

СОДЕРЖАНИЕ

Додиг-Крнкович Г. Инфокомпьютационализм и философские аспекты научного исследования в информатике	3
Милоевич С., Лейдесдорф Л. Информационные метрики (iMetrics): научная сфера с социокогнитивной идентичностью	9
Кордые С. Мера сходства между научными журналами и разнообразия списка публикаций	21
Гарфилд Ю. Век создания указателей ссылок	25
Валтман Л., ЭК ван Н. Я., Уотерс П. Подсчет публикаций и ссылок: больше всегда ли лучше ?	29
Шрайбер М. Насколько соответствующей является предсказательная сила h -индекса? Экспериментальное изучение индекса Хирша в зависимости от времени	35

РЕДАКЦИОННАЯ КОЛЛЕГИЯ

Академик РАН **Ю.М. Арский** (Российская Федерация) — *главный редактор*, ВИНТИ РАН, 125190, Москва, ул. Усневича, 20. Телекс 411249

Проф. д-р. **Р.С. Гиляревский** (Российская Федерация) — *заместитель главного редактора*, ВИНТИ РАН, 125190, Москва, ул. Усневича, 20. Телекс 411249

С. Дж. Паркер — *заместитель главного редактора*, IDCR, P.O. Box 8500, Ottawa, Ontario K1G 3H9, Canada

А. Джилкрайст (Великобритания) — CURA Consortium and GAVEL g.e.i.e, 38 Ship Street, Brighton BN1 1AB, UK

М. Дрейк (США) — Технологический институт шт. Джорджия, Библиотечный и информационный центр, 704 Cherry Street, Atlanta, Georgia 30332-0900, USA

Н. Дюсулье (Франция) — 192 Chemin du Cannet, 06220 Vallauris, France

А. де Кемп (Германия) — Издательство “Springer-Verlag”, Postfach 10 52 80, D-69042 Heidelberg, Germany

Д-р **Т. Кеннон** (Великобритания) — Отдел исследований и разработок Британской библиотеки, 2 Sheraton Street, London W1V 4BH, UK

М. Миддлтон (Австралия) — Школа информационных систем, QUT Gardens Point Campus, 2 George Street, Brisbane, 4000 QLD., Australia

Т. Молвиг (Норвегия) — Национальное управление по научной информации, вузовским и специальным библиотекам, P.O. Box 2439 Solli, N-0201, Oslo, Norway

Х. Ринкон Феррейра (Бразилия) — Бразильский институт информации по науке и технике (IBICT), SAS— Quadra 5, Lote 06, Bloco H, 700-70-000 Brasilia D.F., Brazil

С. Феррейро (Чили) — Чилийский университет, Системы информационных и библиотечных служб, Casilla de Correo 10D, Santiago, Chile

Проф. **Ю. Фудзивара** (Япония) — Университет Цукуба, Институт электроники и информатики, Tsukuba-shu, Ibaraki, 305 Japan

Д-р **М. Хименес** (Испания) — Испанское общество по научной документации и информации, Fuencarral, 123-6° dcha., 28010, Madrid, Spain

Д-р **А.И. Чёрный** (Российская Федерация) — ВИНТИ РАН, 125190, Москва, ул. Усневича, 20. Телекс 411249

Инфокомпьютериализм и философские аспекты научного исследования в информатике*

Гордана ДОДИГ-КРНКОВИЧ
(Gordana DODIG-CRNKOVIC)

Мелардаленский университет, Швеция

Историческое развитие привело к разрушению натуральной философии, которая до XIX в. включала всё наше знание о физическом мире в растущем множестве специальных наук в «классическую модель науки». Концентрация внимания на глубинном вопросе, помогающем выпутываться из широкого контекста, ведет к проблеме потери общей мировой точки зрения и невозможности коммуникации между специалистами научных областей из-за различных языков, которые они разработали в изоляции. Необходимость нового унифицированного (единого) подхода становится все очевиднее в связи с информационной технологией, обеспечивающей и интенсифицирующей коммуникацию между различными научными областями, интеллектуальными сообществами и информационными источниками. В настоящее время не только естественные науки, но также все человеческое знание интегрируется в глобальную сеть, такую как Интернет с его разнообразием интеллектуальных и языковых сообществ.

Инфокомпьютериализм (Info-computationalism – ICON) как синтез панкомпьютериализма и панинформационизма представляет единый подход к пониманию естественных феноменов, включая живых существ и их познание, их способы обработки информации и создания знания. В ICON физический мир понимается как сеть вычислительных процессов на информационной структуре. Материя/энергия в этой модели заменяется информацией/вычислением, где информация является структурой, чьи динамики определяются как естественное (натуральное) вычисление.

ICON – это пример философского подхода, непосредственно связанного с соответствующими научными областями, а процесс представляет собой взаимный обмен: научные результаты влияют на философское мышление и наоборот. Исследование заключается в естественном (натуральном) вычислении, моделировании естественных феноменов, включая живые организмы в качестве информационно-вычислительных агентов и в применении принципов натурального (естественного) вычисления к технологическим артефактам. Уроки, полученные из разработки и реализации нашего понимания живых естественных вычислительных агентов, ведут к артефактам, все в большей степени способным моделировать основные возможности живых организмов обрабатывать и структурировать информацию. Среди прочих ICON поддерживает научное понимание разума (восприятие, мышление, умозаключение, воля, чувства, память и т.д.), обеспечивая вычислительные модели познания натуралиста.

*Перевод Dodig-Crnkovic G. Info-computationalism and philosophical aspects of research in information sciences. – <http://arxiv.org/ftp/arxiv/1207.1032.pdf> - 2012.

«Наши самые большие усилия направлены на обнаружение того, почему вещи являются такими, как они есть, или почему события вокруг нас происходят так, как они происходят. Так происходит во всех дисциплинах, и философия в этом отношении не является исключением; что выделяет философию, так это то, что она, будучи более общей, исследует глубже. Она рассматривает предположения, которые другие дисциплины не затрагивают, и пытается разъяснить такие предположения и проследить их взаимосвязи, пытается выявить, как мир связан и как он работает» [1].

Знание, как пропозициональное, так и непропозициональное, является основной составляющей науки, созданной научным процессом. Тем не менее, знание часто рассматривается как идентичное пропозициональному знанию, и наука отождествляется с поиском *истины* о мире; считается, что *истина* должна быть в пропозициональной форме. Хотя история, социология и философия науки – все предлагают достаточные аргументы, чтобы рассматривать науку как целеустремленную человеческую деятельность, направленную на производство *моделей*, которые дают нам возможность предсказывать, коррелировать и структурировать (в соответствии с [2], сжимать) релевантную информацию о мире.

Роль науки как механизма сжатия информации (в смысле алгоритмической информации [2]) особенно заметна в ее поиске простых и универсальных законов, которые, главным образом, характеризуют физические науки. Однако исторически идея естественного закона значительно эволюционировала. Прошлая идея *детерминистского порядка* мира (вселенной) была тесно связана с *принципом причинности*, который указывает на необходимую связь между одним событием (причиной) и другим событием (эффектом), которое является непосредственным следствием (результатом) первого. Однако *индетерминизм* квантовой физики стимулировал новые элементы в картине естественных законов. Позднее *беспорядок* был обнаружен даже в управляемых законом системах, показывая, что детерминистские функции могут порождать непредсказуемые результаты.

Что является решающим для научного знания, так это не его несомненность, иначе не так много можно было бы квалифицировать как знание в истории науки. Даже понимание фундаментальных научных идей, таких как время, пространство, масса и движение, было успешно исторически пересмотрено. Что постоянно характеризовало науки, так это *рациональность* их подходов, предположений и целей. Наука, в первую очередь, это объясняющее и предсказывающее *средство* *имения смысла* [1].

В настоящее время сложность проблем реального мира становится центром внимания наук, в первую очередь, благодаря вычислительным возможностям информационно-коммуникационной технологии (ИКТ). Вместо статичного, симметричного и установившегося твердого статичного порядка возникает новая динамичная картина, в которой все изменяется, и порядок создается как модель через слои пульсирующих основных физических процессов, и очень простые основные правила могут привести к эволюции сложных систем.

КЛАССИЧЕСКАЯ МОДЕЛЬ НАУКИ И СЛОЖНОСТЬ РЕАЛЬНОГО МИРА

Анализируя связь между философией и наукой, важно признать, что сдвиг парадигмы продолжается как в науках, так и в философии, как следствие недавнего роста трансдисциплинарного, междисциплинарного и перекрестно-дисциплинарного научного исследования, не знакомого классическому философскому анализу. Научные практики и появляющиеся в результате этого науки сегодня не являются тем, чем они были до революции ИКТ. Эффективная обработка информации, хранение и возможности обмена между современными, сетевыми, глобальными научными сообществами позволяют рассматривать сложные проблемы, что прежде было неосуществимо из-за их информационной и вычислительной сложности. Такие проблемы обычно определяются на нескольких уровнях абстракции (уровни описания, уровни функциональности) и, таким образом, они, как правило, охватывают несколько классических научных дисциплин.

Это развитие в направлении сложных проблем прояснило ряд предположений существующих наук. Среди других их области обоснованности становятся видимыми, делая различия и связи между разными областями более четкими и легкими для узнавания. Типичный пример, где зависимость от области различных наук становится очевидной, есть в моделировании живых организмов.

На основе нашего общего интуитивного понимания существует идеализированная точка зрения на науки, классическая модель науки, которая, согласно [3], является системой S предложений и понятий, удовлетворяющей следующие условия:

1. Все предложения и все понятия (или термины) S касаются специфического набора объектов или являются чем-то относящимся к определенной области живых существ.

2а. В S существует ряд так называемых *фундаментальных понятий* (или терминов).

2б. Все другие понятия (или термины), встречающиеся в S , состоят (или являются *определяемыми*) из этих фундаментальных понятий (или терминов).

3а. В S существует ряд так называемых *фундаментальных предложений*.

3б. Все другие предложения S *происходят от* или *основываются* (или являются *доказуемыми* или *очевидными*) на этих фундаментальных предложениях.

4. Все предложения S являются *истинными*.

5. Все предложения S являются *универсальными* и *необходимыми* в том или ином смысле.

6. Все предложения S *известны как истинные*. Нефундаментальное предложение известно как истинное через его *доказательство* в S .

7. Все понятия или термины S являются *достаточно известными*. Нефундаментальное понятие достаточно известно через его структуру (или определение).

Классическая модель науки является последующей реконструкцией и суммирует идеал научного объяснения исторического философа [3].

Фундаментальное означает, что «все предложения и все понятия (или термины) касаются *определенного ряда объектов* или являются *чем-то относящимся к определенной области живых существ*».

Эта точка зрения науки вместе с экспоненциальным ростом научного знания ведет к высшей степени изолированности и специализации наук, которые могут хо-

рошо работать в одних случаях, тогда как в других случаях ее (науки) цель и значение могут подвергаться сомнению. Одним примером может быть медицина с ее многими узкими специализациями, где пациент подвергается действию различных специалистов, которые видят состояние его здоровья только в одной определенной области и назначают лекарства так, как если бы пациент представлял собой сущность разделенных областей, а не один единый организм, в котором все различные области перекрываются и взаимодействуют. Такая изолированность имеет свои исторические корни в ограничениях человеческих возможностей для обработки информации. Современные информационно-вычислительные сети представляют надлежащие (надежные) средства для обработки и обмена информацией, и они дают возможность создавать сложные структуры знания, в которых эксперты различных специальностей могут адаптивно взаимодействовать и принимать хорошо информированные решения, учитывая существующее знание из других областей.

Информационно-коммуникационная технология (ИКТ) уже влияет на способ выполнения исследования, и производится научное знание, которое сделает «классическую науку» просто элементом сложной структуры производства знания.

После периодов идеализаций и изолированности науки начинают изучать свои собственные корни, предположения и взаимные связи (контексты, осведомленность). Наука, использующая компьютер, походит на электрифицированную индустрию – новый мир возможностей противопоставляет себя миру, предшествующему ИКТ. Вместе с современными технологиями управления информацией и знанием науки могут позволить себе принимать во внимание и моделировать сложность реального мира с огромным разнообразием параметров, структур и сложных динамик. Вместо принятия идеализированных, замороженных слоев реальности мы можем делать реалистичные модели, моделировать и изучать динамику сложных систем и их взаимодействий. Шаг от «старого», «довычислительного» мира к новому миру ИКТ не является обычным. Часто возникают концептуальные путаницы относительно различных видов знания, области их применения, основных предположений, взаимосвязей с другим возможным знанием и т. п.

ИНФОКОМПЬЮТАЦИОНАЛИЗМ КАК СТРУКТУРА ДЛЯ СЛИЯНИЯ ЗНАНИЯ

Инфокомпьютационализм (ICON) – это точка зрения, в соответствии с которой физический мир (вселенная) на фундаментальном уровне может пониматься как информационная структура, динамики которой являются вычислительным процессом. Материя/энергия в этой модели заменяется информацией/вычислением; материя (структура) соответствует информации, в то время как динамики – постоянные изменения в информационной структуре – являются вычислительными процессами. С этой точки зрения, мир представляет собой огромную компьютерную сеть, которая с помощью физических законов «вычисляет» свое собственное последующее состояние (панинфокомпьютационализм [2]). Информация – это материя мира. Моментальный «снимок» мира раскрывает структуру. Изменения являются вычислительными процессами. Вычисление – это просто обработка информации.

ICON сочетает панкомпьютационализм (натуральный компьютеризм) с панинформационализмом. Короче говоря, ICON – это состоящий из двух аспектов подход, основанный на двух фундаментальных понятиях: информация и вычисление. Согласно этой точке зрения, как считают многие представители вычислительной сферы [2, 4-7], мир представляет огромную вычислительную систему (или сеть вычислительных процессов), которая посредством физических законов вычисляет свое собственное состояние. *Нужно подчеркнуть, что вычислительный мир не является идентичным (или сводимым к) современному компьютеру.* Скорее вычисление – это то, что делает весь мир во время обработки своей собственной информации, просто следуя естественным законам («естественное вычисление» в его различных формах). (Насколько «естественные законы» правильны и насколько случайной является отдельная дискуссия, см. в [2].)

Можно сомневаться, что идея представителя вычислительной сферы бессмысленна, и если всё есть информационное вычисление, тогда это ничего не говорит о мире. Однако заявление представителя вычислительной сферы должно восприниматься как очень полезное понятие, которое помогает пониманию мира во многих областях. Таким является информационно-вычислительный подход ICON. Мир (вселенная) является «ни чем иным, как атомами», но по некоторой точке зрения (уровень организации, уровень абстракции) может рассматриваться как атомы.

В отличие от многих других представителей (пан)компьютационализма я не полагаю, что компьютеризм обязательно подразумевает цифровое вычисление. Как указывает автор работы [7], на основном квантово-механическом уровне происходит дискретное и аналоговое, цифровое и непрерывное вычисление. Более подробно о вопросе цифрового/аналогового мира см. в [8].

Как уже подчеркивалось, физическая реальность может рассматриваться на многих различных уровнях организации. Жизнь и интеллект являются феноменами, особенно характеризующимися информационно-вычислительными структурами и процессами. Живые системы имеют способность действовать самостоятельно и хранить информацию, осуществлять поиск информации (помнить), предвидеть будущее поведение в окружающей среде с помощью накопленной информации (учиться), адаптироваться к окружающей среде (в надежде выжить). В работе [9] я представляю модель, которая помогает пониманию механизмов обработки информации и порождения знания в организме. Мнение о нас и вселенной как о сети вычислительных устройств/процессов позволяет легче подойти в вопросу о границах между живыми и неживыми существами.

INCO рассматривает нас как продвинутые вычислительные машины в постоянном взаимодействии с окружающими вычислительными процессами. Наши мозги являются информационными архитектурами, производящими вычислительные процессы на многих уровнях организации. Вычисление происходит на уровнях основных физических законов. Все, что физика может концептуализировать, вычислять, моделировать и предсказывать, может быть выражено в информационно-вычислительных терминах. Вычислительные процессы происходят на уровне молекул (с атомами и элементарными частицами в качестве структурных элементов). Уровень нервной клетки может пониматься как следующий уровень релевантности в нашем понимании

вычислительного характера умственных процессов. Нейроны организованы в сети, и с нейронами в качестве строительных блоков появляются новые вычислительные феномены на уровне нервной сети. Сложная архитектура информационных структур в мозге, применяющая различные уровни механизмов контроля, не похожа на виртуальные машины на высоком уровне, действующие на структуру ниже. То, что мы называем «информационной архитектурой», является жидкой и взаимодействующей, не очень твердой конструкцией с кристаллической решеткой, но более похожей на сети агентов.

Развитие идет в двух направлениях: анализ живых организмов как информационно-вычислительных систем/агентов, и реализация стратегий естественного вычисления в артефактах. Уроки, полученные из разработки и реализации нашего понимания живых естественных вычислительных агентов через сложный процесс совершенствований, приведут к артефактам, которые в еще более высокой степени будут способны моделировать характеристики живых организмов.

ИНФОРМАЦИОННАЯ И ВЫЧИСЛИТЕЛЬНАЯ НАУКИ И ФИЛОСОФИЯ

Обсуждая связь между философией и науками, можно найти поучительный пример в области вычисления и философии (CAP – Computing and Philosophy). Далее приводится перечень по некоторым научным областям, представленным на конференциях CAP: *философия информации; философия вычисления; вычислительные подходы к проблеме разума; философия вычислительной технологии; реальное и виртуальное, моделирование, имитации, эмуляции; вычислительная и информационная этика; социальные аспекты вычисления и информационной технологии; философия сложности; вычислительная метафизика; вычислительная эпистемология; основанное на компьютере обучение и преподавание*. Из этого перечня очевидно, что область CAP представляет форум для обмена перекрестно-дисциплинарным, междисциплинарным и многодисциплинарным знанием и установления связей между существующими областями знания и философского размышления относительно этих областей. Это развитие нового корпуса знания сопровождается определенным сдвигом парадигмы в механизмах производства знания [10]. Глобализация, организация информационной сети, плюрализм и разнообразие, выраженные в перекрестно-дисциплинарном исследовании в сложной сети мирового порождения знания, являются феноменами, которые необходимо рассмотреть на высоком уровне абстракции, предлагаемом философским дискурсом. Примеры философских подходов, тесно связанных с происходящим сдвигом парадигмы, можно найти в [2,4,7, 11-14].

Для понимания разнообразных важных аспектов происходящего информационно-вычислительного поворота и способности развивать знание и технологии центральными являются диалог и исследование различных аспектов вычислительных и информационных феноменов. Принимая информацию как фундаментальную структуру, а вычисление как обработку информации (информационная динамика), можно их рассматривать как дополняющие друг друга, взаимно определяющие феномены. Никакая информация не возможна без вычисления (информационная динамика), и нет вычисления без информации [9, 15].

ЗНАНИЕ КАК СЛОЖНАЯ ИНФОРМАЦИОННАЯ АРХИТЕКТУРА. НЕОБХОДИМОСТЬ МНОГОДИСЦИПЛИНАРНОГО ДИАЛОГА

Почему важно развивать многодисциплинарный дискурс, который будет представлять отход от монолитной «классической модели науки», вызванный разнообразием областей, методов и уровней организации/уровней абстракции? Основной причиной является эпистемология – многодисциплинарность обеспечивает фундаментальную структуру, подходящую для общего понимания и коммуникации между несопоставимыми в настоящее время областями. Этот аргумент строится на представлении знания как информационной конструкции. Согласно работе [16], данные представляют ряды разрозненных фактов и наблюдений, которые превращаются в информацию посредством анализа, перекрестного цитирования, отбора, сортировки и суммирования данных. В свою очередь, модели информации могут быть превращены в знание, которое состоит из организованного корпуса информации. Конструктивистская точка зрения автора работы [16] подчеркивает два важных факта:

- продвижение от данных к информации, к знанию включает на каждой ступени ввод работы,
- на каждой ступени этот ввод работы ведет к росту организации, таким образом производя иерархию организации.

Исследование сложных феноменов [14] привело к пониманию, что научные проблемы имеют много различных аспектов, к которым можно подойти по-разному на разных уровнях абстракции, и что каждая область знания имеет определенную сферу обоснованности.

Это новое понимание многомерного, многослойного, интеллектуального пространства знания феноменов среди прочих стало результатом видения вселенского вывода относительно научных войн путем признания необходимости полной и сложной архитектуры знания, которая признает важность многообразия подходов и типов знания (см., например, [17]). Основываясь на источниках философии, социологии, теории сложности, теории систем, когнитивной науки, эволюционной биологии и нечеткой логики, авторы работы [17] представляют новую междисциплинарную перспективу для *самоорганизующихся сложных систем*. Они анализируют связь между процессом самоорганизации и его средой/экологией. Два центральных фактора – это роль информации в создании сложных структур и разработка топологий пространств возможного выхода. Авторы приводят доводы в пользу продолжающегося развития от внезапно появляющихся сложных порядков (последовательностей) в физических системах до когнитивных способностей живых организмов, сложных структур человеческой мысли и до культур. Это новое понимание единства междисциплинарного знания, единства в структурном многообразии также можно найти в [14].

В сложной информационной архитектуре знания, логики, математики, квантовой механики, термодинамики, теории хаоса, космологии, комплексности, происхождения жизни, эволюции, познания, адаптивных систем, интеллекта, сознания, интеллектуальных обществ и их производства знания и других артефактов ... вместе существуют два основных феномена: *информация* и *вычисление*, которые обеспечивают структуру для тех фрагментов составных картинок-загадок знания, чтобы собрать их вместе в сложное и динамичное единое информационно-вычислительное представление.

Корпус знания и практики вычислительных и информационно-технологических наук как новая научная область выросла вокруг артефакта — компьютера. В отличие от старых научных дисциплин, особенно физики, которая имеет глубокие исторические корни в естественной философии, научная традиция в вычислительном сообществе до сих пор была, в первую очередь, сконцентрирована на решении проблем и не развила очень тесных связей с философией.* Открытие философского значения вычисления как в философских, так и в вычислительных сообществах привело к разнообразию новых и интересных пониманий с обеих сторон.

Точка зрения, что информация — это центральная идея вычисления/информатики, является показательной как в научном плане, так и в социологическом. В научном плане она предлагает рассматривать информатику как обобщение информационной теории, которая касается не только передачи/коммуникации информации, но также ее трансформации и интерпретации. В социологическом плане она предполагает параллель между индустриальной революцией, которая связана с использованием энергии, и информационной революцией, которая связана с использованием информации [10].

РЕЛЕВАНТНОСТЬ ФИЛОСОФИИ ДЛЯ НАУК И НАУК ДЛЯ ФИЛОСОФИИ

Развитие философии иногда понимается как ее опреледеление новых научных областей, а затем предоставление их наукам для дальнейшего исследования (Лекция Флориды в шведском национальном курсе «Философия информации» о развитии философии [19]). Одновременно философия традиционно учится (черпает знания) у наук и технологий, используя их как средства для производства наиболее надежного знания о фактическом положении дел в мире. Мы можем привести недавний пример современного прогресса в моделировании и имитации мозга и познания, что жизненно важно для философии разума. Как уже много раз было в истории, первый самый лучший подход это тогда, когда существует дефицит эмпирического знания, интуитивный подход совсем не обязательно должен быть самым лучшим. Например, автор работы [20] указывает, что наука является *не естественной формой мышления*, и она очень часто производит контринтуитивное знание, порождаемое экспериментами с миром, осуществляемыми с помощью средств, отличающихся от повседневных, опытов, экспериментов в микрокосмосе, макрокосмосе и других областях, скрытых для нашего лишенного помощи познания.

Полезным примером «неестественного» характера научного знания является полностью контринтуитивное открытие астрономии, что Земля вращается вокруг Солнца, а не наоборот, как подсказала бы нам наша интуиция. В настоящее время революция, подобная революции Коперника, кажется, продолжается в философии разума, эпистемологии (понимаемой в информационно-технологических терминах), в философии информации и философии вычисления. Недавно опубликованная книга [21] приводит справедливые доводы в пользу философии (и особенно метафизики), информированной последними разработками в специальных науках, вместо априорных

интуиций философов и здравого смысла. Специальные науки, такие как когнитивная наука и нейронаука, например, накопили ценное знание, которое должно быть принято философией разума вместо того, чтобы полагаться на исторические идеи, основанные на здравом смысле. Процесс идет в обратном направлении, от философии к специальным наукам, что несколько раз упоминалось в данной статье в примере ICON. Кроме того, должна быть упомянута этика как отрасль растущей важности философии для наук.

Сегодня науки понимаются как рациональные средства для сжатия информации, которые дают нам возможность предсказывать и эффективно обращаться с различными феноменами, а не с источником (как полагали ранее) *абсолютной истины* относительно мира, открывающей пророческий план во время чтения «книги природы». Вместо этого они (науки) говорят нам, по крайней мере, так много о том, кем мы являемся по отношению к миру — в таком новом свете сознательное размышление над нашими ценностями, мотивами, приоритетами и способами является жизненно важным. Особенно, когда это касается прикладной науки и технологии, очевидна релевантность этического анализа. Компьютеры как наши основные средства обработки информации в научном исследовании в других случаях могут использоваться во благо или во вред, а вновь разработанные технологии могут быть неправильно использованы. Вызов радикально новых прикладных наук и технологий состоит в том, что Джеймс Мур [22] называет «политическим вакуумом», — т.е. часто невозможно предсказать, каким образом технология может быть неправильно использована и какие конечные результаты это может иметь. Этическое суждение, без сомнения, необходимо, и в конце концов даже в образовательной системе научная этика, вычислительная этика, информационная этика и другие специфические ветви прикладной этики, сопутствующие специальным наукам и технологиям, должны стать неотъемлемой частью системы образования и науки.

ОТВЕТЫ НА НЕКОТОРЫЕ ОБЫЧНОГО РОДА КРИТИКИ ОТНОСИТЕЛЬНО ICON

Точка зрения представителя информационно-вычислительной сферы может интерпретироваться как заявление, что весь мир — это «не что иное, как (вычислительная) машина, и что мы, люди, являемся основными машинами без свободы воли или чувств. Очевидно, это не тот случай. Представление, что мир есть информационно-вычислительный феномен, означает, что мир в том виде, в каком он есть, можно понимать и моделировать на информационно-вычислительном языке. Чувства, качества (характеристики, определения) и другие ментальные феномены являются внезапно появляющимися свойствами физического мира, которые являются информационно-вычисляемыми.

Роль различных парадигм в нашем понимании мира может быть проанализирована на исторических примерах. В прошлом произошло несколько основных переходов парадигм: от мифическо-поэтических к механистическим, к появлению вычислительного понимания мира. Следовательно, мы можем задать тот же самый вопрос относительно мифическо-поэтического и механистического мира: было ли такое понимание мира истинным? было ли оно реальным или просто метафорическим? Для мифическо-поэтического мира ответ является простым — это была метафора. Хотя механицизм в первую очередь был точкой зрения на неорганическое

* Алан Тьюринг (Alan Turing) являлся одним из заметных исключений из правил. Другими были Вейзенбаум, Виноград и Флорес [18]. Следует отметить, что вычисление всегда имело сильные связи с логикой, и что особенно искусственный интеллект всегда признавал философские аспекты.

вещество, и механистические подходы к роботам (механическим квазилюдям) не работали для любой другой цели, кроме развлечения, но механистическое мировоззрение, тем не менее, помогло нам много узнать о мире, в первую очередь неорганическом, а также много основных фактов о живом мире (таких как тот, что *не существует никаких жизненных энергий*, но те же самые физические механистические законы управляют всем физическим миром, как живым, так и неживым).

Параллельное развитие идет теперь в направлении компьютеризации. Мы узнаем об информационных и вычислительных ресурсах и возможностях мира, и мы будем развивать даже более мощные способы изучения, через интеллектуальные системы мы будем успешно их совершенствовать.

Знание, что биологические организмы (включая людей) являются «машинами», обрабатывающими информацию, не делают их менее привлекательными. Таким же образом, знание того, что мы все состоим из атомов, не означает, что мы не обладаем свободой воли, воображением и реальными чувствами. Понимание процессов фундаментального уровня не делает музыку, искусство и философию устаревшими.

ICON помогает нам как предоставлением средств для знания и производства артефактов, так и похожих средств для понимания природных феноменов и артефактов на многих различных уровнях. Вот почему философия обращается к наукам, основанным на информационно-вычислительном знании. Холистическое, высокого уровня абстракции представление необходимо как самоотражающий процесс знания.

Итак, ICON совсем не обязательно является *окончательным ответом* на «жизнь»! «мир (вселенную)»! и «всё вообще»! (который является 42-м по счету, как мы узнали от Адамса, цитируемого Винсентом Мюллером в его книге), а служит средством познания, которое снова поможет нам достичь единства знания на одном определенном уровне абстракции – информационно-вычислительном.

Таким образом, ответ (ICON) не является *окончательным ответом*, но все равно кажется приемлемым ответом на разумный вопрос: как достичь общего языка для несопоставимых специальных областей, который может обеспечить взаимное понимание и создание нового знания?

ЛИТЕРАТУРА

1. *Pivčević E.* The reason why: A theory of philosophical explanation. - KruZak, 2007.
2. *Chaitin G.J.* Epistemology as information theory. Alan Turing Lecture given at E-CAP 2005 /G. Dodig-Crnkovic, S. Stuart (eds.) // Computation, information, cognition – The Nexus and The Liminal. – Cambridge Scholars Publishing, Cambridge UK, 2007. – <http://www.cs.auckland.ac.nz/CDMTCS/chaitin/ecap.html>
3. *Betti A., De Jong W.R.* (guest edit). The classical model of scientific rationality, forthcoming in *synthese*, Special issue. – <https://www.surfgroupen.nl/sites/cmsone/com-seneweb/index.html>

4. *Zuse K.* Rechnender Raum //Elektronische Datenverarbeitung. – 1967. – Vol. 8. – P. 336-344.
5. *Fredkin E.* Digital philosophy (and references therein). – <http://www.digitalphilosophy.org/Home/Papers/tabid/61/Default.aspx>
6. *Wolfram S.* A new kind of science. - Wolfram Science, 2002.
7. *Lloyd S.* Programming the Universe: A quantum computer scientist takes on the Cosmos. – Jonathan Cape, 2006.
8. *Dodig-Crnkovic G.* Investigations into information semantics and ethics of computing. – Mälardalen University, 2006. – http://www.idt.mdh.se/personal/gdc/work/shifting_paradigm_singlespace.pdf
9. *Dodig-Crnkovic G.* Epistemology naturalized: The info-computationalist approach// APA Newsletter on philosophy and computers. – Spring 2007. – Vol. 06, No. 2. – <http://www.apa.udel.edu/apa/publications/newsletter/v06n2/Computers/04.asp>
10. *Dodig-Crnkovic G.* Shifting the paradigm of the philosophy of science: The philosophy of information and a New Renaissance// Minds and Machines. – 2003. – Vol. 13, No. 4. – <http://www.springerlink.com/content/g14t483510156726>; http://www.idt.mdh.se/personal/gdc/work/shifting_paradigm_singlespace/pdf
11. *Floridi L.* Open problems in the philosophy of information// Metaphilosophy. – 2004. –Vol. 35, No. 4.
12. *Floridi L.* (ed.). The Blackwell guide to the philosophy of computing and information (Blackwell philosophy guides). - 2004.
13. *Mainzer K.* Computerphilosophie zur Einführung. – Junius, Hamburg, 2003.
14. *Mainzer K.* Thinking in complexity. – Springer, Berlin, 2004.
15. *Dodig-Crnkovic G., Stuart S.* (eds.). Computation, information, cognition – The Nexus and the Liminal. – Cambridge Scholars Publishing, Cambridge, UK, 2007. – <http://www.idt.mdh.se/ECAP-2005/Intro-Preface-ComputationInformationCognition.pdf>
16. *Stonier T.* Information and meaning. An evolutionary perspective. – Springer, Berlin, New York, 1997.
17. *Smith J., Jenks C.* Qualitative complexity ecology, cognitive processes and the re-emergence of structures in post-humanist social theory. – Routledge, 2006.
18. *Lanzarone G.A.* In search of a new agenda// APA Newsletter Computing and Philosophy. – Fall 2007.
19. PI, Swedish National Course, 2004. – http://www.idt.mdh.se/~gdc/PI_04/index.html
20. *Wolpert L.* The unnatural nature of science. – Harvard University Press, 1993.
21. *Ladyman J., Ross D., Spurrett D., Collier J.* Every thing must go: Metaphysics naturalized. – Oxford University Press, 2007.
22. *Moor J.* What is computer ethics? // Metaphilosophy. – 1985. – Vol. 16, No. 4. – P. 266-275.

Информационные метрики (iMetrics): научная сфера с социокогнитивной идентичностью?*

Сташа МИЛОЕВИЧ

(Staša MILOJEVIĆ)

Школа библиотековедения и информатики, Университет шт. Индиана, г. Блумингтон, США

Лут ЛЕЙДЕСДОРФ

(Loet LEYDESDORFF)

Амстердамская школа коммуникационных исследований, Амстердамский университет, г. Амстердам, Нидерланды

«Библиометрия», «наукометрия», «информетрия» и «вебометрия» могут все рассматриваться как примеры одной научной сферы с одинаковыми задачами и методами, которые мы называем «информационными метриками», или iMetrics. Это исследование изучает когнитивное и социальное отличие iMetrics от общей информатики, фокусируясь на ядре исследователей, совместно использующих словарную и литературную/интеллектуальную базу. Наш анализ исследует сходства и различия между четырьмя массивами документов. Массивы документов получаются из трех ядерных журналов для исследования iMetrics (Scientometrics, Journal of the American Society for Information Science and Technology (JASIST) и Journal of Informetrics). Мы распределили JASIST на массивы документов, содержащие статьи по темам iMetrics и статьи по общей информатике. Объем публикаций в таком представлении этой сферы быстро увеличивался в течение последнего десятилетия. Таким образом, может быть определено ядро исследователей, изначально сфокусированных на темах iMetrics. Эта ядерная группа разработала совместный словарь, что выражается в высокой степени сходства слов названий, и в том, что используется интеллектуальная база. Научный авангард этой области движется быстрее, чем в информатике в целом, подводя ее ближе к мечте Прайса.

ВВЕДЕНИЕ

Такие термины, как «библиометрия», «наукометрия», «информетрия» и «вебометрия», использовались для описания количественных исследований библиографий (книг и библиотек), науки, информационных феноменов и всемирной паутины. Хотя эти термины появились в различных контекстах и возникли из различных дисциплинарных предпосылок, они явно быстро стали использоваться как взаимозаменяемые. Библиометрия, например, имеет корни в библиотековедении и информатике. Термин «библиометрия» сам был впервые введен Причардом [1] для описания «применения математических и статистических методов к книгам и другим средствам коммуникации» [1, с. 348]. Другие авторы [2-4] дали свои собственные определения, некоторые из них (Уайт и Маккейн

и Боргман и Фурнер, [2,4]) связывали библиометрию с исследованием науки и научной коммуникации.

Наукометрия, определяемая как количественные исследования науки [5] или «количественное исследование науки, коммуникации в науке и научной политики» [6, с. 75], имеет свои корни в 1950-х и 1960-х гг. и возникает из работы историка науки Дерекы де Солла Прайса [7,8] параллельно с развитием индексов цитирования Юджина Гарфила [9,10]. Первый международный журнал *Scientometrics*, специализирующийся на библиометрии и количественных исследованиях науки, появился в 1978 г. В ее ранние годы наукометрия рассматривалась как количественный аспект исследований науки и технологии, появившийся в то же самое время [11]. Тем не менее, основной фокус в исследовании науки и технологии сосредоточивался на качественной социологии науки и научно-политическом анализе [12]. Куртиаль [13], например, называл наукометрию «гибридной областью, созданной из незримого колледжа и множества пользователей» [13, с. 251]. Однако с дальнейшим развитием индекса научного цитирования как мощного нового инструмента наукометрия становилась все больше частью информа-

* Перевод Milojević S., Leydesdorff L. Information metrics (iMetrics): A research specialty with a socio-cognitive identity?—<http://arxiv.org/ftp/arxiv/papers/1209/1209/3406.pdf>

тики на протяжении 1980-х и 1990-х гг., а в ряде недавних исследований наукометрия рассматривалась как часть информатики [14-16].

Информетрия, определенная Эгге [17, с. 1311] как научная область, «включающая все метрические исследования, относящиеся к информатике», стала использоваться как термин в конце 1980-х гг. [18]. Информетрия может рассматриваться шире, чем две другие области, так как она включает исследования «количественных аспектов информации в любой форме, не только записей или библиографий, и не только ученых в любой социальной группе» [19, с. 1]. Распирение интереса к этим темам было засвидетельствовано выходом нового журнала *Journal of Informetrics* в 2007 г. Наконец, *вебометрия* может рассматриваться как «применение методов информетрии к всемирной паутине» [20, с.404]; это наиболее новая ветвь из четырех. В 1997 г. был основан электронный журнал *Cybermetrics*, охватывающий, в первую очередь, вебометрическое исследование.

Подробная дискуссия относительно сходств и различий среди этих научных областей уведет нас от цели этого исследования и она широко охвачена другими авторами [21-22]. Наше общее впечатление является таким, что хотя эти области исследования имели различные корни, они развивались, чтобы охватить многие цели, и сегодня имеют множество общих методов. Де Беллис [23], например, утверждал, что они являются часто «неразличимыми». Другие исследования [24-25] используют эти термины как взаимозаменяемые. Учитывая эти фундаментальные сходства и общий акцент на *документах* как единицах анализа, мы считаем эти четыре научные области как различные обозначения, представляющие одну область исследования, и называем ее «*информационными метриками*» (*information metrics*) и сокращаем ее как «*iMetrics*»^{*}.

Учитывая это разнообразие (в происхождении и акценте), не удивительно, что многие ученые пытались расположить научную область, которую мы называем *iMetrics*, в ряду с другими областями или дисциплинами. Таким образом, в ряде исследований [12,15-16,26,27] обнаружилось сильные связи между наукометрией, чему примером служит журнал *Scientometrics*, и информатикой, чему примером служат такие журналы, как *Journal of the American society for information science and technology* (JASIST), *Journal of Documentation* (JDOC) и *Information Processing and Management* (IPM). Эти связи часто приводят к характеристике *iMetrics* как неотъемлемой части информатики [16,28], но иногда также как части «исследований науки» [29-30] или «пересечения между исследованиями науки и информатикой» [14, с. 2492]. Однако другие недавние исследования отмечали, что, по крайней мере, в *когнитивном* смысле *iMetrics* может рассматриваться как отдельная от информатики или от объединенной категории библиотекведения и информатики. Например, используя иерархическую кластеризацию терминов, определенных из названий 16 журналов по библиотековедению и информатике Милоевич и др. [31] обнаружили сильную область *iMetrics* (примером послужил журнал *Scientometrics*), которая стоит

рядом (а не внутри) с областью информатики и библиотековедения. Наблюдение Янссенса и др. [32] подтверждает, что журнал *Scientometrics* может быть в значительной степени отделен от других журналов по библиотековедению и информатике на основе «различного терминологического профиля» [32, с. 1622]. Данное исследование кроме того изучает различие *iMetrics* в когнитивном смысле, а также добавляет очень важный аспект *социального* отличия.

Для любой области исследования при рассмотрении сферы в социологическом смысле [33,34] недостаточно того, чтобы она имела определенный когнитивный профиль, ей также необходимо иметь *социальную* идентичность, т.е. ее практические деятели должны представлять собой сообщество, чьи внутренние связи гораздо сильнее, чем связи вне сообщества, даже если они изначально принадлежат к таким «внешним» социальным структурам в административном смысле (например, принадлежат к отделениям или школам по библиотековедению и информатике, вычислительной технике и т.д.). Бёрнер и др. [35] считают научно-исследовательскую сферу «самой крупной однородной единицей науки, в которой каждая область имеет свое собственное множество проблем, ядро исследователей, совместное знание, словарь и литературу» [35, с. 21]. С использованием библиометрических методов научно-исследовательская сфера может быть введена в эксплуатацию через изучение «эволюционирующего множества родственных документов» [36]. Наше исследование изучает социальную и когнитивную идентичность *iMetrics* как научно-исследовательской сферы с особым акцентом на ядре исследователей, использующих совместный словарь и литературу.

Ряд исследований фокусируется на природе того, что мы называем *iMetrics*. И хотя некоторые исследователи [24] считают ее областью в кризисе, наступившем из-за отсутствия консенсуса, вызванного, помимо других вещей, «потерей интегрирующих личностей», другие ученые [14,37] используют эмпирические данные, чтобы показать, что *iMetrics*, по-видимому, имеет социальную идентичность. Например, в их сетевом библиометрическом и социальном анализе журнала *Scientometrics* за его первые 25 лет (1978-1993 гг.) Уотерс и Лейдесдорф [37] нашли согласованную, хорошо интегрированную группу ученых со связанным дискурсом. Ван ден Бесселар [27] в анализе сводного журнала к журнальным ссылкам для трех журналов: *Social Studies of Science*, *Scientometrics* и *Research Policy* обнаружил, что кластеризация вокруг *Scientometrics* является очень неоднородной и изменяется год от года, таким образом побуждая его не согласиться с требованием Уотерса и Лейдесдорфа, что *iMetrics*, по-видимому, сформировалась в стабильную область. Совсем недавно в их исследовании четырех журналов (*JASIST*, *JDOC*, *IPM* и *Scientometrics*), применяя анализ особой комбинации слов названия статей и ссылок, Лусио-Ариас и Лейдесдорф [14] обнаружили, что все эти журналы, исключая *JDOC*, показывают признак «взаимодействия на уровне области» [14, с. 2495], который проявляется сходством в «тематическом пространстве».

Некоторые различия в этих выводах относительно статуса того, что мы называем *iMetrics*, возникают из-за неадекватного различия между когнитивными и социальными аспектами. Другие различия могут относиться к предположениям отдельных исследований. А именно, некоторые исследователи использовали *Scientometrics* как первоисточник или как единственный журнал для исследования природы *iMetrics*, не связывая его с какой-либо окружающей дисциплиной априори [37], хотя другие

^{*} Мы будем также использовать термин *iMetrics* для обсуждения результатов предыдущих исследований. Мы осознаем, что авторы не могли использовать этот термин, поскольку он вводится в этой статье. Тем не менее, предыдущие исследования часто охватывают точно такую научную область, которую мы предлагаем называть здесь *iMetrics* для целей краткости и ясности.

предполагали, что iMetrics является частью исследования науки и технологии [15,27] или информатики [14]. Наконец, некоторые различия возникают из-за использования журналов в качестве единиц анализа, и, таким образом, неполного рассмотрения того, что журналы могут часто и в разной степени охватывать несколько научных областей, из которых одной является iMetrics [38].

Целью настоящего исследования является изучить социальные и когнитивные отличия iMetrics от общей информатики. Так как различия являются относительным свойством, необходим образец естественного уровня неоднородности для представления в научно-исследовательской области. Таким образом, наш анализ вращается вокруг изучения сходств и различий между четырьмя массивами документов (т.е. статей) – три принадлежат iMetrics и один состоит из статей по общей информатике без iMetrics. Массивы документов iMetrics получаются из трех журналов, которые публикуют большую часть исследований по iMetrics. Затем мы изучаем сходства между этими тремя массивами документов iMetrics, чтобы установить внутренний уровень неоднородности исследования iMetrics, и потом сравниваем каждый массив с документами не по iMetrics, чтобы определить, если да, то в какой степени исследование iMetrics может рассматриваться отличным от информатики. Наш подход пользуется преимуществом того факта, что большая доля первоначального и представительного исследования iMetrics опубликована в очень малом числе издаваемых выпусков – которые мы называем ядерными журналами по iMetrics – из-за преобладающей ассиметричности в наукометрических распределениях [39]. Также мы используем относительно простой все еще эффективный метод определения литературы по iMetrics и не iMetrics, который облегчает вопросы, представленные в некоторых предыдущих исследованиях природы iMetrics.

ДАННЫЕ И МЕТОДЫ

Понятие ядерных журналов по iMetrics и массивы документов

Научная область iMetrics испытала быстрый рост публикаций с 1990-х гг. [21,40]. Эта область также характеризуется специализированными журналами [24]. Ими являются, в первую очередь, *Scientometrics* (SCI) и *Journal of Informetrics* (JOI). *Scientometrics* начал издаваться в 1978 г. и был первым журналом, посвященным исключительно количественным исследованиям науки. Журнал *Journal of Informetrics* является более поздним (2007 г.). Специализированные журналы служат не только для передачи и архивирования исследований, но и являются способом определения границ дисциплинарных или исследовательских областей, фактом, который мы будем использовать в этом исследовании. Более того, значительное число статей по iMetrics опубликовано в журнале JASIST, который также охватывает более общие темы по информатике. JASIST начал печататься в 1950 г. (первоначально издавался под названием *Journal of the American Society for Information Science*)*.

Ключом нашего метода является определение нескольких выпусков, публикующих большое число статей по iMetrics, которое будет определять массивы данных iMetrics, чью согласованность мы изучим путем сравнения их друг с другом, а также с массивом данных не по

iMetrics. Мы получаем эти наборы данных из трех ядерных журналов по исследованию iMetrics: *Scientometrics*, *Journal of Informetrics* и JASIST. Ядерные журналы очень важны для формирования областей, позволяя «координацию коммуникации и доступ к репутации... интеллектуальному взаимодействию и творчеству» [41, с. 775]. Мы рассматриваем эти три журнала как ядерные по iMetrics, так как они публикуют большую часть оригинального iMetrics-исследования. Журналы SCI, JASIST и JOI насчитывают 3/4 всех статей по iMetrics, опубликованных в журналах, классифицируемых как журналы по библиотековедению и информатике в *Journal Citation Reports* 2010 г. в базе данных Web of Science (WoS).

Мы признаем тот факт, что имеются другие журналы, публикующие статьи по iMetrics, наиболее известные из которых: *Research Evaluation*, *Information Processing and Management*, *Journal of Information Science* и *Research Policy*. Кроме того, они все имеют тесные когнитивные связи с тремя журналами по iMetrics, которые мы считаем ядерными [42]. Однако объем статей по iMetrics, публикуемых в каждом из этих журналов, значительно ниже, чем в SCI, JASIST и JOI. Применяя тот же самый метод различения статей по iMetrics, который мы использовали в JASIST (описанный в следующем разделе), мы обнаружили, что *Research Evaluation* имеет в два раза меньше статей по iMetrics, чем самый маленький из ядерных журналов, JOI.

Наш выбор ядерных журналов также поддерживается наблюдениями в других исследованиях. Например, в своем обзоре литературы по «информетрии» (что мы называем iMetrics) Бар-Илан [25] обнаружила, что *Scientometrics* и JASIST имеют самое большое число информетрических статей, а *Research Policy* находится на третьем месте (JOI не мог быть в это время включен в ее исследование). Более того, мы основываемся на трех ядерных журналах вместо большего числа журналов, так как метод, нами применяемый, базируется на попарных сравнениях, так что дополнительные многие массивы данных сделают сравнения тяжелыми.

Кроме того, мы рассматриваем SCI и JOI полностью специализированными в iMetrics и поэтому используем их для операционного определения когнитивной области iMetrics. Используя это определение, мы затем разделим статьи в JASIST путем двухъярусной процедуры, подробно объясняемой в следующем разделе, на статьи, принадлежащие iMetrics и не принадлежащие ей. Статьи по iMetrics из JASIST, все статьи из SCI и из JOI определяют три массива документов iMetrics. Статьи не по iMetrics из JASIST определяют четвертый массив документов. В этом исследовании мы будем сравнивать массивы документов iMetrics между собой (3 сравнения) и каждый из них с массивом документов не по iMetrics (другие три сравнения). Заметим, что многие исследования используют журналы целиком как единицу анализа, а не массивы статей, что затем приводит к результатам, трудным для интерпретации, когда включены журналы с различной шириной фокуса.

Для нашего исследования достаточно, чтобы у нас имелось несколько, разумно крупных, наборов статей по iMetrics с тем, чтобы установить внутренний уровень неоднородности iMetrics. В чем нет необходимости или что неопосредованно с нашим методом, так это определить всевозможные статьи по iMetrics из журналов, выходящих за рамки ядерных. Соответствующее предположение состоит в том, что большинство тем по iMetrics представлено в ядерных

* Между 1950 и 1970 гг. журнал JASIS(T) выходил под названием *American Documentation*.

журналах и что большинство активных исследователей в сфере iMetrics публикуется, по крайней мере иногда, в ядерных журналах этой сферы.

Данные, определяющие массивы документов

Мы взяли из базы данных Thomson Reuters' Web of Science полные характеристики для всех публикаций в SCI, JASIST и JOI – 8280 записей.* Из этого массива мы выбрали только научные статьи, т.е. публикации, классифицируемые как «статья» или «доклад конференции», так как эти два типа документов содержат оригинальные результаты исследований. Получилось 6092 таких записей. По причинам, описанным ниже, только статьи, опубликованные после 1982 г., подошли для анализа. JOI начал издаваться в 2007 г. и учитывался по необходимости. Всего было 2159 и 189 научно-исследовательских статей в SCI и JOI, соответственно, и они были предварительно определены как два из трех массивов документов по iMetrics.

Чтобы определить статьи по iMetrics в JASIST (третьем массиве документов по iMetrics), мы применили следующую двухъярусную процедуру. Во-первых, мы рассматривали любую статью журнала JASIST, которая содержала ссылки или на SCI, или на JOI как статью по iMetrics, т.е. статьи в SCI и JOI использовались как критерий для выявления статей JASIST. Это кажется разумной процедурой, но возникает вопрос, является ли это надежным в случаях, когда, скажем, только одна ссылка указывает на SCI или JOI. Мы проверили это, изучая темы каждой десятой статьи JASIST, ссылающейся на SCI или JOI только один раз. Все эти статьи однозначно имели отношение к iMetrics. Этот основанный на ссылках отбор дал 511 статей по iMetrics из JASIST.

Вышеуказанный метод мог описать только статьи JASIST, которые были опубликованы, так как SCI начал печататься в 1978 г. Самая ранняя статья JASIST, которая содержит ссылку на SCI, относится к 1982 г. Несомненно, прошло несколько лет, пока SCI стал «видимым» для аудитории, печатающейся в JASIST. Поэтому мы ограничили анализ всех данных периодом после 1982 г., т.е. периодом, после которого в описании статей JASIST стало возможным использование этого метода.

Хотя этот метод цитирования обеспечил очень чистую выборку статей по iMetrics в JASIST, он опускает статьи, которые не ссылаются на статьи SCI или JOI, но которые, тем не менее, могут рассматриваться как относящиеся к iMetrics. Чтобы найти упущенные статьи-кандидаты, мы дополнительно отобрали статьи JASIST после 1982 г., которые содержат одно из следующих семи частых, свойственных iMetrics слов или два префикса в названии: «цитирование», «библиометрический», «наукOMETрический», «показатель», «продуктивность», «отображение» или «цитировать», а также два префикса: «h-» или «со-». Мы определили значимость этих слов и префиксов путем анализа наиболее часто встречающихся слов в названиях статей из SCI и JOI после того, как были опущены нехарактерные слова и запрещенные слова. По этому критерию отбора найдена 81 дополнительная статья в JASIST, опубликованная после 1982 г. После ручной проверки того, какие из статей-кандидатов действительно относятся к iMetrics, мы удалили 19 статей, которые к ней не относятся.** Поэтому окончательный массив ста-

тей по iMetrics, опубликованных в JASIST, содержал 573 статьи, которые очень точно отобраны с использованием нашего большей частью неконтролируемого метода. Чтобы отличать этот массив статей от общего массива JASIST, мы обозначили его JASIST-iM.

Для определения уровня социокогнитивного отличия iMetrics нам необходима представительная и относительно четкая выборка статей не по iMetrics. Мы определяем, что этот окончательный массив документов содержит 2104 статьи JASIST, опубликованные после 1982 г., которые не были отобраны как касающиеся iMetrics. Это предположительно статьи, охватывающие другие аспекты информатики.*** Мы назвали массив статей не по iMetrics JASIST-O, где «O» подразумевает «другие».

Хотя наш массив данных охватывает три десятилетия (1982-2011 гг.), направления за этот период представлены только для того, чтобы зафиксировать этап остальной части анализа. Остальная часть анализа будет основываться на последних пяти годах (2007-2011).**** Другими словами, мы заинтересованы в свойствах современного периода, а не в динамике. Построение статистической картины, тем не менее, требует достаточно длинного временного интервала для того, чтобы все основные действующие лица имели возможность быть представленными в структуре, однако период не должен быть настолько длинным, чтобы подвергнуться изменениям. Пятилетний период отвечает этим критериям. Период после 2007 г. соответствует периоду, в течение которого все три ядерных журнала издаются, и, таким образом, позволяет нам исследовать их последовательно.

В итоге, 2921 научная статья, опубликованная между 1982 и 2011 гг., была определена как относящаяся к iMetrics и включена в анализ (573 в JASIST-iM, 2159 в SCI и 189 в JOI). Для анализа периода 2007-2011 гг. мы использовали 1221 из этих 2921 статьи по iMetrics: 265 из JASIST-iM, 767 из SCI и 189 из JOI. Кроме того, было 569 статей из JASIST-O за этот период.

Другая обработка данных

Чтобы изучить социальную идентичность iMetrics, мы устранили неясности в именах авторов, используя фамилии и первые инициалы. Число авторов, определенных в различных массивах документов (без учета места в списке авторов и только первые авторы) для периода 2007-2011 гг., представлено в табл.1.

Наконец, чтобы проанализировать темы статей на основе их названий, проводилась следующая процедура. Во-первых, мы отбрасывали пунктуацию из названий и затем использовали программное обеспечение WordStat для

являются: «Представление национальной политической свободы в разработке сетевого интерфейса: показатель», «Имеет ли значение область знания: проявление опыта пользователей в их информационных взаимодействиях» и «Облегчение поисковой неточности через понятийные ассоциации: автоматическое индексирование, анализ совместной встречаемости и параллельное вычисление».

*** Этот массив документов может страдать от некоторой «засоренности», от неопределенных статей по iMetrics, т.е. таких, которые не ссылаются на SCI или JOI и не содержат в себе девяти характерных для iMetrics слов. Мы имели возможность оценить процент засоренности массива не по iMetrics в 4%. Это приемлемый уровень, который не может, по-видимому, компрометировать результаты этого исследования.

**** Отметим, что данные для финального года (2011) были неполными ко времени написания этого исследования и не будут учитываться (т.е. данные 2011 г. будут пропущены) в направлениях, включая абсолютные величины.

* Данные загружены 20 августа 2011 г.

** Некоторыми примерами статей, отобранных с использованием ключевых слов названий, но не принадлежащих iMetrics,

объединения вариантов слов (мн. ч. и т.д.). Затем мы использовали WordStat для определения всех фраз, встречающихся три или более раз. Фразы могут быть длиной до пяти слов. Мы затем создали совместный список частоты слов или фраз (т.е. терминов), исключались запрещенные слова и общие слова. Эта процедура объясняется в работе [31]. Чтобы определить, доминирует ли некоторый термин в одном массиве документов над другими, мы просили, чтобы его частота в одном массиве документов в сравнении с другими была выше 50%. Если термин однозначно характеризовался во всех трех массивах документов, его вклад в каждый массив документов составлял 33%.

Таблица 1

Массив документов	Статьи	Авторы	Первые авторы
JASIST-iM	265	359	159
SCI	767	1245	530
JOI	189	279	124
JASIST-O	569	1054	435
iMetrics	1221	1589	686

Примечание. Приводится число различных авторов (всех и первых) в четырех массивах документов (JASIST-iM, SCI, JOI, JASIST-O) за 2007-2011 гг. Числа также даны для трех массивов документов по iMetrics вместе.

Меры сходства (различия)

Для исследования социальной и когнитивной идентичности сферы iMetrics и степени ее отличия от общего сообщества по информатике мы полагаемся на различные меры сходства. Для авторов мы используем простые доли авторов, которые являются авторами в массивах документов по iMetrics или не по iMetrics, отличных от исследуемого. Для других характеристик (терминов названий и источников ссылок) мы используем косинусное сходство. Косинусное сходство является очень эффективным способом определения уровня сходства (раз-

личия) среди сложных объектов [43]. Оно, по существу, измеряет геометрическое разделение между мульти-пространственными векторами, каждый из которых представляет некоторое свойство. Чем меньше угол между векторами (тем *косинус* ближе к 1), тем более схожими они являются. Если два вектора перпендикулярны (*косинус* = 0), свойства не имеют ничего общего.

РЕЗУЛЬТАТЫ

Публикационные направления в документальных наборах по iMetrics

Перед тем как исследовать вопрос о социальной и когнитивной идентичности iMetrics, давайте сначала изучим общие публикационные направления в трех массивах ядерных документов по iMetrics. Предыдущие исследования [24,37] обнаружили рост числа публикаций в JASIST (не только статей по iMetrics) и в SCI. Мы обновляем эти направления особенно для статей по iMetrics и показываем изменения после 1982 г. (рис. 1).

На рис. 1 демонстрируется рост числа статей во всех трех массивах документов по iMetrics, особенно с середины 90-х гг. Кроме того, мы видим, что SCI и JASIST-iM увеличили долю издания статей по iMetrics с 2004 г. В целом мы наблюдаем взрыв литературы по iMetrics в ядерных журналах с числом статей в 2010 г. приблизительно в четыре раза выше, чем десятью годами ранее.

На рис. 2 сравниваются доли, которые каждый из трех массивов документов имел в массиве статей по iMetrics в трех объединенных массивах документов. Доля статей JASIST-iM упала после 1982 г. – вероятно, как следствие появления SCI, – достигнув минимума в 1988 г. Это соответствует рис. 6 в [14, с.2494]. С тех пор, тем не менее, JASIST вернул свою долю, которая сегодня почти такая же, как когда появился SCI. JOI появился в 2007 г.; это привело к падению доли SCI. В настоящее время SCI публикует примерно половину статей по iMetrics в трех массивах документов, тогда как JASIST-iM и JOI делят оставшуюся половину.

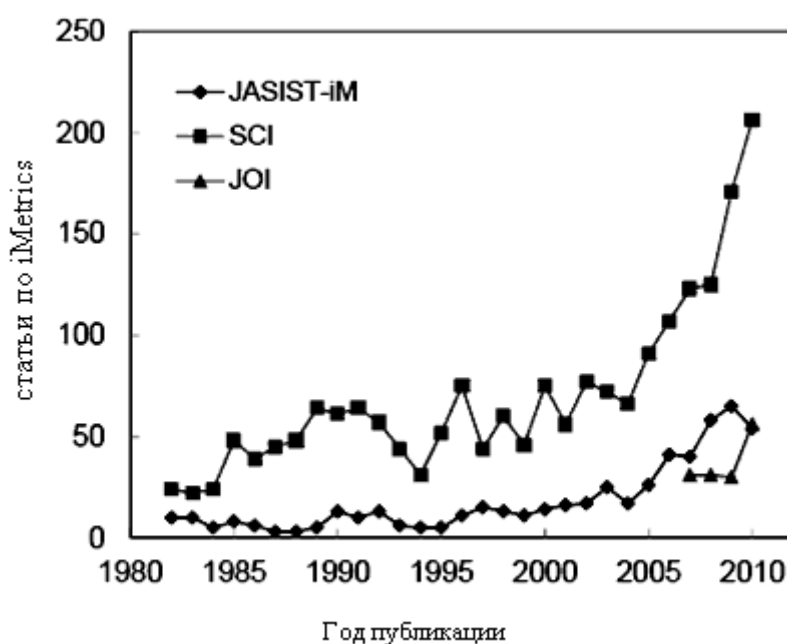


Рис. 1. Число статей по iMetrics, публикуемых в JASIST-iM, SCI и JOI ежегодно с 1982 г. JOI начал издаваться только в 2007 г.

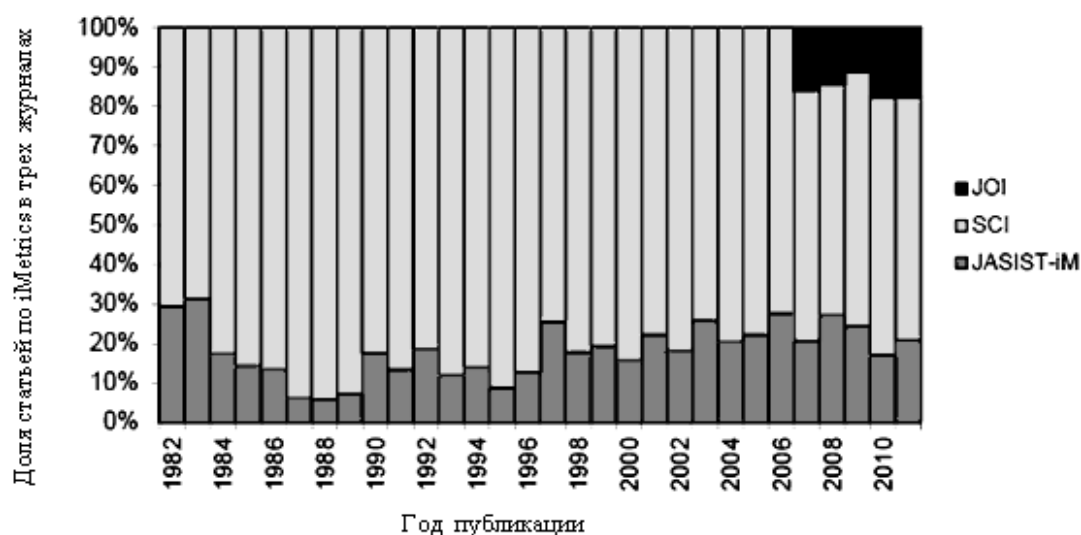


Рис. 2. Доля статей по iMetrics, опубликованных в JASIST-iM, SCI и JOI

Социальная идентичность и различие iMetrics

Социальная идентичность iMetrics как научной области может быть изучена через связи между учеными. Как указывали Крейн [44] и Прайс [7], наука практикуется в явно тесно связанных группах ученых, работающих над одними и теми же проблемами и регулярно обменивающимися информацией друг с другом. Наиболее наглядной формой формального взаимодействия является публикация научных статей в журналах. В этом отношении «взаимодействие хорошо определенных групп однородных исследователей, сконцентрированных вокруг отдельных наборов журналов, приводит к формированию сплоченных (под)групп, связанных друг с другом» [41]. По уже объясненным причинам мы не будем основывать наш анализ на отношениях между авторами и журналами, а будем концентрироваться на авторах и четырех различных массивах документов, три из которых считаются ядром в сфере iMetrics, а один представляет информатику. В анализе мы объясним, насколько сильно взаимосвязаны авторы, участвующие в исследовании по iMetrics, и отличаются ли они от авторов, публикующихся в сфере информатики.

Для определения того, есть ли у iMetrics социальная идентичность, т.е. отличаются ли авторы, публикующие исследование в ядерных журналах по iMetrics, от тех, что печатаются по темам общей информатики, мы сравниваем для каждого массива документов по iMetrics долю авторов, публикующихся в других массивах документов по iMetrics, с долей авторов, публикующихся вне сферы iMetrics. В табл. 2 мы используем только данные для последнего пятилетнего периода (2007-2011).

Примерно половина первых авторов, публикующих статьи по iMetrics в JASIST или JOI, также публикуется (снова как первые авторы) в двух других местах сбора iMetrics. Таким образом, для большей доли авторов JASIST-iMetrics и JOI эти места являются не единственными*. С другой стороны, только 17% первых авторов SCI также публикуются в JASIST-iM или JOI. Такая низ-

кая доля является естественным следствием того факта, что SCI является местом гораздо большего сбора исследований по iMetrics, чем JASIST или JOI. Следовательно, он будет единственным местом большого числа авторов. Гораздо важнее в контексте этого исследования определить, какая доля авторов из трех массивов документов по iMetrics также публикуется (опять как ведущие авторы) в четвертом массиве документов не по iMetrics. Теперь доли будут значительно ниже (между 4% и 14%).

Мы можем сравнить две части, чтобы определить коэффициент различия, т.е. как наиболее вероятно авторы должны публиковаться в другом месте по iMetrics, чем в массиве документов не по iMetrics (т.е. JASIST-O). Этот коэффициент составляет 3,1 и 3,9 для авторов JASIST-iM и SCI, соответственно, и является таким же высоким, как 7,4 для авторов JOI. Отсюда вывод, что авторы исследований по iMetrics происходят из одного и того же соответствующего пула авторов, которые публикуются в группе ядерных журналов по iMetrics. С другой стороны, эти исследователи, по-видимому, менее вероятно должны публиковаться в журналах по общей информатике. Этот результат свидетельствует о высоком уровне социальной идентичности и различия iMetrics при сравнении с общей информатикой.

Когнитивная идентичность и различие iMetrics

Для определения когнитивной идентичности и отличия iMetrics от информатики мы исследуем слова названий статей и интеллектуальную базу, выраженную через ссылки.

Слова названий статей и когнитивные основы массивов документов

Во-первых, мы определяем уровень сходства в когнитивных основах статей по iMetrics, опубликованных в трех массивах документов, путем анализа слов, которые появляются в названиях статей (за 2007-2011 гг.). Значения косинуса между массивами документов основаны на частотах терминов, появляющихся в трех множествах названий, за исключением общих английских слов (запрещенные слова). Результаты данных представлены в табл. 3.

* Поскольку некоторые авторы имели только одну публикацию за пятилетний период, они будут отражены как единственные авторы этого массива документов.

Таблица 2

Массив документов	Число первых авторов	Доля авторов, публикующихся в двух других массивах документов по iMetrics, %	Доля авторов, публикующихся в JASIST-O, %	Соотношение долей (коэффициент отличия)
JASIST-iM	159	43	14	3,1
SCI	530	17	4	3,9
JOI	124	54	7	7,4

Примечание. Отображаются авторы в каждом из трех массивов документов по iMetrics и доля, публикующихся в двух других массивах по iMetrics, а также в массиве документов не по iMetrics (JASIST-O). Авторы гораздо вероятнее публикуются в других местах iMetrics, чем в сфере общей информатики (коэффициент отличия)

Таблица 3

Массивы документов	Косинус сходства (погрешность)
JASIST-iM & SCI	0,807 (0,020)
JASIST-iM & JOI	0,830 (0,030)
JOI & SCI	0,779 (0,026)

Примечание. Даются значения косинуса сходства между тремя массивами документов по iMetrics на основе терминов, используемых в названиях. Значения в скобках — стандартные погрешности отклонения

Высокие значения косинуса (около 0,8) показывают, что когда это касается понятий, используемых в названиях, когнитивные основы трех мест сбора по iMetrics являются очень схожими. В терминах значений косинуса JASIST-iM и JOI являются в некотором смысле более схожими, чем два, сравниваемых с SCI, но отличия являются статистически не значительными, как видно из погрешностей значений косинуса, полученных из загруженной повторной выборки.

Должны ли такие высокие значения сходства косинуса также наблюдаться между названиями трех массивов документов по iMetrics и названиями статей по общей информатике (JASIST-O)? Результаты представлены в табл. 4. Сходство относительно JASIST-O гораздо ниже (около 0,3), чем среди массивов документов по iMetrics. Это подтверждает, что iMetrics отличается от информатики в терминах тематик, представленных в названиях.

Таблица 4

Массивы документов	Косинус сходства
JASIST-iM & JASIST-O	0,315
JOI & JASIST-O	0,297
SCI & JASIST-O	0,301

Примечание. Дается косинус индекса сходства между тремя массивами документов по iMetrics и документальным набором не по iMetrics (JASIST-O) на основе терминов, используемых в названиях.

Хотя имеется общий высокий уровень сходства среди трех массивов документов по iMetrics, нас интересует открытие любых особенностей в основе. Мы подходим к этой проблеме двумя способами. Во-первых, мы определили 50 самых часто используемых терминов в полном массиве данных по iMetrics за период 2007-2011 гг. Мы обнаружили, что большинство (32) из наиболее часто

используемых терминов не являются доминирующими в любом приведенном массиве документов. Из тех, которые являются особыми, большая часть принадлежит SCI, так как он имеет самую большую долю статей, поэтому имеет самый высокий вклад в список наиболее часто используемых слов. Затем мы исследуем 20 наиболее частотных терминов, которые служат характеристикой каждого массива документов (табл. 5). Термины, которые являются сильно доминирующими (т.е. более 67% их встречаемости могут принадлежать определенному массиву документов), выделены жирным шрифтом.

Учитывая все, а не только сильно доминирующие термины, мы предполагаем, что особая база JASIST-iM может характеризоваться как темы, связанные с научной коммуникацией. В SCI особый акцент сделан на географических направлениях, тогда как в JOI — на показателях. Тем не менее, как утверждалось ранее, сходства между когнитивным охватом множеств данных являются гораздо большими, чем различия, а специфические термины, определенные нами, указывают скорее на некоторую более высокую тенденцию появления статей по указанным выше темам в этих журналах, чем на единственный охват этих тем любым из ядерных журналов.

Характеристики интеллектуальной базы в различных массивах документов

Среднее число ссылок на статью в SCI и JOI является схожим (27 и 30, соответственно), но это число значительно выше в JASIST-iM (40 ссылок). Вообще говоря, массивы документов по iMetrics имеют очень схожие распределения возрастов ссылок (рис. 3) с пиком в возрасте 2 лет для всех трех массивов документов, т.е. большинство ссылок являются современными. Индексы Прайса* — 45, 43 и 51 для JASIST-iM, SCI и JOI, соответственно. Хотя в JASIST-iM и JOI число ссылок однолетнего возраста

* Индекс Прайса [46] — доля ссылок (из всех статей) в пределах пяти лет.

является почти таким же, как и в пике, для SCI это число является гораздо меньшим. Фактически полное распределение ссылок SCI, по-видимому, сдвигается приблизительно на один год при их сравнении с JASIST и JOI. Скорее, чем объяснять это различие разными практиками авторов, альтернативное объяснение состоит в том, что оно (различие) отражает, вероятно, более длительное время между датами представления рукописи и публикации в SCI по сравнению с другими массивами документов.

Интересно, что ссылки в статьях JASIST-O являются даже более старыми, чем в SCI. Поэтому, даже если они происходят из одного журнала, индекс Прайса статей JASIST-iM составляет 45, тогда как статей JASIST-O – 38. Вообще, мы принимаем это за обозначение того, что научный фронт в iMetrics движется гораздо быстрее, чем в общей информатике. Это служит другим показателем того, что iMetrics растет когнитивно отлично от информатики.

Таблица 5

Наиболее частые характерные термины из JASIST-iM	Наиболее частые характерные термины из SCI	Наиболее частые характерные термины из JOI
цитирование	патент	подход
автор	эффективность	оценка
сеть	сотрудничество	тип
сравнение	китайский	применение
научный	университет	распределение
информация	наукометрический	ядерный
доступ	международный	эмпирический
против	страна	обзор
открытый	производительность	g(-индекс)
метод	направление	Хирш (h-индекс)
крупный	нанотехнология	средство
совместное цитирование	авторство	коллега
термин	оценка	результат
динамический	технологический	теория
оценивать	национальный	вариант
коммуникация	мировой	библиографическая_пара
предмет	оценивать	слово
использование	технология	информетрический
результат	НИОКР	исследование
глобальный	европейский	уровень

Примечание. Приводится список 20 наиболее специальных терминов для каждого массива документов (2007-2011 гг.). Термины, которые весьма доминировали, выделены жирным шрифтом

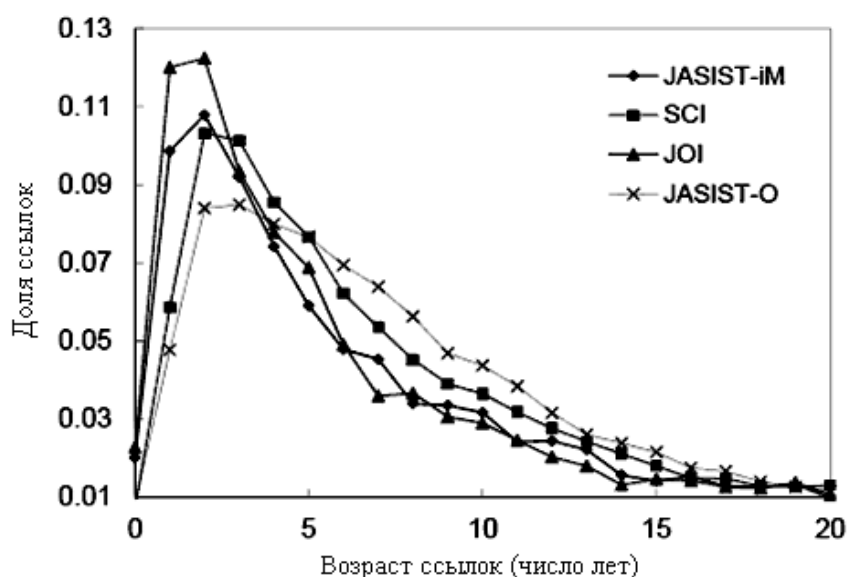


Рис.3. Распределение возраста ссылок в статьях (2007-2011 гг.) определенного массива документов

Уотерс и Лейдесдорф [37] задали себе вопрос, стала ли явью мечта Прайса [45] о «наукометрии» как о «твердой» науке спустя 25 лет существования журнала SCI, и ответили тогда на него отрицательно. По прошествии двух десятилетий эта ситуация не изменилась, по крайней мере, когда стало известно о научной сфере журнала SCI – его индекс Прайса был равен 43, как обнаружили Уотерс и Лейдесдорф [37]. Даже JOI, с индексом Прайса 51, находится ниже значений для «твердых наук» из указанных выше 60 [46].

Далее мы исследуем комплект источников, на которые ссылались в массивах документов по iMetrics (2007-2011 гг.). Всего имеется только девять источников, которые вносят вклад более 1% ссылок в любой из массивов документов. Список этих источников, распределенных по общему числу ссылок во всех трех массивах документов, представлен в табл. 6. Результаты, которые мы получили, похожи на те, что получили Перитц и Бар-Илан [47, с. 282], которые проанализировали ссылки на статьи, опубликованные в журнале *Scientometrics* (1990 – 2000), и сделали вывод, что «область полагается в большей степени на себя, на библиотечное и информатическое и на социологию, историю и философию науки» [с. 282].

Таблица 6

Источник	Доля в JASIST-iM, %	Доля в SCI, %	Доля в JOI, %
JASIS(T)	16,8	20,0	17,2
Scientometrics	12,3	18,0	14,8
Res Policy	1,5	3,9	2,3
J Informetr	1,8	1,0	4,1
Science	1,3	1,5	1,7
P Natl Acad Sci USA	1,3	1,2	2,4
Nature	1,3	1,4	1,4
Inform Process Manag	1,8	0,8	1,8
J Doc	1,5	0,7	0,9

Примечание. Приводится список наиболее часто цитируемых источников. Долями, выделенными курсивом, являются источники, в преобладающей степени цитируемые статьями в одном из трех массивов документов

Статьи во всех трех массивах документов по iMetrics цитировали статьи JASIS(T) любого типа в наибольшей степени. Интересно то, что статьи, опубликованные в SCI, цитируют самое большее три статьи из JASIST. Статьи в SCI наиболее цитируются статьями в SCI, а статьи JOI – статьями JOI. Вклад ссылок на JOI в SCI и JASIST, вероятно, ниже, чем его нынешнее значение, так как JOI начал издаваться в 2007 г., поэтому его полная значимость еще не может быть исследована.

Мы применяем величину косинуса, чтобы увидеть, насколько схожи/различны интеллектуальные базы этих трех массивов документов. Результаты данных показаны в табл. 7. Все значения являются довольно высокими. Самое большое различие существует между

JASIST-iM и SCI. JOI, имеющий более высокие значения относительно обоих, может также рассматриваться промежуточным между JASIST-iM и SCI.

Табл. 8 показывает косинусы между тремя массивами документов по iMetrics и JASIST-O. Это сходство является опять значительно ниже, чем оно было среди массивов документов по iMetrics. В терминах источников ссылок JASIST-iM является наиболее похожим на JASIST-O, тогда как SCI является менее похожим. Интересно то, что JASIST-iM, по-видимому, происходит из той же интеллектуальной базы, что и общая информатика, хотя мы видим, что когда дело касается актуальных тем в названиях, он также не похож на JASIST –O, как и на SCI или JOI.

Таблица 7

Массивы документов	Косинус сходства
JASIST-iM & SCI	0,879
JASIST-iM & JOI	0,950
JOI & SCI	0,947

Примечание. Дается косинус сходства между тремя массивами документов по iMetrics на основе источников, использованных в ссылках

Таблица 8

Массивы документов	Косинус сходства
JASIST-iM & JASIST-O	0,589
JOI & JASIST-O	0,387
SCI & JASIST-O	0,223

Примечание. Дается косинус сходства между тремя массивами документов по iMetrics и массивом документов не по iMetrics (JASIST-O) на основе источников, использованных в ссылках

В итоге, три массива документов по iMetrics берут начало из одной интеллектуальной базы, которая отличается от таковой, состоящей из статей по общей информатике. В качестве иллюстрации этих сходств в табл. 9 приводятся десять наиболее цитируемых имен первых авторов (в течение 2007-2011 гг.) в каждом из трех массивов документов по iMetrics.* Эти списки схожи по пяти из десяти имен (выделены жирным шрифтом), появляющихся во всех трех списках, и по двум, появляющимся в двух списках. Все авторы в этом списке также появились в одно или другое время как первые авторы в ядерных журналах по iMetrics. Это указывает на то, что основные вкладчики в интеллектуальную базу этой области являются в то же время активными деятелями в ядерной литературе по iMetrics.

* Цитируемые ссылки в документах, загруженных из WoS, предоставляют только имена первых авторов. Список в табл. 9 поэтому не указывает на влияние или воздействие этих авторов.

JASIST-iM:		SCI:		JOI:	
LEYDESDORFF L	314	GLANZEL W	387	EGGHE L	185
GARFIELD E	210	GARFIELD E	280	GLANZEL W	117
EGGHE L	187	LEYDERDORFF L	259	BORMANN L	111
WHITE HD	126	EGGHE L	215	LEYDESDORFF L	99
GLANZEL W	124	BRAUN T	170	HIRSCH JE	85
BORMANN L	118	MOED HF	165	MOED HF	75
CRONIN B	111	HIRSCH JE	140	GARFIELD E	75
SMALL H	104	SCHUBERT A	132	ROUSSEAU R	68
THELWALL M	91	NARIN A	124	SCHREIBER M	59
MOED HF	90	MEYER M	123	BURRELL QL	48

Примечание. Сравниваются десять наиболее часто цитируемых первых имен авторов в трех массивах документов по iMetrics (2007–2011 гг.). Жирным шрифтом указаны имена, появляющиеся во всех трех массивах. Числа цитирований основаны на первых именах авторов, как они появляются в рекомендательных списках записей WoS, и их не следует интерпретировать как ранги оценок

ЗАКЛЮЧЕНИЕ

iMetrics является очень активной научной областью, имеющей рост, который подтверждает разговоры о взрыве литературы по iMetrics в последнее десятилетие. Число статей по iMetrics в ядерных журналах в 2010 г. было примерно в четыре раза выше, чем в предыдущие десять лет. Хотя в течение 1980—1990-х гг. сфера iMetrics формировалась и искала свою идентичность где-то между научно-технологическими исследованиями и информатикой, эта научная область стала более определенной по мере сближения с информатикой в течение 1990-х гг. В последнее десятилетие это слияние достигло расцвета, и теперь пришло время изучить, имеет ли iMetrics социокогнитивную идентичность.

Хотя явный рост научной области и установление новых мест для публикации могут предполагать образование сферы, для того чтобы стать предметом практиков научной области, необходимо также показать признаки социальной идентичности, т.е. им нужно оформиться в общество со связями, более сильными внутренне, чем внешне. Чтобы изучить, выполняются ли эти критерии в случае iMetrics, нам потребовался соответствующий массив данных, который мы получили из трех ядерных журналов по iMetrics. Полагая, что все документы, публикуемые в SCI и JOI, попадают в категорию исследования с помощью iMetrics, мы использовали эти два журнала как критерий для двухъярусной процедуры определения статей по iMetrics в JASIST. Это разделение между статьями по iMetrics и не по iMetrics в JASIST обеспечило нас не только четкой выборкой документов по iMetrics, но также и сравнительной группой, которую можно рассматривать как представляющую исследование по информатике вообще.

Подход, использованный нами для тестирования социального различия авторов iMetrics, должен был сравнить по каждому журналу iMetrics долю тех авторов, которые публикуются в двух других массивах документов по iMetrics относительно доли тех авторов, которые публикуются в массиве документов не по iMetrics. Мы обнаружили, что большинство авторов по iMetrics вероятнее всего (от 3 до 7 раз) должны публиковаться в журналах по iMetrics, а не в журналах не по iMetrics (т.е. JASIST-O). С другой стороны, они, как правило, не связаны с одним журналом по iMetrics. Эти результаты показывают, что авторы статей по iMetrics являются социально отличны-

ми от авторов статей по более общей информатике, с очень малой долей тех, кто работает по обоим темам как по iMetrics, так и не по iMetrics.

Анализ тем, отраженных в трех массивах документов по iMetrics, с использованием слов названий статей, показал, что различия между терминами являются небольшими. С другой стороны, различия массива документов не по iMetrics в этом отношении являются сравнительно большими. Этот результат указывает на очень высокий уровень когнитивного различия. Анализ наиболее часто используемых терминов, которые являются характерными для данного массива документов, показал, что помимо по большей части общих тем, каждый журнал имеет также некий специальный фокус: для JASIST — это научная коммуникация, для SCI — исследования по отдельным географическим областям, для JOI — показатели эффективности. Эти различия в фокусах вносят некоторую неоднородность в исследование по iMetrics, публикуемое в различных журналах, но различий гораздо меньше, чем сходств.

Анализ практик цитирования и интеллектуальной базы указал на следующие сходства между журналами по iMetrics. Не только авторы по iMetrics, публикующиеся в различных журналах, имеют одинаковые практики цитирования, но и сама область как множество ссылок, по-видимому, движется быстрее, чем информатика. А именно, большинство ссылок в статьях по iMetrics являются недавними, они достигают пика в возрасте двух лет для всех трех массивов документов. Средний возраст ссылок для статей не по iMetrics в JASIST более старый.

JASIST является наиболее цитируемым источником во всех трех местах. Наличие общей интеллектуальной базы далее поддерживается очень высокими значениями сходств косинуса среди ссылок по трем журналам и сравнении списков ведущих десяти наиболее цитируемых авторов (которые выделяют пять авторов).

Хотя предыдущие работы решительно показывают, что сфера iMetrics является областью с четко описанным когнитивным фокусом, теперь мы видим, что iMetrics представляет научную сферу со связанной социальной и когнитивной идентичностью, которая отличается от общей информатики. Методология, которую мы применили или ввели в этом исследовании (описание политематических журналов, основанное на цитировании монотематических журналов, ядерных и сравниваемых массивов документов), и связанные понятия (понятие сходства

между ядерными массивами документов и отличие сравнимых массивов) могут быть также использованы для изучения того, достиг ли другой кандидат или первичные научные области социальной и когнитивной идентичности на пути развития во вполне сформировавшиеся научные области или даже дисциплины.

Благодарность. Благодарим Блейза Кронина и анонимных рецензентов за предоставление полной обратной связи по предыдущим вариантам статьи.

ЛИТЕРАТУРА

1. Pritchard A. Statistical bibliography or bibliometrics? // Journal of Documentation. — 1969. — Vol. 25. — P. 348-349.
2. Borgman C. L., Furner J. Scholarly communication and bibliometrics/ B. Cronin (Ed.) Annual Review of Information Science and Technology.—Vol. 36. — P. 3-72. — Medford, N.J.: Information Today, 2002.
3. Broadus R. N. Toward a definition of "bibliometrics" // Scientometrics. — 1987. — Vol. 12. — P. 373-379.
4. White H. D., McCain K. W. Bibliometrics/ M. E. Williams (Ed.) Annual Review of Information Science and Technology. — Vol. 24. — P. 119-186. — Amsterdam: Elsevier Science Publishers, 1989.
5. Elkana Y., Lederberg J., Merton R. K., Thackray A., Zuckerman H. (Eds.). Toward a metric of science: The advent of science indicators. — New York: John Wiley & Sons, 1978.
6. Hess D. J. Science studies: An advanced introduction. — New York: New York University Press, 1997.
7. Price D. J. de S. Little science, big science. — New York: Columbia University Press, 1963.
8. Price D. J. de S. Networks of scientific papers // Science. — 1965. — Vol. 149, No.30 July. — P. 510-515.
9. Garfield E. Citation indexes for science: A new dimension in documentation through association of ideas // Science. — 1955. — Vol. 122, No. 3159. — P. 108-111.
10. Garfield E. Citation indexes in sociological and historical research // American Documentation. — 1963. — Vol.14, No.4. — P. 289-291.
11. Spiegel-Rösing I., Price, D. J. d. S. (Eds.). Science, technology and society. A cross-disciplinary perspective. — London: Sage, 1977.
12. Leydesdorff L., Van den Besselaar P. Scientometrics and communication theory: Towards theoretically informed indicators // Scientometrics. — 1997. — Vol. 38, No.1. — P. 155-174.
13. Courtial J. P. A cword analysis of scientometrics // Scientometrics. — 1994. — Vol. 31, No.3. — P. 251-260.
14. Lucio-Arias D., Leydesdorff L. An indicator of research front activity: Measuring intellectual organization as uncertainty reduction in document sets // Journal of the American Society for Information Science and Technology. — 2009. — Vol. 60, No.12. — P. 2488-2498.
15. Van den Besselaar P. The cognitive and the social structure of STS // Scientometrics. — 2001. — Vol. 51, No.2. — P. 441-460.
16. Van den Besselaar P., Heimeriks G. Mapping research topics using word-reference co-occurrences: A method and an exploratory case study // Scientometrics. — 2006. — Vol. 68, No.3. — P. 377-393.
17. Egghe L. Expansion of the field of informetrics: Origins and consequences // Information Processing & Management. — 2005. — Vol. 41, No.6. — P. 1311-1316.
18. Egghe L., Rousseau R. (Eds.). Informetrics 87/88. — Amsterdam: Elsevier, 1988.

19. Tague-Sutcliffe J. An introduction to informetrics // Information Processing & Management. — 1992. — Vol. 28, No.1. — P. 1-3.
20. Almind T. C., Ingwersen P. Informetric analyses on the World Wide Web Methodological approaches to 'webometrics' // Journal of Documentation. — 1997. — Vol. 53, No.4. — P. 404-426.
21. Hood W. W., Wilson C. S. The literature of bibliometrics, scientometrics, and informetrics // Scientometrics. — 2001. — Vol.52, No.2. — P. 291-314.
22. Sengupta I. N. Bibliometrics, informetrics, scientometrics and librametrics: An overview // Libri. — 1992. — Vol. 42, No.2. — P. 75-98.
23. De Bellis N. Bibliometrics and citation analysis: From the Science Citation Index to cybermetrics. — Lanham: The Scarecrow Press, 2009.
24. Glänzel W., Schoepflin U. Little scientometrics, big scientometrics ... and beyond? // Scientometrics. — 1994. — Vol.30, No.2-3. — P. 375-384.
25. Bar-Ilan J. Informetrics at the beginning of the 21st century - A review // Journal of Informetrics. — 2008. — Vol. 2. — P.1-52.
26. Leydesdorff L. Mapping interdisciplinarity at the interfaces between the science citation index and the social science citation index // Scientometrics. — 2007. — Vol.71, No.3. — P. 391-405.
27. Van den Besselaar P. Communication between science and technology studies journals: A case study in differentiation and integration in scientific fields // Scientometrics. — 2000. — Vol. 47. — P.169-193.
28. Åström F. Visualizing library and information science concept spaces through keyword and citation based maps and clusters /H. Bruce, R. Fidel, P. Ingwersen, P. Vakkari (Eds.). Emerging frameworks and methods // CoLIS4: Proceedings of the Fourth International Conference on Conceptions of Library and Information Science, Seattle, WA, July 21-25.— 2002.— P. 185-197.— Greenwood Village: Libraries Unlimited, 2002.
29. Leydesdorff L. The relations between qualitative theory and scientometric methods in science and technology studies // Scientometrics. — 1989. — Vol.15, No.5-6. — P. 333-347.
30. Moya-Aneón F., Herrero-Solana V., Jiménez-Contreras E. A connectionist and multivariate approach to science maps: The SOM, clustering and MDS applied to library and information science research // Journal of Information Science. — 2006. — Vol. 32, No.1. — P. 63-77.
31. Milojević S., Sugimoto C. R., Yan E., Ding Y. The cognitive structure of library and information science: Analysis of article title words // Journal of the American Society for Information Science and Technology. — 2011. — Vol. 62, No.10. — P.1933-1953.
32. Janssens F., Leta J., Glänzel W., De Moor B. Towards mapping library and information science // Information Processing & Management. — 2006. — Vol.42, No.6. — P. 1614-1642.
33. Law J. The development of specialties in science: The case of x-ray protein crystallography /G. Lemaire, R. Macleod, M. J. Mulvey & P. Weingart (Eds.). Perspectives on the emergence of scientific disciplines. — P. 123-152.— Chicago: Aldine, 1976.
34. Mullins N. C. The development of a scientific specialty: The Phage group and the origins of molecular biology // Minerva. — 1972. — Vol. 10. — P. 52-82.
35. Börner K., Boyack K. W., Milojević S., Morris S. A. An introduction to modeling science: Basic model types, key definitions, and a general framework for the comparison of proc-

esses /A. Scharnhorst, K. Börner, P. Van den Besselaar (Eds.). Models of science dynamics: Encounters between complexity theory and information science. — P. 3- 22.— Springer, 2012.

36. *Lucio-Arias D., Leydesdorff L.* The dynamics of exchanges and references among scientific texts, and the auto-poiesis of discursive knowledge// Journal of Informetrics. — 2009. — Vol. 3, No.3. — P. 261-271.

37. *Wouters P., Leydesdorff L.* Has Price's dream come true: Is scientometrics a hard science? // Scientometrics. — 1994. — Vol. 31, No.2. — P. 193-222.

38. *Boyack K. W., Klavans R.* Multiple dimensions of journal specificity: Why journals can't be assigned to disciplines/ E. Noyons, P. Ngulube, J. Leta (Eds.)// The 13th Conference of the International Society for Scientometrics and Informetrics (Vol. I, pp. 123-133). Durban, South Africa — ISSI, Leiden University and the University of Zululand, 2011.

39. *Seglen P. O.* The skewness of science// Journal of the American Society for Information Science. — 1992.— Vol. 43, No.9. — P. 628-638.

40. *Van Noorden R.* A profusion of measures// Nature. — 2010. — Vol. 465.— P. 864-866.

41. *Minguillo D.* Toward a new way of mapping scientific fields: Authors' competence for publishing in scholarly

journals// Journal of the American Society for Information Science and Technology. — 2010. — Vol. 61, No.4.— P. 772-786.

42. *Leydesdorff L.* Visualization of the citation impact environments of scientific journals: An online mapping exercise// Journal of the American Society for Information Science and Technology. — 2007. — Vol. 58, No.1. — P. 25-38.

43. *Ahlgren P., Jarneving B., Rousseau R.* Requirement for a cocitation similarity measure, with special reference to Pearson's correlation coefficient// Journal of the American Society for Information Science and Technology. — 2003.—Vol. 54, No.6. — P. 550-560.

44. *Crane D.* Invisible colleges: Diffusion of knowledge in scientific communities. — Chicago: University of Chicago Press, 1972.

45. *Price D. J. de S.* Editorial statement// Scientometrics. — 1978. — Vol.1, No.1. — P. 7-8.

46. *Price D. J. de S.* Citation measures of hard science, soft science, technology, and nonscience /C. E. Nelson & D. K. Pollock (Eds.). Communication among Scientists and Engineers. — P. 3-22. — Lexington, MA: Heath, 1970.

47. *Peritz B. C., Bar-Ilan J.* The sources used by bibliometrics-scientometrics as reflected in references// Scientometrics. — 2002. — Vol.54, No.2.— P. 269-284.

Мера сходства между научными журналами и разнообразия списка публикаций*

Стефан КОРДЬЕ
(Stéphane CORDIER)

Орлеанский университет, г. Орлеан, Франция

Предлагается определение научного разнообразия и, следовательно, мера «междисциплинарности» сотрудничества. В соответствии с предыдущими изучениями предлагаемый подход состоит из двух шагов: 1) определение сходства между журналами, 2) использование этого сходства для характеристики однородности (или, напротив, разнообразия) списка публикаций (который касается одного лица или группы лиц).

ВВЕДЕНИЕ

Сегодня междисциплинарность представляет интерес по нескольким причинам и для многих людей и учреждений. Сошлемся только на две недавние инициативы во Франции: создание «междисциплинарной миссии» в Национальном центре научных исследований – CNRS (<http://www.cnrs.fr/mi/>) или отчет AERES, который предлагает интересное направление оценки междисциплинарности [1] на основе качественного анализа. Мы не намерены обсуждать причины такого интереса и ссылаемся на работу [2], содержащую недавний подробный обзор по междисциплинарности.

Здесь мы предлагаем метод количественного определения междисциплинарности, основанный только на библиометрических данных без какой-либо предварительной классификации научных областей и/или произвольного знания об их близости. Полученные результаты должны сравниваться с существующей классификацией и анализироваться учеными для признания (или непризнания) их значительного интереса.

Настоящая заметка дает лишь предварительное описание идеи, которая не была проверена на библиографических данных. Автор не является экспертом в области наукометрии и не имеет доступа к большой базе данных, что необходимо для проверки этого подхода. Много исследований было проведено по соавторству (см., например, [3] и ссылку, приведенную в сноске на с.159). Однако настоящий подход (с двумя шагами, которые детализируются ниже) до меня еще не был предложен. В настоящей версии эту заметку следует считать предварительной публикацией, содержащей предложения, хорошо подходящие, в частности, для того, чтобы быть информированным о предварительных разработках в этом направлении.

Цель состоит в определении меры междисциплинарности внутри списка публикаций (одного человека, группы, лаборатории, учреждения). Такая количественная

информация должна быть дополнена более тщательным анализом со стороны ученых, чтобы определить соответствующую релевантность научных сотрудничеств. Мы только пытаемся предложить подход, чтобы увидеть его возможность и в перспективе доказать его способность характеризовать междисциплинарные исследования.

Предлагаемый подход основан на двух шагах: на первом шаге мы определяем на основе библиографической базы данных меру сходства (подобия) между научными журналами, основанную на соавторстве, т. е. чем больше два журнала имеют соавторов, тем более близкими они являются (далее это будет уточняться). Он должен быть нацелен на то, что мы измеряем не научную «близость», а действительные практики публикаций. Второй шаг состоит в использовании этих сходств для характеристики того, является или не является публикационный список научно «однородным» массивом.

Предполагается, что информация о соавторстве считается более надежной для оценки междисциплинарных сотрудничеств, чем использование ссылок. На самом деле, общим является скорее то, что статья, например в области математики, цитирует некоторые статьи в прикладной области, чтобы показать происхождение научной проблемы или объяснить сделанный выбор моделирования, но суть статьи может быть полностью сосредоточена на математическом анализе. С другой стороны, подписание статьи вместе с коллегой означает (надемся, что так) взаимный интерес и работу над статьей.

Заметим также, что публикация статьи в так называемом междисциплинарном журнале (что является определением таких журналов?) не означает, что сама статья является результатом сотрудничества между несколькими научными областями.

Можно сказать, что предлагаемый подход больше измеряет оригинальность массива публикаций, что является причиной того, почему это называется научным разнообразием, поскольку сходство имеет значение для существующих сотрудничеств, даже если они уже являются междисциплинарными.

Несколько сетевых сайтов действительно обеспечивают информацию о научных сотрудничествах, напри-

* Перевод Cordier S. A measure of similarity between scientific journals and of diversity of a list of publications. – <http://arxiv.org/pdf/1210.6510> v 1. pdf - 2012.

мер *researchgate*, *researchid*, *googlescholar*, *sciencewatch* (не полный перечень) и могут представлять интерес в обеспечении новых услуг и информации для их посетителей (см. в упоминании примеров).

Предлагаемый анализ может быть также интересен редакторам научных журналов (они могут уже иметь подобные средства, которые не известны автору). Метод представляется алгоритмическим способом, чтобы облегчить его реализацию. Повторяю, что эксперимент обратной связи приветствуется.

ДАННЫЕ, ПОНЯТИЯ

Метод полагается на библиографические данные (использование очень крупной базы данных обеспечит более релевантную информацию).

Уточним понятия. База данных входит в перечень статей, который будет обозначаться уникальным идентификатором (рассматриваемым как целое число) с использованием буквы i . Каждая статья i (где $i \in 1 \dots N$) будет описываться следующим образом:

- Журнал публикации обозначается уникальным идентификатором j . Точнее, $j(i)$ – это журнал, где статья i опубликована. Список журналов считается законченным (даже если его длина увеличивается с созданием новых журналов каждый год). В качестве примера возьмем базу данных издательства Thomson – ISI, включающую около 13 тыс. журналов.

- $y(i)$ – год публикации статьи i .

- $K(i)$ – перечень соавторов статьи i . Авторы должны быть идентифицированы, т.е. каждый автор должен иметь уникальный идентификатор, который может быть представлен целым числом. Мы будем использовать k для авторов. Итак, $k \in K(i)$ означает, что автор k является одним из авторов статьи i^* .

- $p(i)$ – число страниц в статье. Полезно различать короткую заметку и более подробное изучение, хотя это подлежит обсуждению. Интерес к статье, естественно, не пропорционален ее длине, но она (длина) может рассматриваться как полезный показатель, однажды перенормализованный для данного журнала (или данного автора).

В этой заметке мы будем использовать прописные буквы для представления списков. Так, J представляет (законченный) список всех журналов, а I – список всех статей. Например, мы должны заметить, что $I(j)$ – это перечень статей, опубликованных журналом j , $I(j,k)$ – перечень статей, опубликованных автором k в журнале j , а $I(j,k,y)$ – перечень статей, опубликованных автором k в журнале j в году y . Аналогично, $J(k,y)$ представляет перечень журналов, где автор k опубликовался в году y .

Предположим, что N – кардинал множества. Например, $N(I(k))$ – это общее число статей автора k . P – общее число страниц, например, $P(j,y) = \sum_{i \in I(j,y)} p(i')$ является общим числом страниц в журнале j за год y .

Согласно обычному принципу в области математических наук (которая является областью автора) вес данной статьи должен быть распределен одинаково между всеми авторами статьи. Этот момент, естественно, является сомнительным, но и утверждение, что важность

статьи пропорциональна числу ее авторов, как, например, при подсчете ссылок или фактора влияния, также должно подлежать обсуждению. Предлагаем обратиться к работе [4] для обсуждения вопроса многоавторства.

СХОДСТВО ЖУРНАЛОВ

Используя эти понятия, определим теперь сходство между журналами путем учета всех статей и (со)авторов этой статьи, всех других статей по одному и тому же автору.

Точнее, для всех $i \in I$, всех $k \in K(I)$ и всех $i' \in I(k)$ сходство между журналами $j(i)$ и $j(i')$ увеличивается в соответствии с формулой

$$S(j(i), j(i')) = \min \left(\frac{p(i)}{N(K(i))}, \frac{p(i')}{N(K(i'))} \right), \quad (1)$$

причем $S(j(i), j(i'))$ увеличивается на ту же величину (когда меняется роль i на i'). Мы предлагаем увеличить сходство между двумя журналами на минимальную величину «веса» двух статей (число страниц, деленное на число авторов), вместо использования арифметического значения, например, поскольку полагается, что научная близость сильнее, если две статьи имеют одинаковый вес (с тем же арифметическим весом). Другие выборы, как, например, геометрическое значение $\sqrt{ab}$, могут дать лучшие результаты. Единственным способом выбрать правильную формулу должно быть испытание нескольких выборов и сравнение полученной матрицы сходства (см. ниже некоторые идеи для помощи в выборе или в обосновании релевантного определения сходства).

Кто-то может обсуждать нормализацию числа страниц, т.е. делить число страниц статьи на общее число страниц в журнале j . Другими словами, мы предлагаем заменить $p(i)$ в выше приведенном уравнении на

$$\tilde{p}(i) = p(i) / P(j, i).$$

Эти варианты должны быть проверены по мере доступности данных. Очевидно, что ненормализованный выбор будет увеличивать влияние журналов, которые дают много статей и/или страниц, тогда как сходство должно измеряться близостью между журналами, которая не должна быть связана с «размером» журнала. Следовательно, нормализация с использованием нормализованных по общему числу страниц статей, должна быть более подходящей.

Замечания

- В результате $S(j, j) > 1$, поскольку $i \in I(k)$, $\forall k \in K(i)$ (или $S(j, j) > P(j)$, если мы не используем нормализованную версию). Эффективное значение будет измерено, если одни и те же авторы имеют много своих публикаций в журнале j .

Сходство может быть вычислено за определенный год (или период) путем учета только статьи за соответствующий год (или период).

- Ясно, что матрица S будет грубой, поэтому необходимо принимать в расчет соавторство второго порядка (т.е. соавтора соавторов, следуя идее ссылки на число Эрдёша?). Определим путем суммирования бинарные взаимодействия как следующие

$$S_2(j, j') = \sum_{j''} S(j, j'') S(j'', j').$$

* Это является причиной того, почему я не могу проверить предлагаемый подход на данных хранилища французских публикаций HAL.

В результате мы видим, что $S_2 = S^2$. Затем мы можем использовать $\tilde{S} = S + \theta S_2$, где θ - постоянная величина, которая представляет относительный вес вторичного соавторства. И опять, это должно быть проверено на реальной крупной базе данных (см. ниже).

Фаза обоснования

На этой стадии необходимо проверить, соответствует ли предлагаемое сходство обычно используемой классификации в научной области. Точнее, используя данную дисциплинарную классификацию, интересно проверить, является ли усредненное сходство внутри научной области меньше (или нет и до какой степени), чем та же средняя величина для всех журналов. Это может также служить для вычисления среднего сходства между двумя выбранными различными областями путем вычисления среднего значения $S(j, j')$ для любого журнала j в области 1 и журнала j' в области 2. Это обеспечит матрицу сходства между научными областями, и она должна быть проанализирована на соответствие обычной классификации научной области.

Заметим, что, когда рассматривается статья из «многодисциплинарного» журнала (с использованием произвольного списка/классификации), ее ссылки влияют на область в соответствии со ссылками на статью (<http://sciencewatch.com/about/met/classpappmultijour/>).

Другое исследование, такое как кластеризация, может быть проведено с использованием сходства между журналами [5,6].

Следует также проверить, зависит ли сходство журналов «общего списка» от их специфичности. Кто-то может, например, взять 22 так называемые широкие области (см. <http://www.in-cites.com/journal-list/>) в базе данных Essential Science Indicators издательства Thomson-ISI.

Возможные услуги – сервисные программы

Однажды установленное сходство между журналами может представлять интерес для редакторов при оценке влияния их редакционного выбора на научное позиционирование. Например, если делается попытка охватить статью в близкой области (которая соответствует данному подмассиву журналов), можно обнаружить, что среднее сходство журнала с журналами данного подмассива растет со временем (путем вычисления усредненного сходства, ограниченного до последующих лет). Это также помогает редакторам видеть, развивается ли журнал в сторону большей специализации, или, наоборот, становится все более многодисциплинарным.

МЕЖДИСЦИПЛИНАРНОСТЬ, ИЛИ ИНДЕКС НАУЧНОГО РАЗНООБРАЗИЯ

Теперь допустим, что матрица сходства S известна (и обоснована).

В этом разделе мы будем строить индекс для некоторого произвольного списка L публикаций (для одного лица, группы, лаборатории, учреждения, журнала, редактора).

Определим так называемый индекс научного разнообразия SD (scientific diversity) списка L как усредненное сходство между журналами в списке, ранжированном в соответствии с весами статей.

Другими словами,

$$SD(L) = \frac{1}{N(L)^2} \sum_{i \in L} \sum_{i' \in L} S(j(i), j(i')) p(i) / N(K(i)). \quad (2)$$

Заметим, что этот индекс не связан с количеством бумаги (если кто-то дублирует перечень, число элементов в сумме дублирования умножается на 4, а $N(L)$ умножается на 2, и значение не изменяется).

Индекс научного разнообразия SD должен рассматриваться не как показатель качества статей в списке L , а, наоборот, как качественный показатель для этого перечня статей. Заметим, что этот индекс создается с использованием статистических/усредненных количественных определителей. Поэтому очень сомнительным является его использование на небольшом списке статей и, таким образом, он, вероятно, больше подходит для характеристики коллективного списка публикаций, чем для отдельных авторов, за исключением ученых с достаточно длинным перечнем публикаций, чтобы результат был значительным (для таких ученых будет интересно увидеть, как их SD согласуется с числом статей – индекс I , т.е. $N(I(k))$ с нашим определением – или ссылками (например, b -индексом).

Существует много исследований, которые могут быть сделаны с использованием этого индекса, но он опять же измеряет не «качество», «важность» или «влияние» (в смысле влияния на других ученых), а лишь относительное разнообразие, разнородность или оригинальность списка публикаций по отношению к другим. Такой показатель должен использоваться только с большой осторожностью и имеет смысл только для относительного сравнения (когда точное значение не представляет интереса). Например, кто-то может решить вычислить перечень определенного автора (обозначенный $I(k)$ в наших определениях), а кто-то может сравнить значение $SD(I(k))$ с соответствующим значением для всех соавторов этого k .

Если такая сетевая услуга реализована, можно спросить автора, кажется ли ему предлагаемое ранжирование в отношении одного из его соавторов в терминах «научного разнообразия» соответствующим или нет? Поэтому, это будет хорошим способом оценить, дает ли предлагаемый показатель информацию, которая совпадает с общим мнением. Если какой-либо вариант из приведенного выше определения предлагается, должно быть проверено, какое определение лучше показывает научное разнообразие.

ЦЕНТРАЛЬНЫЙ ЖУРНАЛ СПИСКА ПУБЛИКАЦИЙ

Кто-то также может использовать матрицу сходства, чтобы определить для какого-либо списка L центральный журнал, отыскивая журнал, который в списке $J(L)$ имеет самое большое сходство с

$$\begin{aligned} \bar{j}(L) &= \{j \in J(L) \text{ s.t. } \sum_{i \in L} S(j, j(i)) \frac{p(i)}{N(K(i))} = \\ &= \max_{j' \in J(L)} \sum_{i \in L} S(j', j(i)) \frac{p(i)}{N(K(i))} \}. \end{aligned} \quad (3)$$

Иначе говоря, если мы интерпретируем сходство как инверсию псевдорасстояния, центральным журналом является тот, который будет минимизировать среднее расстояние с другими в списке $J(L)$. Это должно соотноситься с точкой Ферма – Вебера значения Фреше.

Вероятно, интересно ранжировать журналы списка по уменьшению значения их усредненного сходства в

сравнении с другими (как определено выше). Это должно дать в верхней части списка журналы, которые соответствуют принципиальной области автора или списку, а в конце – журналы, которые весьма далеки от его специальности.

Что касается сравнения научного разнообразия (SD) его соавторов, любой человек может спросить такую сетевую службу, соответствует ли ранжирование тому, что обычно признается (использование пула и, в конце концов, сравнение результатов другого определения).

Еще одна возможная услуга, которая может быть полезной для ученых, - предложение некоторых журналов, в которых они никогда не публиковались, но которые являются «близкими» в том смысле, что их сходство высоко с их центральным журналом (некоторые могут ограничить предложения статей в списке журналов, где публиковались их соавторы. Она (услуга) может предложить увеличить их список журналов и избежать научной концентрации.

Можно также дать информацию о развитии их научного разнообразия за годы. Заметим, что определение «центральный журнал» является только примером использования индекса сходства между журналами, и может быть предложено множество других понятий в случае использования теории графов, сетевого анализа, кластеризации.

Еще раз приглашаю к обратной связи!

Благодарность. Автор благодарит в хронологическом порядке С. Манчини (Орлеанский университет), А. Каппелли (Центр управления научной коммуникацией, г. Лион), В. Мьела (Национальный центр научных исследований, г. Лион).

ЛИТЕРАТУРА

1. Critères d'évaluation des entités de recherche : Le référentiel de l'AERES, partie III, AERES, 2012. - <http://tinyurl.com/9zzjwn2>
2. *Wagner C.S., Roessner J.D., Bobb K., Thompson Klein J., Boyack K.W., Keyton J., Rafols I., Borner K.* Approaches to understanding and measuring interdisciplinary scientific research : A review of the literature // *Journal of Informetrics.* - 2011.
3. *Fortunato S.* Community detection in graphs // *Physics Reports.* – 2010. – No.486. – P. 75-174.
4. *Price D.* Multiple authorship // *Science.* – 1981. – No.212 - P. 986.
5. *Newman M.E.J.* Networks: An Introduction. - Oxford University Press, 2010.
6. *Girvan M., Newman M.E.J.* // *Proc. Natl. Acad. Sci. USA.* – 2002. – Vol. 99, No.12. – P.7821.

Век создания указателей ссылок*

Юджин ГАРФИЛД
(Eugene GARFIELD)

Thomson Reuters – Professional
(бывший Институт научной информации), г. Филадельфия, США

Прошло 60 лет с тех пор, как я начал мою карьеру в информатике. С того времени произошли огромные изменения в технологии, которые выражаются в способе, с помощью которого научная информация распространяется, отыскивается и оценивается. Молодое поколение ученых в Турции и повсюду даже не помнит «Current Contents» (специализированное сигнально-библиографическое издание, помещающее оглавления текущих выпусков первичных периодических изданий по определенной тематике) или огромные печатные тома Science Citation Index (Указатель цитированной литературы по точным, естественным наукам, США - SCI). В первом периоде существования издание «Current Contents» стало для ученых всего мира обязательной привычкой для чтения. Затем, во второй половине жизни этого издания мы начали публиковать в «Current Contents» серии под названием «This week's Citation Classic». В последние 15 лет мы просили тысячи авторов писать комментарии об их высокоцитируемых статьях. Четыре тысячи таких одностраничных комментариев можно найти на сайте (<http://www.citationclassics.org/>). Очень часто авторы говорят, что это не обязательно были их наиболее важные статьи. Так, когда меня попросили написать комментарий о моей статье 1955 г. в журнале «Science» [1], я должен был указать, что эта исходная статья о создании указателей ссылок не является моей самой цитируемой работой, а также не была самой важной. Моя статья 1972 г. по использованию анализа цитирования для оценки журналов в «Science» [2] привлекла гораздо больше внимания. Моей самой цитируемой работой, на самом деле, является моя книга 1979 г. «Citation Indexing» [3]. Книги, как и обзорные статьи, являются удобными для ссылок на работу автора вообще. Однако прослеживание генеалогии ссылок на мою статью 1955 г. показывает эволюцию идеи создания указателя ссылок — от системы информационного поиска до средства научной оценки. В статье, которую я подготовил по просьбе тогдашнего редактора журнала «Science» в 1995 г., я намекал на то, что теперь хвост виляет собакой [4]. Позже я скажу об этом больше. Действительно, это была тема моего основного выступления на последней конференции COLLNET в г. Далянь (Китай). Я польщен тем, что снова приглашен на конференцию COLLNET.

В первые несколько десятилетий после появления моей статьи 1955 г. и последующей статьи 1964 г. [5] большинство цитирующих их статей было связано с

доводами за и против создания указателя ссылок для информационного поиска. Тогда библиотекари и ученые были заняты традиционным, основанным на контролируемом словаре индексированием, или каталогизацией. Поэтому мы создали Пермутационный предметный указатель (Permuterm Subject Index) как приложение на естественном языке к указателю ссылок. Го-

*Перевод Garfield E. A century of citation indexing. – 2012. - http://www.tarupublications.com/journals/cjsim/04_CJSIM_6_1_pp.001-006.pdf

раздо позже Генри Смол формализует роль ссылок как символов понятий [6].

Ранним предвестником использования ссылок для оценки науки была статья 1967 г. австралийца Джозефа Марголиса [7]. К тому времени Ирвинг Шер и я уже упростили сортировку *Science Citation Index* (SCI) для создания списка 50 самых цитируемых авторов. Около трети из них оказались Нобелевскими лауреатами и почти все были авторами Нобелевского класса [8]. В то время, когда моя статья 1955 г. была опубликована, компьютеров не было. Методы с использованием перфокарт считались революционными. Даже спустя 10 лет, когда мы начали издавать SCI, перфокарты использовались в качестве ввода данных в первые примитивные компьютеры фирмы IBM для производства печатного указателя SCI. В те дни концепция «Memex» Ванневера Буша была такой близкой, что мы стали задумываться об идее Интернета [9]. Но топографические и сетевые свойства ссылок были полностью признаны и получили формальные описания, сделанные моим коллегой Ральфом Гарнером [10] и Дерекком Прайсом [11]. Идея представления науки, основанная на связывании свойств ссылок, была достаточно понятна и использована для исследования историографии архитектуры цифровой сети [12].

Раньше небольшая группа людей увидела в SCI его важный потенциал для библиометрических оценок. Было бы заманчиво описать многочисленные важные статьи и отчеты, которые, в конечном счете, сделали SCI и его преемницу базу данных Web of Science стандартным средством в руках аналитиков научной стратегии и других специалистов, заинтересованных в научной оценке. И она (база данных) также является основой для ежегодных предсказаний или прогнозов относительно Нобелевских премий, осуществляемых Дэвидом Пендлберри. Однако база данных Web of Science в настоящее время считается не только основной в библиотеках и повсюду, но также достаточно популярной, чтобы порождать конкуренцию между Google Scholar, Scopus и другими базами данных. Действительно, следует отметить, что связывание ссылок находится в центре успеха Google как поискового средства, основанного, фактически, на процессе упорядочивания страниц, берущем начало от *Journal Citation Reports*.

Повторное чтение моей статьи 1955 г. напоминает мне о вдохновении по поводу того, что идея возникла из моего раннего интереса к энциклопедизму. В 1970 г. профессор Манфред Коэн (Мичиганский университет) прокомментировал роль этой идеи в мировом энциклопедическом движении [13]. Сегодня Интернет дает возможность развитию Википедии и других грандиозных схем, которые сделают мечту Г. Дж. Уэллса о мировом разуме реальностью. Иногда я называю это «библиографической нирваной». Я должен отметить, что Г. Дж. Уэллс, возможно, был вдохновлен бельгийским документалистом Полем Отле. Относительно небольшой объем памяти компьютеров в начале их появления должен был препятствовать их применению для такого использования в течение двух или более десятилетий. С тех пор сетевые свойства указателей ссылок изучались многочисленными исследователями. Их возглавлял Дерек де Солла Прайс в своей работе о сетях научных статей 1965 г. [11]. Она появилась вскоре после моей статьи 1964 г. в журнале «Science» [5], которая описывала SCI не только как новое измерение в индексировании, но указывала на его использование в оценке науки. Сегодня легко забыть, что даже спустя десять лет после того, как первый печатный ежегодный SCI был опубликован, библиотеки горячо

обсуждали, следует ли закупать его для дополнения или даже замены комбинации традиционных служб индексирования. В конце концов, основной консерватизм ученых и библиотекарей был преодолен. Эта эволюция проходила параллельно с ростом объема памяти компьютера – от памяти в 16 килобайт компьютера IBM 1401, который мы тогда использовали, до гигабайтной или даже терабайтной мощности, которую мы сегодня считаем само собой разумеющейся. Было сделано даже больше, когда ISI добавил авторские рефераты к этой базе данных. Когда это произошло, многие библиотеки больше не испытывали необходимости в сохранении традиционных служб реферирования.

Поскольку теперь по установившейся практике база данных Web of Science используется в индустрии для целей оповещения, одним из моих величайших разочарований является неудача большинства ученых в том, чтобы использовать ее в качестве средства оповещения, т.е. для избирательного распространения информации, или SDI. Сегодня SDI осуществляется с помощью еженедельных или ежедневных оповещений о ссылках, но первая такая служба – автоматизированная система текущего оповещения (Automatic Subject Citation Alert – ASCA) – [14] начала действовать в 1965 г., спустя год после того, как мы начали издавать SCI. Для большинства пользователей все еще трудно проявлять сознательность относительно ссылок, которая является еще одним способом выражения мнения о том, почему они сопротивляются подготовке поисковых профилей, включающих цитируемых авторов или ссылки, а также ключевые слова. Молодое поколение использует систему оповещения Google, не зная, что электронные службы оповещения были доступны в течение более 45 лет.

Я думаю, что имеется более, чем достаточно, оснований, чтобы показать, что количественные исследования науки и корпуса знаний достигли зрелости. Пятьдесят лет назад, после выхода указателя *Science Citation Index*, охватывающего литературу 1961 г., уместно спросить, виляет ли хвост наукометрии собакой информационного поиска [15]. Новому поколению надо регулярно напоминать, что SCI был создан не для проведения количественных исследований, не для вычисления факторов влияния журналов, не для облегчения изучения истории науки. Профессор Корнельского университета ошибочно высказал это требование в статье, озаглавленной «Награда или убеждение? Битва за определение значения ссылки» [16]. В сообщении по электронной почте профессор Майкл Кёниг напомнил нам, что *Science Citation Index* был разработан как вспомогательное средство для изучения истории науки, а также в помощь при информационном поиске [17]. Эта амнезия поколения не является необычной. На той же неделе в послании почтового реестра открытого доступа (listserv) Стива Харнада читателям должны были напомнить, что ASCA (Автоматизированная система избирательного распространения информации на основе текущих изданий, в настоящее время известная как научное оповещение – Research Alert), система составления персональных профилей, основанная на указателе ссылок и ключевых словах, существовала в Институте научной информации США (ISI) с 1965 г.

Итак, чтобы расширить историческое описание, SCI и создание указателя ссылок предназначались не только для помощи при информационном поиске, но также для облегчения SDI (избирательного распространения информации), как описывалось ранее. SDI со временем эволюционировало в персональное оповещение – Personal

Alert. И, конечно, это стало также частью системы Web of Science. Служба оповещения о ссылках Google (Google Citation Alerts) не представляет ничего нового. Совершенно не удивительно, что некоторые люди предполагают, что SCI был создан как инструмент для историков. С начала 1960-х гг. я был связан с Робертом Мертоном, Дерекком де Солла Прайсом, В.В. Налимовым и среди других с Дж. Д. Берналом, на чью работу я часто ссылался. Бернал был прародителем «науковедения», которое позже развилось в наукометрию, тем, кто заложил основу этой области. Я описал всю эту историю на 11-й конференции ISSI в Мадриде несколько лет назад [18, 19]. Эта страница из фонда статей HistCite, цитирующая его книгу 1939 г. «The social function of science» («Социальная функция науки»), показывает влияние Бернала на наших предшественников (<http://garfield.library.upenn.edu/histcomp/index-bernal.html>).

Связь SCI с историей науки, возможно, далее упростилась, когда Ирвинг Шер и я опубликовали в 1964 г. наш отчет об использовании анализа ссылок для написания истории науки. Этот проект, в конечном счете, привел к созданию вместе с Александром Пудовкиным и Владимиром Истоминным программного обеспечения HistCite для создания историографии как побочного продукта поисков в Web of Science [20].

Пожалуйста, обращайтесь к моему сетевому сайту (<http://garfield.library.upenn.edu/histcomp/>) за дюжиной других мини-историй науки, которые освещаются хронологической историографией для ученых Нобелевского класса, таких как Альберт Эйнштейн, Дж. Д. Уотсон, Дж. Х. Хаббел, Вернер Гейзенберг, Макс Планк, Джошуа Ледерберг, Рональд Хофман и многие другие. И я должен добавить, что расширенная база данных Web of Science делает эти исторические анализы возможными, поскольку мы вовремя возвращаемся назад, чтобы обработать важную литературу двадцатого века. Короче говоря, я всегда верил, что проследивание истории и эволюции тематик тесно связано с задачей документирования не только новых научных журнальных статей и книг, но также и с проследиванием источников технологий в анализе патентных ссылок.

Многие ученые и редакторы часто совершают даже более грубую ошибку, предполагая, что SCI был создан только для того, чтобы производить его побочную базу данных, названную *Journal Citation Reports* (JCR). Эта база данных в настоящее время обеспечивает сведения о факторах влияния примерно 15 тыс. журналов ежегодно. При рассмотрении разнообразного круга редакционных комментариев в защиту и против факторов влияния журналов, становится понятно, что научные редакторы и администраторы, являющиеся новыми для области библиотекостроения, полагают, что SCI и его преемница база данных Web of Science были созданы для вычисления факторов влияния журналов. Рост числа конференций COLLNET и ISSI и этот симпозиум показывают, что стоит говорить о цитационологии (citationology), или анализе ссылок, которая не только является хвостом, виляющим собакой, но хвостом, ставшим огромным животным, которое быстро превращается в страдающего множественным параличом монстра. Действительно, научная оценка и создание научных показателей - это индустрия, которая управляет современным научно-исследовательским предприятием через его влияние на администраторов и тех, кто делает политику. Будь то отчет NSF Biennial Science & Engineering Indicators в США, ежегодные отчеты JCR от ThomsonReuters или Research Assessment Exercise (RAE) в Великобри-

тании или их аналоги в других странах, я полагаю, что нет возврата в прошлое.

Рассматривая влияние публикационных исследований на личную карьеру и решения относительно финансирования, всегда крайне важно, чтобы научное сообщество и особенно редакторы журналов применяли «мягкое искусство цитирования» [20]. Мы не можем позволить ученым поколения Google повторить то, что Гинзбург назвал «синдромом пренебрежения» [21]. Каждый автор, как и каждый потенциальный владелец патента, должен гарантировать, что соответствующие исторические прецеденты цитируются правильно, чтобы обеспечить точность. Когда это произойдет, мы, возможно, достигнем библиографической nirваны.

ПРИЛОЖЕНИЕ

(Если позволит время.)

Мне бы хотелось повторить мудрые высказывания Питера Суорта и Поля Карлинга, касающиеся этической и критической роли цитирования, как они опубликованы в журнале «*Sedimentology*», журнале, который большинство из нас не читает. Их высказывания констатируют данный аспект очень хорошо, и поэтому я буду цитировать непосредственно их вышеупомянутую редакционную статью ... «Ссылки, которые используют авторы, раскрывают много важного относительно того, что авторы думают о предмете, как статья будет оцениваться рецензентами и, в конце концов, сколько раз сама статья будет цитироваться. Первая проблема, которую мы как редакторы видим неоднократно, относится к цитированию первичной ссылки на определенный феномен. Часто авторы скорее ссылаются на более позднюю статью, а не на оригинальную работу».

«Некоторые авторы не хотят выполнять трудную работу по выявлению того, кто первым выступил с определенной идеей, поэтому они используют легкий способ, ссылаясь на совокупную группу статей. Эта ссылка «защищает ваш тыл», таким образом вы можете отдать долг многим лицам, включая тех, кто, может быть, не имел ничего общего с созданием этой идеи первоначально. Эта проблема особенно вредна для некоторых статей, в которых каждая отдельная ссылка сопровождается списком статей, делаая цитирование, по существу, бесполезным. Наконец, мы имеем проблему чрезмерного самоцитирования. Люди, которые чувствуют, что на них недостаточно ссылаются, стараются втиснуть как можно большее число своих статей в библиографию. Естественно, надо цитировать себя в статье, как правило, по вполне обоснованным причинам, но когда коэффициент самоцитирования становится чрезмерным, это плохо отражается на статье, авторе и журнале. Мы пытались скорректировать эти проблемы, но это все же является делом авторов, которые знают свою область лучше, чем кто-либо еще, чтобы обоснованно включать ссылки. Найдите, кто первоначально отвечает за идею, и не приводите ссылки неразборчиво. Надлежащее использование ссылок принесет пользу каждому и отсрочит, что, вероятно, неизбежно, инфляцию цитирования».

ЛИТЕРАТУРА

1. Garfield E. Citation indexes for science: A new dimension in documentation through association of ideas // Science. – 1955. – Vol. 122. – P. 108-111. – http://www.garfield.library.upenn.edu/papers/science_v122v3159h108y1955.html.

(Reprinted Int J Epidemiol. – 2006. – Vol. 35. – P. 1123-1127).

2. *Garfield E.* Citation analysis as a tool in journal evaluation // Science. – 1972. – Vol. 178. – P. 471-479. (Reprinted in Current Contents, 1973; in Essays of an Information Scientist. – Vol. 1. – P. 528-544. – Philadelphia: ISI Press, 1977. – <http://www.garfield.library.upenn.edu/essays/V1p527y1962-73.pdf>

3. *Garfield E.* Citation indexing: Its theory and application in science, technology, and humanities. – Philadelphia: ISI Press, 1977. – P. 79. – <http://www.garfield.library.upenn.edu/ci/title.pdf>

4. *Garfield E.* From citation indexes to informetrics: Is the tail wagging the dog? // Libri. – 1998. – Vol. 48. – P. 67-80. – [http://www.garfield.library.upenn.edu/papers/libriv48\(2\)p67-80y1998.pdf](http://www.garfield.library.upenn.edu/papers/libriv48(2)p67-80y1998.pdf)

5. *Garfield E.* Science Citation Index: A new dimension in indexing // Science. – 1964. – Vol. 144. – P. 649-654. – <http://www.garfield.library.upenn.edu/essays/v7p525y1984.pdf>

6. *Small H.G.* Cited documents as concept symbols // Soc Stud Sci. – 1978. – Vol. 8. – P. 327-340. – <http://www.garfield.library.upenn.edu/small/hsmallsoctuds civ8y1978.pdf>

7. *Margolis J.* Citation indexing and evaluation of scientific papers // Science. – 1967. – Vol. 155. – P. 12-13. – <http://www.sciencemag.org/cgi/content/abstract/155/3767/1213>.

8. *Garfield E.* Citation indexing for studying science // Nature. – 1977. – Vol. 227. – P. 669-671. (Reprinted in Current Contents, 1970; Essays of an Information Scientist. – Vol. 1. – P. 132-138. – Philadelphia: ISI Press, 1970. – <http://garfield.library.upenn.edu/papers/naturev227p669y1970.pdf>

9. *Bush V.* As we may think // Atlantic Monthly. – 1945. – Vol. 176. – P.101-108. – <http://www.theatlantic.com/doc/194507/bish>.

10. *Garner R.* A computer oriented, graph theoretic analysis of citation index structures. – Philadelphia: Drexel University Press, 1967. – <http://www.garfield.library.upenn.edu/rgarner.pdf>

11. *Price D.J.D.* Networks of scientific papers: The pattern of bibliographic references indicates the nature of the scientific research front // Science. – 1965. – Vol. 149. – P. 510-515. – <http://garfield.library.upenn.edu/papers/pricenetworks1965.pdf>

12. *Garfield E., Sher I.H., Torpie R.J.* The use of citation data in writing the history of science. – <http://www.garfield.library.upenn.edu/papers/useofcitdata writinghistofsci.pdf>

13. *Kochen M.* WISE: Word information synthesis and encyclopedia // J Document. – 1972. – Vol. 28. – P. 322-343.

14. *Garfield E., Sher I.H.* ISI's experiences with ASCA: A selective dissemination system // J Chem Document. – 1967. – Vol. 7. – P. 147-153. (Reprinted in Essays of an Information Scientist. – Vol. 6. – P. 533. – Philadelphia: ISI Press, 1984). – <http://www.garfield.library.upenn.edu/essays/v6p533y1983.pdf>

15. *Garfield E.* From citation indexes to informetrics: Is the tail wagging the dog? // Libri. – June 1998. – Vol. 48, No. 2. – P. 67-80. – [http://www.garfield.library.upenn.edu/papers/libriv48\(2\)p67-80y1998.pdf](http://www.garfield.library.upenn.edu/papers/libriv48(2)p67-80y1998.pdf)

16. *Davis P.M.* Reward or persuasion? The battle to define the meaning of a citation // Learned Publishing. – January 2009. – Vol. 22, No. 1. – P. 5-11.

17. *Koenig M.* E-message dated February. – 2009. – Vol. 9.

18. *Garfield E.* From The Science of Science to Scientometrics: Visualizing the history of science with HistCite software // Presented at the 11th ISSI International Conference, Madrid, Spain – June 25, 2007. – <http://garfield.library.upenn.edu/papers/issispain2007.pdf>

19. *Garfield E.* From The Science of Science to Scientometrics: Visualizing the history of science with HistCite software // Presented at the 11th ISSI International Conference, Madrid, Spain – June 25, 2007. (Reprinted in Special Issue of Journal of Informetrics. – 2009. – Vol. 3. – P. 173-179).

20. *Swart P., Carling P.* Sedimentology. – October 2008. – Vol. 55. – P. 115-116.

21. *Ginsburg I.* The disregard syndrome: A menace to honest science // The Scientist. – December 10, 2001. – Vol. 15, No. 24. – P. 51.

Подсчет публикаций и ссылок: больше всегда ли лучше?*

Людो ВАЛТМАН
(Ludo WALTMAN),

Нес Ян ван ЭК
(Nees Jan van ECK),

Пол УОТЕРС (Paul WOUTERS)

Центр исследований науки и технологий,
Лейденский университет, г. Лейден, Нидерланды

Больше всегда ли лучше? Мы исследуем этот вопрос в контексте библиометрических индексов с целью оценить научное влияние отдельных ученых путем подсчета числа их высокоцитируемых публикаций. Мы предлагаем простую модель, в которой число ссылок на публикацию соответствует научному влиянию публикации, а также зависит от других «случайных» факторов. Наша модель показывает, что больше не всегда значит лучше. Получается, что самые влиятельные ученые могут иметь систематически низкую продуктивность с точки зрения высокоцитируемых публикаций, чем некоторые из их менее влиятельных коллег. Эта модель также предполагает улучшенный способ подсчета высокоцитируемых публикаций.

ВВЕДЕНИЕ

Когда библиометрия применяется для целей научной оценки, общее предположение кажется таким, что больше всегда лучше. Чем больше публикаций, тем лучше; чем больше ссылок, тем лучше. В то же самое время все больше осознается, что «больше всегда лучше» не должно приниматься слишком буквально. Например, интерпретация числа ссылок на публикацию как *приблизительная* мера научного влияния публикации, имеющей больше ссылок, не всегда соответствует наличию большего влияния. Публикации с большим числом ссылок могут в *среднем* иметь больше влияния, но *отдельные* публикации могут отклоняться от этой модели. Можно предположить, например, что авторы публикации стремятся скопировать значительную часть их библиографического списка из списка ссылок более ранних публикаций, часто не обращая особого внимания на содержимое цитируемых работ [1,2]. Если, в самом деле, имеется доля истины в этой идее, то не кажется маловероятным, что публикации иногда становятся высокоцитируемыми без действительного большого влияния на прогресс науки. Это иллюстрирует то, что не существует идеальной связи между научным влиянием и ссылками. Действительно, связь между научным влиянием и ссылками является избыточной. Помимо научного влияния имеется много других факторов, влияющих на число ссылок на публикацию [3-5]. В этой статье мы будем интерпретировать эти другие факторы как «случайные эффекты» в отношениях между научным влиянием и ссылками.

Также при оценке научного влияния совокупности публикаций, а не одной отдельной работы, идея того, что

больше значит лучше, должна рассматриваться с осторожностью. Не очевидно, например, является ли сравнение шедевров двух ученых, основанное на общем числе ссылок каждого ученого, хорошим подходом. Один ученый может иметь больше ссылок, чем другой, а может быть так, что второй ученый имеет авторизованное число высокоцитируемых публикаций, тогда как первый ученый заслужил свои ссылки, создавая обширный шедевр, содержащий исключительно низко- и среднецитируемые работы. В такой ситуации ученый с высокоцитируемыми публикациями может быть действительно более влиятельным, несмотря на его малое общее число ссылок. При оценке научного влияния ученого на основе общего числа ссылок на его публикации, *приблизительное* предположение является таким, что число ссылок на публикацию пропорционально научному влиянию публикации. Это вполне строгое предположение. Как утверждают авторы [6], истинная связь между научным влиянием и ссылками может быть действительно нелинейной.

В последние годы было введено большое число библиометрических индексов (показателей), которые могут служить альтернативой подсчету общего числа ссылок ученого. Самым известным примером, конечно, является *h*-индекс [7]. Этот индекс надежен как для публикаций с очень малым числом ссылок, так и публикаций с очень большим числом ссылок. Эта надежность часто считается сильным свойством *h*-индекса. Тем не менее, к сожалению, *h*-индекс обладает другими свойствами, которые трудно объяснить и которые могут вызывать противоречия в результатах, производимых с помощью индекса [8]. Привлекательной альтернативой *h*-индексу является индекс высокоцитируемых публикаций [8, 9]. Этот индекс учитывает число публикаций ученого, которые получили, по крайней мере, некоторое минимальное число ссылок [10,11]. Индекс высокоцитируемых публикаций имеет похожее свойство надежности, что и *h*-индекс, но он не подвержен противоречиям *h*-индекса.

* Перевод Waltman L., van Eck N.J., Wouters P. Counting publications and citations: Is more always better? — 2013.— <http://arxiv.org/ftp/arxiv/papers/1301/1301.4597/pdf>

В этой статье наше внимание обращено к индексу высокоцитируемых публикаций. Изучаемый вопрос, рассматриваемый нами, состоит в том, является ли больше всегда лучшим при подсчете высокоцитируемых публикаций. Для изучения этого вопроса мы вводим простую модель связи между научным влиянием и ссылками. Эта модель показывает, что в результате вышеупомянутых случайных эффектов в связи между научным влиянием и ссылками, ответ на исследуемый нами вопрос отрицателен. Само по себе это не может быть удивительным. При работе с небольшими количествами публикаций ожидается, что случайные эффекты могут вызывать отклонения от принципа, что больше — лучше. Например, ученый с одной высокоцитируемой публикацией не обязательно всегда будет более влиятельным, чем ученый, который не имеет ни одной высокоцитируемой публикации. Тем не менее, наша модель открывает то, что случайные эффекты могут привести к отклонениям от принципа больше значит лучше, который носит систематический характер. Эти отклонения происходят, даже когда мы имеем дело с большими количествами публикаций. В конкретных терминах эта модель демонстрирует, как случайные эффекты в связи между научным влиянием и ссылками могут приводить к парадоксальным ситуациям, в которых самые влиятельные ученые имеют систематически более низкую продуктивность с точки зрения высокоцитируемых публикаций, чем некоторые из их менее влиятельных коллег. Эта модель также предполагает, как индекс высокоцитируемых публикаций должен быть модифицирован, чтобы избежать парадоксальных ситуаций.

Перед тем как продолжить наш анализ, важно отметить, что проблема, изучаемая в этой статье, не относится конкретно к индексу высокоцитируемых публикаций. Мы сосредотачиваемся на индексе высокоцитируемых публикаций, так как он является важным библиометрическим индексом, который благодаря простоте своего определения может быть проанализирован соответствующим образом. Однако наблюдения, подобные нашим наблюдениям, могут быть сделаны также для других библиометрических индексов.

БОЛЬШЕ НЕ ВСЕГДА ОБЯЗАТЕЛЬНО ЛУЧШЕ

Решающее отличие нашего анализа лежит между числом ссылок на публикацию и научным влиянием публикации. Для цели нашего анализа то, что действительно понимается под термином «научное влияние», не является таким важным, хотя должно быть ясно, что это влияние не является просто синонимичным ссылкой. Одна интерпретация, которую можно дать понятию научного влияния, — это вклад публикации в долговременный прогресс науки. Другая интерпретация может больше фокусироваться на краткосрочном влиянии публикации. Краткосрочное научное влияние должно, вероятно, быть сильнее связано со ссылками, чем влияние долговременное. В нашем анализе мы будем использовать понятие научного влияния вообще, без особого акцента на типе влияния, который мы подразумеваем.

Сейчас мы представляем простую модель связи между научным влиянием и ссылками. Предположим, мы хотим измерить общее научное влияние ученого, и полагаем, что общее научное влияние ученого определяется числом высоковлиятельных публикаций, произведенных ученым. Мы определяем высоковлиятельную публикацию как публикацию, принадлежащую к ведущим 10% публикаций ее области с точки зрения ее научного влияния. Публикации, которые по их научному влиянию не принадлежат к 10% ведущим публикаций их области, относятся к низковлиятельным пуб-

ликациям. Мы полагаем, что научное влияние низковлиятельных публикаций должно игнорироваться.

Научное влияние публикации не может быть непосредственно наблюдаемым, и поэтому мы рассматриваем число ссылок на публикацию. Мы различаем два класса публикаций: публикации, которые принадлежат к ведущим 10% публикаций в своей области с точки зрения ссылок и публикаций, которые, основываясь на их числе ссылок, не принадлежат к ведущим 10% публикаций в своей области. Мы относим публикации, принадлежащие к ведущим 10% публикаций, наиболее часто цитируемых своей областью, к высокоцитируемым публикациям.* Публикации, которые не принадлежат к ведущим 10% публикаций, наиболее цитируемых своей областью, относятся к низкоцитируемым публикациям. Подсчет числа высокоцитируемых публикаций ученого приводит к вышеупомянутому индексу высокоцитируемых публикаций.

В идеальном мире, в котором существует совершенная (идеальная) корреляция между научным влиянием публикации и числом ссылок на публикацию, высокоцитируемая публикация соотносится с наличием высокого влияния. Другими словами, каждая высокоцитируемая публикация является также высоковлиятельной публикацией, и наоборот. В таком идеальном мире индекс высокоцитируемых публикаций точно указывает на число высоковлиятельных публикаций ученого, и поэтому этот индекс всегда обеспечивает правильную оценку научного влияния ученого.

Однако, как мы уже обсудили, идея совершенной связи научного влияния и ссылок является трудно подтверждаемой. Некоторые публикации являются высокоцитируемыми, даже если они имеют только ограниченное научное влияние. Наоборот, некоторые публикации принадлежат к ведущим 10% высоковлиятельных публикаций своей области, даже если они не принадлежат к ведущим 10% самых высокоцитируемых публикаций своей области. Возможный сценарий показан в табл. 1. В этом сценарии 3% публикаций в области имеют высокое влияние и являются также высокоцитируемыми, тогда как 7% публикаций имеют высокое влияние, но не являются высокоцитируемыми, и еще 7% публикаций являются высокоцитируемыми, но не имеют высокого влияния. Остальные 83% публикаций имеют низкое влияние и также являются низкоцитируемыми.

В сценарии, показанном в табл. 1, мы можем иметь следующую интересную ситуацию. Предположим, у нас имеется два ученых, ученый А и ученый В (см. табл. 2). Ученый А создал 100 публикаций, и все они высоковлиятельные. Ученый В создал 500 публикаций, т.е. в пять раз больше, чем ученый А, но ни одна из этих публикаций не имеет высокого влияния. Учитывая наше предположение, что общее научное влияние ученого определяется числом высоковлиятельных публикаций, созданных ученым, мы должны сделать вывод, что ученый А является высоковлиятельным, тогда как научное влияние ученого В незначительно, несмотря на большое число публикаций этого ученого.

Интересным вопросом является, подтверждает ли индекс высокоцитируемых публикаций этот вывод. Учитывая доли, представленные в табл. 1, мы можем ожидать, что ученый А имеет $(3\%/10\%) \times 100 = 30$ высокоцитируемых публикаций. Что касается ученого В, то ожидаемое

*Для целей нашего анализа практические трудности в определении того, принадлежит ли публикация к ведущим 10% публикаций, наиболее часто цитируемых своей областью [12], могут игнорироваться.

число высокоцитируемых публикаций равняется $(7\%/90\%) \times 500 \approx 39$. Если ученые А и В действительно каждый имеют свое статистически ожидаемое число высокоцитируемых публикаций, мы оказываемся в парадоксальной ситуации, в которой индекс высокоцитируемых публикаций показывает, что ученый В со значением индекса высокоцитируемых публикаций, равным 39, вероятно, является более влиятельным, чем ученый А со значением индекса высокоцитируемых публикаций, равным 30. Следовательно, индекс высокоцитируемых публикаций обеспечивает неправильную оценку научного влияния двух ученых. Более того, эта неправильная оценка не вызвана случайным статистическим колебанием. Так как ученые А и В каждый имеют свое статистически ожидаемое число высокоцитируемых публикаций, то индекс высокоцитируемых публикаций систематически неверен в ситуациях, подобных нашей.

Итак, почему индекс высокоцитируемых публикаций в некоторых ситуациях дает систематически неверные оценки научного влияния ученых? Это потому, что пока нет идеальной связи между научным влиянием и ссылками, ученый с данным числом высоковлиятельных публикаций может всегда уступать с точки зрения высокоцитируемых публикаций другому ученому с достаточно большим числом низковлиятельных публикаций. Низковлиятельные публикации менее вероятно становятся высокоцитируемыми, чем высоковлиятельные публикации, но благодаря созданию множества низковлиятельных публикаций все же возможно получить большое число высокоцитируемых публикаций. Что демонстрирует вышеупомянутый сценарий, так это то, что больше не всегда значит лучше при подсчете высоко-

цитируемых публикаций. Здесь могут быть систематические отклонения от принципа больше значит лучше. В частности, индекс высокоцитируемых публикаций может переоценивать научное влияние ученых, которые сосредоточились на создании множества публикаций, не уделяя большого внимания влиянию их работы.

Табл. 3 дает обобщенное представление о табл. 1. Параметр α определяет степень соответствия научного влияния и ссылок. Идеальная корреляция получается установкой α на нуль. Другая крайность – установка значения α , равного 0,09, в таком случае научное влияние и ссылки вообще не соотносятся, и число ссылок на публикацию совсем не указывает на научное влияние публикации. Отсутствие любой корреляции между научным влиянием и ссылками для $\alpha=0,09$ следует из того факта, что установка значения α , равного 0,09 заставляет каждую ячейку табл. 3 быть равной суммарному произведению соответствующего ряда и колонки. Вероятность установки значения α , равного выше 0,09, может игнорироваться. Это приведет к неправдоподобной ситуации отрицательной корреляции между научным влиянием и ссылками. Установка значения α , равного 0,07, приведет к сценарию, отраженному в табл.1. Наконец, значение α , которое можно считать наиболее реальным, зависит от того, как много надежды возлагается на способность ссылок показывать научное влияние публикации. Оно также зависит от точной интерпретации того, что придается понятию научного влияния (например, краткосрочное влияние против долгосрочного влияния).

Таблица 1

Иллюстрация сценария, в котором нет идеальной связи между научным влиянием публикации и числом ссылок на публикацию, %

	Низкоцитируемые публикации	Высокоцитируемые публикации	Всего
Низковлиятельные публикации	83	7	90
Высоковлиятельные публикации	7	3	10
Всего	90	10	100

Таблица 2

Использование некоторых ученых для иллюстрации последствий различных подходов к подсчету высокоцитируемых публикаций

	Число публикаций		Число публикаций	
	низковлиятельных	высоковлиятельных	низковлиятельных	высоковлиятельных
Ученый А	0	100	70	30
Ученый В	500	0	461	39
Ученый С	50	200	186	64
Ученый D	270	70	298	42

Таблица 3

Научное влияние и ссылки. Параметр α определяет степень корреляции ($0 \leq \alpha \leq 0,09$)

	Низкоцитируемые публикации	Высокоцитируемые публикации	Всего
Низковлиятельные публикации	$0,9-\alpha$	α	0,9
Высоковлиятельные публикации	α	$0,1-\alpha$	0,1
Всего	0,9	0,1	1

На основе табл. 3 можно видеть, что создание n_{HI} высококачественных публикаций в среднем дает результат в $[(0,1-\alpha) / 0,1] \times n_{HI}$ высокоцитируемых публикаций. Подобным образом, создание n_{LI} низкокачественных публикаций в среднем дает результат $[\alpha / 0,9] \times n_{LI}$ высокоцитируемых публикаций. Из этого следует, что получение одной высокоцитируемой публикации в среднем потребует $1 / [(0,1-\alpha) / 0,1]$ высококачественных публикаций или $1 / [\alpha / 0,9]$ низкокачественных публикаций. Ясно, что чем ниже значение α , тем больше индекс высокоцитируемых публикаций поощряет создание высококачественных публикаций. Тем не менее, для любого ненулевого значения α ученый с определенным числом высококачественных публикаций может систематически быть превзойденным с точки зрения высокоцитируемых публикаций ученым со множеством низкокачественных публикаций. Точнее, ученый, создающий больше, чем $[(0,1-\alpha) / 0,1] / [\alpha / 0,9] \times n_{HI} = (0,9-9\alpha) / \alpha \times n_{HI}$ низкокачественных публикаций в среднем превосходит коллегу, создающего n_{HI} высококачественных публикаций. Безусловно, если значение α близко к нулю, то число низкокачественных публикаций, необходимое, чтобы превзойти ученого с n_{HI} высококачественных публикаций, становится очень большим, и на практике может быть невозможным иметь столь большую совокупность публикаций.

УЛУЧШЕННЫЙ ПОДХОД ПОДСЧЕТА

Очевидным вопросом является, может ли быть изменен индекс высокоцитируемых публикаций так, чтобы он больше не подвергался систематическим ошибкам в оценке научного влияния ученых. Другими словами, возможна ли разработка улучшенного способа подсчета высокоцитируемых публикаций?

Одной возможностью может быть движение от величинно независимого индекса высокоцитируемых публикаций к независимому от величины. В таком случае вместо подсчета *числа* высокоцитируемых публикаций ученого будет подсчет *доли* высокоцитируемых публикаций ученого. В некоторых ситуациях это в самом деле приведет к улучшенным результатам. Например, рассмотрим сценарий, отраженный в табл. 1, и возьмем ситуацию ученых А и В, обсуждавшуюся в предыдущем разделе (см. табл. 2). Ученый А создал 100 высококачественных публикаций, из которых 30 являются высокоцитируемыми. Ученый В создал 500 низкокачественных публикаций, из которых 39 являются высокоцитируемыми. Как мы видели, если смотреть на число высокоцитируемых публикаций ученого, ученый В превосходит ученого А, даже если научное влияние ученого В является незначительным по сравнению с ученым А. Теперь предположим, что мы смотрим на долю высокоцитируемых публикаций ученого, т.е. число высокоцитируемых публикаций ученого, деленное на общее число его публикаций. Ученый А имеет $30 / 100 = 30\%$ высокоцитируемых публикаций, тогда как ученый В имеет только $39 / 500 = 7,8\%$ высокоцитируемых публикаций. Следовательно, если смотреть на долю высокоцитируемых публикаций ученого, то ученые А и В ранжируются правильно относительно друг друга.

К сожалению, величинно зависимый индекс высокоцитируемых публикаций также имеет недостатки. Чтобы показать это, мы вводим третьего ученого, ученого С (см. табл. 2). Предположим, что ученый С создал 200 высококачественных публикаций и 50 низкокачественных публикаций. В соответствии с отраженными в

табл. 1 долями это дает результат в $(3\%/10\%) \times 200 + (7\%/90\%) \times 50 \approx 64$ высокоцитируемых публикаций. Так как ученый С создал в два раза больше высококачественных публикаций, чем ученый А, то научное влияние ученого С также в два раза больше, чем ученого А. Однако ученый А имеет 30% высокоцитируемых публикаций, тогда как ученый С имеет только $64 / (200+50) = 25,6\%$ высокоцитируемых публикаций. Следовательно, согласно величинно независимому индексу высокоцитируемых публикаций, ученый А превосходит ученого С. Ясно, что это является неправильной оценкой научного влияния двух ученых.

Фундаментальной проблемой величинно зависимо-го индекса высокоцитируемых публикаций является то, что продуктивность не поощряется. Если два ученых имеют одинаковую долю высокоцитируемых публикаций, то их научное влияние оценивается также одинаково. Это не имеет смысла, если ученый, например, имеет в два раза большую совокупность публикаций, чем другой ученый. При остальных равных фактах научное влияние ученого должно оцениваться пропорционально его совокупности публикаций. Если один ученый имеет в два раза больше высокоцитируемых и в два раза больше низкоцитируемых публикаций, чем другой ученый, научное влияние первого ученого должно оцениваться в два раза больше, чем научное влияние второго ученого. Величинно независимый индекс высокоцитируемых публикаций терпит неудачу в учете таких случаев продуктивности.

Оказывается, есть лучший способ, в котором индекс высокоцитируемых публикаций может быть модифицирован, чтобы гарантировать, что он обеспечивает правильную оценку научного влияния ученых. Индекс высокоцитируемых публикаций может рассматриваться как взвешенная сумма публикаций ученого, в которой высокоцитируемая публикация имеет свой вес, а низкоцитируемая публикация имеет вес, равный нулю. Теперь мы видим, что веса, используемые в индексе высокоцитируемых публикаций, могут быть изменены таким образом, чтобы в среднем значение индекса высокоцитируемых публикаций ученого точно равнялось числу высококачественных публикаций, созданных ученым.

Нашей отправной точкой является общий сценарий, показанный в табл. 3, с параметром α ($0 \leq \alpha \leq 0,09$), определяющим степень соотношения научного влияния и ссылок. Мы предлагаем взвешивать высокоцитируемые публикации следующим образом:

$$w_{HC} = \frac{0,1\alpha - 0,09}{\alpha - 0,09}, \quad (1)$$

а низкоцитируемые публикации таким образом:

$$w_{LC} = \frac{0,1\alpha}{\alpha - 0,09}. \quad (2)$$

Значит, значение индекса высокоцитируемых публикаций ученого задается так:

$$HCP = n_{LC} w_{LC} + n_{HC} w_{HC}, \quad (3)$$

где n_{LC} и n_{HC} подразумевают число низко- и высокоцитируемых публикаций ученого. Заметим, что задание значения α , равного нулю, приводит к $w_{HC} = 1$ и $w_{LC} = 0$, что означает, что формула (3) сводится к стандартному индексу высокоцитируемых публикаций, обсуждавшемуся в предыдущем разделе. Отметим также, что w_{HC} и w_{LC} не определяются, если α имеет значение, равное 0,09. Как мы видим в предыдущем разделе, если α имеет значение, равное 0,09, то число ссылок на публикацию никак не указывает на научное влияние публикации.

Предположим, ученый создал n_{HI} высококовлятельных публикаций и n_{LI} низкоковлятельных публикаций. Ожидаемое значение индекса высокоцитируемых публикаций ученого, подсчитываемое по формулам (1)–(3), тогда равняется n_{HC} . Это может рассматриваться как следующее. На основе табл. 3 мы получаем

$$E(n_{HC}) = \frac{0,1-\alpha}{0,1}n_{HI} + \frac{\alpha}{0,9}n_{LI} \quad (4)$$

$$\text{и } E(n_{LC}) = \frac{\alpha}{0,1}n_{HI} + \frac{0,9-\alpha}{0,9}n_{LI}, \quad (5)$$

где $E(\bullet)$ означает ожидаемый значимый оператор. Из формулы (3) следует, что

$$E(HCP) = E(n_{LC})w_{LC} + E(n_{HC})w_{HC}. \quad (6)$$

Подстановка формул (1), (2), (4) и (5) в формулу (6) приводит к

$$E(HCP) = n_{HI}. \quad (7)$$

Это доказывает, что в среднем значение индекса высокоцитируемых публикаций ученого, подсчитанное с использованием формул (1) – (3), точно равняется числу высококовлятельных публикаций, созданных ученым. В отличие от стандартного индекса высокоцитируемых публикаций наш измененный индекс высокоцитируемых публикаций не подвержен систематическим ошибкам в оценке научного влияния ученого.

Чтобы понять механизм нашего модифицированного индекса высокоцитируемых публикаций, важно видеть, что вес w_{LC} в формуле (2) всегда отрицательный (за исключением, когда значение α равно нулю). Значит, низкоцитируемым публикациям придается отрицательный вес в нашем измененном индексе высокоцитируемых публикаций. При равенстве других вещей, чем больше имеется низкоцитируемых публикаций, тем ниже значение индекса высокоцитируемых публикаций. Почему мы придаем отрицательный вес низкоцитируемым публикациям? Учитывая наше предположение о том, что научное влияние низкоковлятельных публикаций является незначительным, мы хотим свести в среднем к нулю вклад низкоковлятельной публикации в значение индекса высокоцитируемых публикаций ученого. Однако из-за случайных эффектов в связи между научным влиянием и ссылками некоторые низкоковлятельные публикации станут высокоцитируемыми, и эти публикации вносят положительный вклад в значение индекса высокоцитируемых публикаций ученого. Чтобы компенсировать это, мы придаем отрицательный вес низкоцитируемым публикациям. Этот отрицательный вес выбирается таким образом, чтобы в среднем вклад низкоковлятельной публикации в значение индекса высокоцитируемых публикаций ученого был равен нулю. Что касается высококовлятельной публикации, то мы хотим, чтобы ее вклад в значение индекса высокоцитируемых публикаций ученого был в среднем высококовлятельным. Используя веса в формулах (1) и (2), мы выполняем обе наши цели: низкоковлятельные публикации вносят в среднем нулевой вклад, а высококовлятельные публикации вносят в среднем высококовлятельный.

Наконец, имеется интересное свойство нашего модифицированного индекса высокоцитируемых публикаций, которое мы хотим продемонстрировать. Снова рассматриваем сценарий, отраженный в табл. 1. Давайте введем нового ученого, ученого D (см. табл. 2). Предположим, что этот ученый создал 70 высококовлятельных и 270 низкоковлятельных публикаций. В таком случае, он получает ожидаемое число $(3\% / 10\%) \times 70 + (7\% / 90\%) \times$

$270 = 42$ высокоцитируемых публикаций. Его остальные $70 + 270 - 42 = 298$ публикации являются низкоцитируемыми. Задавая значение α , равное 0,07, в формулах (1) и (2), мы получаем $w_{HC} = 4,15$ и $w_{LC} = -0,35$. Используя формулу (3), тогда находим, что значение индекса высокоцитируемых публикаций ученого D равняется $298 \times (-0,35) + 42 \times 4,15 = 70$. Значит, как ожидалось, значение индекса высокоцитируемых публикаций ученого D равняется числу его высококовлятельных публикаций. Подобный подсчет может быть произведен для ученого A, введенного ранее (см. табл. 2). Напомним, что этот ученый создал 100 высококовлятельных публикаций, которые привели к 30 высокоцитируемым и 70 низкоцитируемым публикациям. На основе числа высоко- и низкоцитируемых публикаций мы получаем значение индекса высокоцитируемых публикаций, равное 100, для ученого A, которое точно равняется числу высококовлятельных публикаций, созданных этим ученым. Сравнивая ученых A и D, наш измененный индекс высокоцитируемых публикаций правильно определяет ученого A как ученого с большим научным влиянием.

Что интересно в сравнении ученых A и D так это то, что ученый A уступает ученому D с точки зрения как высокоцитируемых публикаций (30 против 42), так и низкоцитируемых публикаций (70 против 298). Интуитивно это может казаться достаточным свидетельством, чтобы прийти к выводу, что ученый D должен иметь большее научное влияние, чем ученый A. Однако, как мы видим, ученый A является ученым с более высоким научным влиянием. Значит, на основе простой логики, что больше значит лучше, можно легко сделать неправильный вывод в сравнении ученых A и D. Отходом от логики, что больше значит лучше, наш измененный индекс высокоцитируемых публикаций достигает правильного вывода.

ОБСУЖДЕНИЕ И ЗАКЛЮЧЕНИЕ

Принцип, что больше значит лучше, играет центральную роль в оценочной библиометрии. В этой статье мы привели примеры ситуаций, в которых больше не всегда значит лучше. Когда общее научное влияние ученых определяется числом их высококовлятельных публикаций, потребность в наличии более высокоцитируемых публикаций не всегда соотносится с наличием более высокого научного влияния. Это вызывается случайными эффектами в связях между научным влиянием и ссылками. Чем сильнее эти случайные эффекты, тем более трудным становится поддерживать принцип, что больше значит лучше. Важно, что отклонения от принципа больше значит лучше, которые мы изучаем, носят систематический характер. Они не просто происходят из случайных статистических колебаний. Это показывает, что в противоположность тому, что иногда кажется требуемым [13], случайные эффекты в ссылках нельзя сводить на нет. В самом деле, случайные эффекты могут иметь систематические последствия, по крайней мере, при использовании некоторых типов библиометрических индексов.

Модель, анализируемая нами в этой статье, является чрезвычайно стилизованной. С одной стороны, это делает эту модель легкой для изучения, но, с другой стороны, также означает, что данная модель имеет значительные недостатки. Самым важным недостатком может быть то, что научное влияние публикации предполагается быть бинарной переменной: публикация либо имеет, либо не имеет научного влияния. Хотя это, безусловно, нереальное предположение, оно хорошо соотносится с идеей подсчета высокоцитируемых публикаций, которая также

полагается на бинарное различие, хотя основанное скорее на ссылках, чем на влиянии. Дальнейшая работа должна фокусироваться на создании более детальных моделей связи между научным влиянием и ссылками, чтобы найти, в каких условиях наши результаты остаются или не остаются обоснованными.

Мы подчеркиваем, что рассматриваем модифицированный индекс высокоцитируемых публикаций, введенный в разделе «Улучшенный подход подсчета», в основном для поддержки теоретического интереса. Для получения соответствующих весов для низко- и высокоцитируемых публикаций потребуется иметь реальное значение параметра α . Не ясно, как такое значение может быть определено эмпирически. Более того, наш модифицированный индекс высокоцитируемых публикаций полностью основан на нашей очень простой модели связи между научным влиянием и ссылками. Это делает данный индекс уязвимым перед недостатками этой модели.

Тем не менее, мы верим, что наш модифицированный индекс высокоцитируемых публикаций обеспечивает некоторые интересные результаты. Данный индекс показывает, как случайные эффекты в связи между научным влиянием и ссылками могут быть скорректированы, чтобы на время остаться в рамках простых дополнительных индексов с их многими привлекательными свойствами [6,14]. Кроме того, наш модифицированный индекс высокоцитируемых публикаций вводит идею присвоения отрицательного веса некоторым публикациям, не потому что эти публикации имеют негативное влияние, а просто как тип корректирующего фактора, чтобы гарантировать, что данный индекс в среднем дает правильные результаты. Мы подчеркиваем, что результаты, полученные нами для индексов высокоцитируемых публикаций, могут быть также применимы к другим библиометрическим индексам.

Мы надеемся, что эта статья стимулирует большее исследование в развитии библиометрических индексов с подходом, основанным на модели, в частности, с подходом, в котором связь между ссылками, с одной стороны, и понятиями, такими как научное влияние и научное качество, — с другой стороны, становится явной (см. также [6]).

ЛИТЕРАТУРА

1. *Simkin M.V., Roychowdhury V.P.* Read before you cite! // *Complex Systems*. — 2003. — Vol. 14, No.3. — P. 269–274.
2. *Simkin M.V., Roychowdhury V.P.* Stochastic modeling of citation slips // *Scientometrics*. — 2005. — Vol.62, No.3. — P. 367–384.
3. *Bornmann L., Daniel H.-D.* What do citation counts measure? A review of studies on citing behavior // *Journal of Documentation*. — 2008. — Vol. 64, No.1. — P. 45–80.
4. *Moed H.F.* Citation analysis in research evaluation. — Springer, 2005.
5. *Nicolaisen J.* Citation analysis // *Annual Review of Information Science and Technology*. — 2007. — Vol. 41. — P. 609–641.
6. *Ravallion M., Wagstaff A.* On measuring scholarly influence by citations // *Scientometrics*. — 2011. — Vol. 88, No.1. — P. 321–337.
7. *Hirsch J.E.* An index to quantify an individual's scientific research output // *Proceedings of the National Academy of Sciences*. — 2005. — Vol.102, No. 46. — P. 16569–16572.
8. *Waltman L., Van Eck N.J.* (2012). The inconsistency of the *h*-index // *Journal of the American Society for Information Science and Technology*. — 2012.—Vol. 63, No.2.—P. 406–415.
9. *Bornmann L.* A better alternative to the *h* index // *Journal of Informetrics*. — 2013. — Vol. 7, No. 1. — P. 100.
10. *Plomp R.* The significance of the number of highly cited papers as an indicator of scientific prolificacy // *Scientometrics*. — 1990. — Vol. 19, No.3–4.—P. 185–197.
11. *Plomp R.* The highly cited papers of professors as an indicator of a research group's scientific performance // *Scientometrics*. — 1994. — Vol. 29, No. 3. — P. 377–393.
12. *Waltman L., Schreiber M.* On the calculation of percentile-based bibliometric indicators // *Journal of the American Society for Information Science and Technology*. (в печати)
13. *Van Raan A.F.J.* In matters of quantitative studies of science the fault of theorists is offering too little and asking too much // *Scientometrics*— 1998. — Vol. 43, No. 1.—P. 129– 139.
14. *Marchant T.* Score-based bibliometric rankings of authors // *Journal of the American Society for Information Science and Technology*. — 2009.—Vol. 60, No.6.—P. 1132–1137.

Насколько соответствующей является предсказательная сила h -индекса? Экспериментальное изучение индекса Хирша в зависимости от времени*

Михель ШПРАЙБЕР
(Michael SCHREIBER)

Институт физики, Хемницкий технический университет, г. Хемниц, Германия

Кажется, h -индекс должен иметь предсказательную силу. Здесь сообщаются результаты эмпирического изучения, показывающие, что рост h -индекса часто зависит от продолжительности ссылок во времени на довольно старые публикации. Это инертное поведение h -индекса означает, что он является трудным для использования в качестве меры предсказания будущего научного результата.

ВВЕДЕНИЕ

Индекс Хирша (или h -индекс) ученого определен Хиршем [1] как наибольшее число h публикаций ученого, которое получило, по крайней мере, по h ссылок каждая. Он стал популярным и часто используется для оценки научного достижения ученого. Но является спорным, действительно ли h -индекс считается представительной мерой [2-4]. Хирш [5] показал, что h -индекс имеет предсказательную силу в том смысле, что существует высокая корреляция между значениями индекса Хирша после 12 лет и после 24 лет карьеры ученого. Короче, « h -индекс сам по себе является лучшим предсказателем, чем любые другие показатели», исследуемые в этом изучении, и что « h -индекс имеет наибольшую способность предсказывать будущее научное достижение». Хирш приходит к выводу, что h -индекс «может полезно использоваться в процессах назначения на научную должность и для распределения научных ресурсов».

Цель настоящего исследования – усугубить сомнения относительно этого вывода. Используя эмпирические данные, я показываю, что рост h -индекса со временем после определенного момента времени (например, времени вступления в должность или времени распределения ресурсов) не обязательно связан с научными достижениями после этой даты. Я специально привожу примеры, где рост h -индекса остается тем же самым независимо от того, является ли исследуемый ученый таким же продуктивным или он не опубликовал больше ни одной другой работы.

Разумеется, вышеупомянутое соображение, что h -индекс используется в процессах найма ученых и процессах распределения научных фондов, – это только

один аспект взгляда на развитие h -индекса во времени. Другой аспект состоит в том, что значительный рост h -индекса даже после того, как ученый перестал публиковаться, свидетельствует о важности достижений этого ученого.

Недавно авторы работы [6] применили зависящий от времени h -индекс для нейроученых, используя большое число параметров (18), и смогли предсказать будущий h -индекс более точно за несколько лет. Даже сокращенный набор из пяти параметров и, следовательно, шести коэффициентов дал удивительно хорошее соответствие. Однако предостережение, что «с четырьмя параметрами я могу соответствовать слону, а с пятью параметрами я могу заставить его покачивать хоботом», которое Энрико Ферми приписывал Нейману [7], показывает, что такое соответствие может легко предполагать точность предсказания, которая не может быть значительной. Авторы работы [8] выступили с возражениями против такого подхода к предсказанию h -индекса по сложному соответствию.

ПЕРВЫЙ ПРИМЕР

Для последующих исследований записи ссылок определялись в базе данных ISI Web of Science за июль – октябрь 2012 г. Большое внимание уделялось установлению целостности множеств данных в отношении омонимов, исключению других авторов с одинаковой фамилией как исследуемых ученых. Замечу, что средство ISI для поиска авторов не оказалось полезным, по крайней мере, для моего собственного множества данных. Результаты выгружались из отчета цитирования в электронную таблицу, где непосредственно суммировались ссылки за определенный год и также отдельно подсчитывались статьи с высокой частотой ссылок до величины h , т.е. в зависимости от желаемого интервала после года публикации. В таком случае h -индекс можно легко определить за год y , учитывая (только) публикации

* Перевод Schreiber M. How relevant is the predictive power of the h -index? A case study of the time-dependent Hirsch index. – <http://arxiv.org/ftp/arxiv/papers/1301/1301.2060.pdf> - 2013.

до определенного года $s \leq y$. В дальнейшем я буду обозначать, таким образом, полученный индекс как $b_s(y)$. Он означает b -индекс за год y , если ученый перестал публиковаться в году s . Очевидно, что за $s = y$ ученый получает обычный b -индекс: $b = b_y(y)$. Значения $s > y$ не являются существенными.

В личных интересах я определил развитие b -индекса во времени для себя, как представлено на рис.1, показывающем более или менее стабильный рост с отклонением на 1 за 30 лет, который является довольно удовлетворительным значением. В ранние годы после моей первой публикации как студента-дипломника в 1976 г. рост был несколько ниже.

Тогда я задал себе вопрос, как мой b -индекс мог бы развиваться, если бы я перестал публиковаться, например, в 1983 г., вскоре после защиты моей докторской диссертации. Как видно на рис.1, индекс $b_{1983}(y)$ был бы таким же, как $b_y(y)$ за следующие четыре года y , и имел бы отклонение только на два индексных пункта после шести лет. Аналогично, если бы я стал ленивым после получения моей университетской преподавательской степени в 1987 г., тогда бы $b_{1987}(y) = b_y(y)$ до 1991 г. и $b_{1987}(1993) = b_{1993}(1993) - 2$. Также, если бы я прекратил публиковаться в 1990 г., году моего первого назначения на должность профессора, тогда после шести лет моя величина $b_{1990}(y)$ также была бы только на два пункта меньше, чем $b_{1996}(1996)$. Несколько более длительное время инертного поведения моего b -индекса можно наблюдать, начиная с 1994 г. Если бы я прекратил исследование в 1994 г., вскоре после моего повышения в звании, означающего достижение высшей степени в области преподавательской деятельности, шесть лет спустя мой индекс $b_{1994}(2000)$ был бы тем же самым, что и $b_{2000}(2000)$. Даже более сильным является эффект, если бы я оставил должность уже 1998 г. За первые четыре года рост моего b -индекса не пострадал бы вообще, и даже после 12 лет он уменьшился бы только на один индексный пункт, т.е. $b_s(y) = b_y(y)$ за $s \geq 2006$.

В результате за все упомянутые годы s потребовалось бы несколько лет, чтобы можно было обнаружить посредством b -индекса, что я полностью прекратил публиковаться. Таким образом, посредством b -индекса нельзя долгое время после решающего года s понять (различить), жил ли я согласно ожиданиям в этом году s , или я прекратил работать. После года s b -индекс увеличивался не благодаря моей продуктивности в течение этого периода времени, а благодаря прошлым достижениям, а именно тому, что мои предыдущие публикации все еще часто цитировались.

СЛЕДУЮЩИЕ ПРИМЕРЫ

Чтобы посмотреть, не может ли быть мой собственный случай исключением, я проанализировал запись цитирования Дж. Хирша (J.Hirsch), представленную на рис. 2. Он начал публиковаться также в 1976 г. Значения его индекса в два раза выше моих, а инертность индекса даже сильнее. Это могло быть связано с тем фактом, что дополнительные публикации труднее вводить в b -индекс, если этот индекс уже довольно высок. Прежде всего, на рис. 2 показывается, что рост индекса был бы тем же самым спустя три года, если бы Хирш перестал публиковаться в 1983 г., но уже через год наблюдается отклонение на пять индексных пунктов. Однако, допус-

кая, что Хирш не опубликовал больше ни одной статьи после 1989 г., даже через десять лет рост индекса был бы неизменным, т.е. $b_{1989}(1999) = b_{1999}(1999)$, а спустя 16 лет только $b_{1989}(2005) = b_{2005}(2005) - 2$. Для $s = 1992$ даже через 13 лет наблюдался бы тот же самый рост b -индекса: $b_{1992}(2005) = b_{2005}(2005)$, а спустя 16 лет $b_{1992}(2008) = b_{2008}(2008) - 2$. Если бы Хирш прекратил работать в 2001 г., его индекс не претерпел бы изменений в 2010 г. и даже в 2012 г. отклонился бы только на один индексный пункт. После 2005 г. изменения не отмечалось, за исключением отклонения на один индексный пункт в 2009 г.

В качестве следующих примеров я представляю на рис. 3 и 4 соответствующие результаты для М. Кардоны (M.Cardona) и Э. Уиттена (E. Witten). Я выбрал этих двух физиков, поскольку Хирш [5] был также одним из этих ученых. Между прочим, Уиттен также начал публиковаться в 1976 г. В его случае инертное поведение b -индекса сравнительно короткое, даже хотя значения его b -индекса в два раза выше, чем значения Хирша, и почти в четыре раза выше, чем мои значения. Как показано на рис. 3, для $s = 1986$ рост индекса оставался неизменным только три года, в случае $s = 1991$ - только пять лет и в случае $s = 1999$ - семь лет. Только в последнее время задержка в ответе стала бы больше: если бы он прекратил публиковаться в 2002 г., индекс в 2008 г. был бы незатронутым и отклонился бы только на один индексный пункт в 2012 г. После $s = 2005$ вообще не наблюдалось никаких изменений. В сравнении с рис. 1 и 2 интервалы времени, в течение которых b -индекс остается неизменным вне зависимости от публикационной продуктивности, становятся в целом короче. Это наблюдение показывает, что предположение, сделанное при обсуждении данных Хирша, а именно то, что индекс является более инертным при больших значениях индекса, не может быть общепринятым.

Глядя на данные для М. Кардоны на рис. 4, можно также увидеть поразительно высокие значения b -индекса, даже учитывая, что он начал публиковаться на 18 лет раньше, чем другие ученые, обсуждаемые в предыдущих параграфах. Если бы он перестал публиковаться в 1980 г., его b -индекс рос бы постоянно до 1988 г. В этот год можно видеть явное отклонение в кривой b -индекса. Интересно заметить, что кривая $s = 1980$ продолжает свой рост до 1988 г., с тем же уклоном до 1996 г. и с двумя исключениями даже до 2001 г. Это значит, что весьма инертное поведение $b_{1980}(y)$ как критерия достижения, основанного на продуктивности до 1980 г., наблюдается до 2001 г., т.е. в течение 20 лет. Довольно инертное поведение можно также наблюдать для $s = 1990$, поскольку после десяти лет $b_{1990}(2000) = b_{2000}(2000)$ и даже спустя 16 лет только $b_{1990}(2006) = b_{2006}(2006) - 2$. Аналогично, если Кардона прекратил бы публиковаться в 1994 г., b -индекс рос бы постоянно в течение 12 лет и уменьшился бы только на один индексный пункт в последующие два года. Также для $s = 1997$ спустя 12 лет мы обнаружим $b_{1997}(2009) = b_{2009}(2009)$. Прекращение публикаций в году его официальной отставки (2000 г.) также не повлияло бы на кривую его b -индекса в течение 11 лет до 2011 г. и дало бы только слабое снижение на один индексный пункт в 2012 г. Для $s \geq 2005$ вовсе не наблюдается изменения, что очень похоже на другие примеры.

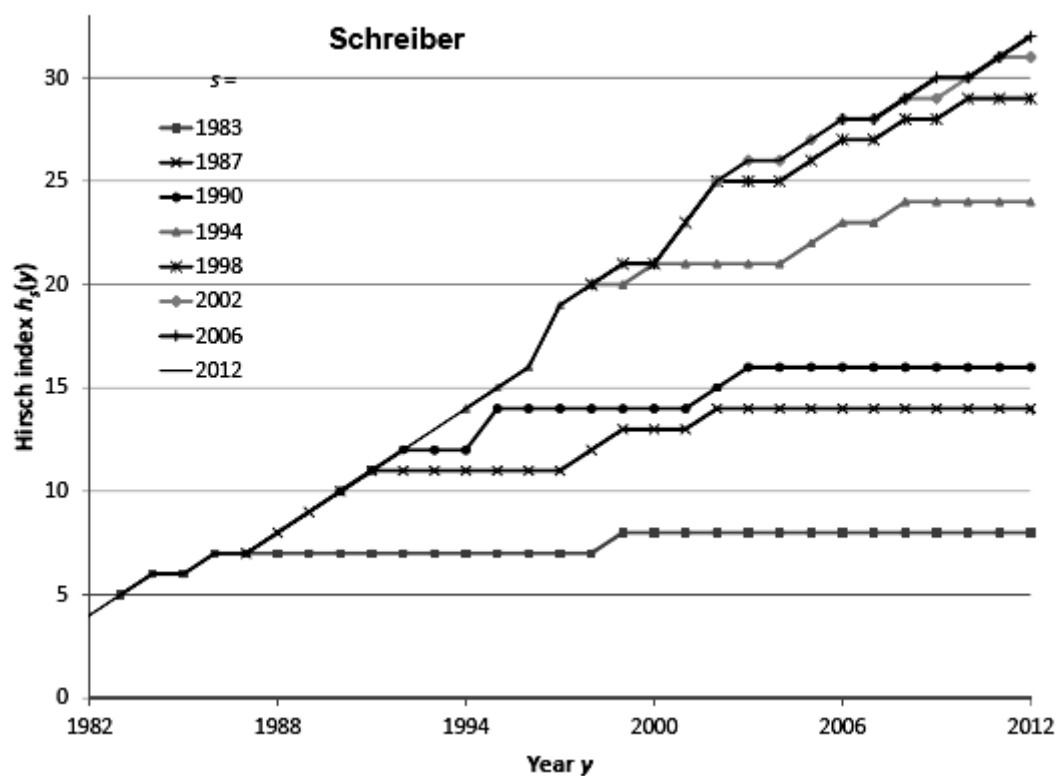


Рис. 1. Развитие во времени h -индекса для публикаций Шрайбера (верхняя линия). Дополнительно зависимость $h_s(y)$ показана для выбранных годов (см. легенду), начиная с года s , который предполагается как год, в который автор перестал публиковаться. (Эта подрисуночная подпись относится также к рис. 2, 3, 4, на которых представлены кривые h -индекса Хирша, Уиттена и Кардоны, соответственно)

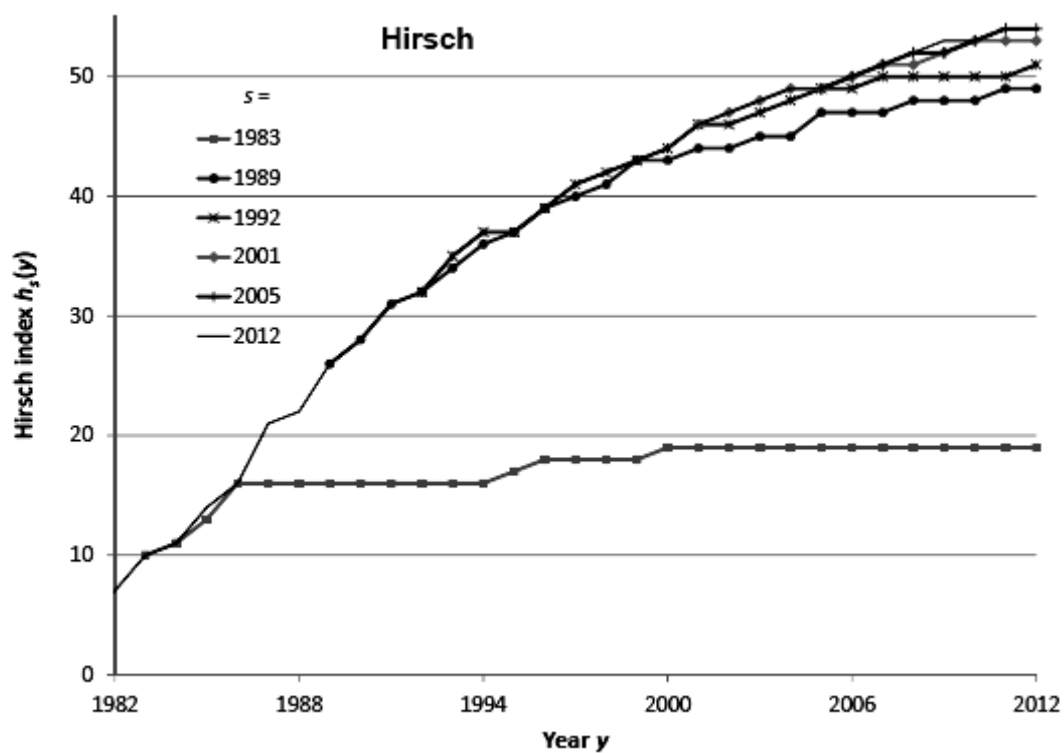


Рис. 2

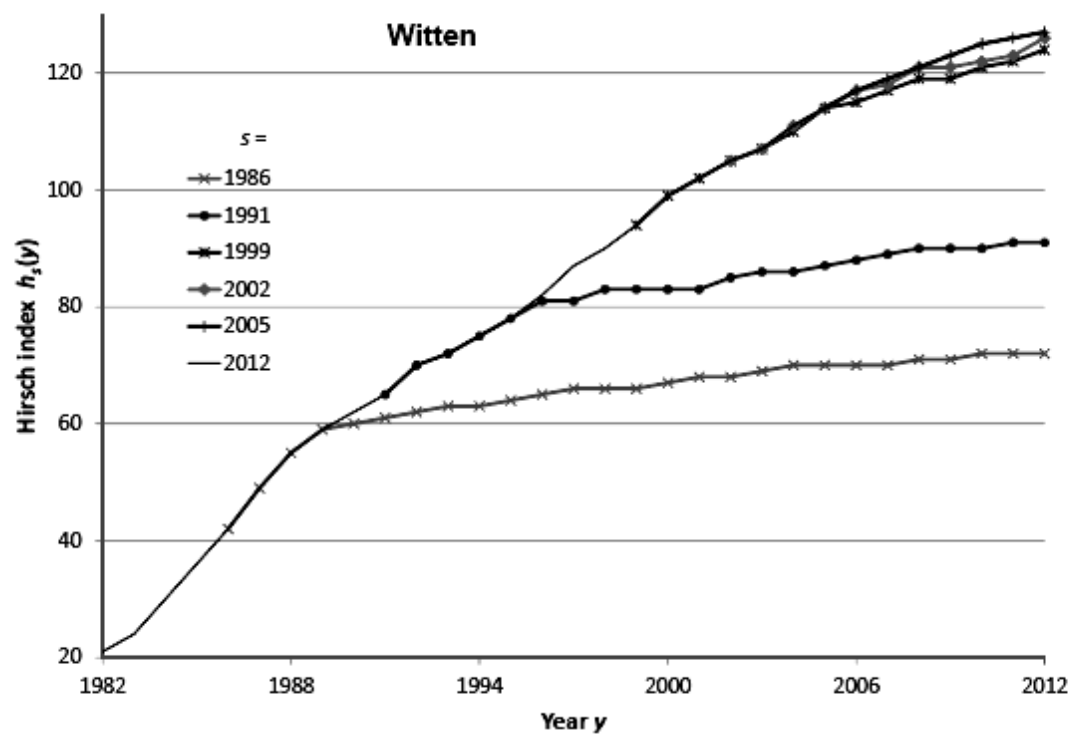


Рис. 3

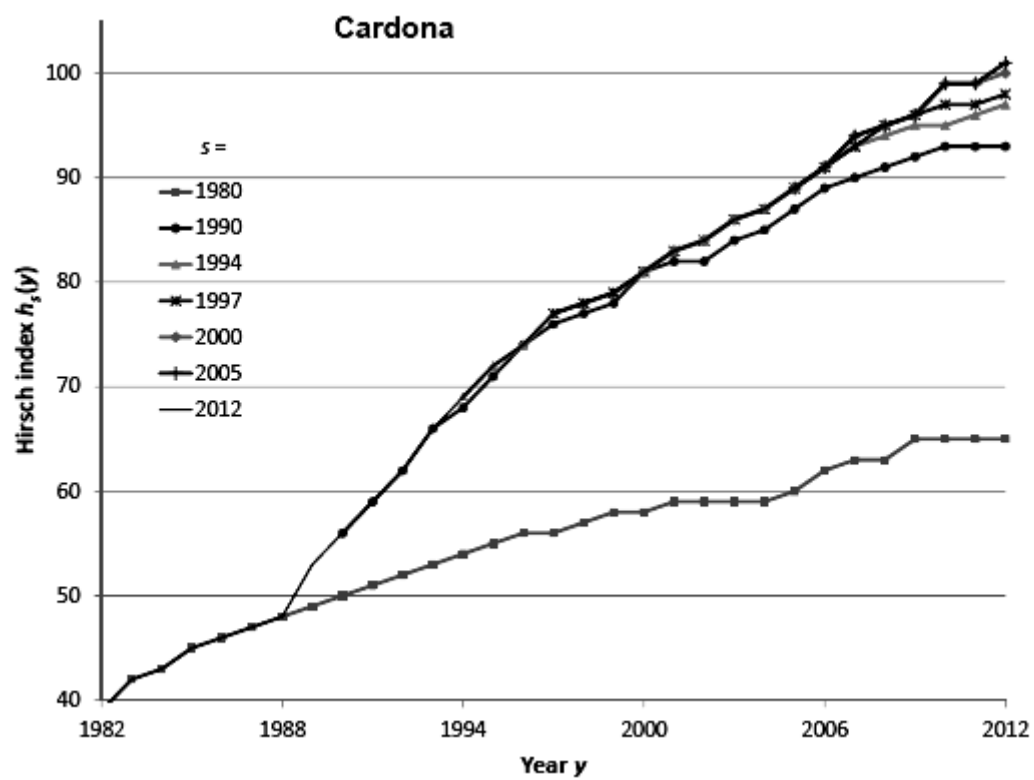


Рис. 4

ЗАКЛЮЧЕНИЕ

Хирш [5] продемонстрировал предсказательную силу h -индекса, показывая, что «ученый с высоким h -индексом после 12 лет, вполне вероятно, должен иметь высокий индекс и после 24 лет». В настоящем исследовании я выбрал различные временные интервалы. В представленных примерах h -индекс растет довольно плавно во времени (с одним исключением: Кардона, 1988). Это подтверждает заявление, что h -индекс сам по себе является хорошим предсказателем. Поэтому заманчиво применять h -индекс в процессах назначения на должность или в целях распределения ресурсов. Такая привлекательность основана на предположении, что предсказываемый рост h -индекса соответствует ожидаемой дальнейшей продуктивности кандидата. Однако это может не оправдаться.

Как показывают четыре примера, рост h -индекса не обязательно зависит от фактической продуктивности в течение нескольких лет в будущем, но, вероятнее, от результата предыдущих, часто довольно старых публикаций, т.е. от прошлых достижений. Таким образом, даже если какой-то ученый не оправдал ожиданий, он не разочарует тех оценщиков, которые измеряют продуктивность в терминах h -индекса, через 3-5 и даже больше лет. Я признаю, что для вышеприведенного обсуждения я выбрал самые выразительные примеры для года s , так что интервалы инерции являются определенно длинными и не представительными. Тем не менее, эти примеры показывают, что опасно делать выводы из предсказательного значения h -индекса в отношении будущей продуктивности.

Цинично было бы считать достоинством h -индекса то, что кто-то может перестать работать без руководителей, замечаящих это, если они смотрят только на h -индекс. Но я могу даже представить себе бюрократов, которые даже не заботятся о том, сколь долго они могут ждать выполнения их ожиданий, если оценка ожидаемого роста h -индекса не зависит от того, действительно была ли достигнута ожидаемая продуктивность или нет. Хотя это определенно не новая находка, я прихожу к выводу, что любой человек должен быть очень осторожен, когда использует показатели вроде h -индекса для предсказания будущей научной продуктивности.

Только четыре примера были приведены в настоящем исследовании, но, по-моему, этого достаточно, чтобы поднять дебаты, рассматривающие идею, что h -индекс может быть применим для предсказания будущих научных достижений. Я не исследовал еще другие примеры, но ожидаю подобных результатов во многих, вероятно, в большинстве случаев. Конечно, необходимы

более полные дальнейшие изучения для подтверждения результатов не только с использованием крупных множеств данных, но также и сравнения ученых из различных дисциплин и разного научного возраста. Для этой цели необходимо также определить количественный показатель, чтобы вычислять инерцию h -индекса для определенного временного периода. Например, можно использовать наибольшую продолжительность $d(y) = y - s$, в течение которой h -индекс остается неизменным или только слабо измененным, подобно $\Delta = b_s(y) - b_s(y) = 0$ или 1, или 2. Это должна быть функция y , и необходимо усреднение (возможно, включение начального временного периода во избежание проблем с очень небольшими числами индекса и проблем со случайным поведением индекса в начале карьеры) или некоторого другого рода сжатия информации. Другая возможность должна фиксировать время продолжительности $d \equiv y - s$, скажем, за пять или десять лет, и затем определять отклонение $\Delta(y)$ и анализировать эту функцию снова после некоторого рода усреднения или иного сжатия информации. В итоге, только разнообразные дальнейшие изучения смогут точно оценить предсказательную силу h -индекса.

ЛИТЕРАТУРА

1. Hirsch J.E. An index to quantify an individual's scientific research output // *Proceedings of the National Academy of Science*. - 2005. - Vol.105, No.46, - P.16569-16572.
2. Lehmann S., Jackson A.D., Lautrup B.E. A quantitative analysis of indicators of scientific performance // *Scientometrics*. - 2008. - Vol.76, No.2. - P.369-390.
3. Leeuwen T. Testing the validity of the Hirsch-index for research assessment purposes // *Research Evaluation*. - 2008. - Vol.17, No.2. - P.157-160.
4. Bornmann L. Redundancies in H -index variants and the proposal of the number of top-cited papers as an attractive indicator // *Measurement*. - 2012. - Vol.10. - P.149-153.
5. Hirsch J.E. Does the h -index have predictive power? // *Proceedings of the National Academy of Sciences*. - 2007. - Vol.104, No.49. - P.19193-19198.
6. Acuna D.E., Allesina S., Kording K.P. Predicting scientific success // *Nature*. - 2012. - Vol.489. - P.201-202.
7. Dyson F. A meeting with Enrico Fermi // *Nature*. - 2004. Vol.427. - P.297.
8. Rousseau R., Hu X. Predicting the h -index // *ISSI Newsletter*. - 2012. - Vol.8, No.3. - P.53-55.

Приглашаем российских и зарубежных авторов к сотрудничеству
в журнале «Международный форум по информации».
Оригинальные статьи и другие материалы (рецензии, письма)
можно присылать на русском или английском языке
по почтовому адресу, указанному в «Памятке для авторов»
или по электронной почте: mfi@viniti.ru.

Ответственный за выпуск *Л. В. Кобзева*

Компьютерная верстка *М. А. Филимонова*

ИД № 04689 от 28.04.2001 г.

Подписано в печать 18.03.2013 г.

Бумага “Хегох”. Формат 60x841/8. Гарн. литер. Печать цифровая

Усл. печ. л. 5,00 Уч.-изд. л. 4,42 Тираж 55 экз.

Адрес редакции: 125190, Россия, г. Москва, ул. Усиевича, д. 20

Тел. (499) 155-44-95