». Мне представлялось, что я вполне в теме, поэтому и участвовал в написании статей, в которых публиковались результаты исследований нового для меня коллектива о вкладе российских авторов в мировой поток публикаций по различным отраслям и разделам науки.

Мы были убеждены, что на платформе *WoS* по ее 254 тематическим категориям индексируются статьи, как вероятно думают сотни исследователей, ведущих различные вычисления, связанные с этими категориями. Но это оказалось не так, о чем неоднократно упоминается во втором издании «Руководства по наукометрии» [2], написанном коллективом российских авторов с участием руководства и специалистов *ISI*.

Вот некоторые из этих упоминаний. «В указателях цитирования большей частью классификация осуществляется на уровне индексируемых в них журналов, а не на уровне отдельных документов… Поскольку для детального библиометрического анализа часто недостаточно основываться на используемых в указателях цитирования журнальных классификаторах, проводятся многочисленные исследования, позволяющие осуществлять анализ направлений на уровне отдельно взятых статей» [2, с.164-166]. «В базах научного цитирования каждый журнал имеет «тематическую привязку», он отнесен к той или иной научной дисциплине (может быть отнесен более чем к одной). Что важно: при расчете практиче-

* Работа выполнена в рамках исследования по теме 0003-2022-0001 Государственного задания ВИНИТИ РАН при поддержке Российского фонда фундаментальных исследований – проект № 20-07-00014.

ски всех индикаторов, о которых пойдет речь в настоящей главе, тематическая рубрика *статьи* определяется в базе данных цитирования на основании тематической рубрики *журнала*, где она опубликована. Все статьи одного журнала имеют одну и ту же рубрику (рубрики)» [2, с. 181-182].

ОСОБЕННОСТИ ОБЪЕКТОВ КЛАССИФИКАЦИИ

Как объекты классификации статья и книга, с одной стороны, и журнал, с другой, имеют совершенно разную природу и проявляют разный характер. Статья и книга – это однократно возникшие сущности; будучи опубликованы в определенном году и месте, они уже не могут их изменить, меняется только отношение к ним – их могут читать, цитировать или никак не реагировать на их появление. Другое дело журнал – это живой социальный организм, он может расширять или сужать свой тематический профиль, разделяться на серии, объединяться с другими журналами, прекращать выходить, как это происходит со многими новыми журналами. Поэтому для статей и книг применяют иерархические и предметные классификации, которые консервативны и стабильны, а для журналов – чаще всего списки, что мы и видим на платформах индексации и цитирования.

Список журналов – инструмент коварный, ему нужно постоянное библиографическое обслуживание. В большинстве таких списков, например, в так называемом ваковском, много журналов – жертв «детской смертности». Вообще, если список журналов составлен не *de visu*, то он, выражаясь юридическим языком, ничтожен – в нем уже при его составлении присутствуют не только давно прекратившиеся, но даже и никогда не существовавшие журналы. Работники Высшей аттестационной комиссии Министерства науки и высшего образования РФ ограничивают использование журналов для публикации содержания диссертаций не только номенклатурой научных специальностей, но и факультетскими дисциплинами, по которым присуждаются степени (математические, технические, филологические и др. науки). Таким образом к списку журналов они привязывают не только классификацию людей-специалистов, но и менее стабильную дисциплинарную направленность журнала, что профессионально ошибочно.

Особой заботы требуют журналы многопрофильные или с быстро меняющимся профилем. Об этом также написано в цитированном Руководстве: «В первую очередь это касается журналов, которые относятся к категории мультидисциплинарных. В случае этих журналов анализ осуществляется по отдельным статьям на основании списков процитированной литературы и профиля журналов, из которых статьи получили цитирования, поэтому такие журналы в *Essential Science Indicator*s могут попадать в несколько из выделенных 22 широких областей, при этом их ранжирование определяется количеством статей и цитирований в соответствующей области» [2, с. 173-174].

Следует иметь в виду, что названная в цитате информационно-аналитическая служба работает с 1% высокоцитируемых статей, а 22 «широких области» – это специальная классификация для статей, отличающаяся от 254 тематических категорий для журналов. Другая подобная служба этой платформы *InCites* в конце 2020 г. создала функцию *citation topics* для кластеризации и анализа цитируемости на уровне отдельных статей, а не журналов.

ЯВЛЕНИЕ РАССЕИВАНИЯ СТАТЕЙ ОПРЕДЕЛЕННОЙ ТЕМАТИКИ ПО ЖУРНАЛАМ

Необоснованность многих вычислений по тематическим категориям журналов для определения наукометрических параметров не исчерпывается различиями классифицируемых объектов. Гораздо серьезнее то, что при этом не учитывается явление рассеивания статей определенной тематики по журналам, объективно существующее с момента возникновения этого инструмента научной и социальной коммуникации в 1665 г. Это явление заключается в том, что в каждом журнале любого конкретного профиля помимо соответствующих ему статей публикуются статьи смежной, а подчас и вовсе другой тематики. Оно было случайно открыто в 1934 г. английским документалистом С. Брэдфордом, который спустя полтора десятилетия опубликовал окончательную формулировку закона, известного под его именем: «Если научные журналы расположить в порядке убывания числа помещенных в них статей по какому-либо заданному предмету, то в полученном списке можно выделить ядро журналов, посвященных непосредственно этому предмету, и несколько групп или зон, каждая из которых содержит столько же статей, что и ядро. Тогда количества журналов в ядре и в последующих зонах будут относиться как $1 : a : a^2$» [3, 4].

Спустя два десятилетия об этом уже писали в справочниках: «С. Брэдфорд установил, что если совокупность всех публикаций, посвященных какому-либо вопросу, принять за единицу, то в специальных периодических изданиях по данному профилю, число которых сравнительно невелико, помещается лишь около одной трети этих публикаций. Вторая треть статей по данному вопросу оказывается опубликованной в значительно большем числе тематически родственных журналов другого профиля. И, наконец, последняя треть этих публикаций рассеяна в огромном числе периодических изданий, в которых появление статей по данному вопросу нельзя предвидеть, так как эти периодические издания носят слишком общий характер или тематически не связаны с данным вопросом.

Эта зависимость, получившая название закона рассеивания Брэдфорда, в аналитическом виде может быть записана следующим образом:

$$T_1 : T_2 : T_3 = 1 : n : n^2,$$

где T_1, T_2 , T_3 – число периодических изданий, содержащих соответственно x, $2x$ и $3x$ статей по тому или иному вопросу, а n – число, находящееся в прямой зависимости от x и имеющее разное значение для различных отраслей знания» [5]. Отсюда следует, что полный охват научно-технической литературы по какому-либо вопросу не может быть обеспечен, если ограничиться лишь просмотром журналов по соот-

ветствующему профилю и журналов по родственной тематике. Для обеспечения исчерпывающей полноты охвата литературы, например, по химии, информационная служба должна просматривать практически все научно-технические журналы.

Из закона и явления, на котором он основан, вытекают два важных следствия. Во-первых, при подсчете статей в журналах, относящихся к определенной тематической категории, учитываются далеко не все статьи этой тематики. Если предположить, что в категории *WoS* приводится число журналов близкое к брэдфордовскому ядру (что возможно, но не очевидно), то это значит, что отражается только одна треть статей, а оставшиеся другие две трети рассеяны по журналам во многих других категориях. Во-вторых, что еще хуже, по тематике этой категории мы учитываем некоторое количество статей, к ней не относящихся. Много их или мало мы еще попробуем оценить. Для этого нужно иметь в виду, что профиль журнала, его тематика – это довольно расплывчатое понятие с размытыми границами, тогда как «предмет» статьи в законе Брэдфорда более четок и конкретен, поскольку одним из условий выполнения этого закона является несомненность отнесения основного смыслового содержания статьи к этому предмету.

Закон этот – объективный социальный закон, действующий в любой совокупности журналов, имеющих или не имеющих определенного порядка их расположения. Это означает, что когда мы указываем количество статей и/или ссылок в одной из тематических категорий *WoS*, имеются в виду только статьи и ссылки в журналах, приписанных к этой категории, причем далеко не все они имеют тематику этой категории, а те, которые относятся к ней, это возможно только треть всех статей этой тематики, рассеянных по всей совокупности журналов в 254 тематических категориях.

Следовательно, все рассуждения, основанные на вычислениях по разделам науки в индексах цитирования не могут использоваться для каких-либо достоверных выводов о соотношениях вклада ученых или старения публикаций. Более того, полагаю, что и усилия специалистов *ISI* по замене импакт-фактора журнала на более сопоставимый индикатор цитирования журнала неадекватны полученному результату, поскольку не учитывают явления рассеивания статей определенной тематики по журналам.

ВОЗМОЖНЫЕ ВЫХОДЫ ИЗ СИТУАЦИИ

Размеры бедствия можно было бы исследовать, и попробовать оценить степень неточности вычислений. Если просчитать количество журналов в тематической категории и в близком ей по тематике и предмету ядру брэдфордовского распределения, и окажется, что их объемы не очень отличаются друг от друга, а в содержании их статей заметно сильное сходство, то неполнотой результатов вычислений без учета закона Брэдфорда можно было бы пренебречь (особенно при их сопоставлении и с соответствующими оговорками). Однако подобное предположение приходится отвергнуть, так как вычисления лежат в разных плоскостях и при ближайшем рассмотрении оказываются несопоставимыми. Степень неточности вычислений при игнорировании явления рассеивания статей по журналам каждый раз зависит от выбора предмета (которых множество) и периода вычисления (который по Брэдфорду не может превышать 3-х лет).

Результаты, полученные в свое время М.В. Араповым и А.Н. Либкиндом, показали, что введенное ими понятие «хронологической замкнутости» для различных тематик, может охватывать разные периоды от 2-х до 5-ти лет. Причем для динамичных, относительно молодых тематик (ядерные реакторы; авиационные и ракетные двигатели) период хронологической замкнутости составлял около 2-х лет, а для давно сложившихся, достаточно консервативных тематик (двигатели внутреннего сгорания; котлостроение) – больше 5-ти лет. Эти результаты были получены на основании данных соответствующих тетрадей РЖ ВИНИТИ РАН. Хронологическая замкнутость – это период, на котором наилучшим образом выполняется закон Ципфа, который, как известно, является одним из математических вариантов описания закона рассеивания [6, 7].

Естественным выходом из этой ситуации был бы переход на постатейную индексацию, что пока невозможно в полном объеме. Свыше 20 тыс. названий журналов в *WoS CC* дает около 3 млн статей в год, что исключает их ручную обработку. Системы автоматического индексирования находятся в экспериментальной стадии и пока не дают надежных результатов даже при классификации по десятку признаков.

Между тем интерес к тематическим сопоставлениям в потоке научных публикаций все время возрастает, что подогревается и кризисом в системе научной коммуникации. В этом кризисе большое место занимают экономические проблемы. Стоимость годовой подписки на средний журнал настолько возросла, а количество издающихся журналов так быстро увеличивается, что ни сами ученые, ни библиотеки не могут оформить подписку на необходимые им журналы. Интеллектуальные усилия по распространению новых научных достижений оказываются напрасными, а знания менее доступными.

Для противодействия этой тенденции в среде крупных зарубежных ученых распространилось движение так называемого открытого доступа к научным журналам. Его смысл заключается в том, чтобы издательские расходы несли авторы статей в этих журналах, а для читателей они были бы бесплатны. В мире существуют уже десятки тысяч таких журналов, публикация одной статьи в которых стоит несколько тысяч долларов. Еще одной опасностью для научной журналистики является так называемое стимулирование публикационной активности, заключающееся в необходимости для ученого подтверждать свою значимость не качеством и результатами своих исследований, а количеством опубликованных статей и числом библиографических ссылок на них. Этим пользуются недобросовестные издатели, которые обещают авторам опубликовать их статьи без серьезного рецензирования в своих журналах, получивших название хищнических или мусорных.

Неоправданная публикационная активность ведет к некорректности вычислений при необдуманной методике использования исходных данных, к заведомой неточности вычислений вклада ученых, организаций и стран, периода полужизни публикаций, количества статей и их цитирований по разделам науки при использовании данных платформ индексации и цитирования. Объясняется некорректность тем, что на этих платформах по тематическим категориям распределяются журналы, тогда как объекты вычислений требуют классификации по статьям. Распределение статей конкретной тематики по журналам неравномерно и подчиняется закону рассеивания Брэдфорда, который при подобных вычислениях не учитывается. Результаты таких вычислений трудно выявить и невозможно исправить.

Как и любое социальное явление, наука ищет выход из кризисной ситуации в изменении редакционной политики при публикации численных результатов проведенных экспериментов. В статье сохраняют все ее обычные разделы – историю вопроса, концепцию исследования, его методику, экспериментальную базу, обсуждение результатов, перспективы их применения, а сами численные данные передают в банк данных, который принимает их в стандартных форматах, готовыми для последующего бесплатного применения другими учеными и специалистами.

Продолжающие играть важную роль в системе научной коммуникации реферативные журналы остаются единственными органами этой системы, осуществляющими постатейную индексацию статей в объемах, сопоставимых с обсуждаемыми. Для этого они располагают соответствующими инструментами в виде рубрикаторов и необходимой организационной структурой. Именно реферативные журналы и нужны для удовлетворения растущего спроса на такие сопоставления. Что касается вычислений по разделам науки с использованием тематических категорий индексов цитирования, то они, конечно, не прекратятся, хотя и дают бессмысленные результаты из-за невозможности оценить их неточность.

ЗАКЛЮЧЕНИЕ

В этой небольшой статье не сообщается ничего нового. В ней просто обращено внимание на необходимость считаться с закономерностями социальных явлений, как мы вынуждены подчиняться законам природы. Как бы ни соблазнительно было использовать для вычислений надежные базы данных *WoS CC*, этого нельзя делать для ряда вычислений по сопоставлению разделов науки, поскольку тематические категории, в которых размещены журналы, для этого не предназначены. И никакие оговорки об учете этого обстоятельства ничему не помогут – вычисления некорректны, и степень неточности результатов определить нельзя.

* * *

Автор благодарит А.Н. Либкинда и И.А. Либкинда за прочтение рукописи статьи, ценные замечания и дополнения.

СПИСОК ЛИТЕРАТУРЫ

1. Михайлов А.И. Основы научной информации / А.И. Михайлов, А.И. Черный, Р.С. Гиляревский. – Москва: Наука, 1965. – С. 90.
2. Акоев М.А., Маркусова В.А., Москалева О.В., Писляков В.В. Руководство по наукометрии: индикаторы развития науки и технологии; 2-е изд. / под. ред. М.А. Акоева. – Екатеринбург: Изд-во Уральского ун-та, 2021. – 358 с. –ISBN 978-5-7996-3154-3.
3. Bradford S.C. Documentation. – London: Lockwood, 1948. – 2nd ed. – 1953. – P. 154.
4. Vickery B.C. Bradford's law of scattering // Journal of Documentation. – 1948. – Vol. 4, № 2-3. – P. 198-203.
5. Handbook of special librarianship and information work; 2nd ed. / Gen. ed W. Ashworth. – London: Aslib, 1962. – P. 3.
6. Арапов М.В., Либкинд А.Н. Концепция замкнутости информационного потока // Научно-техническая информация. Сер. 2. – 1977. – № 6. – С. 1-15; Arapov M.V., Libkind A.N. The Concept of the Closed Information Flow // Automatic Documentation and Mathematical Linguistics. – 1977. – Vol. 11, № 2. – P. 77-94.
7. Libkind A.N. One approach to study communication in science // Scientometrics – 1985. – Vol. 8, № 3-4. – P. 217-223.

Материал поступил в редакцию 10.01.22.

Сведения об авторе

ГИЛЯРЕВСКИЙ Руджеро Сергеевич – доктор филологических наук, профессор, заведующий Отделением теоретических и прикладных проблем информатики ВИНИТИ РАН; профессор факультета журналистики Московского государственного университета им. М.В. Ломоносова
e-mail: giliarevski@viniti.ru