ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ УЧРЕЖДЕНИЕ НАУКИ
ИНСТИТУТ ВЫЧИСЛИТЕЛЬНОЙ МАТЕМАТИКИ
И МАТЕМАТИЧЕСКОЙ ГЕОФИЗИКИ
СИБИРСКОГО ОТДЕЛЕНИЯ РОССИЙСКОЙ АКАДЕМИИ НАУК

НЕКОММЕРЧЕСКОЕ ПАРТНЕРСТВО
“НАЦИОНАЛЬНЫЙ ЭЛЕКТРОННО-ИНФОРМАЦИОННЫЙ КОНСОРЦИУМ”

С. В. Бредихин, А. Ю. Кузнецов

МЕТОДЫ БИБЛИОМЕТРИИ И РЫНОК ЭЛЕКТРОННОЙ НАУЧНОЙ ПЕРИОДИКИ

Новосибирск, Москва, 2012

УДК 001.12+303.2
ББК 72+78

Бредихин С. В., Кузнецов А. Ю. Методы библиометрии и рынок электронной научной периодики. Новосибирск: ИВМиМГ СО РАН, НЭИКОН, 2012. 256 с.

Книга посвящена методам оценки научной деятельности путем анализа результатов цитирования научных публикаций. Рассмотрена история развития библиометрии как научной дисциплины, представлены гиперболические законы, сыгравшие значительную роль в ее становлении. Сформулированы понятия теории цитирования и подходы к формальной оценке результативности научной деятельности, основанные на арифметике цитирований. Определены метрики измерения производительности научных журналов и авторов научных публикаций. Представлен обзор исследования рынка научной периодики и предложен метод определения оптимальных комплектов подписки на электронные журналы.

Книга снабжена современным библиографическим материалом, некоторые главы могут быть использованы в виде учебного пособия и (или) справочника по курсу “библиометрия”. Предназначена библиотечным работникам, комплектующим фонды научной периодики; ученым и администраторам, занимающимся планированием научного труда, а также всем специалистам, интересующимся приложениями математики.

Bredikhin S. V., Kuznetsov A. Yu. Bibliometrics Methods and Market for Scientific Periodicals. Novosibirsk: ICM&MG SB RAS, NEIKON, 2012. 256 p.

The book is devoted the methods of the evaluation of scientific activity that rely on citation index analysis. It considers the history of Bibliometrics and presents the hyperbolic laws that played the significant role in the development of the scientific discipline. In the book one formulates the concepts of a citation theory as well as approaches to the scientific research output evaluation based on citation arithmetic. It also describes the metrics to measure the productivity of scientific journals and authors of publications. Along with the theoretical part, the book presents a survey of a market for scientific periodicals and suggests an optimal method for one to subscribe for electronic journals.

The references of the book are modern and up-to-date, so some of the chapters could be used as a tutorial or a manual for a study course into Bibliometrics. It is intended for the librarians stockpiling the collections of scientific libraries; scholars and administrators planning scientific activities and for every professional interested in Applied Mathematics.

ISBN 978-5-91907-007-8

Работа выполнена в рамках государственного контракта № 07.551.11.4002 между Министерством образования и науки Российской Федерации и Некоммерческим партнерством “Национальный электронно-информационный консорциум” по теме «Поддержка и расширение системы обеспечения новыми информационными технологиями участников Федеральной целевой программы “Исследования и разработки по приоритетным направлениям развития научно-технологического комплекса России на 2007–2013 гг.”».

ПРЕДИСЛОВИЕ

Эмпирические основы “Библиометрии” были заложены в первой половине XX в. В 1926 г. математик А. Лотка описал закономерность распределения частот публикаций по авторам в любой области исследований. В 1932 г. математик и лингвист Д. Ципф открыл закономерность распределения частоты слов естественного языка. Математик и библиотекарь лондонского “Музея науки” С. Бредфорд в 1934 г. установил эмпирическую закономерность распределения публикаций по изданиям.

В 1963 г. британский историк науки D. Price опубликовал книгу “Little Science, Big Science”, которая уже в 1966 г. была переведена на русский язык и вышла под названием “Малая наука, большая наука”. В своей работе Д. Прайс делит историю науки на два периода. Первый период − “малая наука” − отражает разрозненные усилия ученых по наблюдению за окружающим миром, выведению закономерностей и постулатов, описывающих функционирование природы и человека. Со второй половины XVII в., с появлением научных обществ и научных учреждений, начался новый период в истории науки, который положил начало “большой науке”. Наука становится управляемым и профессиональным видом деятельности. Д. Прайс предрек экспоненциальный рост научных исследований на ближайшие столетия. В связи с этим необходимость создания понятного измерительного механизма для оценки научной деятельности становится почти очевидной. Такой механизм должен как минимум генерировать и публиковать проверяемые результаты измерений, поскольку эти результаты должны обладать высоким уровнем доверия к ним со стороны научной общественности, для которой они предназначены.

Вклад ученых СССР и России в развитие библиометрии трудно переоценить. В 1969 г. вышла книга В. В. Налимова и З. М. Муль-

ченко “Наукометрия. Изучение развития науки”, в которой представлены методы и результаты исследований количественных показателей развития науки, рассматриваемой как информационный процесс. В работе проведен анализ числа публикаций, количества журналов, числа научных работников и ассигнований на науку. Теоретические и методологические проблемы рассмотрены в трудах Ю. В. Грановского, а позднее в работах С. Д. Хайтуна, А. И. Яблонского, И. В. Маршаковой-Шайкевич и др.

Термин “bibliometrics” (библиометрия) введен в 1969 г. в работе [Pritchard, 1969]. Библиометрические показатели (число научных публикаций по различным отраслям знания и их цитируемость) позволяют строить суждения о вкладе ученых различных стран как в общемировой прогресс науки, так и в развитие отдельных научных дисциплин. Термины “наукометрия” и “библиометрия” были определены одновременно и независимо друг от друга. Учитывая близость заявленных определений, в 2003 г. автор учебника по теоретической и прикладной библиометрии [Glanzel, 2003] предложил использовать эти термины как синонимы. В российском обиходе прижились оба термина.

Основоположником “прикладной” библиометрии следует считать Ю. Гарфилда. В 1960 г. он разработал систему индексирования для научной литературы, основанную на анализе цитирования; изобрел “*impact factor*” – меру количественной оценки научного журнала; организовал прозрачную систему вычисления IF на базах данных “Science Citation Index”, учитывающую характер и особенности использования профессиональной литературы научным сообществом. Это послужило толчком к созданию национальных индексов цитирования: Китай (1988 г.), Япония (1995 г.). В РФ в 2005 г. “Научная электронная библиотека” приступила к работам по реализации проекта “Российский индекс научного цитирования”.

Следует отметить, что современные исследования в области библиометрии имеют междисциплинарный характер с математическим уклоном. Процесс “математизации” знаний обусловлен двумя причинами. Первая причина заключается в том, что происходит накопление фактического материала, позволяющее обнаруживать закономерности; вторая − в том, что взаимосвязи между наблюдаемыми явлениями достаточно сложны, и для их анализа необходимо использовать математические методы [Яблонский, 1986]. Процесс анализа результатов научных исследований непрост ввиду отсутствия количественных характеристик, отражающих понятия “научная продуктивность’’, “ценность научного результата” и др. Например, массив научных публикаций является конечным результатом, который количественно характеризует работу ученого. Однако вопрос адекватности соотношения количества научных публикаций и качества научной деятельности остается открытым.

Данная книга предназначена для научных сотрудников и работников научных библиотек. Авторы старались придерживаться двух ограничений. Во-первых, подать материал с той степенью формальности, которая не позволяет трактовать его двояко. Во-вторых, для привлечения широкой читательской аудитории строгость и глубина изложения не должна превышать уровень научных статей английского варианта Википедии. Наконец, авторы придерживаются афоризма И. Канта: “В каждой естественной науке заключено столько истины, сколько в ней есть математики”.

Материал книги можно условно разделить на четыре части. Первая часть содержит общие сведения и определения, необходимые в дальнейшем, а также определения эмпирических законов. Во второй части представлен обзор зарубежных работ 1990–2007 гг., посвященных анализу состояния рынка научной периодики. Определены основные игроки рынка, их цели и модели по-

ведения. Третья часть содержит формальную постановку задачи оптимальной подписки на научные журналы и ее решение. Четвертая часть посвящена измерениям и количественным оценкам научной деятельности с использованием библиометрических методов. В качестве объектов измерения выступают автор научной публикации и научный журнал.

Авторы выражают благодарность А. Б. Хуторецкому, принимавшему участие в работах по данной тематике, Н. Г. Щербаковой за подготовку материала четвертой части книги, а также Н. А. Мазову за ценные замечания на этапе подготовки рукописи.

С. В. Бредихин, А. Ю. Кузнецов

Глава 1. Научная периодика

Наука представляет собой вид познавательной деятельности, направленной на получение, уточнение и производство объективных, системно-организованных и обоснованных знаний о природе, обществе и мышлении. Эта деятельность основана на сборе научных фактов, их постоянном обновлении и систематизации, критическом анализе и на этой базе синтезе новых научных знаний и прогнозов. Естественно-научные теории и гипотезы, которые подтверждаются фактами или опытами, формулируются в виде законов природы или общества. Наукой занимаются ученые. Термины “наука” и “ученый” введены [Википедия] У. Уэвеллом (1794−1866).

Научная литература является результатом научной деятельности и представляет собой совокупность письменных трудов, которые созданы в результате исследований, теоретических обобщений, сделанных в рамках научного метода. Научная литература предназначена для информирования ученых и специалистов о последних достижениях науки, а также для закрепления приоритета на научные открытия. Как правило, научная работа не считается завершенной, если она не была опубликована.

Первые научные произведения создавались в различных жанрах: трактатов, рассуждений, поучений, диалогов, путешествий, жизнеописаний − и даже в стихотворных формах. В настоящее время жанры научной литературы стандартизованы: монография, обзор, статья, доклад (в т. ч. тезисы докладов), автореферат, реферат и рецензия.

С начала XX в. наблюдается экспоненциальное увеличение объема публикуемой научной литературы. В настоящее время одними из главных носителей научной литературы являются периодические издания, главным образом рецензируемые научные журналы.

Журналом называют печатное периодическое издание. В соответствии с ГОСТ 7.60-2003 “Печатные издания” это “…периодическое журнальное издание, имеющее постоянную рубрикацию и содержащее статьи или рефераты по различным общественно-политическим, научным, производственным и др. вопросам, литературно-художественные произведения”. Научный журнал − это журнал, в котором статьи перед публикацией в обязательном порядке рецензируются независимыми специалистами. Рецензенты, как правило, не входят в состав редакции журнала и ведут исследования в областях, близких тематике статьи. Научные журналы являются одной из главных составляющих научной литературы. Рецензирование материалов выполняется для того, чтобы оградить читателей от возможных ошибок или фальсификаций, а также для того, чтобы гарантировать, что публикуемые работы выполнены на основе научного метода.

Во многих странах, в том числе в России, научные журналы проходят аттестацию в правительственных или общественных организациях, которая удостоверяет научность издания и соблюдение правил рецензирования. В РФ эти функции выполняет Высшая аттестационная комиссия (ВАК).

Справка [Wikipedia]. История научных журналов начинается в 1665 г., когда французский *“Journal des sçavans”* и английский *“Philosophical Transactions of the Royal Society”* впервые начали систематически публиковать результаты исследований. В XVII в. мнение, что наука может двигаться вперед только за счет прозрачного и открытого обмена идеями, подтвержденными опытными данными, было крайне непопулярным. Сам факт публикации научного исследования был противоречивым и широко осмеивался. О новом открытии было принято объявлять в виде анаграммы, которая охраняла приоритет открывателя и не поддавалась расшифровке. Ньютон и Лейбниц пользовались этим методом, однако выяснилось, что он работал недостаточно хорошо. Социолог Р. К. Мертон [Wikipedia] обнаружил, что в XVII в. одновременные открытия в 92 % случаев заканчивались спорами. Число споров снизилось до 72 % в XVIII в., до

59 % во второй половине XIX в. и до 33 % в первой половине XX в. Уменьшение числа оспаривающих приоритет в научных открытиях может быть отнесено на счет роста числа обращений к публикациям статей в современных научных журналах.

Научной статьей называют законченное и логически цельное произведение, посвященное конкретному вопросу. Научная статья раскрывает значимые результаты, требующие развернутого изложения и аргументации. Публикации статьи в научном журнале может предшествовать выпуск препринта. Препринт − научное издание, посвященное какой-либо теме, с которой автор хочет ознакомить заинтересованных лиц и специалистов (для обсуждения и (или) уточнения полученных результатов работы), выпускаемое в свет до публикации статьи в рецензируемом научном журнале или до выхода полноценной монографии. Как правило, препринты перед выходом в свет не рецензируются, вследствие чего они могут содержать ошибки и поэтому часто не учитываются в отчетах в качестве публикаций.

1.1. Рецензирование

Научные журналы являются одной из главных составляющих научной литературы. Статьи, публикуемые в научных журналах, в обязательном порядке рецензируются независимыми специалистами. Для этого издательство или редакция научного журнала перед публикацией новой научной работы направляет ее нескольким (не менее двух) рецензентам, считающимся специалистами в данной области. Рецензенты, как правило, не входят в состав редакции журнала и ведут исследования в областях, близких тематике статьи. На основе представленных рецензий редколлегия журнала принимает решение о публикации статьи. Процесс рецензирования призван исключить публикации тех материалов, которые содержат грубые методологические ошибки или прямые фальсификации.

Рецензия − анализ, разбор, некоторая оценка публикации, произведения или продукта. Рецензия является жанром журнальной публицистики. Рецензирование − процесс, благодаря которому ученые оценивают работы своих коллег. Это процедура рассмотрения научных статей и монографий учеными-специалистами в той же области. Цель рецензирования до публикации − удостовериться и в необходимых случаях добиться от автора следования стандартам, принятым в конкретной области или науке в целом. Публикации произведений, не прошедших рецензирование, профессионалами в различных областях часто воспринимаются с недоверием. Рецензирование используется издателями для отбора и оценки представленных рукописей.

В процессе рецензирования участвуют следующие стороны: автор, представитель издателя (как правило, редактор журнала) и рецензенты. Особенности взаимодействия сторон в процессе подготовки научной статьи к изданию изложены в лекции К. И. Сонина “Оценка научной статьи: взгляд рецензента” [Сонин, 2010].

1.2. Полезность

Предположим, что библиотека A имеет множество $I^A=\{i_1,\dots i_n\}$ годовых комплектов журналов, полученных в результате подписки. Библиотека A определяет свое значение полезности $v^A(i_k) \geq 0$ каждого журнала i_k из множества I^A. Далее предположим, что библиотека B имеет множество $I^B = \{i_1, \dots i_n\}$ годовых комплектов журналов и $I^A = I^B$, т. е. подписки библиотек были одинаковы. Библиотека B самостоятельно определяет значение полезности $v^B(i_k) \geq 0$ каждого журнала из множества I^B. Очевидно, что в общем случае $v^A(I^A) \neq v^B(I^B)$, поскольку оценки полезности журналов, выполненные библиотеками A и B, субъективны.

Сформулируем три следствия. Во-первых, обладая навыком вычисления полезности журналов, библиотекарь может оценивать

(в условных единицах) текущую стоимость библиотечного фонда периодических изданий. Во-вторых, умение вычислять полезность может неоценимо пригодиться на этапе составления плана подписки на следующий год. В-третьих, при проведении библиометрических исследований знание функции полезности дает возможность использовать методы микроэкономики. В свою очередь, применяя эти методы, мы надеемся приблизиться к ответам на фундаментальные вопросы о функционировании агентов рынка научной периодики, а именно: почему агенты-потребители выбирают такие наборы благ; как и почему агенты-производители формируют наборы благ таким образом; как и почему так, а не иначе формируются цены на товары и услуги; при каких условиях существует равновесие на этом рынке; как асимметрия информации влияет на агентов рынка. Знание ответов на эти вопросы придает агентам рынка уверенности в поведении и делает предсказуемыми результаты сделок.

Справка. Микроэкономика [Википедия] изучает функционирование экономических агентов в ходе их производственной, распределительной, потребительской и обменной деятельности. Исследования проводятся по следующим основным направлениям: а) проблема потребителя − почему агенты выбирают именно такие наборы благ (как правило, для конечного потребления); б) проблема производителя − как и почему агенты-производители выбирают именно такие наборы факторов производства и структуры выпуска; в) рыночное равновесие и структура рынка; г) общее равновесие − как и почему формируются цены на товары и услуги, как происходит обмен при различных предположениях; когда рынок экономически эффективен; д) асимметрия информации − как и почему несовпадение информационных множеств экономических агентов может привести к экономической неэффективности; е) внешние эффекты (экстерналии) − как и почему возможность своим выбором косвенно повлиять на решения других агентов может привести к экономической неэффективности; ж) общественные блага − как и почему существование некоторых типов экономических благ может привести к экономической неэффективности.

В экономической теории [Википедия] рынок − это совокупность основанных на взаимном согласии, эквивалентности и конкуренции экономических отношений между субъектами рынка по поводу движения товаров и денег. Функции рынка: а) информационная − рынок дает его участникам информацию о необходимом количестве товаров и услуг, их ассортименте и качестве; б) посредническая − рынок выступает посредником между производителем и потребителем; в) ценообразующая − цена на рынке складывается на основе взаимодействия спроса и предложения, с учетом конкуренции; г) регулирующая − рынок приводит в равновесие спрос и предложение; д) стимулирующая − рынок побуждает производителей создавать нужные обществу экономические блага с наименьшими затратами и получать достаточную прибыль. Рыночные отношения регулируются с помощью рыночного механизма, который состоит из четырех законов: закон стоимости, закон спроса и предложения, закон труда и закон конкуренции.

В микроэкономике проблему полезности изучает теория потребления и спроса, а именно какой товар или набор товаров выбирает потребитель при заданных ограничениях. Ниже приведены основные положения очерка [Сторчевой, 2010] на эту тему.

Изучение выбора отдельно взятого потребителя имеет своей целью выведение индивидуальной функции спроса. Конечной целью анализа данной проблемы является теория спроса, устанавливающая основные зависимости между ценовыми и неценовыми факторами и рыночным спросом. Была развита идея, согласно которой при покупке товара потребитель максимизирует полезность − психологическое удовлетворение от использования того или иного блага. При этом действует принцип убывания предельной полезности, согласно которому полезность потребляемого блага убывает по мере увеличения его потребления.

Идея полезности как основы мотивации потребителя привела к постановке проблемы измерения полезности. Л. Вальрас, сначала считавший существование меры полезности самоочевидным фактом, под влиянием работ выдающегося математика Ж. Анри Пуанкаре постепенно пришел к выводу о том, что полезность неизмерима, несмотря на то что является количественной величиной. У. С. Джевонс предложил косвенный метод измерения полезности в виде количества денег, которое потребитель готов заплатить за товар. Впоследствии данный метод был взят на вооружение и усовершенствован А. Маршаллом.

Постепенно экономисты заметили, что подход к анализу потребительского выбора с точки зрения принципа максимизации полезности нисколько не страдает от невозможности измерять полезность. Достаточно просто ранжировать различные альтернативы, что и делает потребитель. Как правило, он сравнивает различные товары или товарные наборы, и выбирает наиболее предпочтительный вариант. В этом случае для изучения потребительского выбора на основе принципа максимизации измерение не требуется.

Таким образом, сложился ординалистский (порядковый) подход, состоящий в том, чтобы определить потребительский выбор не путем прямого или косвенного измерения полезности, а с помощью сравнения различных товаров или товарных наборов. Основателями данного подхода являются британский экономист и математик Ф. Эджуорт и итало-швейцарский экономист В. Парето. В рамках этого подхода особую популярность получил графический аналитический аппарат кривых безразличия. Последний представляет собой график, на осях которого отложены количества двух товаров, а кривые безразличия объединяют множества наборов благ, дающие потребителю равную полезность. Другим аналитическим инструментом являются индексные функции полез-

ности, в которых значения функции отражают ранжирование потребителем своих предпочтений, имеющих порядковый смысл.

В реальной жизни потребителю приходится выбирать среди неопределенных альтернатив: покупая товар, потребитель, как правило, точно не знает, какую полезность он ему принесет. Это замечание достаточно точно характеризует подписку на журналы.

Глава 2. История развития библиометрии

Термин “библиометрия” (Bibliometrics) имеет несколько определений. Приведем примеры. Согласно первому определению [Bibliometrics, Def01] это “…набор методов, используемых для изучения или измерения текстов и информации, в который входят анализ цитирования и контент-анализ”. Второе [Bibliometrics, Def02] определяет библиометрию как “…научное направление, основанное на методах количественного анализа библиографических характеристик документов, дающих основу для их качественной оценки”. Третье определение [Bibliometrics, Def03] гласит, что это “…вспомогательная книговедческая научная дисциплина, разрабатывающая теорию и практику применения математических и статистических методов в приложении к письменным и печатным средствам коммуникации”. Одни исследователи считают, что этот термин введен А. Притчардом в 1969 г. [Pritchard, 1969], другие отдают пальму первенства П. Отле и называют 1934 г.

Автор работы [Glänzel, 2006] высказывает мнение, что термины “библиометрия” (далее − БМ) и “наукометрия” (далее – НМ) были введены независимо и одновременно в 1969 г. А. Притчардом и В. В. Налимовым. А. Притчард пояснял термин БМ как “…применение математических и статистических методов для исследования печатных коммуникаций…” [Pritchard, 1969], а В. В. Налимов определил термин НМ как “…использование таких количественных методов, которые имеют дело с анализом научных исследований, рассматриваемых как информационный процесс…” [Налимов, Мульченко, 1969]. В соответствии с этими интерпретациями БМ ограничивается измерением научных коммуникаций, в то время как НМ имеет дело с более общими информационными процессами. Так или иначе, размытые границы между этими двумя понятиями почти исчезли в течение трех по-

следних десятилетий, и в настоящее время оба термина используются практически как синонимы. Мы считаем, что разночтения в определениях и датах свидетельствуют о молодости рассматриваемых терминов.

Следует отметить, что в конце XX в. появился ряд аналогичных терминов. В 1988 г. в ВИНИТИ [Горькова, 1988] был введен в оборот термин “инфометрия” (Infometrics). Он обозначает подобласть информационной науки, занимающуюся статистическим анализом процесса научных коммуникаций. Инфометрия имеет дело с электронными средствами информации и включает такие темы, как статистический анализ (научных) текстовых и гипертекстовых систем, библиотечных циркуляций, информационных средств в электронных библиотеках, моделей для процессов производства информации. В работе [Brooks, 1990] представлен полемический обзор по теме “Библиометрия, наукометрия и инфометрия? О чем мы говорим”. В 1997 г. появился термин “вебометрия” (Webometrics).

> **Справка**. Термин “вебометрия” (далее – ВМ) обозначает раздел информатики, в рамках которого исследуются количественные аспекты конструирования и использования информационных ресурсов, структур и технологий применительно к World Wide Web (далее – 3W). Его авторами принято считать Т. Алминда и П. Ингверсена [Almind, Ingwersen, 1997]. Сегодня ВМ включает четыре направления исследований: 3W-индикаторы (индексы цитирования, наблюдаемость сайтов ...), социальные феномены в 3W (социальные сети, сообщества сайтов ...), сбор данных о 3W (роботы, поисковые машины, информационный поиск ...), анализ гиперссылок (в частности, связи между сайтами...). Обзор работ по ВМ за десятилетие представлен в [Ingwersen, 2006]. В качестве резюме приведем цитату из работы [Пенькова, Тютюнник, 2001]: “…терминологическая путаница привела к тому, что в одни и те же термины авторы вкладывают различный смысл, а идентичная суть описывается разными понятиями”.

В настоящей работе под термином БМ будем понимать комплекс методов и средств, предназначенных для измерений процессов производства, коммуникаций и использования научной ин-

формации. Целью применения измерительных методов является анализ процесса научных исследований. Исходя из этого, основное назначение БМ становится очевидным. БМ предназначена для оценки процесса исследований, но не для оценки результатов исследований. По определению БМ не нацелена на замену качественных методов количественными оценками. Результаты библиометрических измерений и их последующий анализ никоим образом не доминируют и не заменяют экспертных заключений. Оценки качества научных исследований и результаты библиометрического анализа процесса научных исследований дополняют друг друга.

Начало статистических исследований научных библиографий может быть отнесено к 20-м гг. XX в. В 1926 г. А. Лотка опубликовал результаты пионерского анализа частотного распределения научной производительности. Спустя восемь лет появилась работа С. Бредфорда, в которой было представлено частотное распределение научных статей по журналам. Эти работы оставались незамеченными до начала 60-х гг. XX в., когда D. Price опубликовал работу “Little science, big science” [Price, 1963], благодаря которой вопросы, относящиеся к количественным аспектам процесса научных исследований, стали интересовать ученых и руководителей научных исследований. Д. Прайс был одним из основных пропагандистов использования индекса научного цитирования, предложенного Институтом научной информации (ISI, Филадельфия, США) в качестве инструмента количественного анализа процесса научных исследований. Книга появилась в то время, когда необходимость оценки продуктивности и эффективности научного исследования стала настоятельной. Пришло время глобализации средств научного общения, роста объемов знаний и опубликованных результатов. Увеличился масштаб междисциплинарных исследований. Резко усложнились системы финансирования, основанные на экспертных оценках.

Рост академической активности в 60-х гг. XX в. тесно связан с прогрессивной информационной технологией, развитием компьютерных наук, сетевых технологий и особенно со всемирной доступностью больших библиографических баз данных, служащих основой для БМ-исследований. Здесь прежде всего следует отметить базы данных ISI. В 70-е гг. XX в., когда сбор данных часто представлял собой ручную работу, благодаря персональным качествам исследователей-энтузиастов и могучим традициям библиотечной науки БМ прошла путь от "хобби" до математических моделей и осознания необходимости реализации и внедрения технологичных решений. В начале 80-х гг. XX в. БМ становится самостоятельной научной дисциплиной со специфическим профилем исследований. С 1978 г. выпускается журнал "Scientometrics" [Scientometrics] − первое периодическое издание, специализирующееся на БМ и НМ. Основные публикации по БМ также содержаться в журналах: Journal of the American Society for Information science and Technology, Journal of Informetrics, Journal of Information Science и др. Ведущей международная конференция по этой тематике является The International Conference on Scientometrics and Informetrics.

В свет выходит несколько всеобъемлющих монографий по БМ, в которых применяется современный математический аппарат, например [Хайтун, 1983], [Van Raan, 1988]. Тот факт, что БМ-методы используются для самой БМ, также указывает на развитие данной дисциплины.

2.1. Основные даты и события

Хронология основных событий на этапе становления БМ представлена на домашней странице Р. Руссо [Rousseau, 1992]. Мы снабдили ее ссылками, добавили отечественные события и представили в виде табл. 2.1.

Таблица 2.1

Библиометрия: события и даты

Год	**Событие**
1913	F. Auerbach открыл гиперболическую зависимость между численностью населения и площадью немецких городов. В настоящее время эту и подобные закономерности называют законом Ципфа
1916	J. Estoup установил гиперболическую “природу” использования слов в языке. “Закон Ципфа”
1922	W. Hulme ввел термин “статистическая библиография”
1926	A. Lotka опубликовал статью “Частотное распределение научной продуктивности”. [Lotka, 1926]
1927	P. Gross и E. Gross разработали “Анализ цитирования”. [Gross, Gross, 1927]
1928	E. Condon опубликовал статью **“**Statistics of Vocabulary”. [Condon, 1928]
1929	G. Zipf защитил диссертацию “Relative frequency as a determinant of phonetic change”. Harvard Studies in Classical Philology
1934	P. Otlet использовал термин “Bibliometrie” в своем “Трактате о документации”
1934	S. Bradford вывел эмпирический закон распределения статей в научных журналах. Это распределение получило название “Закон рассеяния Бредфорда”
1935	G. Zipf опубликовал работу “The Psycho-Biology of Language. An introduction to dynamic philology” (Cambridge, Mass.), в которой впервые дал определение закона Ципфа
1948	H. Fussler ввел термин “ключевой журнал” для научной периодики в области химии и физики
1948	S. Bradford ввел термин “документация”

1948	C. Shannon создал математическую теорию связи. [Shannon, 1948]
1955	E. Garfield публикует статью “Citation Indexes for Science”. [Garfield, 1955]
1956	R. Fano высказывает идею оценки значимости научных журналов, которая базируется на анализе библиографических сочетаний (Bibliographic Coupling). [Fano, 1956]
1960	M. Raisig определяет термин RPR index (Index of Realized Research Potential). [Raisig, 1960]
1963	M. Kessler определяет термин “Bibliographic Coupling” и на его основе разрабатывает метод анализа сетей цитирования. [Kessler,1963a; b]
1963	В филадельфийском институте научной информации (ISI, USA) под руководством Ю. Гарфилда начаты работы по созданию системы баз данных для различных областей науки. С помощью этой системы вычисляется индекс цитирования научных работ (Science Citation Index)
1963	Вышла в свет монография D. Price “Little science, big science”, заложившая основы современной наукометрии. [Price, 1963]
1963	Издан первый номер “бумажной” версии журнала “Journal impact factor” под редакцией E. Garfield, I. Sher
1964	Опубликована работа по теории эпидемий. [Goffman, Newill, 1964]
1965	Опубликована работа D. Price, в которой рассматриваются “сети научных публикаций”. [Price, 1965]
1966	Опубликован русский перевод монографии Д. Прайса “Малая наука, большая наука”. [Прайс, 1966]
1966	K. Rosengren на этапе изучения социальных аспектов цитирования ввел термин *Comentioning*, который в дальнейшем преобразовался в “коцитирование” (*co-citation*)

1967	F. Leimkuhler вывел важное следствие из закона Бредфорда, позволяющее вычислять распределение статей по журналам, которое получило имя “кривая Леймкулера”. Современная трактовка приведена в статье. [Sarabia, 2008]
1968	Социолог R. Merton проанализировал “эффект Матфея”. Этот феномен заключается, во-первых, в том, что известные ученые получают высокое признание за исследования, которые не всегда могут считаться значимыми, тогда как неизвестные специалисты за аналогичные результаты получают гораздо меньшее признание, во-вторых, в том, что работа, получившая признание, превращается в “прецедентный текст”, воспринимающийся не с точки зрения его содержания, а с точки зрения его конституированного “значения”. [Мертон, 1968]
1969	Вышла в свет книга В. В. Налимова, З. М. Мульченко “Наукометрия: Изучение развития науки как информационного процесса”. [Налимов, Мульченко, 1969]
1969	A. Pritchard вводит термины “*статистическая библиография*” (Statistical Bibliography) и “*библиометрия*” (Bibliometrics)
1969	R. Fairthorne публикует статью о применении степенных законов распределения для библиометрических описаний и прогнозов. [Fairthorne, 1969]
1972	Опубликована работа J. R. Cole и S. Cole, в которой на основе анализа цитирований статей по физике утверждается, что лишь немногие ученые вносят вклад в научный прогресс. Однако это противоречит гипотезе испанского философа Хосе Ортега-и-Гассет о том, что большое число ученых средней квалификации вносит существенный вклад в развитие науки. Остается неизвестным, смогут ли появиться ученые “высокой” квалификации вне окружения “средних” ученых. [Cole, Cole, 1972; Ortega, 1932]

1973	В ISI разработана и внедрена методика вычисления индекса цитирования для социальных наук
1973	И. В. Маршакова и G. Small независимо друг от друга разработали метод измерения взаимных связей между научными статьями, основанный на анализе ссылок. Этот метод, получивший название “коцитирование” (*Co-citation*), используется при построении карт науки в Институте научной информации (ISI) США, а также в базах Web of Knowledge. [Маршакова, 1973; Small, 1972]
1975	В работе M. J. Moravcsik, P. Murugesan для проведения анализа научных статей предлагается классифицировать множество цитат по категориям. Например: глубокое или поверхностное, подтверждение или отрицание. [Moravcsik, Murugesan, 1975]
1975	В Лондоне состоялся первый международный форум по научной информации (IRFIS)
1976	D. Price публикует работу, в которой обсуждается феномен “предпочтительного присоединения” (*preferential attachment*). Например, в сетях новые узлы чаще присоединяются к тем узлам, которые уже имеют наибольшее число связей, т. е. обладают наибольшей известностью и популярностью. Подтверждает “эффект Матфея”. [Price, 1976]
1976	Основан журнал “Journal Citation Reports”
1977	B. Mandelbrot создал фрактальную геометрию
1978	Основан журнал “Scientometrics”
1978	В ISI разработана и внедрена методика вычисления индекса по БД “Arts & Humanities”
1983	Опубликована статья J. Irvine, B. R. Martin, в которой приведен пример “большой науки”. [Irvine, Martin, 1976]

1983	Вышла монография С. Д. Хайтуна “Наукометрия: состояние и перспективы”**. [**Хайтун, 1983]
1984	Ю. Гарфилд награжден медалью Д. Прайса
1986	Опубликована статья B. C. Brooks, в которой обсуждаются вопросы мотивации цитирования. [Brooks, 1986]
1986	Вышла монография А. И. Яблонского “Математические модели в исследовании науки”**. [**Яблонский, 1986]
1987	В. В. Налимов награжден медалью Д. Прайса
1987	В Бельгии прошла Первая международная конференция по библиометрии и теоретическим аспектам информационного поиска. [ISSI Society]
1988–1992	Создание национального индекса цитирования в Китае
1988	Вышла монография И. В. Маршаковой “Система цитирования научной литературы как средство слежения за развитием науки”**. [**Маршакова, 1988]
1989	Вышла монография С. Д. Хайтуна “Проблемы количественного анализа науки”**. [**Хайтун, 1989]
1990	Разработан и запущен в эксплуатацию сервис “Web of Science” – научный индекс цитирования, работающий в реальном времени. Провайдером услуг является Thomson Reuters
1995	“Национальный институт информатики” Японии приступил к работе по созданию собственного индекса цитирования “Citation Database for Japanese Papers”
2004	Вышел сборник статей “Handbook of Quantitative Science and Technology Research: The Use of Publication and Patent Statistics in Studies of S&T Systems” Henk F. Moed (Editor), Wolfgang Glänzel (Editor), Ulrich Schmoch (Editor)

2005	“Научная электронная библиотека” стала головным исполнителем проекта по созданию Российского индекса научного цитирования (РИНЦ)
2009	Вышла монография N. De Bellis “Bibliometrics and Citation Analysis: From the Science Citation Index to Cybermetrics”
2009	Вышла ставшая впоследствии популярной книга A. Andrés “Measuring Academic Research: How to Undertake a Bibliometric Study”
2010	В РФ создан электронный журнал “Informetrics.ru”, который является “…независимым научным электронным изданием, обобщающим опыт исследований в области наукометрии, инфометрии, библиометрии, вебометрии, киберметрии, нобелистики” (см. http://informetrics.ru/)
2010	Вышла монография Anne-Wil Harzing “The Publish or Perish Book: Your Guide to Effective and Responsible Citation Analysis”
2010	Вышла монография Chun Wei Choo, B. Detlor, D. Turnbull “Web Work: Information Seeking and Knowledge Work on the World Wide Web”
2010	Вышла монография H. F. Moed “Citation Analysis in Research Evaluation”

Рис. 2.1. Аверс медали Д. Прайса

Справка. Медаль Д. Прайса.

Медалью награждаются ученые, получившие выдающиеся результаты в области исследований “наука о науке”, “коммуникации в науке” и “стратегии развития науки”. С 1984 по 1989 гг. награждения медалью проходили ежегодно, а с 1993 г. раз в два года. Аверс медали представлен на рис. 2.1.

К настоящему времени медалью Д. Прайса награждены:

1984 Eugene Garfield (USA)

1985 Michael J. Moravcsik (USA)

1986 Tibor Braun (Hungary)
1987 Vasily V. Nalimov (Soviet Union), Henry Small (USA)
1988 Francis Narin (USA)
1989 Bertram C. Brookes (UK), Jan Vlachy (Czechoslovakia)
1993 András Schubert (Hungary)
1995 Anthony F. J. Van Raan (The Netherlands), Robert K. Merton (USA)
1997 John Irvine (UK) and Ben Martin (UK), Belver C. Griffith (USA)
1999 Wolfgang Glänzel (Germany/Hungary), Henk F. Moed (Netherlands)
2001 Ronald Rousseau (Belgium), Leo Egghe (Belgium)
2003 Loet Leydesdorff (The Netherlands)
2005 Peter Ingwersen (Denmark), Howard D. White (USA)
2007 Katherine W. McCain (USA)
2009 Pter Vinkler (Hungary), Michel Zitt (France)
2011 Olle Persson (Sweden)

Краткую информацию о достижениях обладателей медали можно получить на сайте [Price Medal, 2012].

За сравнительно короткий период БМ переросла “мануфактурную” форму “малой науки” и стала развиваться в направлении “большой науки” – многонациональных исследовательских центров, обеспеченных серьезной правительственной и промышленной поддержкой. На этом пути самой БМ было необходимо осуществить переход от “малой” формы на “большую” с огромными компьютеризированными базами данных, развитой и высокопроизводительной телекоммуникационной структурой, большим квалифицированным персоналом, способным развивать, поддерживать и предоставлять современный БМ-сервис. В 90-е гг. XX в. БМ превратилась в стандартный инструмент научной политики и управления исследованиями. В частности, многие важные индикаторы научной деятельности используют статистику публикаций и цитирований.

В настоящее время БМ-исследования ведутся по трем основным направлениям. Во-первых, для научных целей (поиск и анализ научной информации). Ученые, осуществляющие исследо-

вания в различных научных дисциплинах, образуют самую крупную, но также и наиболее разнонаправленную по интересам группу в БМ. Вследствие первичной научной ориентации интересы специалистов этой группы, как правило, соответствуют их специальности. Вместе с тем, наблюдается увеличение количества междисциплинарных исследований. БМ-методы и средства позволяют наблюдать этот рост и “измерять” его. Во-вторых, БМ для осуществления научной политики и управления (БМ для оценки процесса исследований). В настоящее время эти оценки имеют очень большое значение при сравнительном анализе, например национальных, региональных областей и отраслей науки. В-третьих, библиометрия для библиометрии (проведение фундаментальных исследований в БМ). Эти работы необходимы непосредственно БМ и, как правило, финансируются обычными грантами.

Итак, БМ имеет междисциплинарную сущность. Фундаментом БМ является библиотечное дело, она тесно связана с информационным поиском и социологией науки. Результаты исследований применяются в качестве научного сервиса, а также для оценки процесса научных исследований с целью принятия политических решений в области науки.

Глава 3. Эмпирические законы

Высказывание “Ничто и никогда не является абсолютно верным” (Nothing is always absolutely so) приписывают американскому писателю-фантасту Т. Старджону (1918–1985), утверждавшему, что “90 % фантастики − полная чушь”. К сожалению, данное утверждение не содержит указаний о том, каким образом выбрать те самые 10 %. Для решения задач выбора, как правило, применяются вероятностно-статистические методы.

Библиометрия изначально предполагает наличие результатов измерения каких-либо величин. Эти измерения неизбежно содержат погрешности, обусловленные разнообразными причинами. Следует различать погрешности систематические и случайные. Систематические ошибки обусловливаются причинами, действующими вполне определенным образом, и всегда могут быть устранены или достаточно точно учтены. Случайные ошибки вызываются большим числом причин, не поддающихся точному учету и действующих в каждом отдельном измерении различным образом. Ошибки невозможно совершенно исключить; учесть же их можно только в среднем, для чего необходимо знать законы, которым подчиняются случайные ошибки. Результат опыта является случайной величиной, значения которой до ее измерения нельзя точно предсказать. Обозначим измеряемую величину через A, а случайную ошибку при измерении через x. Так как ошибка x может принимать любые значения, то она является (непрерывной либо дискретной) случайной величиной. Область значений случайной величины и вероятности их принятия описывает закон распределения вероятностей. Аксиоматическое определение случайной величины дано А. Н. Колмогоровым в 1933 г.

Многие статистические закономерности, наблюдающиеся в библиометрии, удобно формулировать в терминах ранговых рас-

пределений случайных величин. Приведем четыре примера из работы [Арапов и др., 1975].

Пример 3.1. Пусть T – множество всех слов некоторого связного законченного текста, а $V = \{x\}$ – множество различных слов в этом тексте (словарь данного текста). Под $F(x)$ понимается число вхождений слова x в текст T (или, иначе, частота слова x).

Словарь V задает разбиение текста T на надмножества в каком-то смысле одинаковых слов. Каждому слову $x \in V$ соответствует подмножество $T(x)$ всех вхождений этого слова в текст T. Очевидно, что $F(x) = |T(x)|$. Обозначим длину (объем) текста через $L = |T|$, а объем его словаря – через $N = |V|$. Перенумеруем элементы словаря $V = \{x_1, x_2, ..., x_N\}$ таким образом, чтобы частота слова была невозрастающей функцией его номера:

$$F(x_1) > F(x_2) > F(x_N).$$

Ранговым распределением называется функция $\Phi(n) = F(x_N)$, которая ставит в соответствие номеру или рангу $n(x)$ слова $x \in V$ частоту $F(x)$ этого слова.

Пример 3.2. Пусть T – множество всех статей по определенной тематике, опубликованных за данный период в некотором множестве журналов V. Тогда для каждого журнала $x \in V$ $F(x)$ – это количество статей, опубликованных в данном журнале по данной тематике. Здесь ранговое распределение характеризует степень близости того или иного журнала по данной тематике, или, в соответствии с другой точкой зрения, насколько публикации в данной области рассеяны по журналам. В этом случае параметр V задает разбиение множества статей на подмножества статей, опубликованных в одном журнале.

Пример 3.3. Пусть V – коллектив ученых, а T – множество выполненных в этом коллективе работ. Тогда $F(x)$ – это число работ ученого x, и ранговое распределение характеризует распределение ученых по продуктивности или по престижу. Здесь множество статей T разбивается по авторам, но эти классы пересекаются, поскольку имеются общие статьи. В случае если существенно выполнение условия $T = \sum F(x)$, "автором" можно считать любую группу соавторов.

Пример 3.4. Пусть V – множество журналов, T – множество имеющихся в некотором массиве документов по определенной тематике библиографических ссылок на статьи в этих журналах. Тогда $F(x)$ – число ссылок на работы, помещенные в журнале x, которое характеризует "авторитетность" этого журнала в данной тематической области. Множество V задает разбиение множества всех ссылок T.

В приведенных примерах функции $\Phi(n)$ имеют настолько много общего, что часто считают возможным говорить о едином законе рангового распределения, присваивая этому закону имена Ципфа, Бредфорда, Лотки, Мандельброта и т. д. Присваиваемая закону фамилия отражает некоторые различия в его формулировке или в области применения.

Далее нам потребуется специальный тип математической зависимости, называемый степенным законом (Power law) или, в некоторых случаях, правилом 20/80. Предположим, что имеются две переменные, одна из которых представляет собой частоту появления некоторого события, а другая − размер этого события. Говорят, что зависимость между этими событиями представляет собой распределение по степенному закону, если частота появления событий снижается со скоростью, превышающей скорость увеличе-

ния размера событий. Например, вдвое мощное землетрясение встречается в четыре раза реже. График распределения по степенному закону представляет собой гиперболу, справа расположен длинный “хвост”, а слева небольшое число доминирующих событий. Примером распределения по степенному закону является распределение Парето.

3.1. Закон Парето

Закон Парето (Pareto's law) − это универсальное эмпирическое правило, введенное в научный оборот в 1897 г. итальянским экономистом и социологом Вильфредо Парето [Wikipedia]. В просторечии закон Парето формулируется следующим образом: “20 % усилий дают 80 % результата, а остальные 80 % усилий − лишь 20 % результата”. Закон используется как базовый принцип для оптимизации какой-либо деятельности, т. е. правильно выбрав минимум важных действий, можно быстро получить значительную часть планируемого результата, при этом дальнейшие улучшения неэффективны и могут быть неоправданными. Выполнение закона наблюдается в разных областях, например 20 % людей обладают 80 % капитала, 20 % покупателей или постоянных клиентов приносят 80 % прибыли. Закон подтверждает наблюдаемые общественные и научные явления.

3.1.1. Распределение Парето. Распределение Парето (Pareto distribution) представляет собой двухпараметрическое семейство абсолютно непрерывных распределений вероятности по степенному закону [Wikipedia]. Вне области экономики его называют распределением Бредфорда.

Если X является случайной величиной с распределением Парето, тогда вероятность того, что X больше x, задается равенством

$$\Pr\ (X > x) = \begin{cases} (x_m / x)^{\alpha} & \text{для всех } x \geq x_m, \\ 1 & \text{для всех } x < x_m \end{cases}$$

где x_m – минимально возможное значение X; α – положительный параметр. Семейство распределений характеризуется двумя параметрами – x_m и α. Отсюда следует, что функция распределения (CDF) имеет вид (рис. 3.1)

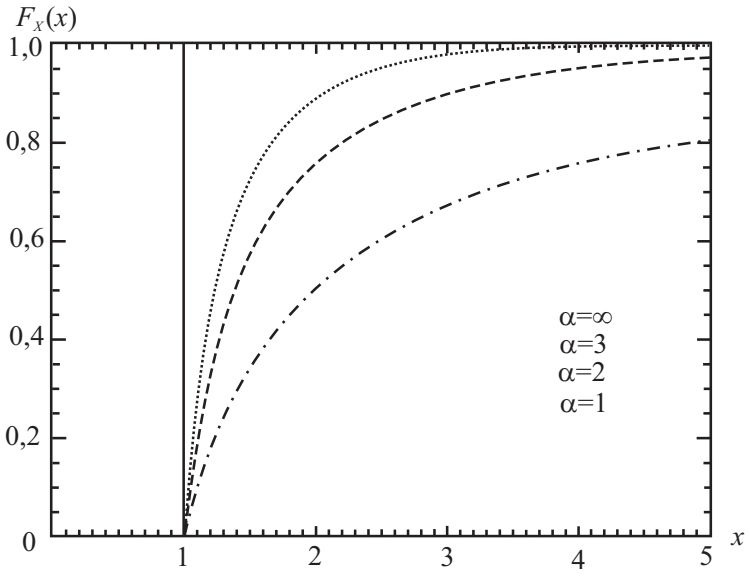


Рис. 3.1. Функция распределения вероятности (CDF)

Плотность вероятности (PDF) имеет вид (рис. 3.2)

$$f_X(x) = \begin{cases} \alpha\ x_m{}^{\alpha} / x^{\alpha+1} & \text{для всех } x \geq x_m, \\ 0 & \text{для всех } x < x_m. \end{cases}$$

Основные характеристики распределения Парето представлены в табл. 3.1

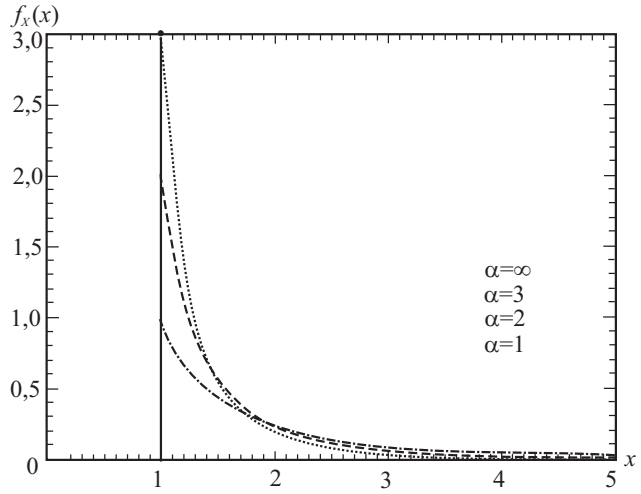


Рис. 3.2. Функция плотности вероятности (PDF)

Таблица 3.1

Основные характеристики распределения Парето
с параметрами $x_m > 0$, $\alpha > 0$, $x \in [x_m; +\infty)$

Характеристики	**Значения**
Функция плотности вероятности (PDF)	$\frac{\alpha x_m^{\alpha}}{x^{\alpha+1}}$ для $x \geq x_m$
Функция распределения вероятности (CDF)	$1-\left(\frac{x_m}{x}\right)^{\alpha}$ $x \geq x_m$
Математическое ожидание	$\frac{\alpha x_m}{\alpha-1}$ для $\alpha > 1$
Медиана	$x_m \sqrt[\alpha]{2}$
Мода	x_m
Дисперсия	$\frac{x_m^2 \alpha}{(\alpha-1)^2(\alpha-2)}$ для $\alpha > 2$

3.2. Закон Бенфорда

Данный закон назван по имени физика Ф. Бенфорда [Wikipedia], который сообщил об этом в 1938 г., хотя ранее, в 1881 г., это утверждение сделал С. Ньюкомб.

Закон Бенфорда (Benford's law) гласит, что в любой последовательности чисел, описывающей динамику какого-либо процесса или множество каких-либо объектов, числа, начинающиеся в записи с единицы, встречаются много чаще всех других. Ф. Бенфорд не только сформулировал этот закон, но и вывел формулы, которые позволяют рассчитать частоту появления каждой цифры в начале числа в том или ином числовом массиве. Подобная тематика рассмотрена в работе [Арнольд, 1998].

Согласно закону Бенфорда первая цифра оказывается цифрой 1 (единица) почти в одной трети случаев, и бóльшие цифры оказываются первой цифрой со все более низкой частотой вплоть до цифры 9, которая оказывается первой цифрой менее чем один раз из двадцати. Эти противоречащие интуиции результаты оказались пригодными для разнообразных данных включая счета за электроэнергию, уличные адреса, курсы акций, смертность населения, физические и математические константы, а также процессы, описываемые степенным законом. Заметим, что результат сохраняется вне зависимости от системы счисления, в которой представлены данные, хотя точные соотношения меняются.

В общем виде закон Бенфорда утверждает, что если основание системы счисления равно b $(b > 2)$, то для цифры d $(d \in \{1, \ldots, b-1\})$ вероятность быть первой значащей цифрой в любой достаточно большой выборке статистических данных составляет $P(d) = \log_b(d+1) - \log_b(d) = \log_b(1 + \frac{1}{d})$. Величина $P(d)$ пропорциональна разности между d и $d+1$ в логарифмиче-

ской шкале. При основании 10 первые цифры по закону Бенфорда имеют следующее распределение, представленное в табл. 3.2, где d – первая цифра, p – вероятность ее появления в числе на первом месте.

Таблица 3.2

Распределение Бенфорда

d	1	2	3	4	5	6	7	8	9
p	30,1 %	17,6 %	12,5 %	9,7 %	7,9 %	6,7 %	5,8 %	5,1 %	4,6 %

3.3. Закон Бредфорда

Закон Бредфорда (Bradford's law) оценивает экспоненциальное снижение получаемых результатов при продолжении поиска ссылок в научных журналах [Wikipedia; Писляков, 2007b; Редькина, 2005]. Одна из формулировок гласит, что при сортировке журналов в определенной области по числу статей в три группы, в каждой из которых имеется приблизительно 30 % статей, в каждой группе число журналов будет находиться в соотношении $1 : n : n^2$. В экономике это явление называют распределением Парето.

В качестве практического примера предположим, что у исследователя имеется пять основных научных журналов по его предмету и за месяц в этих журналах появляется 12 работ, представляющих интерес. Предположим также, что для поиска еще 12 представляющих интерес статей исследователю нужно обратиться к 10 дополнительным журналам. Тогда для этого исследователя множитель Бредфорда b_m равен 2 (т. е. 10/5). Для каждых новых 12 статей этому исследователю придется просмотреть в b_m раз больше журналов. После просмотра 5, 10, 20, 40 и т. д. журналов большинство ученых понимает, что продолжать поиск нецелесообразно.

Области научных исследований различаются числом основных журналов и множителем Бредфорда. Однако во многих областях общая картина остается достаточно постоянной. Как и для закона Ципфа, с которым закон связан, не существует надежного объяснения, почему он работает. Но знание его оказывается очень полезным для библиотечных работников. Согласно закону Бредфорда, в каждой области достаточно определить "основные публикации" и направить в фонды только их. Исследователям очень редко приходится выходить за пределы этого набора.

Однако значение данного эмпирического закона намного больше описанного. В 1960-х гг. Ю. Гарфилд разработал указатель научного цитирования (SCI), который позволял достаточно точно определять, какой ученый оказал влияние на развитие данной научной дисциплины и в каком журнале появились его публикации. Это привело к неожиданному открытию. Оказывается, лишь немногие журналы типа "Nature" и "Science" являются основными для всех естественных наук. Такое явление не наблюдается в гуманитарных и социальных науках, возможно, потому, что в них значительно труднее установить объективную истину, или потому, что в этих областях используют более обширную литературу с меньшим упором на журналы. Вследствие этого возникло давление на ученых с требованием публиковаться в лучших журналах и давление на университеты с требованием обеспечить доступ к этому основному комплекту журналов.

3.3.1. Параметры *p* и *k*. Для применения закона Бредфорда на практике необходимо уметь определять значения параметра p – общее количество групп, на которые разбита рассматриваемая коллекция журналов, и параметра k, показывающего, как растет количество журналов от одной группы к другой. Знание этих параметров позволит отвечать на вопросы типа: удовлетворяет ли анализируемое множество журналов закону Бредфорда и какие

журналы составляют “ядро” коллекции. Далее в отношении 1:n:n^2 заменим n на k, как это принято в современных публикациях.

На практике закон Бредфорда редко выполняется точно, поэтому используется процедура подбора параметра k, причем количество групп не обязательно выбирается равным трем. Согласно работе [Egghe, 1990] процедура состоит из трех шагов:

Шаг 1. Выберем произвольное количество групп p (обычно в пределах от 4 до 10).

Шаг 2. Вычислим параметр k по формуле [Egghe, 1986]

$$k = \left(e^{\gamma} \times y_m\right)^{1/p};$$

где γ – постоянная Эйлера (γ=0,5772, e^{γ}=1,781…); y_m – количество публикаций в наиболее продуктивном журнале (т. е. в журнале ранга 1); p – количество групп.

Шаг 3. Определим размер первой группы r_0 по формуле

$$r_0 = T \times ((k-1)/(k^p-1)),$$

где T – число всех журналов, публикующих статьи в данной предметной области. Тогда теоретические размеры групп r_1, r_2, r_3 при p=4 будут следующими:

$$r_1 = r_0 \times k; \quad r_2 = r_0 \times k^2; \quad r_3 = r_0 \times k^3.$$

Теперь проверим, соответствуют ли исследуемые данные закону Бредфорда с выбранными параметрами. Путем перебора выберем такое p, чтобы значение r_0 было целым. Если это не получается, то следует в качестве значения r_0 выбрать округленное значение, целую часть от r_0+0,5. При подсчете значений для остальных групп следует брать неокругленное значение r_0 и только после вычисления произведения округлить полученное значение.

Бредфорд дал только словесную формулировку своего закона. Математические описания сделаны в работах [Leimkuhler, 1967;

Brookes, 1969; Rousseau, Leimkuhler, 1987; Egghe, 1990; Rousseau, 1994]. Однако популярность получила модель, представленная в работе [Leimkuhler, 1967]. Если издания упорядочить по убыванию производительности, то совокупное множество статей *R(r)*, опубликованных в журналах рангов 1, 2, 3,…,*r*, может быть представлено в виде формулы:

$$R(r) = \alpha \times \log_e(1 + b \times r)$$

где a и b – константы. Эта функция получила название кривой Леймулера.

В работе [Egghe, 1985] представлены формулы для вычисления констант a и b:

$$a = y_0 / \log_e k.$$

Здесь y_0 – количество статей в каждой группе (в предположении, что каждая группа содержит одинаковое количество статей);

$$y_0 = A / p,$$

A – общее количество статей; p – число групп;

$$b = (k - 1) / r_0 .$$

Если придерживаться модели Леймкулера, то, зная значения k и r_0, можно подсчитать ожидаемое количество публикаций для любого совокупного количества журналов r.

В монографии [Andres, 2009] приведен пример использования описанной методики для нахождения значений k и r_0 с последующей проверкой исходных данных на соответствие закону Бредфорда для четырех групп. Подобный пример, в котором рассмотрены кривые Леймкулера, приводится в работе [Sudhier, 2010].

В заключение следует отметить тесную связь представленного метода определения параметров с экономическими понятиями “кривая Лоренца” и “коэффициент Джини”. В работе [Burrell,

2005] показано, что кривые Леймкулера являются вариантом кривой Лоренца (зеркальная симметрия относительно диагонали *OC* (рис. 3.3)). В работе [Burrell, 1991] изучено влияние коэффициента Джини на распределения, соответствующие закону Бредфорда.

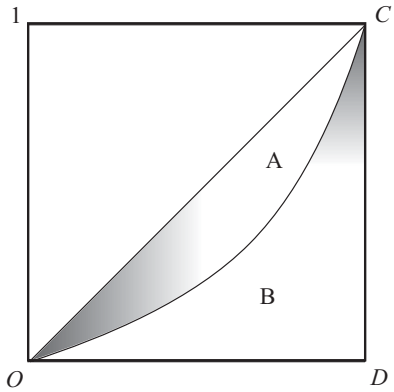


Рис. 3.3. Кривая Лоренца

Справка. Кривая Лоренца (Lorenz curve) – это функция отображения совокупной пропорции упорядоченных индивидов на совокупную пропорцию рассматриваемых признаков [WolframMW, Lorenz]. Пусть дано n упорядоченных индивидов, пусть x'_i признак индивида i и $x'_1 < x'_2 < \ldots < x'_n$. Тогда для этой выборки кривая Лоренца представляет собой многоугольник, соединяющий точки (h/n, L_h/L_n), где $h = 0,1,..,n,\ L_0 = 0,\ L_h = \sum_{i=1}^{h} x'_i$. Термин “кривая Лоренца” был введен в 1905 г. американским экономистом Максом Отто Лоренцем (1876–1959) как показатель неравенства в доходах населения.

Справка. Коэффициент Джини (Gini ratio, G) является оценкой неравномерности распределения изучаемого признака [WolframMW, Gini]. Коэффициент G связан с кривой Лоренца следующим образом. Рассмотрим треугольник $0CD$ (рис. 3.3), который состоит из двух областей A и B. По определению $G=A/(A+B)$. Поскольку $A+B=0{,}5$, то $G = 2A = 1 - 2B$. Если кривую Лоренца представить в виде функции $Y=L(x)$, то

$$G = 1 - 2\int_0^1 L(x)dx.$$

В дискретном случае коэффициент Джини можно рассчитать по формуле

$$G = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}\left| x_i - x_j \right|}{2n^2\mu},$$

где $\mu = \sum xP(x)$.

Данный метод оценки разработан итальянским статистиком и демографом Коррадо Джини (1884–1965) и впервые опубликован в 1912 г.

3.4. Закон Ципфа

Лингвист Дж. Ципф установил, что частота использования *n*-го наиболее часто используемого слова в естественных языках приблизительно обратно пропорциональна *n*. Эта эмпирическая закономерность получила название закона Ципфа (Zipf's law). Данный закон устанавливает следующее распределение частоты слов естественного языка: если все слова языка (или достаточно длинного текста) упорядочить по убыванию частоты их использования, то частота n-го слова в таком списке окажется приблизительно обратно пропорциональной его порядковому номеру n (так называемому рангу этого слова) [Wikipedia]. Например, второе по используемости слово встречается примерно в два раза реже, чем первое, третье − в три раза реже, чем первое, и т. д.

Таким образом, произведение частоты на ранг является константой. Соответствующее уравнение имеет простой вид: $k \times f = C$, где k – ранг слова, f − частота, C – константа. Ципф проиллюстрировал действие своего закона на анализе романа Дж. Джойса “Улисс” (James Joyce, “Ulysses”), показав, что десятое по рангу

слово появлялось 2653 раза, сотое по рангу появлялось 265 раз, двухсотое − 133 раза и т. д. Таким образом, константа C в этом случае приблизительно равна 26500. Закон Ципфа не является статистически безупречным, однако это не мешает достаточно широко применять его на практике, например для сравнения достаточно больших текстов (массивов статей).

3.4.1. Распределение Ципфа. Распределение Ципфа (Zipf distribution) является дискретным распределением вероятности степенного типа. Пусть N − число элементов, k − ранг элемента, s − значение экспоненты, характеризующей данное распределение. Тогда согласно закону Ципфа в популяции из N элементов частота появления элементов ранга k, $f(k; s, N)$ равна

$$f(k;s,N) = \frac{1/k^s}{\sum_{n=1}^{N}\left(1/n^s\right)}.$$

Данный закон можно также записать в виде

$$f(k;s,N) = \frac{1}{k^s H_{N,s}},$$

где $H_{N,s}$ – N-е гармоническое число.

Справка. В математике n-м гармоническим числом называется сумма обратных величин первых n последовательных чисел натурального ряда $H_n = \sum_{i=1}^{n} 1/i = 1+1/2+1/3+...+1/n$. Гармонические числа являются частичными суммами гармонического ряда.

Функции CDF и PMF представлены на рис. 3.4 и 3.5 соответственно. Заметим, что функции CMF и PMF определены только для целых значений k. Соединительные линии не указывают на непрерывность. Основные характеристики распределения Ципфа приведены в табл. 3.3.

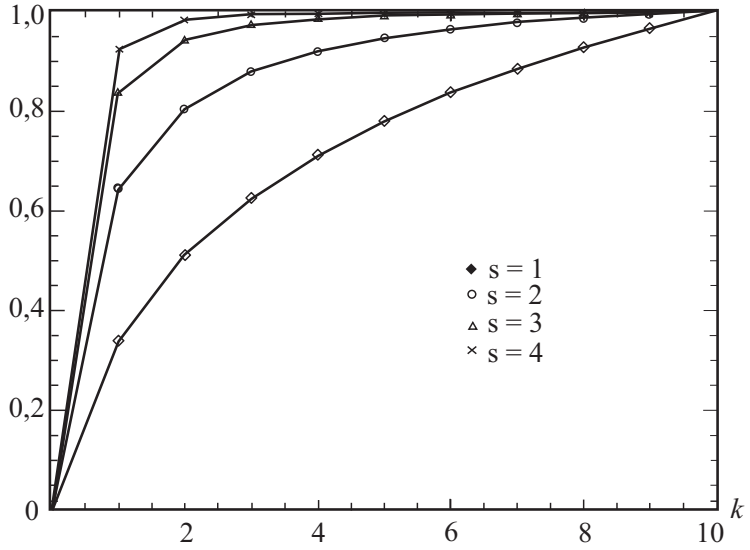


Рис. 3.4. Функция распределения вероятности (CDF), $N = 10$

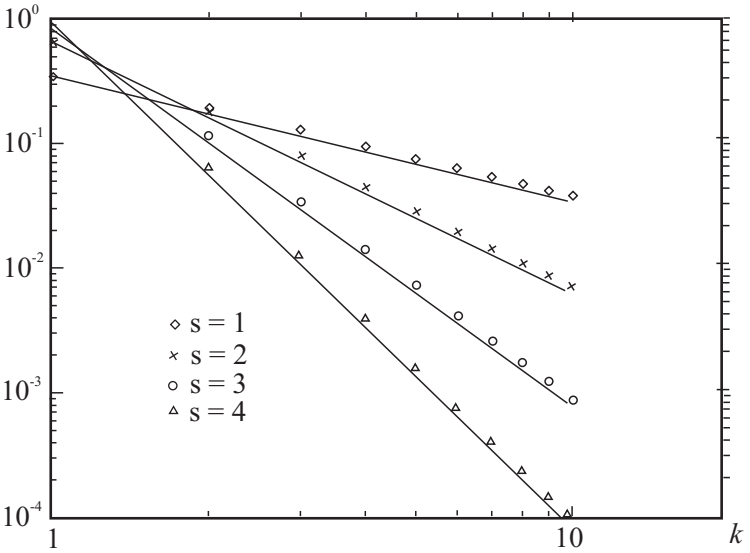


Рис. 3.5. Функция распределения масс (PMF), $N = 10$

Таблица 3.3

Основные характеристики распределения Ципфа
с параметрами $s > 0$, $N \in \{1,2,...\}$, $k \in \{1,2,\ldots,N\}$

Характеристики	**Значения**
Функция распределения вероятности (CDF)	$\frac{H_{k,s}}{H_{N,s}}$
Функция распределения масс (PMF)	$\frac{1/k^s}{H_{N,s}}$
Математическое ожидание	$\frac{H_{N,s-1}}{H_{N,s}}$
Мода	1

3.5. Закон Лотки

Закон Лотки (Lotka's law), названный по имени А. Дж. Лотки, представляет собой вариант закона Ципфа. Закон Лотки определяет частоту публикаций по авторам в некоторой фиксированной научной дисциплине. Утверждается, что число авторов, опубликовавших n статей, приблизительно равно $1/n^a$ от числа авторов с одной публикацией, где значение константы a практически всегда равно двум. Проще говоря, существует 1/4 авторов, опубликовавших две статьи за заданный период времени, от числа авторов с одной публикацией, 1/9 – опубликовавших три статьи, 1/16 – опубликовавших четыре статьи, и т. д.

В общем случае верна формула $Y = C / X^n$, где X – число публикаций, Y – относительная частота участия автора в X публикациях, n и C – константы, зависящие от научной дисциплины.

Пример 3.5. Рассмотрим вопрос о частоте публикаций авторов в фиксированной научной области. А. Лотка проанализировал

массив публикаций в реферативном журнале “Chemical Abstracts” за 10 лет (1907−1916) и определил, что доля авторов, которые публикуют одну статью, составляет приблизительно 60 %. Согласно закону Лотки 15 % будут иметь две публикации ($1/2^2$ от 60), 7 % авторов будут иметь три публикации ($1/3^2$ от 60) и т. д. Итак, только 0,6 % авторов в данной области будут публиковать более 10 статей.

Согласно работе [Яблонский, 1986] задачу частотного распределения продуктивности авторов сформулируем следующим образом. Пусть

$$n_i = n_1 / i^2 ,$$

где i = 1, 2, ..., $i_{\max}$; n_1 − число ученых, написавших одну статью; $i_{\max}$ − максимальная продуктивность ученого в этой области. Очевидно, что

$$\sum_i^{\max i} n_i = n_1 \sum_i^{\max i} 1 / i^2 = L ,$$

где L – общее число ученых – авторов в данной области. Полагая $i_{\max} \to \infty$ и учитывая, что

$$\sum_i^{\infty} 1 / i^2 = \pi^2 / 6 , \qquad (3.1)$$

получаем предельное (максимально возможное) значение доли ученых с минимальной продуктивностью в одну статью:

$$p_1 = n_1 / L = 6 / \pi^2 \approx 0,6 ,$$

что хорошо согласуется с эмпирическим законом Лотки: “…доля авторов, которые публикуют одну статью, составляет около 60 %”.

Заметим, что равенство (3.1) является решением задачи о сумме ряда обратных квадратов, так называемой “Basel problem”, которая сформулирована в 1644 г. и решена Л. Эйлером в 1734 г.

Закон Лотки неоднократно подвергался проверке на различных информационных массивах библиографий, реферативных журналах и т. п., относящихся к различным научным дисциплинам. Инвариантность и устойчивость этого закона позволяет говорить о нем как об одной из основных закономерностей распределения научной продуктивности. Данный закон позволяет получить достаточно точные оценки, когда он используется для больших объемов информации, накопленной в течение длительного времени, но не является точным в смысле математической статистики.

Пример 3.6. Пусть 100 авторов написали по одной статье за некоторый период времени. В табл. 3.4 представлено распределение числа авторов и числа статей, соответствующее закону Лотки, в предположении, что $C = 100$, $n = 2$.

Таблица 3.4

Пример распределения Лотки

Число статей	**Число авторов**
10	$100/10^2 = 1$
9	$100/9^2 \approx 1$ (1,23)
8	$100/8^2 \approx 2$ (1,56)
7	$100/7^2 \approx 2$ (2,04)
6	$100/6^2 \approx 3$ (2,77)
5	$100/5^2 = 4$
4	$100/4^2 \approx 6$ (6,25)
3	$100/3^2 \approx 11$ (11,111...)
2	$100/2^2 = 25$
1	100

Завершим данную главу цитатой из работы [Петров, Яблонский, 1980]: “…закон Бредфорда и все другие рассмотренные выше эмпирические примеры гиперболических распределений отражают, вообще говоря, два разных метода анализа соответствующего статистического материала: частотный (распределение Парето, закон Лотки) и ранговый (законы Ципфа, Бредфорда). Являясь, по сути, двумя разными “проекциями” одной и той же закономерности, характеризующей структуру сложных систем самого различного плана, частотный и ранговый подходы находятся в некотором смысле в отношении дополнительности и взаимосвязи между собой”.

Глава 4. Модели рынка электронной научной периодики

Рассматривается рынок, на котором в качестве товара выступают электронные научные издания. Агенты рынка – это издатели и владельцы баз данных, выступающие в роли производителей товара, с одной стороны, и библиотеки научных и учебных учреждений, выступающие в роли потребителей товара, с другой стороны. Потребитель представляет интересы читателей. Взаимодействие между производителями и потребителями обеспечивают посредники.

Минимальной единицей товара на рассматриваемом рынке является одна научная статья. Журналы, труды конференций и коллективные монографии представляют собой наборы статей на определенную тему, которые в обязательном порядке проходят процедуру рецензирования. Эта процедура осуществляется независимыми специалистами в данной области. Услуга получения одной статьи доступна только по индивидуальному запросу читателя. Для библиотек минимальной единицей товара является один из наборов журналов, сформированный издателем. Доступ к подписанным журналам осуществляется по сети передачи данных. Процедура доступа оговаривается в контракте, заключаемом между издателями и библиотеками. Для заключения контракта агенты рынка, как правило, пользуются услугами посредника.

Справка. С ноября 2002 г. в качестве посредника на рынке электронных журналов РФ выступает Некоммерческое партнерство “Национальный электронно-информационный консорциум” (НЭИКОН). Основная цель его деятельности − обеспечить доступ российских библиотек, университетов и институтов к научной периодической информации в электронном формате; кроме того, упорядочить и оптимизировать условия подписки на различные ресурсы по гуманитарным и естественным наукам, предлагаемые как зарубежными издательствами и агентствами, так и российскими поставщиками. В 2004 г. консорциум НЭИКОН и издательство “Elsevier” открыли

проект для университетов России, в рамках которого любая российская образовательная организация имеет возможность подписаться по своему выбору на любое количество из 21 предметной (тематической) коллекции издательства и реферативную базу данных по инженерным дисциплинам Compendex (см. [NEICON]).

Далее представлен краткий обзор моделей рынка электронных научных журналов. Особенность электронного журнала (и любого информационного ресурса) как товара состоит в том, что затраты на его создание (единовременные затраты), как правило, велики, а предельные затраты поставщика (стоимость обслуживания еще одного потребителя) близки к нулю. Если при такой структуре затрат конкуренция вынуждает производителей снижать цены (до уровня, близкого к предельным затратам), то за приемлемое для производителя время единовременные затраты не покрываются операционной прибылью. Поэтому производители уходят с рынка, пока не останется только один производитель каждого товара (информационного ресурса), ценовая политика которого уже не ограничена конкуренцией [Varian, 1995].

Присутствие на рынке электронных журналов нескольких поставщиков не делает его конкурентным, так как каждый журнал предлагается только одним поставщиком (иначе говоря, каждый поставщик является монополистом в отношении “своих” журналов). Монопольное положение поставщика дает ему рыночную власть (market power), позволяет формировать цены, учитывая в первую очередь не поведение конкурентов, а денежный эквивалент полезности журнала для потребителя, готовность потребителя платить (willingness to pay).

Производитель, ощущающий рыночную власть, всегда заинтересован в осуществлении ценовой дискриминации, т. е. хотел бы назначать разные цены на разные единицы товара. Существует множество форм ценовой дискриминации (см.: [Тироль, 2000, гл. 3]); основные ее виды − количественная (стоимость единицы

товара зависит от объема покупки) и межгрупповая (цена товара зависит от типа потребителя) дискриминация. На рынке электронных журналов возможности количественной дискриминации незначительны, поскольку каждый подписчик покупает, как правило, одну подписку. Межгрупповая дискриминация возможна, только если поставщик способен отличить "богатого" подписчика от "бедного". В [Bannerman, 1998] описана стратегия подписки на комплекты журналов для консорциумов потребителей, которая является, на самом деле, завуалированной формой межгрупповой дискриминации. Это прекрасно иллюстрирует следующий пример из работы [Varian, 1995].

Предположим, что есть два профессора и два журнала, "Сложение" и "Вычитание". Профессор А − специалист по сложению, поэтому он готов подписаться на журнал "Сложение" за $ 120, а на журнал "Вычитание" − только за $ 100. Профессор Б − специалист по вычитанию, поэтому его предпочтения противоположны: $ 100 за "Сложение" и $ 120 за "Вычитание". При единой цене на каждый журнал продавец подписки может получить не более $ 400 (по $ 100 за журнал). Но объединив журналы в комплект, он может продать два комплекта по $ 220. Оба профессора согласятся на такую цену комплекта, поскольку заплатят за каждый журнал ровно столько, сколько были готовы заплатить. Продавец увеличит свой доход на $ 40, вследствие того что фактически продаст журналы профессорам по разным ценам.

Наблюдаемая реальность подтверждает данные рассуждения. Именно межгрупповая дискриминация посредством формирования комплектов журналов является сейчас доминирующей схемой подписки на научные журналы. Причем комплекты, предлагаемые различными поставщиками, как правило, попарно не пересекаются, что, конечно, ослабляет конкуренцию между поставщиками на рынке электронных журналов.

Авторы статьи [Fishburn, et al., 1997] отмечают, что скрытую конкуренцию, которая все же существует на этом рынке, трудно изучать и моделировать. Они вводят в модель рынка конкуренцию, предполагая, что два поставщика совершенно заменяемых информационных ресурсов применяют различные способы распространения и ценообразования: одна фирма продает подписку на фиксированный срок с неограниченным использованием, а другая взимает плату за каждое обращение к ресурсу. В зависимости от структуры предпочтений потребителей на таком рынке либо может установиться равновесие, либо одна из фирм будет вытеснена с рынка, либо фирмы вступят в сговор и согласуют цены, либо между ними возникнет “ценовая война”, сводящая прибыль к нулю.

Более естественную модель конкуренции на рынке научных журналов изучают М. МакКейба и К. Снайдера в серии статей; последней по времени является работа [McCabe, Snyder, 2007]. Основная идея состоит в том, что журнал является посредником между авторами и читателями статей и поэтому может взимать плату с обеих сторон рынка. Конкуренция возникает между журналами − за авторов и читателей. Тип рынка (монополия или свободный вход), структура затрат на издание журналов, предпочтения авторов и читателей определяют результат. Могут установиться положительные цены и за публикацию, и за подписку (отрицательные цены не рассматриваются); может оказаться, что журналу выгодно бесплатно публиковать статьи или бесплатно их распространять; может, наконец, возникнуть расслоение авторов и читателей между журналами.

В последнее десятилетие рынок печатных научных журналов нестабилен. Издатели повышают цены, некоторые библиотеки отказываются от подписки, а издатели компенсируют снижение доходов дальнейшим повышением цен [Bannerman, 1998; Bot, et al., 1998; Fishburn, et al., 1997]. То, что такие журналы еще су-

ществуют, можно объяснить только неэластичностью спроса: цены растут быстрее, чем сокращается число подписчиков. Причинами неэластичности спроса являются специфическая структура рынка и особенности мотивации агентов рынка. Спрос исходит в большей степени от авторов, которые стремятся к публикации, чем от читателей. Читатель научного журнала, как правило, является сотрудником учебного или научно-исследовательского учреждения; журнал для него − "рабочий инструмент", за который должен платить работодатель. Поэтому подписка на журналы проходит через библиотеки, которые финансируются за счет накладных расходов, грантов, бюджетных и спонсорских средств, а не из исследовательских бюджетов авторов и читателей [Fishwick, et al., 1998; Ginsparg, 1996]. Авторы стремятся к публикациям в "престижных", а значит, самых дорогих журналах и не имеют стимулов выбирать дешевые журналы [Butler, 1999]. Ситуация не изменится до тех пор, пока ученые не будут вовлечены в финансирование того, что они публикуют и читают [Odlyzko, 1997].

Все сказанное в предыдущем абзаце справедливо и для рынка электронных журналов. Поэтому можно ожидать проявления неэластичности спроса и на этом рынке. Следствием будет рост подписных цен, не сопровождающийся соответствующим сокращением числа подписчиков. Эффект будет усилен тем, что подписчиками являются в основном консорциумы организаций, частично финансируемые из внешних источников.

Предсказываемые многими авторами ([Butler, 1999; Getz, 1992; Odlyzko, 1999; Varian, 1995; Varian, 1996]) изменения структуры рынка научных журналов (как печатных, так и электронных) могут произойти только вследствие рационального поведения поставщиков и потребителей. Сейчас поставщики считают выгодным распространение электронных версий журналов, объединен-

ных в непересекающиеся комплекты. Это стимулирует объединение подписчиков в консорциумы.

4.1. Взаимодействие монопольного поставщика с потребителями

4.1.1. Обозначения и терминология. Пусть I – множество (номеров) потребителей, J – множество (номеров) товаров, $J = \{1,\ldots,n\}$ (на рынке n товаров). Пусть x_{ij} – величина потребления товара $j \in J$ потребителем $i \in I$, а $\mathbf{x}^i = (x_{i1},\ldots,x_{in})$ – его потребительский набор.

Предположение 4.1.1. Потребители имеют бинарный спрос (каждый потребитель хочет купить не более единицы каждого товара). Тогда $x_{ij} \in \{0,1\}$, $\mathbf{x}^i \in X_i = \{0,1\}^n$. Пусть $V_i(\mathbf{x}^i)$ для $\mathbf{x}^i \in \mathrm{X}_i$ – денежная оценка полезности набора $\mathbf{x}^i$ для потребителя i (его готовность платить за этот набор),

$$V_i(\mathbf{x}^i) = \sum_j w_{ij}(\mathbf{x}^i)\, x_{ij} ,$$

где $w_{ij}(\mathbf{x})$ – денежная оценка полезности единицы товара j для потребителя i, если она входит в набор $\mathbf{x}$.

Для общества в целом результатом и оценкой работы рынка принято считать суммарный излишек: это суммарная денежная оценка полезности, полученной потребителями, минус затраты поставщика. Суммарный излишек складывается из излишка потребителей (денежная оценка полученной ими полезности минус суммарный платеж) и излишка поставщика (это прибыль, равная суммарному платежу потребителей за вычетом затрат поставщика).

Допустим, что все n товаров предлагает монопольный поставщик. У многопродуктового монополиста есть три ценовые стратегии [McAfee, et al., 1989]. Во-первых, он может продавать каждый товар отдельно (раздельная продажа). Во-вторых, он может пред-

лагать товары только в комплектах, включающих с учетом бинарности спроса единицу каждого товара, и устанавливать цену на комплект (полное пакетирование). В-третьих, он может предлагать товары на продажу как раздельно, так и в комплектах (полных или частичных), при этом цена комплекта может отличаться от суммы цен входящих в него товаров (смешанное пакетирование). Для монополиста результатом применения стратегии является прибыль, а оценкой стратегии − величина прибыли. Выбор между раздельной продажей и смешанным пакетированием зависит от ситуации.

При полном пакетировании каждый вектор $\mathbf{x}^i$ состоит либо из n нулей (если потребитель i не покупает набор), либо из n единиц. Поэтому денежная оценка единицы товара j зависит только от n: $w_{ij}(\mathbf{x}^i) = v_{ij}(n)$.

Пусть для каждого n $v_j(n)$ − случайная величина со значениями $v_{ij}(n)$, распределенная на множестве потребителей I. Обозначим через $a_j(n)$ и $\sigma_j^2(n)$ соответственно математическое ожидание и дисперсию случайной величины $v_j(n)$. В такой постановке следует говорить об ожидаемых, а не точных значениях спроса, предложения, полезности, прибыли, излишка и пр.

Предположение 4.1.2. Предельные затраты (на производство следующей единицы товара) равны нулю для каждого товара.

Товары с нулевыми (или близкими к нулю) предельными затратами называют “информационными”. Для таких товаров переменные затраты поставщика равны нулю, поэтому суммарный излишек равен суммарной денежной оценке полезности, полученной потребителями.

4.1.2. Независимые оценки товаров

Предположение 4.1.3. Случайные величины $v_j(n)$ (оценки товаров) имеют непрерывные функции плотности, неотрицательны,

независимы в совокупности при фиксированном n и равномерно ограничены по n.

Последнее условие означает, что существует конечный промежуток, в который почти наверняка попадает готовность платить за товар j (при любом числе товаров в комплекте) для случайно выбранного потребителя.

Возможность распространения информационных товаров через Сеть радикально снижает предельные производственные затраты, делая их близкими к нулю. Вследствие этого информационные товары приобретают свойства, нехарактерные для физических товаров. Это, в свою очередь, влияет на структуру рынков таких товаров. В частности, во многих случаях поставщику выгодно осуществлять полное пакетирование [Bakos, Brynjolfsson, 2000a, с. 17].

В некотором смысле пакетирование выгодно и обществу, так как оно увеличивает суммарный излишек, порождаемый рынком. Однако доля суммарного излишка, присваиваемая потребителями, при этом сокращается [Bakos, Brynjolfsson, 2000a, с. 18].

Предположение 4.1.4. Для всех $n > 1$ с вероятностью 1 выполняется неравенство

$$\sum_{j=1}^{n} v_j(n) \geq \sum_{j=1}^{n-1} v_j(n-1). \qquad (4.1)$$

Данное предположение эквивалентно условию *free disposal*, бесплатного избавления от товаров: если можно бесплатно отказаться от использования товара, входящего в приобретенный комплект, то оценка полезности каждого товара неотрицательна при любом составе комплекта.

Пусть $P^*(n)$ – цена комплекта из n товаров, которая максимизирует прибыль монополиста при полной комплектации, $q^*(n)$ – вероятность того, что случайно выбранный потребитель купит комплект при такой цене; $q^*(n)$ можно интерпретировать

как ожидаемую величину спроса на комплекты в долях от общего числа потребителей. Положим $p^*(n) = P^*(n)/n$ (средняя цена единицы товара в комплекте), $\eta_n = (1/n) \sum_{j=1}^{n} v_j(n)$ и $a(n) = (1/n) \sum_j a_j(n)$ (математическое ожидание случайной величины η_n).

Предположение 4.1.2 исключает из модели текущие производственные затраты. Единовременные затраты на производство первой единицы (оригинала) информационного товара учитываются, как правило, внемодельно, путем сравнения с величиной ожидаемого дохода. Существуют также затраты на распространение товара (distribution costs) и организацию сделки (transaction costs). Затраты первого типа зависят, вообще говоря, от числа продаж, затраты второго типа − от числа покупателей. Из предположения 4.1.1 следует, что число продаж равно числу покупателей, поэтому оба типа затрат можно объединить, их сумму назовем дополнительными затратами.

4.1.2.1. *Модель, не учитывающая дополнительные затраты.* Если выполнены предположения 4.1.1–4.1.4, то справедливы следующие результаты [Bakos, Brynjolfsson, 1999, Proposition 1]. Последовательности $a(n)$, $p^*(n)$ и $q^*(n)$ имеют пределы. Обозначим эти пределы через a, p^*, q^*. Тогда $p^* = a$, $q^* = 1$. При $n \to \infty$ ожидаемый излишек потребителя (ожидаемая денежная оценка полезности комплекта минус его цена) стремится к нулю, а ожидаемые значения прибыли монополиста и суммарного излишка стремятся к максимальным возможным значениям.

Применительно к рынку электронных журналов это означает, что монопольный поставщик имеет стимул к составлению комплектов, включающих все распространяемые им журналы. Более того, он заинтересован в увеличении числа монопольно пред-

лагаемых журналов и, следовательно, в расширении комплекта. При больших n ожидаемая прибыль поставщика комплектов будет близка к максимальной. Общество в целом получает практически максимальный результат (суммарный излишек), но его почти полностью присваивает монополист, так как ожидаемый излишек потребителя стремится к нулю. В такой ситуации разнородным потребителям выгодно объединяться в консорциумы, поскольку агрегированному потребителю нужен один комплект.

Пусть, например, на рынке продаются комплекты из двух журналов. Денежные оценки полезности журналов для двух потребителей определим следующим образом: $v_{11} = 100$, $v_{12} = 50$, $v_{21} = 50$, $v_{22} = 100$ (здесь v_{ij} − оценка полезности журнала j для потребителя i). Монопольный поставщик получит максимум прибыли, если продаст два комплекта по цене 150. Если же потребители объединятся, то оценка полезности каждого журнала для агрегированного потребителя будет равна 100 и будет продан один комплект за 200. Суммарный излишек на рассматриваемом рынке равен 300. В первом случае его полностью присваивает поставщик, во втором − треть излишка (300–200) получает агрегированный потребитель.

Заметим, что выше говорилось об операционной прибыли, не учитывающей единовременные затраты (на изготовление первого комплекта). Если операционная прибыль меньше единовременных затрат, поставщик уйдет с рынка. Включение в комплект журналов с низкими оценками полезностей для всех потребителей может увеличить единовременные затраты настолько, что продажа комплектов станет невыгодной. Отсюда следует, что чем меньше разброс предпочтений потребителей, тем меньше возможность расширения комплекта у поставщика. В частности, создание консорциумов однородных потребителей уменьшает разброс предпочтений потребителей.

Приближение к максимальным значениям прибыли и суммарного излишка, происходящее при $n \to \infty$, не является, вообще говоря, монотонным. В частности, при двух видах товара смешанное пакетирование дает прибыль, не меньшую (как правило, большую), чем полное пакетирование [McAfee, et al., 1989, c. 374; Adams, Yellen, 1976, c. 483]. Однако по сравнению с раздельной продажей полное пакетирование может быть выгодно [Bakos, Brynjolfsson, 1999, Proposition 3]. Так, если все случайные величины $v_j(n)$ равномерно распределены (что соответствует линейной функции ожидаемого спроса), то пакетирование любого числа информационных товаров увеличивает прибыль монополиста [Там же, Corollary 3a].

Предположение 4.1.5. $\text{Prob}[|\eta_n - a(n)| < \varepsilon] \leq \text{Prob}[|\eta_{n+1} - a(n+1)| < \varepsilon]$ для всех n и $\varepsilon > 0$.

Если выполняются предположения 4.1.1–4.1.5, то для продавца формирование полных комплектов выгодней, чем разбиение множества товаров на частичные комплекты [Там же, Corollary 3b].

Предположение 4.1.2 является существенным. Если предельные затраты на производство хотя бы одного товара j превышают некоторый (определяемый для данного товара) уровень c_j, то при торговле комплектами рынок порождает меньший суммарный излишек и поставщик получает меньшую прибыль, чем при раздельной продаже товаров [Там же, Proposition 2].

В [McAfee, et al., 1989] авторы рассматривают случай $n = 2$. В предположении, что потребительские оценки товаров имеют непрерывное совместное распределение (на множестве потребителей), указано условие, при котором смешанное пакетирование доминирует раздельные продажи. В частности, это условие выполняется, если оценки товаров независимы.

4.1.2.2. *Учет дополнительных затрат.* Изменим предположение 4.1.3 следующим образом.

Предположение 4.1.6. Для всех n случайные величины $v_j(n)$ (оценки товаров) независимы и одинаково распределены, неотрицательны, имеют непрерывные функции плотности, конечное математическое ожидание a и конечную дисперсию σ^2.

Предположение 4.1.7. Дополнительные затраты (сумма затрат на распространение товара и организацию сделки) равны d для любого комплекта.

При выполнении предположений 4.1.1, 4.1.3 и 4.1.6 комплект из n товаров приносит поставщику прибыль, зависящую от параметров n, a, σ, d. Допустим, что у поставщика имеется три стратегии: отказ от продаж (если невозможно получить положительную прибыль), полное пакетирование и раздельные продажи. В работе [Bakos, Brynjolfsson, 2000a, с. 11−12] указаны соотношения между приведенными выше параметрами и предельными затратами на производство (если они ненулевые), которые определяют, какую стратегию выберет поставщик. Этот результат позволяет, в принципе, определить оптимальный для поставщика размер комплекта в зависимости от значений указанных параметров. При ненулевых предельных затратах чем меньше затраты на распространение товара и организацию сделок, тем менее привлекательна для поставщика комплектация [Bakos, Brynjolfsson, 2000a, с. 17]. Однако для информационных товаров с нулевыми предельными затратами пакетирование всегда выгодней раздельных продаж, поэтому поставщик осуществляет полное пакетирование при $an \geq d$ и уходит с рынка в противном случае [Там же, Corollary 2].

4.1.3. Зависимые оценки товаров. Описанные выше результаты получены в условиях предположения 4.1.3, которое означает, что рынок не сегментирован, так как поставщик не может различать типы потребителей. Для рынка электронных журналов это предположение представляется слишком сильным, потому что

предпочтения специалистов коррелируют с тематикой журналов. Если рынок сегментирован, то при единой для всех сегментов цене полное пакетирование не всегда максимизирует суммарный излишек и (в зависимости от распределения потребителей по типам и распределения оценок товаров внутри каждого типа) может оказаться менее выгодным для поставщика, чем раздельная продажа журналов [Bakos, Brynjolfsson, 1999, Proposition 4]. Если поставщик способен дифференцировать цены по типам потребителей (по сегментам рынка), то полное пакетирование дает такой же результат, как и в случае несегментированного рынка [Там же, Proposition 5]. Этот теоретический результат подтверждается наблюдаемой в реальности межгрупповой ценовой дискриминацией на рынке электронных журналов: цена зависит от типа подписчика (коммерческая организация, научно-исследовательский или учебный институт, индивидуальный подписчик и т. д.).

В случае если поставщик знает типы потребителей, но не может их различать, межгрупповая дискриминация при полном пакетировании нереализуема. Однако поставщик может формировать пакеты, различающиеся составом и стоимостью и предназначенные для потребителей различных типов. Достаточная информация о структуре рынка и предпочтениях потребителей позволяет поставщику определить состав и стоимость пакетов таким образом, чтобы каждый потребитель предпочел пакет, предназначенный именно для потребителей соответствующего типа (эффект самовыявления). В этом случае частичное пакетирование может оказаться более прибыльным, чем раздельная продажа и полное пакетирование [Там же, п. 4.4].

Таким образом, ценовая стратегия поставщика существенно зависит от доступной ему информации о структуре рынка. Если он представляет себе покупателей как случайную выборку из однородной генеральной совокупности потребителей, то частичное

пакетирование не менее выгодно для него, чем полное, однако полное пакетирование может быть выгодней, чем раздельная продажа; при достаточно большом числе информационных товаров предпочтительным является полное пакетирование. Если поставщик имеет информацию о типах (предпочтений) потребителей, но не может различать эти типы, то он будет стремиться к использованию эффекта самовыявления при частичном пакетировании. Наконец, если поставщик способен различать типы потребителей, то при достаточно большом числе информационных товаров он будет осуществлять межгрупповую ценовую дискриминацию при полном пакетировании.

В работе [McAfee, et al., 1989] рассмотрен случай двух (не обязательно информационных) товаров, оценки которых являются случайными величинами v_1 и v_2, распределенными на множестве потребителей. Авторы предполагают, что эти случайные величины имеют непрерывное совместное распределение (случай дискретного распределения рассмотрен в [Adams, Yellen, 1976]); коррелированность случайных величин v_1 и v_2 не исключена. Получены достаточные условия (на функцию совместного распределения величин v_1 и v_2), при которых для поставщика смешанное пакетирование более выгодно, чем раздельные продажи [McAfee, et al., 1989, Proposition 1]. В частности, это верно при независимых оценках товаров. Указаны также достаточные условия, при которых для поставщика полное пакетирование более выгодно, чем раздельные продажи [Там же, Proposition 2], если монополист может контролировать продажи, препятствуя раздельной продаже обоих товаров одному потребителю.

4.1.4. Совместное использование информационных товаров. Большинство информационных товаров, в том числе электронные журналы, допускают копирование. Такой товар может быть использован не только покупателем, но и другими потреби-

телями, вследствие чего уменьшается количество покупок. Понимая это, продавец поддерживает объем продаж, повышая цену. В результате цена якобы индивидуально потребляемого товара может стать неприемлемой для индивидуального потребителя. Например, цена в $ 30 за электронную копию журнальной статьи слишком высока для большинства российских ученых. В таком случае потребители вынуждены объединяться не только для использования, но и для покупки товара. Не очевидно, что указанные разнонаправленные процессы приводят к снижению прибыли продавца. В работе [Bakos, et al., 1999], с. 8] предложена следующая модель анализа проблемы.

Предположим, что на рынке один товар (комплект товаров). Каждый пользователь имеет бинарный спрос (см. предположение 4.1.1). Перепродажа товара запрещена (или невыгодна), однако пользователи могут объединяться в группы для совместного использования товара.

Пусть для любого множества $B \subseteq I$ и каждого $i \in B$ $w(i,B)$ – оценка пользователем i полезности потребления единицы товара в составе группы B. Тогда $w(i) = w(i,\{i\})$ – оценка пользователем i полезности индивидуального потребления единицы товара. Пусть $\alpha = \{A_k \mid 1 \le k \le K\}$ – разбиение множества I всех потребителей на K групп: $A_i \cap A_j = \varnothing$ при $i \ne j$ и $\cup_k A_k = I$.

Предположение 4.1.6 изменим следующим образом.

Предположение 4.1.8. Оценки $w(i)$ являются значениями случайной величины w, равномерно распределенной на [0,1].

Отсюда следует линейность ожидаемого спроса: вероятность того, что случайно выбранный пользователь согласится заплатить за единицу товара цену $p \in [0,1]$, равна $1 - p$.

Введем дополнительные предположения.

Предположение 4.1.9. Если $B \subseteq I$ и $i \in B$, то $w(i,B) = w(i)$. Иными словами, для любого потребителя полезность потребления товара в составе группы равна полезности индивидуального потребления. Отсюда, в частности, следует, что копии являются полными заменителями оригинала и копирование внутри группы не требует затрат (транзакционные затраты на использование копии равны нулю).

Предположение 4.1.10**.** В разбиении α все группы имеют одинаковую численность, $|A_k| = n_0$ для всех k.

В предположениях 4.1.1, 4.1.2 и 4.1.8–4.1.10 совместное использование товара группами численностью $n_0 > 1$ увеличивает прибыль поставщика по сравнению с индивидуальным потреблением при $n_0 = 1$ [Bakos, et al., 1999, Proposition 1]. Чем больше n_0, тем больше прибыль [Там же, Corollary 1a]. При этом число потребителей, имеющих доступ к товару (читателей в случае электронных журналов), растет с увеличением n_0 [Там же, см. доказательство Proposition 1].

Откажемся от предположения 4.1.8 о равномерном распределении величины w. Пусть $a = E[w]$. Для поставщика суммарная оценка полезности товара членами группы A численностью m является суммой m случайных величин $w(i)$ (реализаций величины w). Введем случайную величину

$$w_n = (1/m)\sum_{i=1}^{m} w(i).$$

В соответствии с законом больших чисел $\mathrm{Prob}\,(|w_m - a| < \varepsilon) \to 1$ при $m \to \infty$ для любого $\varepsilon > 0$. Следующее предположение состоит в том, что последовательность $\mathrm{Prob}\,(|w_m - a| < \varepsilon)$ сходится к единице, не убывая.

Предположение 4.1.11. Prob ($|w_m - a| < \varepsilon$) ≤ Prob ($|w_{m+1} - a| < \varepsilon$) для всех m и $\varepsilon > 0$.

Если потребители не объединяются и монопольный поставщик установит цену p, то доля потребителей, которые купят товар, равна Prob ($w \geq p$) = $D(p)$. Если общее число потребителей принять за единицу, то $D(p)$ − ожидаемый спрос при цене p. Тогда при нулевых текущих затратах операционная прибыль продавца равна $\pi(p) = pD(p)$. Решением задачи максимизации прибыли при $p \geq 0$ является p^* − оптимальная монопольная цена поставщика.

Предположение 4.1.12. Оптимальная монопольная цена поставщика не превосходит среднюю оценку полезности товара, $p^* \leq a$.

При выполнении предположений 4.1.1, 4.1.2, 4.1.9–4.1.12 сохраняются результаты, сформулированные выше для равномерного распределения: совместное использование товара группами численностью $n_0 > 1$ увеличивает прибыль поставщика по сравнению с индивидуальным потреблением при $n_0 = 1$ [Bakos, et al., 1999, Proposition 2]. Чем больше n_0, тем больше прибыль [Там же, Corollary 2a] и больше число потребителей, имеющих доступ к товару.

Отсюда следует, что в сделанных предположениях потребители заинтересованы в объединении, которое выгодно и поставщику. Будучи монополистом, он может влиять на цену, а повышение цены подталкивает пользователей к созданию консорциумов, консолидирующих индивидуальные бюджеты. Если бы монополист знал все оценки $w(i)$, то он установил бы цену на уровне суммы оценок, что спровоцировало бы возникновение монопсонии на стороне спроса. Такому максимальному повышению цены препятствуют отсутствие у поставщика полной информации о готовности потребителей платить и техническая невозможность объединения всех потребителей.

Описанные выше результаты из работы [Bakos, et al., 1999] получены в предположении (которое явно сформулировано только в доказательствах утверждений), что оценка товара группой потребителей (готовность платить) равна сумме оценок членов группы; отсюда, в частности, следует, что члены группы необязательно поровну платят за использование товара. Если отказаться от этого предположения, то выводы существенно изменятся.

Предположим, что потребители объединяются в группы численностью k, каждая группа покупает единицу товара, и все члены группы поровну платят за совместное использование этого товара [Varian, 2000]. Тогда цена товара не может быть больше минимальной (среди участников групп) индивидуальной готовности платить. Пусть t − транзакционные затраты пользователя, связанные с доступом к товару внутри группы, c − предельные затраты производителя. Сравнение монополистических равновесий при групповом и индивидуальном потреблении товара дает следующие результаты.

Если $t < c\ ((k - 1) / k)$, то при групповом потреблении большее число потребителей получит доступ к товару, потребители будут меньше платить за использование товара, прибыль продавца увеличится. Следовательно, все агенты рынка заинтересованы в создании объединений потребителей. В частности, при нулевых транзакционных и ненулевых предельных затратах ($t = 0$, $c > 0$) условие выполняется, групповое приобретение и потребление товара предпочтительно и для поставщика, и для потребителей. Однако если $t > c\ ((k - 1) / k)$, то продавец получит больше прибыли при индивидуальных продажах и поэтому будет всячески препятствовать совместному использованию товара. Так будет, в частности, если при нулевых предельных затратах транзакционные затраты внутригруппового потребления положительны ($t > 0$, $c = 0$).

Случай “совершенно информационного товара” [Varian, 2000] $t = c = 0$ нейтрален: если k потребителей совместно используют единицу товара без дополнительных затрат, то поставщик просто увеличит цену в k раз и восстановит прибыль. Потребительский излишек не изменится по сравнению с раздельным потреблением.

Заметим, что в работах [Bakos, et al., 1999; Varian, 2000] отсутствуют существенные результаты для случая, когда группы потребителей могут различаться по численности (не выполняется предположение 4.1.10).

В работе [Liebowitz, 1985] показано, что внедрение фотокопирования в начале 1960-х гг. привело к значительному удорожанию журналов. Это согласуется со сформулированными выше результатами. Возможно, быстрый рост цен на научные журналы в последние десятилетия в значительной степени инициирован развитием средств копирования и воспроизведения.

4.2. Монополистическая конкуренция между поставщиками журналов

Научные журналы практически не взаимозаменяемы, и каждый издатель является монопольным поставщиком “своего” журнала. В работе [McCabe, 2000] рассматривается рынок, на котором сторону предложения представляют монопольные издатели научных журналов (каждый издает один журнал), а спрос формируют библиотеки. Каждая библиотека, располагая фиксированным бюджетом, выбирает журналы, ориентируясь на два показателя: стоимость годовой подписки и интенсивность использования (востребованность).

Предположим, что на рынке m библиотек и n журналов. Пусть u_{ij} – показатель востребованности журнала j в библиотеке i. Автор предполагает, что в каждой библиотеке востребованность журналов пропорциональна некоторому показателю качества u_j, зави-

сящему только от номера журнала. Иными словами, $u_{ij} = \lambda_i u_j$, отношение оценок $u_{ij}/u_{ik} = u_j/u_k$ одинаково для всех библиотек.

Допустим, что при ценах p_j каждая библиотека i в рамках бюджета B_i максимизирует суммарную востребованность приобретаемых журналов $\sum_j u_{ij} x_{ij} = \lambda_i \sum_j u_j x_{ij}$, где $x_{ij} \in \{0,1\}$ − число комплектов журнала j, приобретаемых библиотекой i. Это эквивалентно максимизации функции $\sum_j u_j x_{ij}$, коэффициенты которой одинаковы для всех библиотек (не зависят от i). Автор отмечает, что моделью такого выбора может служить целочисленная задача о ранце (о применении данной модели к подписке на комплекты журналов см. [Бредихин и др., 2008]):

$$\sum_j u_j x_{ij} \to \max \text{ при условиях } \sum_j p_j x_{ij} \le B_i, x_{ij} \in \{0,1\}. \quad (4.2)$$

Упорядочим журналы по неубыванию величин p_j/u_j: $p_1/u_1 \le \ldots \le p_n/u_n$. Пусть $k(i)$ − наибольшее k такое, что $p_1 + \ldots + p_k \le B_i$. Если $p_1 + \ldots + p_{k(i)} = B_i$, то решение x^* задачи (4.2) имеет вид

$$x^*_{ij} = 1 \text{ при } j \le k(i), \quad x^*_{ij} = 0 \text{ при } j > k(i).$$

Поставщик стремится максимизировать прибыль. Поскольку предельные затраты на производство журналов равны нулю (выполняется предположение 4.1.2), затраты поставщика сводятся к единовременным затратам, его операционная прибыль равна выручке, а прибыль − выручке за вычетом единовременных затрат. Ценовая дискриминация отсутствует, годовая подписка на журнал всем потребителям предлагается по единой цене.

Упорядочим библиотеки по неубыванию бюджетов: $B_1 \le \ldots \le B_m$. Стратегия поставщика j в работе [McCabe, 2000] определена как

пара чисел (b_j, p_j). Выбор такой стратегии означает, что поставщик предлагает журнал по цене p_j тем и только тем библиотекам, бюджеты которых не ниже b_j. При этом поставщик должен учитывать, что каждой библиотеке, которую он намерен обслуживать, предложены и другие журналы, и она будет выбирать вариант подписки (спрос), решая задачу (4.2).

Пусть $\beta_1 \le \ldots \le \beta_r$ − все различные значения бюджетов ($r \le m$). Предположим, что поставщикам известна функция распределения бюджетов библиотек. Это ступенчатая функция, имеющая в β_k скачок, равный доле библиотек с бюджетом β_k. Тогда, зная общее число библиотек, поставщик может определить число потребителей в каждой бюджетной группе. Число библиотек с бюджетами $B_i \ge b_j$ определяет предложение поставщика j.

В работе [McCabe, 2000] подробно рассмотрен случай $r = 2$ (по утверждению автора, общий случай может быть проанализирован аналогично). В этом случае m_1 библиотек имеют бюджет β_1 и $m_2 = m - m_1$ библиотек имеют бюджет $\beta_2 > \beta_1$. Поставщик может выбрать $b_j \le \beta_1$, $b_j \in (\beta_1, \beta_2]$ или $b_j > \beta_2$. Соответственно он будет обслуживать все библиотеки или только библиотеки с большим бюджетом либо уйдет с рынка. Поставщик выберет третий вариант, если в остальных вариантах его выручка не покрывает постоянные затраты (прибыль отрицательна).

Анализируя случай $r = 2$, автор работы [McCabe, 2000] предполагает, что ни один поставщик не выбрал $b_j > \beta_2$ и не все поставщики выбрали $b_j \le \beta_1$. Это значит, что журналы с номерами из некоторого множества $J \subset \{1, \ldots, n\}$ будут предложены всем библиотекам, а остальные − библиотекам с бюджетом β_2.

Для того чтобы библиотека из первой бюджетной группы смогла приобрести все предложенные ей журналы (т. е. чтобы

спрос был равен предложению), должно выполняться бюджетное ограничение

$$\sum_{j\in J} p_j \le \beta_1. \qquad (4.3)$$

Доказано, что равновесие (ситуация, в которой ни один поставщик $j\in J$ не выигрывает от увеличения цены p_j) достигается при равенстве в (4.3) и $p_j/u_j = \text{const}$ для $j\in J$. Отсюда

$$p_j = \frac{\beta_1 u_j}{\sum_{j\in J} u_j} \text{ для } j\in J, \qquad (4.4)$$

библиотеки первой группы делят бюджет между журналами из множества J пропорционально их качеству.

Библиотеке из второй группы предложены все журналы, и ее бюджетное ограничение, с учетом того что (4.3) выполняется как равенство при ценах, определенных формулой (4.4), имеет вид

$$\sum_{j=1}^{n} p_j = \beta_1 + \sum_{j\notin J} p_j \le \beta_2, \text{ или } \sum_{j\notin J} p_j \le \beta_2 - \beta_1. \qquad (4.5)$$

Равновесие достигается при равенстве в (4.5) и $p_j/u_j = \text{const}$ для $j\notin J$. Поэтому

$$p_j = \frac{(\beta_2 - \beta_1)u_j}{\sum_{j\notin J} u_j} \text{ для } j\notin J,$$

остаток бюджета библиотеки второй группы делится между журналами, не входящими в J, пропорционально их качеству.

4.3. Конкуренция между поставщиками за контент

Если рынок информационных товаров не монополизирован, то конкуренция за потребителя вынуждает поставщиков (применительно к электронным журналам − издателей и посредников) конкурировать за контент. Возникают два взаимосвязанных рынка:

“первичный” рынок контента, из которого формируются информационные товары, и “вторичный” рынок собственно товаров.

На рынке электронных журналов конкуренция за контент происходит в два этапа: издатели журналов конкурируют за статьи, а поставщики электронных копий журналов − за права распространения журналов. Можно предположить, что конкуренция обостряется вследствие того, что электронные научные журналы взаимно дополнительны. Действительно, журнальные статьи не взаимозаменяемы, потому что в научном сообществе не принято публиковать результат более одного раза. Отсюда следует, что даже журналы, относящиеся к одной области знаний, не только не взаимозаменяемы, но строго дополнительны. При этом каждый поставщик является монополистом в отношении распространяемых им журналов.

Модель конкуренции на первичном рынке описана в работе [Bakos, Brynjolfsson, 2000b] следующим образом. Предположим, что на рынке два поставщика. Каждый из n информационных товаров предлагается ровно одним поставщиком. В исходном состоянии поставщик $i \in \{1, 2\}$ предлагает n_i товаров, $n_1 + n_2 = n$. Будем считать, что выполнено предположение 4.1.6, из которого, в частности, следует, что товары не являются заменителями в потреблении (в противном случае потребительские оценки товаров были бы зависимы). Поэтому каждый поставщик является монополистом в отношении “своих” товаров на вторичном рынке, но конкурирует за каждый новый товар на первичном рынке.

Конкуренцию за новый товар (с номером $n + 1$) авторы описывают как двухэтапную игру. На первом этапе поставщики объявляют заявки на товар $n + 1$ в форме (y_i, z_i), где y_i, z_i − суммы, которые поставщик i готов заплатить соответственно за эксклюзивное и совместное (со вторым поставщиком) использование товара.

На втором этапе владелец товара выбирает $u = \max\{y_1, y_2, z_1 + z_2\}$ и разрешает использовать товар эксклюзивно поставщику i, если $u = y_i$, или совместно обоим поставщикам, если $u = z_1 + z_2$.

Заметим, что в работе [Bakos, Brynjolfsson, 2000a] смысл величин y_i и z_i определен недостаточно точно, однако из контекста следует, что это предлагаемые цены покупки в расчете на одного конечного пользователя. Поэтому ожидаемый доход владельца товара $n + 1$ равен Du, где D – ожидаемый спрос на вторичном рынке. Не очевидно (и не доказано в [Bakos, Brynjolfsson, 2000b]), что описанная выше стратегия владельца товара $n + 1$ обеспечивает ему максимальный доход: она вынуждает поставщиков повышать цену на вторичном рынке, вследствие чего уменьшается спрос. Иными словами, отсутствует уверенность в том, что функция выигрыша владельца товара выражает его доход. В явном виде функции выигрышей участников игры не описаны.

Авторы утверждают (Proposition 2), что при достаточно больших значениях n_i и выполнении предположений 4.1.1, 4.1.2, 4.1.4, 4.1.6 эксклюзивное право на использование товара $n + 1$ получит поставщик i с наибольшим значением n_i. Доказательство этого утверждения не кажется убедительным. Если оно все же верно, то позволяет предположить, что в каждом секторе рынка информационных товаров при стабильной структуре на стороне предложения будет либо большое число поставщиков, занимающих приблизительно равные доли рынка, либо одна фирма-лидер в сочетании, возможно, с некоторым количеством мелких поставщиков.

4.4. Модель двустороннего рынка

В работе [McCabe, Snyder, 2007] предложена модель, отражающая “двустороннюю” природу рынка научных журналов. Подписчики, относящиеся к одной стороне рынка, выигрывают от познаний авторов, находящихся на другой стороне рынка. И нао-

борот, авторы выигрывают, если количество их читателей велико. Журналы служат посредниками между этими двумя сторонами.

Модель учитывает следующие особенности рынка научных журналов. Во-первых, читатель может подписываться на несколько журналов, но каждая статья может быть опубликована только в одном из них (журнал заключает с автором эксклюзивный контракт). Во-вторых, журнал объединяет несколько статей, что приводит к экономии постоянных затрат, связанных с обслуживанием читателей. В-третьих, статьи, включенные в один журнал, могут различаться по качеству. Запишем эту модель.

На рынке имеется n^A потенциальных авторов (authors) и n^R потенциальных подписчиков (readers). Журнал получает статьи от авторов, группирует их и распространяет среди подписчиков. Публикация одной статьи требует от журнала затрат c^A (на реферирование, редактирование, типографский набор, корректуру, взаимодействие с автором). Расходы на доставку номера журнала каждому подписчику включают постоянные издержки c^R и переменные издержки, пропорциональные числу статей в номере (c денежных единиц за каждую статью). В постоянные (не зависящие от числа статей) издержки c^R входят затраты на обслуживание счета читателя, а также постоянная часть затрат на транспортировку и погрузочно-разгрузочные работы. Переменные (зависящие от объема журнала) затраты на транспортировку включая плату за канал связи при электронном распространении учтены в величине c.

Предположим, что каждый автор предлагает одну статью. Денежная оценка выгоды, которую получает автор i от публикации, пропорциональна числу читателей: b_i^A за каждого читателя, где b_i^A – это непрерывная случайная величина с функцией распреде-

ления F^A, плотностью f^A и носителем [0, $\overline{b}^A$], распределенная на множестве авторов.

Читатель k может читать статьи только из тех журналов, на которые подписался, получая от каждой прочитанной статьи полезность, денежная оценка которой равна b_k^R. Здесь b_k^R – непрерывная случайная величина с функцией распределения F^R, плотностью f^R и носителем [0, $\overline{b}^R$], распределенная на множестве читателей.

Модель предполагает, что все журналы имеют одинаковые затраты. Авторы могут различаться по полезности, получаемой от публикации статьи, но их статьи не различаются по качеству, т. е. одинаково полезны для читателя. Читатели могут получать разные полезности от прочтения одной и той же статьи, но полезность, получаемая автором от каждого прочтения его статьи, не зависит от того, кто именно ее прочел.

Автор платит журналу за публикацию пропорционально числу читателей, а именно журнал j берет с автора плату $p_j^A \geq 0$ за каждого подписчика; при числе подписчиков n_j^R автор платит $p_j^A\, n_j^R$. Стоимость подписки пропорциональна числу статей. Если журнал j публикует n_j^A статей, то стоимость подписки равна $p_j^R n_j^A$, где $p_j^R \geq 0$ – плата за право прочтения одной статьи.

Замечание 4.4.1. Предположение $p_j^R \geq 0$ отражает существующую ситуацию, но ограничение $p_j^A \geq 0$ выглядит менее правдоподобно. Необходимость в нем отсутствует, так как аналитический аппарат может работать с отрицательными ценами. Предположение $p_j^A \geq 0$ исключает из рассмотрения возможность выплаты авторских гонораров.

В принятых предположениях полные затраты журнала j и его прибыль описываются следующим образом:

$$\mathrm{TC}_j(n_j^A, n_j^R) = c^A n_j^A + c^R n_j^R + c\, n_j^A n_j^R,$$

$$\mathrm{Pr}_j(p_j^A, p_j^R) = n_j^A n_j^R (p_j^A + p_j^R) - \mathrm{TC}_j(n_j^A, n_j^R).$$

От публикации статьи в журнале j автор i получит излишек (чистую выгоду)

$$S_{ij}^A(p_j^A) = n_j^R (b_i^A - p_j^A).$$

Он выберет журнал, приносящий максимальный излишек, при условии что этот излишек неотрицателен. Если $S_{ij}^A(p_j^A) < 0$ для всех j, то автор откажется от публикации. Подписка на журнал j даст читателю k излишек

$$S_{kj}^R(p_j^R) = n_j^A (b_k^R - p_j^R).$$

Предположение 4.4.1. Поскольку читатель k может подписаться на несколько журналов, он подпишется на все журналы, для которых $S_{kj}^R(p_j^R) \geq 0$, т. е. $b_k^R \geq p_j^R$.

Замечание 4.4.2. Из предположения 4.4.1 следует, что читатель способен оплатить подписку на все журналы, приносящие ему неотрицательную выгоду. Иными словами, модель не учитывает бюджетное ограничение читателя. В реальности подписчик (например, библиотека) имеет жесткое бюджетное ограничение, в рамках которого максимизирует суммарную полезность подписки.

Предположим, что на рассматриваемом рынке журналов не существует барьеров входа/выхода (рынок с бесплатным входом, состязательный рынок). Допустим также, что имеется достаточно много потенциальных участников рынка. Издатели журналов не имеют единовременных затрат входа на рынок, однако, как сказано выше, несут затраты c^A на обслуживание каждого автора и c^R – на обслуживание каждого читателя.

Пусть $J-$ число активных (обслуживающих ненулевые доли рынка) журналов. Активные журналы с номерами j $(1 \le j \le J)$ одновременно устанавливают цены p_j^A и p_j^R. Равновесие на рассматриваемом рынке авторы определяют следующим образом: совокупность пар (p_j^A, p_j^R) для $j \in \{1, \ldots, J\}$ такая, что каждый активный журнал имеет неотрицательную прибыль и ни один неактивный журнал не может установить цены так, чтобы получить положительную прибыль.

Замечание 4.4.3. Более точное определение равновесия на состязательном рынке приведено в работе [Mas-Collel, et al., 1995 C. 335].

Авторы работы [McCabe, Snyder, 2007] не доказывают существование равновесия, однако описывают некоторые свойства, которыми должно обладать равновесие, если оно существует. В частности, конкурентная борьба журналов за авторов и читателей приводит к тому, что в равновесии активные (как и неактивные) журналы имеют нулевую прибыль.

Утверждение 4.4.1. [McCabe, Snyder, 2007, с. 11]. В равновесии на рассматриваемом рынке каждый журнал имеет нулевую прибыль.

Структура рынка в равновесии зависит от структуры затрат издателя, что показывают следующие три утверждения.

Утверждение 4.4.2. [McCabe, Snyder, 2007, с. 12]. Пусть $c^R > 0$. Тогда в равновесии существует упорядочение активных журналов, разбиение отрезка $[0, \overline{b}^A]$

$$0 \le B_J^A < B_{J-1}^A < \ldots < B_1^A < B_0^A = \overline{b}^A$$

и разбиение отрезка $[0, \overline{b}^R]$

$$0 \le B_1^R < B_2^R < \ldots < B_{J-1}^R < B_J^R = \overline{b}^R$$

такие, что журнал с номером j (в указанном упорядочении активных журналов) обслуживает всех авторов i, для которых $b_i^A \in (B_j^A, B_{j-1}^A)$, и всех читателей k, для которых $b_k^R \geq B_j^R$. При этом с увеличением номера журнала цены для авторов убывают, а цены для читателей возрастают: $p_j^A < p_{j-1}^A$ и $p_j^R > p_{j-1}^R$ для всех $j \in \{2,\ldots,J\}$.

Таким образом, при $c^R > 0$ в равновесии J активных журналов можно их упорядочить от журнала (с номером 1), устанавливающего максимальную цену для авторов и минимальную цену для читателей, до журнала (с номером J), устанавливающего минимальную цену для авторов и максимальную цену для читателей. Журнал 1 обслуживает наибольшее число читателей и некоторое множество авторов, которые более остальных заинтересованы в публикации (согласны за нее платить). Чем больше номер журнала, тем меньше готовность платить за публикацию у участвующих в нем авторов и тем у́же круг читателей. Если $J > 1$, то некоторые читатели подпишутся более чем на один журнал, так как подписчики журнала j подпишутся и на все более дешевые журналы (см. предположение 4.4.1). Однако утверждение 4.4.2 не позволяет определить число активных журналов, из него даже не следует, что равновесное число журналов больше единицы.

Из утверждения 4.4.2 следует, что при $c^R > 0$ может существовать не более чем один журнал со свободным доступом. Если такой журнал существует, то он обслуживает всех читателей и тех авторов, которые наиболее заинтересованы в публикации. Симметрично, журнал, бесплатный для авторов, может быть только один, и если он существует, то обслуживает самых “богатых” читателей и авторов, наименее заинтересованных в публикации.

Стратегия издателя активного журнала j в соответствии с утверждением 4.4.2 заключается в следующем. Он выбирает группу

авторов, согласных платить от B_j^A до B_{j-1}^A (способ вычисления этих границ в работе не указан), и публикует только их. Этот выбор определяет цену публикации (B_j^A) для автора и число статей (равное числу авторов). Тогда известен доход, полученный издателем от авторов, а также его затраты на работу с авторами и издание одного экземпляра. Цена для подписчиков определяет число подписчиков (всех читателей, у которых готовность платить выше некоторого уровня). Она выбирается таким образом, чтобы покрыть полные затраты издателя (так как прибыль нулевая). В частности, нулевая цена для читателей (свободный доступ) может быть только у одного журнала (с номером *J*). Достаточные условия существования такого журнала сформулированы в работе [McCabe, Snyder, 2007, Propositions 5, 6].

Замечание 4.4.4. Очевидно, что картина, нарисованная утверждением 4.4.2, не соответствует реальности. Это значит, что либо на рассматриваемом рынке существуют барьеры входа/выхода, либо он еще не находится в равновесии, либо $c^R = 0$, либо не выполнено одно из предположений модели (например, предположение 4.4.1).

Равновесие при $c^R=0$ описывается следующим утверждением.

Утверждение 4.4.3. [McCabe, Snyder, 2007, с. 13]. Пусть $c^R = 0$. Тогда в равновесии все авторы с готовностью платить b_i^A (имеющие тип b_i^A) либо обслуживаются одним “нишевым” журналом, либо не обслуживаются. В первом случае журнал устанавливает для читателей и авторов соответственно следующие цены: $p^R(b_i^A) = \text{argmax}_{p^R}\{n^R[1 - F^R(p^R)](b_i^A + p^R - c)\}$ при условии $p^R \in [0,\ c + c^A/n^R[1 - F^R(p^R)]]$ и $p^A(b_i^A) = c - p^R(b_i^A) + c^A/n^R[1 - F^R(p^R)]$. Этот случай реализуется тогда и только тогда, когда $n^R[1 - F^R(p^R)][b_i^A + p^R(b_i^A) - c] \geq c^A$.

Формула цены для подписчика $p^R(b_i^A)$ содержательно означает, что эта цена максимизирует полезность, получаемую автором типа b_i^A при условии безубыточности журнала. Цена автора для $p^A(b_i^A)$ выбирается таким образом, чтобы прибыль издателя была равна нулю.

Из утверждения 4.4.3, с учетом того что b_i^A по предположению есть непрерывная случайная величина, авторы делают следующие выводы относительно равновесной структуры рынка при $c^R = 0$:

а) на рынке может быть континуум активных журналов;

б) интервал типов авторов, в котором $p^R(b_i^A)$ изменяется, обслуживается континуумом журналов;

в) интервал значений b_i^A, в котором цена $p^R(b_i^A)$ постоянна, может обслуживаться любым числом журналов от единицы до континуума.

Замечание 4.4.5. Несмотря на то что распределение величины b_i^A непрерывно, число ее реализаций (авторов) конечно. Поэтому в сформулированных выше выводах лучше говорить не о континууме журналов, а о числе журналов, равном числу наблюдаемых типов авторов.

В работе [McCabe, Snyder, 2007, раздел 5.4] предложена также модель рынка журналов, учитывающая качество статьи и репутацию журнала. Однако в этом случае авторам не удалось получить существенные результаты относительно структуры рынка в равновесии.

4.5. Нелинейное ценообразование

В настоящее время распространены два типа контрактов между поставщиками и потребителями электронных журналов. Потребитель может приобрести право неограниченного доступа к контен-

ту (например, электронному журналу), а может оплачивать каждый доступ пропорционально его интенсивности (измеряемой временем или числом полученных статей). Но возможны и более сложные контракты, например пользователь вносит разовый платеж за право доступа к контенту, а затем оплачивает также каждый доступ пропорционально его интенсивности. Эта нелинейная схема ценообразования получила название “двухставочный тариф” (см. [Тироль, 2000, т. 1, п. 3.3.2). Поскольку при двухставочном тарифе разовый платеж может быть меньше, чем при контракте первого типа, а цена доступа − меньше, чем при контракте второго типа, использование нелинейного ценообразования может быть выгодным для всех участников рынка.

В работе [Varian, 2000, раздел IX] отмечена (но не изучена) возможность нелинейного ценообразования на рынке видеокассет. Применительно к рынку электронных журналов эта идея, насколько известно, не обсуждалась.

4.6. Платит читатель или автор?

В работе [McCabe, Snyder, 2007] предложена схема формального экономического анализа научных журналов. Изучаются две модели деловой активности: традиционная (или “Платит читатель”), с одной стороны, и модель открытого доступа (“Платит автор”) − с другой. Во-первых, приводятся точки зрения основных игроков на рынке (коммерческих издателей, некоммерческих издателей, библиотек) на то, какая модель деловой активности является наилучшей. Во-вторых, обсуждается вопрос о том, какая модель деловой активности является наилучшей для всего общества. Для этого целесообразно рассмотреть, какая модель является наилучшей для ученых − как авторов и как читателей.

4.6.1. Перспективы основных игроков. Крупные коммерческие издатели строго отстаивают статус-кво модели деловой активности “Платит читатель”. Использование ими модели “Пла-

тит читатель", несмотря на то что они имеют возможность выбрать модель "Платит автор" или какую-либо другую альтернативную модель, наводит на мысль, что модель "Платит читатель" рассматривается как наиболее прибыльная. За этим суждением откровенно стоит экономика. Библиотеки должны будут платить высокие цены за журналы, поскольку журналы являются монополистами по статьям, которые они публикуют. Если ученому, которого обслуживает библиотека, для его исследования потребуется статья в данном журнале, то удобная замена ее отсутствует.

В ближайшее время переход от печатного к цифровому распространению может также увеличить прибыльность модели "Платит читатель" (см. [McCabe, 2002]). Модель "Платит автор", в противоположность этому, дает издателям значительно меньше власти на рынке. Авторы могут свободно переходить от журнала к журналу с относительно сходной престижностью, принимая решение, куда представить статью для публикации. В частности, основывая это решение на минимальной стоимости публикации и оказывая тем самым давление на эту стоимость.

Даже если модель открытого доступа (ОД) рассматривать как менее выгодную модель деловой активности, коммерческие издатели могут опасаться ее распространения в силу ряда причин. Первой очевидной причиной является то, что распространение модели ОД означает вход на рынок журналов ОД, а вход любого конкурента, независимо от используемой модели деловой активности, снижает прибыли должностных лиц. Во-вторых, журналы ОД вместе могут быть более сильными конкурентами, а не средними, например если они являются некоммерческими журналами, взявшими на себя обязательства максимизировать импакт-фактор, а не прибыль (безусловно, журнал ОД необязательно должен быть некоммерческим). Третьей возможностью является скоординиро-

ванный бойкот авторов и читателей традиционных коммерческих журналов, если ОД получит популярность.

Некоммерческие издатели открыто признают определенное количество (иногда конфликтующих) целей. Они хотели бы иметь как можно более широкий круг читателей, а также зарабатывать прибыль для финансирования других операций. Статус-кво исторически позволял им выполнять обе эти функции. Таким образом, ОД является привлекательной альтернативой. Круг читателей будет продолжать расширяться для ОД. Кроме того, открытый доступ может быть одним из направлений в борьбе с принимающими угрожающие размеры опасностями, которые обусловлены стратегиями группирования крупных коммерческих издателей (“Большая сделка”) и уменьшением числа индивидуальных (не организаций) подписок. Контракты “Большой сделки” снизили возможности библиотек переводить финансирование подписки, например с издательства “Elsevier”, на других издателей, особенно на небольших некоммерческих издателей. Открытый доступ позволяет обойти эту проблему, полагаясь на оплату авторов, но есть ли авторы (или финансирующие их источники), желающие платить? И сколько их?

4.6.2. Перспективы ученых. Престижность является валютой академической организации. Помимо простого чувства удовольствия с ростом престижа увеличиваются шансы ученого на продвижение, получение новой должности, более высокой зарплаты и т. п. Престиж приходит из выполнения высококачественного исследования. Поскольку измерение качества связано с преодоленными трудностями, используются такие стенографические средства измерения, как репутация журнала, в котором опубликована статья, или число цитирований, которые она сгенерировала. Таким образом, авторы имеют побуждение представлять свои статьи для опубликования в престижные журналы как вследствие

репутации самого журнала, так и вследствие того, что престижные журналы привлекают больше читателей и авторы могут ожидать большего количества цитирований их статей. С учетом такой динамики не очевидно, будут ученые предпочитать модель “Платит автор” или модель “Платит читатель” (или комбинацию этих двух моделей).

Размер платы за публикацию статьи связан с тем, что ОД повышает расходы на проведение исследований и распространение их результатов для авторов. В итоге это снизит объем исследований и количество опубликованных статей, что косвенно повредит читателям. С другой стороны, стоимость подписки, связанная с традиционной моделью деловой активности, уменьшает число читателей, так как библиотеки продолжают отменять подписки, что непосредственно вредит читателям и косвенно вредит авторам (посредством снижения числа читателей). При одновременном рассмотрении ролей “ученый как автор” и “ученый как читатель” определение чистой стоимости ОД для ученых является достаточно сложной задачей.

4.6.3. Общая схема экономического анализа рынка журналов. Упростить сложный вопрос о том, какая модель деловой активности является оптимальной для ученых, может формальная общая схема экономического анализа. Основой такой общей схемы является динамика журнала, связывающая читателей и авторов и называемая в экономической литературе двусторонним рынком [Rochet, Tirole, 2003]. С одной стороны такого рынка авторы получают выгоду от большего воздействия и цитирований и как следствие предпочитают журнал, который имеет больше читателей. С другой стороны, читатели получают выгоду от содержания и, таким образом, предпочитают журналы с большим количеством статей. Определение оптимального баланса между этими двумя группами игроков включает измерение выгод, которые получает

каждая сторона от большего или меньшего участия другой стороны, расчет стоимости увеличения (или уменьшения) числа авторов и читателей и последующую идентификацию ценового набора, т. е. платы автора и стоимости подписки, что будет максимизировать общую чистую прибыль.

На основе предварительного анализа данной проблемы сделано предположение, что оптимальные цены будут зависеть от степени конкуренции на рынке между журналами. При одной крайности – монопольный журнал – цены, выбранные журналом, максимизирующим прибыль, как правило, будут иметь положительное значение как для авторов, так и для читателей, даже если допустить, что стоимость распространения будет нулевой. Положительные цены на обеих сторонах рынка позволяют журналу извлекать определенную прибыль с обеих сторон рынка.

Данный результат подразумевает, что низкие стоимости распространения не будут автоматически приводить к появлению рынка открытого доступа. Данный результат не подразумевает, что ОД не является жизнеспособным при установлении монополии. Если журнал имеет целью максимизировать свой круг читателей, а не прибыль, то отсутствие компетенции будет облегчать его способность экспериментировать с моделями деловой активности, отличающимися от традиционных включая модель открытого доступа.

При другой предельно идеальной конкуренции − между журналами равного качества – возможен континуум равновесий; некоторые из них благоприятствуют читателям, а некоторые – авторам часто включая открытый доступ как равновесие. В случаях, когда стоимостью распространения можно пренебречь, ОД появляется как равновесие и является экономически эффективным (по крайней мере, в случае контрольной задачи, когда выгоды автора и читателя берутся приблизительно равными). Иными словами,

ОД максимизирует общую чистую прибыль для авторов и читателей (т. е. ученых) и для общества в целом. При совместном рассмотрении данный набор результатов предполагает, что ОД может быть жизнеспособным в обстановке конкуренции и эффективным, но его появление не гарантировано.

Интуитивно, ОД представляется эффективным в ситуации с пренебрежимо малыми расходами на распространение и приблизительно равными выгодами автора и читателя, поскольку цены отражают расходы на добавление авторов и читателей. Даже в том случае, когда расходы на распространение равны нулю, добавление автора все еще является дорогостоящим вследствие затрат на выпуск первой копии. Положительная оплата автором своей публикации отражает эти расходы. Добавление читателя является бесплатным, что отражается в нулевой плате читателя, связанной с ОД. Такая гипотеза опирается на предположение о приблизительно равных выгодах автора и читателя. Если читатели получают диспропорциональную выгоду от чтения дополнительных статей, то может быть эффективным введение положительной платы читателя, для того чтобы субсидировать представление авторами статей к публикации.

Рассмотренная общая схема абстрагирована от определенного числа деталей фактического рынка для журналов, которые могут повлиять на жизнеспособность ОД, и целесообразно включить эти детали в более широкую общую схему. Например, решение финансирующих организаций поддерживать оплату авторами публикации их статей будет способствовать появлению ОД. Кроме того, представленный подход, по сути, является статистическим и не учитывает трудность входа на рынок журналов, когда для установления или изменения репутации может потребоваться длительное время. Мы абстрагировались от таких барьеров входа, поскольку они применяются к любому новому журналу, а не

только к журналам открытого доступа. Тем не менее на практике для расширения ОД, вероятно, потребуется формирование новых журналов, и, следовательно, перспективы ОД при данном типе окружающих условий должны быть тщательно рассмотрены.

Представленные в обзоре модели и результаты объясняют основные особенности современных рынков печатных и электронных научных журналов. Рост цен, продажа электронных журналов комплектами, объединение подписчиков в консорциумы, расслоение авторов и читателей между журналами, возникновение большого числа “нишевых” журналов − закономерные явления, обусловленные спецификой рассматриваемых отраслевых рынков. Участники рынка электронных журналов вынуждены приспосабливать свои стратегии к указанным тенденциям.

Многие вопросы, относящиеся к равновесной структуре рынка электронных журналов и оптимальным стратегиям его участников, остаются пока открытыми. В дальнейших исследованиях первоочередного внимания, на наш взгляд, заслуживает модель двустороннего рынка. Она наиболее реалистична и может привести к результатам, имеющим не только теоретическое, но и прикладное значение.

Глава 5. Оптимизация подписки на электронные журналы

> В каждый данный момент существует лишь тонкий слой между “тривиальным” и недоступным. В этом слое и делаются математические открытия. Заказная прикладная задача поэтому в большинстве случаев или решается тривиально, или вообще не решается...
>
> *А. Н. Колмогоров. Запись в дневнике от 14 сентября 1943 г.*
>
> Цит. по: А. Н. Колмогоров (Curriculum Vitae)).

В настоящей главе рассмотрим стратегию подписки на электронные копии научных журналов, т. е. специфический сектор рынка информационных ресурсов с точки зрения потребления.

В сложившейся ситуации, описанной в предисловии к гл. 4, подписчику (консорциуму подписчиков) требуется инструмент для выработки рациональной стратегии. Мы предлагаем две модели, которые могут быть использованы в качестве такого инструмента. Одна из них максимизирует суммарную полезность подписанных журналов в рамках заданного бюджета, другая находит наиболее дешевый вариант подписки на заданный набор журналов. В указанных моделях используется следующая входная информация: бюджет подписки; число, состав и цены предлагаемых комплектов; оценки полезностей журналов, входящих в эти комплекты.

Для того чтобы выбрать (для подписки) некоторый набор комплектов, необходимо уметь сравнивать такие наборы по их полезности для подписчика, т. е. знать функцию полезности подписчика на множестве всех возможных наборов. Предположим, что полезность набора комплектов равна сумме полезностей журналов, входящих в эти комплекты, тогда достаточно знать функцию по-

лезности подписчика на множестве журналов. Заметим, что здесь, как и в общем случае, функция полезности является характеристикой лица, принимающего решения. Иными словами, никто не может знать ее лучше, чем подписчик, поэтому именно его прерогативой является выбор способа оценивания полезностей журналов.

5.1. Постановка задачи

Имеется m журналов (с номерами $1,\ldots,m$) и n комплектов журналов (с номерами $1,\ldots,n$). Состав комплекта j определен величинами a_{ij}, такими что $a_{ij} = 1$, если журнал i входит в комплект j, иначе $a_{ij} = 0$. Потребитель (потенциальный подписчик) хочет подписаться на некоторые журналы, но поставщики предлагают подписку только комплектами. Комплекты могут пересекаться, т. е. один журнал может входить в два и более комплектов. Предполагается, что известна оценка полезности каждого журнала. В данной ситуации проблему выбора рациональной подписки можно сформулировать двумя способами:

1) определить план подписки таким образом, чтобы суммарная стоимость включенных в него комплектов не превосходила некоторой заданной величины (бюджета) и чтобы при таком условии суммарная полезность входящих в эти комплекты журналов была максимальной;

2) определить план подписки на комплекты таким образом, чтобы с минимальными затратами обеспечить подписку на указанные (приоритетные, наиболее полезные) журналы.

5.2. Модель 1

Пусть p_j − цена комплекта j, B − бюджет потребителя, $v_i \geq 0$ − оценка (ретроспективная или экспертная) полезности журнала i для этого потребителя.

Введем переменные:

$x_j = 1$, если потребитель подписался на комплект j, иначе $x_j = 0$;

$z_i = 1$, если журнал i входит в подписку рассматриваемого потребителя, иначе $z_i = 0$.

Замечание 5.1. Значения переменных x_j и z_i должны быть согласованы, т. е. если $x_j = 1$, то $z_i = 1$ для всех i, таких что $a_{ij} = 1$ (подписка на комплект j обеспечивает подписку на все журналы, входящие в этот комплект).

Модель определяет номера комплектов, на которые следует подписаться при фиксированном бюджете.

$$\sum_j p_j x_j \le B; \tag{5.1}$$

$$z_i \le \sum_j a_{ij} x_j \quad \text{для всех } i; \tag{5.2}$$

$$0 \le z_i \le 1,\ 0 \le x_j \le 1; \tag{5.3}$$

$$x_j,\ z_i - \text{целые для всех } i, j; \tag{5.4}$$

$$f(\mathbf{z}) = \sum_i v_i z_i \to \max. \tag{5.5}$$

Ограничение (5.1) − бюджетное. Условие (5.2) означает, что журнал входит в подписку, только если он входит хотя бы в один комплект, на который оформлена подписка. Условия (5.3), (5.4) следуют из определения переменных. Целевая функция (5.5) описывает суммарную полезность журналов, вошедших в подписку.

Утверждение 5.1. Пусть $(\mathbf{x}^0, \mathbf{z}^0)$ − оптимальное решение задачи (5.1)–(5.5). Если $v_i > 0$ и $z_i \le \sum_j a_{ij} x_j$, то $z_i^0 = 1$.

Доказательство. Пусть $z_i^0 < 1$. Определим вектор $\mathbf{z}$ следующим образом: $z_i = 1$, $z_k = z_k^0$ для $k \ne i$. Тогда $(\mathbf{x}^0, \mathbf{z})$ − допустимое решение задачи (5.1)–(5.5) и $f(\mathbf{z}) > f(\mathbf{z}^0)$, что противоречит оптимальности набора $(\mathbf{x}^0, \mathbf{z}^0)$.

Простое утверждение 5.1 показывает, что модель 1 обеспечивает согласованность значений переменных x_j и z_i (см. замечание 5.1). Иными словами, в оптимальный план подписки входят в точности те журналы, которые включены в выбранные комплекты.

Модель 1 приводит к задаче целочисленного линейного программирования (5.1)–(5.5). При большом числе целочисленных переменных решение задач этого класса связано с серьезными техническими проблемами. Если же снять условие целочисленности переменных, т. е. перейти к линейной релаксации (5.1)–(5.3), (5.5) исходной задачи, то решение, вообще говоря, окажется не целочисленным. Это легко показать на следующем примере.

Пример 5.1. Пусть предложение ограничено двумя комплектами, в каждый комплект входит ровно один журнал, эти журналы различны и имеют ненулевую полезность. Предположим, что цена годовой подписки на каждый комплект меньше бюджета, а суммарная стоимость обоих комплектов больше бюджета. Тогда в оптимальный план войдут годовая подписка на журнал с максимальной полезностью и часть годовой подписки на второй журнал. Заметим, что данное решение не противоречит здравому смыслу.

Задача (5.1)–(5.3), (5.5), не включающая условие целочисленности, легко решается с помощью стандартных пакетов линейного программирования. Для нее справедливо следующее утверждение, аналогичное утверждению 5.1.

Утверждение 5.2. Пусть $(\mathbf{x}^0, \mathbf{z}^0)$ − оптимальное решение задачи (5.1)–(5.3), (5.5). Если $v_i > 0$, то $z_i^0 = \min\left\{\sum_j a_{ij}x_j^0, 1\right\}$.

Доказательство. Из условий (5.2), (5.3) следует, что $z_i^0 \le \min\left\{\sum_j a_{ij}x_j^0, 1\right\}$. Пусть $z_i^0 < \min\left\{\sum_j a_{ij}x_j^0, 1\right\}$. Опреде-

лим вектор $\mathbf{z}$ следующим образом: $z_i = \min\left\{ \sum_j a_{ij} x_j^0, 1 \right\}$, $z_k = z_k^0$ для $k \neq i$. Тогда $(\mathbf{x}^0, \mathbf{z})$ – допустимое решение рассматриваемой задачи и $f(\mathbf{z}) > f(\mathbf{z}^0)$, что противоречит оптимальности набора $(\mathbf{x}^0, \mathbf{z}^0)$.

В частности, если оптимальное решение линейной релаксации задачи (5.1)–(5.3), (5.5) рекомендует подписать комплекты j и k на неполный год $(0 < x_j^0 < 1,\ 0 < x_k^0 < 1)$ и журнал i входит только в эти два комплекта, то этот журнал будет подписан на срок $\min\{x_j^0 + x_k^0, 1\}$. Например: если $x_j^0 = 0{,}5$ и $x_k^0 = 0{,}3$, то $z_i^0 = 0{,}8$; если $x_j^0 = 0{,}4$ и $x_k^0 = 0{,}7$, то $z_i^0 = 1$.

Следствие. В условии (5.4) можно ограничиться требованием целочисленности только для переменных x_j.

Доказательство. Если в оптимальном решении $(\mathbf{x}^0, \mathbf{z}^0)$ задачи (5.1)–(5.3), (5.5) переменные x_j принимают целые значения, то значения $z_i^0 = \min\left\{ \sum_j a_{ij} x_j, 1 \right\}$ также являются целыми, так как $a_{ij} \in \{0,1\}$. Следовательно, $(\mathbf{x}^0, \mathbf{z}^0)$ является решением задачи (5.1)–(5.3), (5.5).

Задача (5.1)–(5.3), (5.5) дает хорошо интерпретируемое решение в частном случае, когда комплекты попарно не пересекаются (каждый журнал входит ровно в один комплект). Рассмотрим этот случай.

Пусть $S(j) = \{i \mid a_{ij} = 1\}$ (множество номеров журналов, входящих в комплект j) и $S(j) \cap S(k) = \varnothing$ при $j \neq k$. Положим

$V_j = \sum_{i \in S(j)} v_i$ (суммарная полезность комплекта). Тогда ограничения (5.2) можно записать следующим образом: $z_i \le x_j$ при $i \in S(j)$. Поэтому из (5.3) и требования максимизации следует, что $z_i = x_j$ для всех $i \in S(j)$. Тогда $f(\mathbf{z}) = \sum_i v_i z_i = \sum_j V_j x_j$ и задача (5.1)–(5.3), (5.5) принимает вид

$$\sum_j p_j x_j \le B; \tag{5.6}$$

$$0 \le x_j \le 1; \tag{5.7}$$

$$g(\mathbf{x}) = \sum_j V_j x_j \to \max. \tag{5.8}$$

Исключим из рассмотрения комплекты с нулевой полезностью и будем считать, что $V_j > 0$ для всех j. Положим $R_j = V_j / p_j$. Без ограничения общности допустим, что комплекты упорядочены по невозрастанию величин R_j: $R_1 \ge R_2 \ge \ldots \ge R_n$.

Утверждение 5.3. Существует оптимальное решение $\mathbf{x}$ задачи (5.6)–(5.8), для которого $x_1 = \min\{B/p_1, 1\}$.

Доказательство. Пусть $\mathbf{x}^0$ – оптимальное решение задачи (5.6)–(5.8) и $x_1^0 < \min\{B/p_1, 1\}$. Тогда $B > 0$.

Случай 1. $x_j^0 = 0$ для всех $j > 1$. Положим $x_1 = \min\{B/p_1, 1\} > x_1^0$ и $x_j = 0$ для $j > 1$. Определенный таким образом вектор $\mathbf{x}$ удовлетворяет условиям (5.6), (5.7) и $g(\mathbf{x}) > g(\mathbf{x}^0)$, что противоречит выбору $\mathbf{x}^0$.

Случай 2. $x_k^0 > 0$ для некоторого $k > 1$. Положим $\varepsilon = \min\{1 - x_1^0, x_k^0 p_k / p_1\}$ и определим вектор $\mathbf{x}$ следующим образом:

$$x_1 = x_1^0 + \varepsilon; \quad x_k = x_k^0 - \frac{p_1}{p_k}\varepsilon; \quad x_j = x_j^0 \text{ для } j \notin \{1, k\}.$$

Тогда условия (5.7) выполнены для вектора **x** и

$$\sum_j p_j x_j = \sum_j p_j x_j^0 + p_1\varepsilon - p_k \frac{p_1}{p_k}\varepsilon = \sum_j p_j x_j^0 \le B. \qquad (5.9)$$

При этом

$$g(\mathbf{x}) = g(\mathbf{x}^0) + V_1\,\varepsilon - V_k \frac{p_1}{p_k}\varepsilon = g(\mathbf{x}^0) + \varepsilon\, p_1(R_1 - R_k) \ge g(\mathbf{x}^0),$$

так как $R_1 \ge R_k$. Случай $g(\mathbf{x}) > g(\mathbf{x}^0)$ невозможен по выбору $\mathbf{x}^0$, поэтому **x** − оптимальное решение задачи (5.6)–(5.8). Возможны два случая.

Случай 2.1. Пусть $1 - x_1^0 \le x_k^0 p_k / p_1$. Тогда $\varepsilon = 1 - x_1^0$, $x_1 = 1$ и

$$1 \le (p_1 x_1^0 + p_k x_k^0) / p_1 \le B / p_1.$$

Следовательно, $x_1 = \min\{B/p_1, 1\}$ и утверждение справедливо.

Случай 2.2. Пусть теперь $1 - x_1^0 > x_k^0 p_k / p_1$. Тогда $\varepsilon = x_k^0 p_k / p_1$, $x_1 = x_1^0 + x_k^0 p_k / p_1 < 1$. Из выражения (5.9) следует, что $x_1 \le B / p_1$. Если $x_1 = B / p_1$, то утверждение справедливо. Иначе $x_1 < \min\{B / p_1, 1\}$, и можно выбрать $\mathbf{x}^1 = \mathbf{x}$ в качестве исходного оптимального решения (вместо $\mathbf{x}^0$), вернуться к случаю 1 и повторить рассуждение.

При каждом повторении мы имеем некоторое оптимальное решение $\mathbf{x}^{k-1}$ и находим новое оптимальное решение $\mathbf{x}^k$, которое либо удовлетворяет условию утверждения (см. случаи 2.1 и 2.2), либо имеет меньше ненулевых компонент, чем $\mathbf{x}^{k-1}$. Если у вектора $\mathbf{x}^k$ имеется только одна ненулевая компонента, то выполняется

случай 1, который приводит к противоречию. Следовательно, за конечное число шагов (не более размерности n вектора $\mathbf{x}$) будет построено искомое оптимальное решение.

Утверждение 5.3 позволяет описать оптимальное решение задачи (5.6)–(5.8). Сохраняя введенное выше упорядочение комплектов, для $k \geq 0$ положим

$$x_{k+1}^0 = \min\left\{\frac{B_k}{p_k}, 1\right\}, \tag{5.10}$$

где $B_k = B - \sum_{j=1}^{k} p_j x_j^0$ (очевидно, что $B_0 = B$).

Теорема 5.1. Определенный выше вектор $\mathbf{x}^0$ является оптимальным решением задачи.

Доказательство. Очевидно, что любое оптимальное решение $\mathbf{x}$ задачи (5.6)–(5.8) обладает следующим свойством: вектор $(x_k,\ldots,x_n)$ является оптимальным решением задачи

$$\sum_{j \geq k} p_j x_j \leq B_j; \tag{5.11}$$

$$0 \leq x_j \leq 1 \quad \text{для} \quad j \geq k; \tag{5.12}$$

$$g(\mathbf{x}) = \sum_{J \geq k} V_j x_j \to \max. \tag{5.13}$$

Задача (5.11)–(5.13) является частным случаем задачи (5.6)–(5.8). Поэтому теорема 5.1 следует из утверждения 5.3, примененного к задаче (5.11)–(5.13).

Таким образом, если комплекты не пересекаются, оптимальный план подписки $\mathbf{x}$ легко построить, используя следующий алгоритм.

Шаг 0 (подготовительный). Упорядочиваем комплекты по неубыванию удельной полезности (полезность на единицу стоимости). Полагаем $B_0 = B$.

Шаг $k > 0$ (общий). Полагаем $x_k = \min\{B_k / p_k, 1\}$, $B_{k+1} = B_k - p_k x_k$.

Данный алгоритм строит вектор **x**, все координаты которого, за исключением, может быть, одной, равны единице или нулю. Единичные координаты соответствуют включаемым в подписку комплектам с первыми номерами; дробная координата, если она имеется, соответствует использованию остатка бюджета для неполной подписки на "замыкающий" комплект; остальные комплекты не входят в подписку, и соответствующие координаты вектора **x** равны нулю.

Введем в задачу (5.6)–(5.8) условие целочисленности переменных: заменим ограничение (5.7) условием

$$x_j \in \{0,1\} \tag{5.14}$$

(каждый комплект подписывается полностью либо не подписывается).

Утверждение 5.4. Задача (5.1)–(5.5) NP-полна.

Доказательство. Частным случаем задачи (5.1)–(5.5) (при попарно непересекающихся комплектах) является задача (5.6), (5.14), (5.8). А это − известная задача о рюкзаке, которая NP-полна [Гэри, Джонсон, 1982, с. 87].

Утверждение 5.4 показывает, что эффективного алгоритма решения задачи (5.1)–(5.5), вероятней всего, не существует. Однако можно предположить, что эту задачу удастся решить на реальных данных с помощью стандартных программ целочисленного линейного программирования, так как число целочисленных переменных невелико (по следствию из утверждения 5.2 равно числу комплектов).

Относительно возможности решения задачи (5.6), (5.14), (5.8) (т. е. задачи (5.1)–(5.5) в случае попарно непересекающихся комплектов) приведем авторитетное мнение М. Гэри и Д. Джонсона

[Гэри, Джонсон, 1982, с. 22]: «Алгоритмы ветвей и границ столь успешно решают задачу о рюкзаке, что многие исследователи считают эту задачу “хорошо решаемой”, хотя алгоритмы ветвей и границ имеют экспоненциальную сложность». В работе [Гэри, Джонсон, 1982, с. 171–173] также приведен обзор приближенных алгоритмов полиномиальной трудоемкости для решения задачи о рюкзаке.

5.3. Модель 2

Модель 1 имеет не бросающийся в глаза, но существенный недостаток: она может “предпочесть” комплект, который включает много журналов с небольшой полезностью, комплекту, содержащему небольшое количество очень полезных журналов. Продемонстрируем этот эффект на примере.

Пример 5.2. Имеется только два комплекта. В комплект 1 входит единственный журнал с полезностью 100, а комплект 2 содержит 200 журналов, каждый из которых имеет полезность 1. Цены комплектов одинаковы и равны p. Если бюджет позволяет приобрести только один комплект $\left(p \le B < 2p \right)$, то модель 1 “выберет” комплект 2. Очевидно, что такой выбор можно оспорить.

Причина указанной неадекватности модели заключается в том, что ни один известный способ оценки полезностей журналов не позволяет, строго говоря, суммировать эти полезности: отнюдь не очевидно, что журнал с полезностью $a + b$ равноценен комплекту, включающему два журнала с полезностями a и b. Однако для того чтобы осуществить разумный выбор, не обязательно знать числовые оценки полезностей журналов: достаточно упорядочить журналы по полезности. Представляется целесообразным в первую очередь подписываться на комплекты, содержащие наиболее полезные (авторитетные) журналы, затем − на комплекты, включающие журналы “второго уровня”, и т. д.

Кроме того, если имеются разные (по составу и по стоимости) варианты подписки, обеспечивающие максимальную (в рамках бюджета) суммарную полезность, то модель 1 не обязательно выберет наиболее дешевый вариант.

Еще один стимул для изменения модели состоит в том, что бюджет подписки, как правило, не является жестким. Может оказаться, что небольшое увеличение бюджета приведет к значительному росту целевой функции (5.5) или, напротив, существенное уменьшение бюджета не нанесет существенного ущерба значению целевой функции. В силу этого представляет интерес минимальный бюджет, обеспечивающий подписку на k наиболее приоритетных журналов.

Упорядочим все доступные журналы по невозрастанию полезности: $v_1 \geq v_2, \ldots, v_m$. Заметим, что для такого упорядочения не обязательно иметь числовые оценки полезностей. Модель определяет наиболее дешевый план подписки (на комплекты), включающий первые k журналов.

$$\sum_j a_{ij} x_j \geq 1 \text{ для } i \in \{1, \ldots, k\}; \qquad (5.15)$$

$$x_j \in \{0,1\} \text{ для всех } j; \qquad (5.16)$$

$$\sum_j p_j x_j \to \min. \qquad (5.17)$$

Здесь переменная x_j равна единице, если подписка на комплект j входит в план, и нулю − в противном случае. Поэтому ограничение (5.15) требует, чтобы первые k журналов вошли в подписку. Условие (5.16) связано с приведенной выше интерпретацией переменных x_j. Целевая функция (5.17) описывает суммарную стоимость подписки. Результаты решения задачи (5.15)–(5.17) для различных k дают ориентир для принятия решения о величине бюджета.

Задача (5.15)–(5.17) является задачей целочисленного линейного программирования. В общем случае целочисленность решения ее линейной релаксации не гарантирована, о чем свидетельствует следующий пример.

Пример 5.3. Имеется три журнала и три комплекта. В комплект 1 входят журналы 1 и 2, в комплект 2 − журналы 1 и 3, в комплект 3 − журналы 2 и 3. Полезности всех журналов равны единице ($v_i = 1$), цена каждого комплекта также равна единице ($p_j = 1$). Определим минимальную стоимость подписки на все три журнала, пренебрегая условием целочисленности, т. е. рассмотрим соответствующий вариант задачи (5.15), (5.17) при $k = 3$:

$$x_1 + x_2 \geq 1,\ x_1 + x_3 \geq 1,\ x_2 + x_3 \geq 1,\ 0 \leq x_j \leq 1,\ \ x_1 + x_2 + x_3 \to \min.$$

Отсутствие условия целочисленности переменных несколько изменяет интерпретацию ограничений. Например, первое ограничение требует, чтобы подписки на комплекты, содержащие журнал 1, в совокупности обеспечивали доступ к нему на весь период. Аналогично интерпретируются второе и третье ограничения.

Решением данной задачи является вектор $\mathbf{x} = (0{,}5,\ 0{,}5,\ 0{,}5)$, дающий значение целевой функции (суммарная стоимость подписки) 1,5. Это решение можно интерпретировать следующим образом: комплект 1 подписываем на первое полугодие, комплект 2 − на второе, а комплект 3 − на полгода с дополнительным условием поставки журнала 2 во втором полугодии, а журнала 3 − в первом. Понятно, что количество комплектов и количество журналов в каждом комплекте можно подобрать так, чтобы даже такая сомнительная трактовка дробного решения была невозможна. Заметим, что при учете требования целочисленности переменных доступ ко всем журналам может обеспечить только подписка на все три комплекта с затратами 3.

Задачу (5.15)–(5.17) можно переформулировать следующим образом. Дана совокупность подмножеств $S_1, \ldots, S_n$ множества $M = \{1, \ldots, k\}$ (M − множество номеров журналов, на которые должна быть оформлена подписка, $S_j = \{i \mid i \in M$ и журнал i входит в комплект $j\}$). Каждому подмножеству (комплекту) S_j сопоставлена стоимость p_j. Требуется найти систему подмножеств $S_{j(1)},\ldots,S_{j(r)}$ минимальной суммарной стоимости, покрывающую M, т. е. удовлетворяющую условию

$$\bigcup_{t=1}^{r} S_{j(t)} = M.$$

Утверждение 5.5. Задача (5.15)–(5.17) NP-трудна.

Доказательство. Если $p_j = 1$ для всех j, то в приведенной выше формулировке задача (5.15)–(5.17) эквивалентна задаче о минимальном покрытии, которая NP-полна [Пападимитриу, Стайглиц, 1985, с. 402–403].

Утверждение 5.5 ставит под сомнение существование эффективного алгоритма решения задачи (5.15)–(5.17) в общем случае. Однако можно предположить, что ее удастся решить на реальных данных с помощью стандартных программ целочисленного линейного программирования, так как число целочисленных переменных равно числу комплектов.

Если комплекты попарно не пересекаются, то задача (5.15)–(5.17) тривиальна и может быть решена с использованием следующего алгоритма.

Предположим, что каждый из первых k журналов входит в какой-то комплект (в противном случае задача несовместна). Для $i \in \{1, \ldots, k\}$ пусть $j(i)$ − номер комплекта, включающего журнал i (этот номер однозначно определен, так как комплекты попарно не

пересекаются). Для $j \in \{1, \ldots, n\}$ положим $A_j = \{i \mid$ журнал i входит в комплект $j\}$.

Подготовительный шаг. Полагаем $I = J = \varnothing$ (на каждом шаге I и J – это множества номеров журналов и соответственно комплектов, которые включены в план подписки до текущего шага).

Общий шаг. Если $I = \{1, \ldots, k\}$, то завершаем алгоритм. В противном случае существует i такое, что $1 \le i \le k$ и $i \notin I$. Выбираем наименьшее такое i и полагаем $I := I \cup A_{j(i)}$, $J := J \cup \{j(i)\}$.

Очевидно, что алгоритм заканчивает работу не более чем за k шагов, после чего множество J указывает, какие комплекты следует подписать, а $B_k = \sum_{j \in J} p_j$ – минимальная стоимость подписки, покрывающей первые k журналов.

5.4. Входная информация

Для информационного наполнения моделей необходимо помимо величин m (число журналов) и n (число комплектов) знать следующие параметры:

– бюджет B (для модели 1);

– полезность v_i каждого журнала i;

– стоимость p_j каждого комплекта j.

В заключение сформулируем основные результаты.

5.4.1. Построены две модели формирования плана подписки при условии, что подписываться можно не на отдельные журналы, а только на экзогенно заданные комплекты журналов. Комплекты, вообще говоря, могут пересекаться. Модель 1 максимизирует суммарную полезность подписанных журналов в рамках заданного бюджета. Модель 2 находит наиболее дешевый вариант подписки на заданный набор журналов.

5.4.2. Обе рассмотренные модели приводят к задачам целочисленного линейного программирования. Доказано, что эти задачи NP-полны. Следовательно, отсутствуют основания предполагать существование эффективного алгоритма для решения данных задач.

5.4.3. Доказано, что в модели 1 достаточно наложить условие целочисленности на n переменных, где $n-$ число комплектов. В модели 2 число целочисленных переменных равно n по построению. Отсюда следует, что в реальной ситуации задачи можно решить, используя стандартные средства (например, пакет LINDO [LINDO, soft]), поскольку число комплектов, как правило, невелико.

5.4.4. В случае если комплекты попарно не пересекаются, модели существенно упрощаются. Первая модель сводится к "задаче о рюкзаке", которая хорошо решается методом ветвей и границ даже при сравнительно большой размерности. Задача, соответствующая второй модели, в этом случае эффективно решается с помощью простого алгоритма, приведенного в п. 5.3.

5.4.5. Указанный в п. 5.2 эффективный алгоритм для решения линейной релаксации первой модели (без условий целочисленности переменных) при попарно непересекающихся комплектах журналов дает решение, в котором нецелочисленное значение имеет не более чем одна переменная. Это значение можно интерпретировать как подписку соответствующего комплекта на часть стандартного срока.

Глава 6. Цитирование

Одной из задач библиометрии является изучение межтекстовых связей, поэтому термины “цитата” и “цитирование” должны трактоваться однозначно. Цитатой (*Quotation*) в узком смысле называется дословная выдержка из какого-либо текста, сопровождаемая ссылкой на автора и (или) источник. При этом важно, что цитируемый (вставленный) текст однозначно идентифицируется как вставленный (т. е. как часть другого текста). Термин “цитата” в широком смысле означает любое включение фрагмента чужого текста в авторский текст. В качестве синонимов используются слова “реминисценция”, “аллюзия”, “заимствование”, “подтекст”, “стилизация”, “пародия”.

Основная цель цитирования − проявление интеллектуальной честности по отношению к идеям, которые ранее высказали другие авторы, а не создание у читателя представления, что авторы работы послужили оригинальным первоисточником этих идей. Цитирование представляет собой признание автора, года, названия и места издания источника (журнал, книга или другое издание), использованного в опубликованной работе. Такое цитирование можно считать мерой использования и влияния цитированной работы. Меры цитирования вычисляются для отдельной статьи (как часто она цитируется); автора (общее число цитирований или среднее число цитирований в расчете на статью); журнала (импакт-фактор журнала или среднее число цитирований опубликованных в журнале статей).

Точнее, цитирование представляет собой сокращенное буквенно-цифровое выражение (например, [ABC, 2011]), включенное в структуру творческой работы, которое обозначает запись в разделе библиографических ссылок работы с целью признания важно-

сти работы других авторов для обсуждаемой темы в том месте, где появляется цитирование. Обычное сочетание цитирования в тексте и библиографической записи составляет то, что принято называть цитированием (библиографические записи таковым не являются).

В научной литературе дословное цитирование текста принято для передачи мысли автора без искажений. При условии оформления границ цитаты и ссылки на источник цитирование не является плагиатом. Авторские права на содержание цитаты принадлежат автору цитаты, поэтому человек, который публикует цитату, не несет ответственности за ее содержание. Излишнее цитирование называется оверквотингом.

Формы цитирования, как правило, соответствуют одной из общепринятых систем цитирования типа гарвардской, APA и других, поскольку их синтаксические условные обозначения хорошо известны и понятны читателю. Поскольку каждая система цитирования имеет преимущества и недостатки в отношении своей информативности (без создания слишком больших разрывов), ее следует выбирать в соответствии с потребностями создаваемой публикации. Редакторы часто указывают, какой системой цитирования лучше пользоваться.

Как правило, библиографии и иные компиляции ссылок в виде списка не считаются цитированием, поскольку они не соответствуют истинному толкованию термина: "…преднамеренное признание автором приоритета чьих-то идей".

Обсудим важный вопрос: "Что означает цитирование?" Современная практическая библиометрия собирает данные о количестве ссылок, обрабатывает их и получает статистические результаты. Утверждается, что полученные результаты "объективны". При этом не учитывается, что процесс "автор *A* цитирует работу автора *B*" является сугубо субъективным.

Автор работы [Cozzens, 1989] рассматривает цитирование как результат взаимодействия двух процессов, лежащих в основе выполнения научной публикации, первый из которых представляет собой систему “вознаграждения”, второй − “риторическую” систему. Цитирование, выполненное в первом процессе, обычно подтверждает, что цитирующая статья имеет “интеллектуальный долг” перед цитируемой статьей. Цитирование, выполненное во втором процессе, означает ссылку на статью, которая объясняет определенный результат, возможно вообще не является результатом цитируемого автора. Такие риторические ссылки являются методом научного общения, а не установления интеллектуальной иерархии. В определенных случаях цитирование может иметь оба значения.

В работе [Cozzens, 1989] утверждается, что большинство ссылок являются риторическими, что подтверждается в публикациях по математике. Так, в базе данных цитирования математических обзоров почти 30 % ссылок даны на книги, а не на исследовательские статьи в журналах. Это свидетельствует о том, что в отличие от системы “вознаграждения”, где прослеживается тенденция ссылаться на журналы, содержащие плодотворные идеи, выбор статьи, на которую нужно сослаться риторически, зависит от многих факторов, например от авторитета цитируемого автора или от взаимоотношений цитирующего и цитируемого авторов. Следует также помнить, что существуют ссылки, предупреждающие о некорректных результатах или рассуждениях.

Почему изучают цитирование? Средством измерения цитирований может быть простой подсчет количества цитирований, полученных автором, публикацией или журналом в течение заданного периода времени, и он может быть эффективным инструментом для сравнения продуктивности и взаимодействия исследований между авторами, организациями или странами [Andres, 2009].

По номиналу цитирования трактуются как положительные признания вклада, сделанного цитируемым автором или работой, поэтому чем больше количество цитирований автора, работы или журнала, тем больше их признание. Акт цитирования подразумевает желание автора(ов) включить данную ссылку в статью и дает сведения о характеристиках и научном влиянии. Факт цитирования исследования оказывает также влияние на связь цитируемой и цитирующей статей, например приводя доводы против результатов цитируемой работы. Таким образом, анализ цитирования часто используется для получения информации относительно воздействия и наиболее часто − качества, связанного с автором, публикацией или страной. Тем не менее качество публикации, например, должно определяться не только по числу ее цитирования, но и по результатам других анализов.

Социология цитирования является сложной темой, которая сегодня находится вне области наших интересов. Тем не менее даже беглое обсуждение показывает, что значение цитирования не является простым и что статистика на основе цитирования не является “объективной”, как утверждают ее сторонники. Хорошо известно крылатое выражение: “Есть три разновидности лжи: ложь, гнусная ложь и статистика”, которое приписывается английскому писателю и государственному деятелю, премьер-министру Великобритании от партии консерваторов (1874−1880) Б. Дизраэли. Однако в его работах и высказываниях такая фраза не обнаружена [Академик, словари, а]. Этот факт красноречиво говорит сам за себя.

6.1. Научное цитирование

Научное цитирование представляет собой процесс, с помощью которого выводы предшествующих ученых используются для обоснования целей, положений или экспериментальных процедур. Как правило, такое цитирование устанавливает общие рамки вли-

яния и умонастроение исследования, особенно в зависимости от того, “частью какой науки оно является”, а также помогает определить, кто из коллег может (или должен) проводить рецензирование. В естественных науках нормой цитирования является ссылка на предшествующее авторитетное научное согласие, с помощью цитирования или без него. Например, статья с цитированием $F = ma$ обычно не включает официального цитирования И. Ньютона, хотя это и подразумевается.

Публикуя результаты своего исследования, автор включает в статью ссылки на работы других авторов, имеющие отношение к его публикации. Эти ссылки выявляют связь между авторами, группами исследователей, темами исследований или странами. Более того, воздействие и значимость, оказываемые авторами, исследованиями или журналами на научное сообщество, могут быть измерены с помощью анализа цитирования.

Необходимо различать цитирование и ссылки. Несмотря на то что эти термины используются как взаимозаменяемые, каждый из них представляет различную точку зрения в цитирующей или цитируемой перспективе. Ссылка делается в цитирующем документе и представляет собой признание другого исследования. Цитирование представляет собой признание, полученное цитируемым документом. В общем случае цитируемые документы являются более старыми по сравнению с цитирующим документом, однако в некоторых случаях ссылки могут делаться на документы, которые публикуются одновременно или которые только должны быть опубликованы. Таким образом, факт источника или объекта цитирования подразумевает связь между двумя исследованиями. Тем не менее, эта связь также вовлекает другие ссылки, данные в цитируемом документе, так как они будут представлять основу исследований, на которой базируется третья публикация. Следовательно, анализ цитирования предполагает исследование этих

сложных связей между публикациями, которые связаны ссылкой или цитированием.

Анализ цитирования (*Citation analysis*) представляет собой набор методов и инструментов для выявления связей между авторами или журналами, а также для изучения исследовательских сетей. Изучение частот цитирования научных работ необходимо для понимания взаимосвязей между ними [Пенькова, Тютюнник, 2001]. Данные о цитировании служат основой для вычисления библиометрических метрик (индексов и импакт-факторов). Тем не менее результаты анализа не рекомендуется использовать в качестве единственного (абсолютного) критерия для суждения о важности публикации.

Указатели цитирования позволяют построить иерархию цитирования, т. е. определить, какие более поздние документы цитируют ранние документы. Первыми указателями цитирования были правовые справочники Shepard (1873 г.). Пионером в области индексирования цитирования признан Ю. Гарфилд (Eugene Garfield). В 1960 г. в Институте научной информации (ISI) под его руководством был создан первый указатель цитирования для статей, опубликованных в научных журналах. Сначала был подготовлен Указатель научного цитирования (SCI), позднее Указатель цитирования в общественных науках (SSCI), затем Указатель цитирования в искусстве и гуманитарных науках (AHCI). Основными поставщиками услуг по предоставлению информации о научном цитировании сегодня являются ”Thomson Scientific” и “Elsevier”. Доступ к указателю цитирования ISI, кроме журнала и компакт-дисков, возможен по Сети через сайт *Web of Science*, который, в свою очередь, входит в группу баз данных *Web of Knowledge* (*WoK*). В издательстве “Elsevier” был разработан инструмент *Scopus,* затем появился сетевой инструмент *Scirus*.

Существующие инструменты для анализа цитирования предлагают различные способы доступа к библиографическим БД и разными способами предоставляют результаты. *CiteSeerX* генерирует результаты цитирования по литературе в области вычислительной техники и информатики; *RePec* обеспечивает аналитический сервис по цитированию для экономики; *Google Scholar* (*GS*) также имеет функцию анализа цитирования. Каждый из этих продуктов предоставляет указатель цитирования между публикациями и механизм, позволяющий установить, какой документ цитирует другой документ. Развитие методики анализа цитирования привело к созданию автоматизированных указателей цитирования с открытым доступом. Первой такой системой стала *CiteSeer* (сегодня называется *CiteSeerX*); с ее помощью в 1997 г. разработан первый автоматизированный указатель цитирования SCI. Вскоре была разработана *Cora* (недавно возрожденная как *Rexa*), затем последовал *GS.* Эти автоматизированные системы не идеально выполняют извлечение цитирования и его группирование. В общем случае происходит не более 8 % ошибок. Для получения достоверного результата требуется проведение тщательного статистического анализа. Здесь уместно отметить, что SCI заявляет о создании автоматизированного указателя цитирования с одинаковым уровнем ошибок как для новых, так и для старых публикаций. Из современных инструментов назовем *SciVal* от “Elsevier”. Это комплекс веб-средств, который осуществляет анализ и дает оценку результатов исследований по всем отраслям науки на основе информации, имеющейся в базах данных этого издательства. С помощью *SciVal* можно оптимизировать вложение средств, определять направления исследовательской работы, оценивать варианты при выборе персонала и партнеров.

Цитирование является важной нормой поведения в научном обществе, а ссылки можно рассматривать как средство научной

коммуникации. Ю. Гарфилд считает, что цитирование − это система наград, разменная монета, которой ученые расплачиваются с коллегами. Отсутствие ссылок на источники, используемые в работе, является одной из форм плагиата.

6.2. Самоцитирование

Самоцитирование (СЦ) является общей практикой, когда авторы цитируют свою собственную предыдущую работу. Оно возникает всякий раз, когда группа соавторов цитирующей статьи цитирует статьи, в авторский коллектив которых входит как минимум один автор из группы соавторов. С данными, относящимися к СЦ, необходимо обращаться с осторожностью, поскольку на надежность этого анализа могут повлиять однофамильцы, и, следовательно, показатель СЦ будет искажен. Также на данные о СЦ могут повлиять варианты написания или неправильного написания фамилии автора, что приводит к нераспознаваемости СЦ.

В литературе отсутствует согласие относительно роли СЦ автора. Некоторые авторы считают, что СЦ является потенциальным средством искусственного раздувания уровней цитирования и, следовательно, собственной репутации в научном сообществе. Другие рассматривают СЦ как естественную часть научного общения и доказывают, что полное отсутствие СЦ в течение длительного периода времени должно рассматриваться как патологическое явление [Glanzel, 2004].

Обоснование роли СЦ в научной литературе дано в работе [Van Raan, 2008], главная идея которой заключается в том, что СЦ имеют различную роль и функцию в соответствии с двумя типами исследовательских групп − высокоэффективными и низкоэффективными.

Далее будем использовать термин “накопленное преимущество” (НП), который относится к “эффекту Матфея” (см. справку).

Применительно к настоящему случаю данная концепция предполагает, что крупные исследовательские группы получают не только больше ссылок, но и дополнительные ссылки, зависящие от НП. Цитирование работ в низкоэффективных группах в большей степени зависит от НП. Иными словами, чем больше количество публикаций в группе, тем больше эти публикации продвигаются и тем меньше будет статей, которые не цитируются. Следовательно, высокоэффективные исследовательские группы не нуждаются во внутреннем стимулировании СЦ, тогда как авторы в низкоэффективных группах вынуждены поощрять СЦ. Таким образом, тип исследовательской группы влияет на механизм внутренней стимуляции СЦ.

Справка. Термин “эффект Матфея”, или эффект “накопленного преимущества”, применяется для обозначения ситуации, в которой “богатый становится богаче, а бедный − беднее”. Данный термин в 1968 г. ввел социолог Р. Мертон [Мертон, 1968; Merton, 1988] по аналогии с библейским текстом из евангелия от Матфея: “…всякому имеющему дастся и приумножится…” [Матфей, 25:29]. Следует отметить, что подобные высказывания имеются и в других евангелиях: “…кто имеет, тому дано будет…” [Марк, 4:25], “…всякому имеющему дано будет…” [Лука, 19:26].

Данный феномен проявляется, во-первых, в том, что известные ученые получают дополнительное признание за счет “накопленного преимущества”, во-вторых, работа, получившая признание, превращается в “прецедентный текст”, воспринимаемый не с точки зрения его содержания, а с точки зрения его конституированного “значения” [Мертон, 1968]. Имеется много бесспорных примеров проявления эффекта Матфея в математике, где какая-либо концепция разрабатывается одним математиком (что хорошо отражено в соответствующих публикациях), но приписывается более поздним работам (может быть, даже значительно более поздним) более известного математика, который работал над ней. Например, понятие сложности А. Колмогорова [Wikipedia].

Несмотря на то что в данном разделе рассматривалось СЦ автора, изложенное выше может быть распространено, например, и

на журналы. В работе [Van Raan, 2008] отмечается сходство в группировании СЦ и утверждается, что доля СЦ меньше для тех журналов, которые имеют больший импакт-фактор.

Анализ уровня СЦ становится специальной темой, представляющей интерес благодаря влиянию, которое могут оказывать СЦ на некоторые библиометрические показатели. Широкое использование СЦ, несомненно, влияет на рейтинг ученых, исследовательских групп, журналов и научных дисциплин.

6.3. Время цитирования

Одной из закономерностей, которую Д. Прайс [Price, 1963] установил в поведении науки, является "устаревание научной литературы", т. е. снижение со временем частоты использования документов. Когда документ перестает цитироваться, он считается "устаревшим". Возможно, большая часть документов никогда не будет цитироваться, в то время как на другие будут делаться ссылки сразу после публикации, до того как они устареют. Также возможно, что некоторые документы будут оставаться нецитированными или редко цитируемыми в годы сразу после их опубликования, но впоследствии станут признанными. Это может произойти вследствие того, что со временем все больше исследователей будут осведомлены о ценности документа и, следовательно, число цитирований будет увеличиваться.

В отношении устаревания литературы в анализе цитирования предполагается учитывать "возраст цитирований". Он вычисляется по году цитирования в сравнении с годом опубликования документа. Распределение возраста цитирований представляет собой простой анализ, который показывает, как ссылки, сделанные на статьи, распределяются по времени. Также значимым по отношению ко времени параметром является "показатель оперативности", так как он основан на ссылках, сделанных в первый год после публикации статьи.

В работе [Glanzel, 2004] выполнено обширное библиометрическое исследование документов, проиндексированных в БД WoS с 1992 по 2001 гг. Сравнивая эволюцию СЦ и сторонних цитирований (сделанных другими авторами) во времени, авторы установили, что число СЦ быстро растет в годы, следующие за публикацией, но в течение следующих нескольких лет достаточно быстро сокращается. В противоположность этому, сторонние цитирования достигают своего пика позднее, но удерживают самый высокий уровень цитирования в течение более длительного времени, чем в случае СЦ. Следовательно, можно сказать, что СЦ стареют быстрее, чем это происходит со сторонними цитированиями.

6.4. Анализ сетей цитирования

В библиометрии существует два разных подхода к количественной оценке научной деятельности. Первый подход основан на применении эмпирических законов Лотки, Бредфорда, Ципфа и других корифеев. С помощью этих законов можно, например, оценить частоту публикаций авторов и определить число основных журналов в данной области знаний. Второй подход базируется на изучении механизма цитирования, что позволяет, например, устанавливать связи между авторами или их работами. Напомним два важных определения.

Когда автор K_1 цитирует автора K_2, то между K_1 и K_2 устанавливается связь. Можно выявить много различных связей, таких как связи между авторами, между научными работами, между журналами, между областями исследований или даже между странами. Одним из наиболее общих приемов анализа цитирования является определение влияния одного автора на данную область исследований путем подсчета количества цитирований данного автора другими авторами. Явным недостатком такого подхода является то, что автор K_1 может цитировать автора K_2 в отрица-

тельном контексте (например, говоря о том, что автор K_2 не знает, о чем он пишет, и т. п.). В работе [Moravchik, Murugesan, 1973] рассматриваются вопросы классификации цитирования.

Коцитирование представляет собой метод установления тематического подобия между двумя документами. Если две статьи, *A* и *B,* процитированы в статье *C*, то можно говорить о том, что *A* и *B* связаны друг с другом даже в том случае, когда они не цитируют напрямую друг друга. Если обе статьи, *A* и *B*, цитируются многими другими статьями, то они имеют более сильную связь. Чем больше статей, которые на них ссылаются, тем сильнее связь.

Библиографическое сочетание является зеркальным отображением коцитирования. Считается, что две статьи, *A* и *B,* находятся в состоянии библиографического сочетания, если обе цитируют статью *C,* т. е. даже в том случае, когда *A* и *B* не цитируют непосредственно друг друга. Чем больше статей, которые они обе цитируют, тем сильнее их связь.

Следуя результатам работы [Маршакова, 1973], методы исследования, основанные на цитировании, можно разделить на два класса: анализ статистики цитирования публикаций; анализ сетей цитирования. Статистика цитирования позволяет выявлять закономерности развития науки, прогнозировать темпы развития и “прорывы”. Анализ сетей цитирования дает качественную оценку научных публикаций.

Разработаны два метода анализа сетей цитирования. Первый метод – “библиографические сочетания” (bibliographic coupling, коуплинг), предложенный М. Кесслером в 1963 г., – основан на принципе выделения взаимосвязи между двумя публикациями, в силу того что они (публикации) цитируют один и тот же документ, причем интенсивность их взаимосвязи определяется числом

ссылок, общих для обеих публикаций. С этой целью рассматриваются списки литературы сравниваемых статей. Авторы ссылаются на статьи, которые подтверждают аргументацию или предшествуют данной работе. Таким образом, две статьи *A* и *B*, которые ссылаются на статью *C*, по крайней мере, имеют общую историю и (или) рассматривают сходные темы. Чем обширнее список одних и тех же ссылок, тем более вероятно, что обе статьи посвящены одной и той же проблеме. Метод Кесслера определяет прочные связи между публикациями, которые не меняются при появлении новых публикаций, т. е. не зависят от изменений, происходящих после публикации *A* и *B*. Исходя из этого такую связь между публикациями *A* и *B* называют ретроспективной.

Второй метод − коцитирования (co-citation), или перспективной связи, – разработан в 1973 г. одновременно в СССР (И. В. Маршакова) и в США (Small). В основе данного метода положен принцип выделения взаимосвязи между двумя публикациями *A* и *B* на основе частоты цитирования пары (*A, B*) другими статьями. Две статьи *A* и *B*, на которые имеются ссылки в статье *C* (например, *A* и *B* входят в список литературы статьи *C*), считаются связанными между собой. Частота появления пары (*A, B*) в различных списках литературы характеризует степень взаимосвязи между статьями *A* и *B*. Поскольку с появлением новой пары характеристика взаимосвязи меняется, метод коцитирования назван проспективным, “направленным на будущее”, в отличие от ретроспективного метода Кесслера.

Начиная с 1981 г. метод коцитирования используется в ISI при построении кластеров публикаций, отражающих активные исследовательские направления в различных областях знаний. Оба метода − коуплинг и коцитирование − используются инструментами *Citebase Search* и *Citeseer* для обеспечения навигации по спискам

литературы и цитируемым статьям. Библиографический коуплинг используется в БД WoS для составления отчета “Related Papers”.

Справка. Кластерный анализ (*Data clustering*) − задача разбиения заданной выборки объектов на подмножества, называемые кластерами, таким образом, чтобы каждый кластер состоял из схожих объектов, а объекты разных кластеров существенно различались. Задача кластеризации относится к статистической обработке, а также к широкому классу задач обучения без учителя. Кластерный анализ − это многомерная статистическая процедура, выполняющая сбор данных, содержащих информацию о выборке объектов, и затем упорядочивающая объекты в сравнительно однородные группы. Кластер − группа элементов, характеризуемых общим свойством, главная цель кластерного анализа − нахождение групп схожих объектов в выборке.

Формальная постановка задачи кластеризации выглядит следующим образом. Пусть X − множество объектов, Y − множество номеров (имен, меток) кластеров. Задана функция расстояния между объектами $p(x, x')$. Имеется конечная обучающая выборка объектов $X^m = \{x_1, \ldots, x_m\}$ из X. Требуется разбить выборку на непересекающиеся подмножества, называемые кластерами, так, чтобы каждый кластер состоял из объектов, близких по метрике p, а объекты разных кластеров существенно различались. При этом каждому объекту x_i из множества X^m приписывается номер кластера y_i.

Алгоритм кластеризации − это функция $f : X \to Y$, которая любому объекту x из X ставит в соответствие номер кластера y из Y. Множество Y в некоторых случаях известно заранее, однако чаще ставится задача определить оптимальное число кластеров с точки зрения того или иного критерия качества кластеризации. Решение задачи кластеризации принципиально неоднозначно вследствие ряда причин. Во-первых, не существует однозначно наилучшего критерия качества кластеризации. Известен целый ряд эвристических критериев, а также ряд алгоритмов, не имеющих четко выраженного критерия, но осуществляющих достаточно разумную кластеризацию “по построению”. Все они могут давать разные результаты. Во-вторых, число кластеров, как правило, заранее не известно и устанавливается в соответствии с некоторым субъективным критерием. В-третьих, результат кластеризации существенно зависит от метрики, выбор которой, как правило, также субъективен и определяется экспертом.

6.5. Индекс цитирования

Цитирование отражает связь научных идей. Ссылки, которые авторы статей делают в своих работах, явным образом соединяют их текущие исследования с предыдущими работами в архивах научной литературы. Зная параметры цитирования, для небольшого массива статей легко посчитать основные статистические характеристики: общее количество статей, общее количество цитирований, среднее количество цитирований на одну статью, среднее количество цитирований на одного автора, среднее количество статей на одного автора, среднее количество цитирований в год и т. п. Для вычисления "тонких" статистических параметров необходимы БД с представительными информационными массивами.

Реферативные БД научных публикаций, такие как "Thomson Reuters", индексируют ссылки, перечисляя как цитируемые, так и цитирующие работы. Это позволяет двигаться назад по времени к ранее опубликованным работам и вперед, определяя, кто впоследствии процитировал более ранний исследовательский материал. Необходимая информация о цитируемых источниках извлекается, например, из списков пристатейной литературы этих публикаций. Затем по известным либо собственным алгоритмам вычисляется индекс цитирования (ИЦ) научных статей. Исходными данными для вычисления ИЦ является массив метаданных, которым в данное время располагает БД, и дата (год), когда проводится вычисление. Таким образом, ИЦ зависит от массива метаданных, алгоритма и даты вычислений. Корректная ссылка на значение ИЦ должна содержать эту информацию.

Прозрачность методики вычисления ИЦ и возможность проверки результата являются залогом доверия к официальным результатам, генерируемым роботом БД. Необходимо (сильная постановка) иметь возможность самостоятельно сосчитать свой параметр ИЦ. Для этого нужно иметь доступ к метаданным статей,

которые ссылаются на твои работы (свои ссылки имеются у автора), и знать алгоритм. Этого достаточно для самостоятельных вычислений. Полученный результат, во-первых, нужно сравнить с официальным результатом и понять, почему они не совпадают. Во-вторых, выполненное упражнение (вычислений собственного ИЦ), быть может, развеет широко распространенное мнение о том, что ИЦ указывает на значимость работы. На самом деле ИЦ является результатом арифметических действий с числом ссылок, разнесенных по годам, и не более того. Тем не менее поиск критериев, "наилучшим" образом отражающих результаты авторской научной деятельности (скорее, производительности), продолжается, и заявления: "А в попугаях я гораздо длиннее!" [Сериал "38 попугаев", 1977] имеют место.

6.6. Историческая справка

В русском языке слово "цитата" употребляется с 1820-х гг. В словарях слово "цитата" отмечается с 1861 г. Внимание к цитированию повысилось во второй половине XX в. в связи с появлением понятия "интертекстуальность" [Кристева, 1967] − характеристики "… общего свойства текстов, выражающегося в наличии между ними связей, благодаря которым тексты (или их части) могут многими разнообразными способами явно или неявно ссылаться друг на друга".

Первая работа, в которой рассматривается анализ цитирования, опубликована в 1927 г. [Gross, Gross, 1927]. Авторы изучали ссылки, найденные в течение одного года в выпусках журнала "Journal of American Chemical Society" ("Журнал Американского химического общества").

Пионером индексирования цитирования считают Ю. Гарфилда (Eugene Garfield). Его отчет (1955 г.) рассматривается как инновационная работа, в которой представлены информационные инструменты, ускоряющие процесс исследований и позволяющие оце-

нивать воздействие своей работы, определять научные тенденции и отслеживать историю современной научной мысли [Yancey, 2005].

Применение подсчета цитирований для ранжирования журналов служило методикой, которая использовалась в начале XIX в., однако систематические, непрерывные измерения и вычисления этих подсчетов для научных журналов начал Ю. Гарфилд в Институте научной информации. Он также впервые использовал эти подсчеты для ранжирования авторов и работ. В 1965 г. он вместе с И. Шером продемонстрировал, что количество работ, опубликованных лауреатами Нобелевской премии, в пять раз больше среднего числа работ, и эти работы цитировались в среднем 30−50 раз.

Ю. Гарфилд организовал междисциплинарную научную базу данных и создал для нее указатель цитирования. До этого исследования цитирования были ограничены количеством журналов, областью науки и доступными периодами. Эти факторы, наряду с трудоемкой работой по составлению и обработке вручную огромного количества данных, часто приводили к потере важных данных. Кроме того, до выхода работы Ю. Гарфилда научные показатели были дисциплинарно-ориентированными, поэтому исследователи не могли найти всю информацию, относящуюся к их работе. Анализ научной области исключительно по ее тематике или ключевым словам ограничивает результаты вследствие игнорирования существенных статей из других дисциплин. Исследователи в одной области часто публикуют результаты, которые являются значимыми в другой области. Многодисциплинарный указатель цитирования позволяет строить взаимные связи между исследованиями в различных научных дисциплинах.

В 1963 г. в ISI были созданы три указателя цитирования: SCI, SSCI и AHCI. Первый указатель Гарфилда и Шера (1963 г.) получивший название “Указатель научного цитирования” (*Science Citation Index, SCI*), представлял собой печатное издание из пяти

томов, индексирующих 613 журналов и 1,4 млн цитирований (рис. 6.1). Поскольку в таком формате поиск данных является достаточно трудоемким, два года спустя указатель стал доступен на магнитной ленте. Позднее он был выпущен на компакт-диске вместе с другими двумя указателями цитирования: “Общественные науки” (*Social Science, SSCI*) и “Искусство и гуманитарные науки” (*Arts & Humanities*). После адаптации к сети Интернет был создан сайт *Web of Science* (*WoS*) [Yancey, 2005]. В настоящее время *SCI*, *SSCI и A&HCI* принадлежат “Сети знаний” (*Web of Knowledge*) компании “Thomson Reuters” и являются основными базами данных, используемыми в библиографических поисках. Такая интеграция существенно расширила область доступа и облегчила поиск публикаций и цитирований.

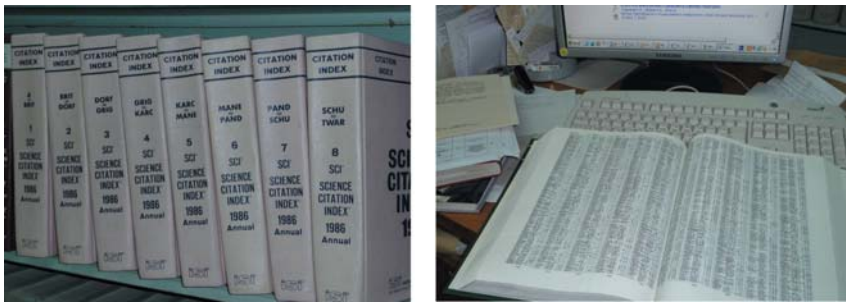


Рис. 6.1. SCI в ГПНТБ СО РАН

В 1965 г. Д. Прайс описал связи между цитированием и цитированными работами. Он назвал их сетями научных работ. Указатель цитирования в общественных науках (1972 г.) стал одной из первых баз данных, доступных в диалоговом режиме. С появлением издания данных о цитировании на компакт-дисках выявление связей существенно упростилось и появилась возможность поиска связанных материалов. Первый автоматизированный указатель цитирования выполнен с помощью *CiteSeer* в 1997 г.

В 1973 г. опубликованы работы по исследованию коцитирования, которое превратилось в самоорганизующуюся систему классификации, позволившую провести эксперименты по группированию документов и в конечном счете выпустить Атлас наук. В это же время технология учета и обработки ссылок достигла уровня, достаточного для появления сервиса оповещения пользователей о том, что вышла в свет новая работа, которая цитирует интересующего автора, статью или книгу.

Библиометрия получила существенное развитие после появления Указателя научного цитирования, который сейчас охватывает литературные источники с 1990 г. Введение в 1998 г. автономных указателей цитирования позволило в автоматическом режиме производить алгоритмическое выделение и группирование цитирования для любых цифровых научных и исследовательских документов. Если ранее процесс выделения проводился вручную, то теперь стало возможным вычисление мер цитирования в большем масштабе для любых областей исследования и различных научных направлений. Появилась возможность выбирать в качестве исходных документов различные коллекции метаданных. Это обеспечило независимость результатов вычисления от результатов, полученных для организаций, отобранных ISI.

В заключение следует отметить, что арифметические манипуляции с числом цитирований в настоящее время широко применяются для обоснованных и необоснованных целей. Необходимо предупредить читателя, что использование только одной этой меры для ранжирования авторов и статей представляется весьма спорным.

Глава 7. Библиографические базы данных и инструменты

По определению [Гражданский кодекс РФ] базой данных (БД) называют совокупность самостоятельных материалов, систематизированных таким образом, чтобы эти материалы могли быть найдены и обработаны с помощью ЭВМ. Существуют другие определения, отражающие, скорее, субъективное мнение тех или иных авторов о том, что означает этот термин в их понимании, однако общепризнанная единая формулировка отсутствует. БД имеет три отличительных признака: 1) БД хранится и обрабатывается в вычислительной системе; 2) данные в БД логически структурированы (индексированы) с целью обеспечения их эффективного поиска и обработки; 3) БД содержит метаданные, описывающие логическую структуру БД в формальном виде [Когаловский, 2002]. Модель данных представляет собой формальное определение объектов, операторов и прочих элементов, в совокупности составляющих абстрактную машину доступа к данным, с которой взаимодействует пользователь. Выделим два основных блока этой машины. Первый поддерживает логическую структуру БД, второй – механизм манипуляции с данными (например, способы помещения и извлечения данных в (из) БД). Перечень современных научных БД и поисковых машин (*academic databases and search engines*) представлен на сайте [Wikipedia].

7.1. Метаданные

Для описания логической структуры БД используются метаданные, содержащие информацию об используемых данных. Метаданные обеспечивают единообразное понимание семантики данных их владельцами и пользователями. Наличие метаданных упрощает выборку, использование информационного ресурса и управление им. Простым примером метаданных могут служить генеральные библиотечные каталоги, которые регистриру-

ют, например, автора, название, предмет и место ресурса на полке.

7.1.1. Дублинское ядро. Хорошим примером определения метаданных является “Дублинское ядро” (*Dublin Core,* DC) − простой и эффективный набор элементов для описания сетевых ресурсов. Семантика “Дублинского ядра” была создана в 1995 г. международной инициативной группой (*Dublin Core Metadata Initiative,* DCMI) профессионалов библиотечного дела, компьютерных наук, кодирования текстов и музейного дела. Для DC характерно использование ограниченного словаря и простой декларативной структуры высказываний в настоящем времени, без сложноподчиненных предложений.

Стандарт DC разделен на два уровня: простой (состоящий из 15 элементов) и компетентный (состоящий из 18 элементов). Простой набор элементов метаданных DC включает: 1) Title − название; 2) Creator − создатель; 3) Subject − тема; 4) Description − описание; 5) Publisher − издатель; 6) Contributor − внесший вклад; 7) Date − дата; 8) Type − тип; 9) Format − формат документа; 10) Identifier − идентификатор; 11) Source − источник; 12) Language − язык; 13) Relation − отношения; 14) Coverage − покрытие; 15) Rights − авторские права. Компетентный набор помимо 15 перечисленных выше элементов, включает 16) Audience − аудитория (зрители); 17) Provenance − происхождение; 18) RightsHolder − правообладатель. Для изучения деталей DC рекомендуем работы [Бешенов, 2007; Манцивода, 2005], в которых представлены полные спецификации этого де-факто стандарта.

7.1.2. MARC (*Machine-Readable Cataloging*) является одним из форматов представления данных в машиночитаемой форме [Wikipedia]. Формат MARC I был разработан Библиотекой Конгресса (БК) США в 1965-1966 гг. с целью получения данных в

машиночитаемой форме. Аналогичная работа выполнялась в Великобритании при подготовке печатного издания Британской национальной библиографии (проект BNB MARC). На основе этих разработок в 1968 г. начал создаваться коммуникативный англо-американский формат MARC (проект MARC II). В 1970-х гг. появились различные версии этого формата, ориентированные на национальные правила каталогизации. Одной из них является RUSMARC − официальная российская версия UNIMARC. В 1999 г. в результате согласования и последующего слияния библиографических форматов США и Канады (USMARC и CANMARC) объявлено об образовании на их основе нового формата ("Формата XXI века") − MARC-21. С этого времени организации, ориентировавшиеся на формат USMARC, должны перейти на формат MARC-21 и отслеживать все его последующие изменения включая новые дополнения к нему (ранее подобные требования отсутствовали). MARC-21 включает форматы: библиографических данных; авторитетных данных; данных о фондах; классификационных данных; общественной информации.

Формальные описания метаданных, разработанные и поддерживаемые организациями стандартизации (ISO, ANSI, W3C и др.) или организациями, взявшими на себя такую ответственность (например, группа DCMI "Дублинского ядра"), называют стандартами метаданных. Современные стандарты используют обобщенный язык разметки SGML или XML. Для библиотечного дела разработаны следующие стандарты представления метаданных: MARC – серия стандартов для представления и передачи библиографической и связанной с ней информации в читаемой машинной форме; METS – стандарт кодирования и передачи метаданных; MODS – схема объектного описания метаданных; XOBIS − XML-схема для моделирования данных MARC.

7.2. Примеры БД

Существует три основных типа библиотечных БД: библиографические БД (*Bibliographical database*), содержащие библиографические записи; реферативные БД (*Referral database*) − библиографические БД, которые дополнительно содержат аннотации, рефераты или иную информацию о документе; фактографические БД, или базы первичных данных (*Source database*), содержащие информацию, относящуюся непосредственно к предметной области.

До середины XX в. индивидуальные поиски литературы полагались на опубликованные библиографические указатели. В начале 60-х гг. применение компьютеров для работы с текстом привело к появлению нового типа информационных ресурсов, известного ныне как библиографические базы данных. Поиск информации в реальном времени стал коммерчески выгодным в начале 70-х гг. Первые службы предлагали небольшие базы указателей и аннотаций научной литературы. Эти базы данных содержали библиографические описания журнальных статей, которые можно было найти по ключевым словам в списке авторов и названии статьи, а иногда по названию журнала или тематическому заголовку. Пользовательские интерфейсы были непродуманными, стоимость доступа была высокой, а сам поиск проводился библиотекарями по поручению заказчиков.

Библиографическая БД является цифровой коллекцией ссылок на опубликованную литературу включая журнальные статьи, труды конференций, патенты, книги и т. п. В отличие от записей библиотечного каталога большая часть библиографических записей описывает статьи, доклады конференций и т. п. в форме ключевых слов или тематических классификационных терминов. Библиографическая БД может быть общей по содержанию или охватывать специальную научную дисциплину. Значительное число библио-

графических БД до сих пор являются частными и доступными только на основании лицензионного соглашения. Многие библиографические БД преобразуются в цифровые библиотеки, которые предоставляют полный текст индексированного содержания. Другие БД объединяются с небиблиографическими научными БД для создания более полных дисциплинарных систем поиска, например Chemical Abstracts.

Индекс цитирования научных статей − реферативная БД публикаций, индексирующая ссылки, указанные в пристатейных списках этих публикаций и предоставляющая количественные показатели этих ссылок. Утверждается [Wikipedia], что первый индекс цитирования разработан в 1873 г. (Shepard's Citations) для нужд юриспруденции.

7.2.1. В качестве примера научной БД широкого профиля рассмотрим *Web of Science* (WoS, *http://scientific.thomson.com*), которая разработана Ю. Гарфилдом в 1960 г. в Институте научной информации. В настоящее время *Web of Science* является одной из БД коллекции *Web of Knowledge* **(**WoK, *http://wokinfo.com*), которой владеет "Thomson Reuters" (USA).

Справка. По состоянию на апрель 2011 г. сайт WoK обеспечивает доступ к следующим БД [Wikipedia]: а) Расширенный указатель научных ссылок. Охватывает более 7100 известных журналов по 150 дисциплинам и предоставляет ссылки с 1900 г. по настоящее время. б) Указатель ссылок по общественным наукам. Охватывает более 2470 журналов по 50 общественным дисциплинам, а также 3500 известных научно-технических журналов. Диапазон охвата по времени: с 1956 г. по сегодняшний день. в) Указатель цитирования по искусству и гуманитарным наукам. Охватывает 1395 искусствоведческих и гуманитарных журналов в дополнение к определенным позициям из более чем 6000 научных и общественно-политических журналов. г) Указатель цитирования трудов конференций. Охватывает более 11000 журналов и трудов в виде монографий в двух редакциях: естественные науки и общественные и гуманитарные науки по 256 дисциплинам. д) Указатель Chemicus. Включает более 2,6 млн

составляющих. Диапазон охвата − 1993 г. по настоящее время.
е) Указатель текущих химических реакций (Current Chemical Reactions). Содержит ссылки более чем на один миллион реакций и охватывает период с 1986 г. по настоящее время. Ежегодно WoK регистрирует 65 млн цитирований, относящихся к ее содержанию.

7.2.2. *Science Direct* (*http://www.sciencedirect.com*) – полнотекстовая БД, содержащая статьи из 2500 журналов начиная с 1823 г. и 11 000 тыс. книг, доступных в режиме онлайн.

7.2.3. *ВИНИТИ РАН* (*http://www*2.*viniti.ru*) − крупнейшая в России БД по естественным, точным и техническим наукам [ВИНИТИ, 2011]. Включает материалы реферативного журнала ВИНИТИ с 1981 г. Общий объем БД − более 30 млн документов. БД формируется по материалам периодических изданий, книг, фирменных изданий, материалов конференций, патентов, нормативных документов, депонированных научных работ, 30 % которых составляют российские источники. БД пополняется ежемесячно. Документы БД содержат библиографию, ключевые слова, рубрики и реферат первоисточника в основном на русском языке. БД включает 28 тематических фрагментов, состоящих из 217 разделов. Для проведения поиска по нескольким тематическим фрагментам, а также с целью обеспечения навигации по БД генерируется единая политематическая БД ВИНИТИ.

7.2.4. *Российский индекс научного цитирования* (*РИНЦ*, *http://elibrary.ru/project_risc.asp*) − это национальная информационно-аналитическая система, предназначенная для оперативного обеспечения научных исследований актуальной справочно-библиографической информацией. РИНЦ также выступает в роли инструмента количественной оценки деятельности научно-исследовательских организаций, ученых и уровня научных журналов. РИНЦ создается в научной электронной библиотеке (НЭБ, eLIBRARY.ru) с 2005 г. Цель проекта заключается в создании отечественной библиографической базы данных по научной перио-

дике. Возможная модель подобных исследований представлена в работе [Редькина, 2006]. Вопросы “Зачем создавать РИНЦ? и как им пользоваться?” рассматриваются в работе [Писляков, 2007а]. Ближайшая перспектива развития РИНЦ представлена в работе [Еременко, 2010]. Критика проекта содержится в работах [Гельфанд, 2010; Каленов, 2011].

7.2.5. Американское математическое общество (AMS) генерирует БД *MathSciNet* (*http://www.ams.org/mathscinet/search.html*), которая содержит краткие обзоры (и иногда оценки) статей по математике, статистике, теории вычислительных машин и систем. Главной задачей этой БД ставилась задача давать обзоры каждой (!) публикации по математике. В основе БД *MathSciNet* лежат материалы реферативного журнала “Mathematical Reviews” (MR), основанного в 1940 г. в качестве альтернативы немецкому журналу “Zentralblatt für Mathematik”. В 1980 г. все содержание журнала MR с 1940 г. стало частью БД MathSciNet, которая имеет информацию о цитировании. Эта БД тщательно следит за правильностью идентификации авторов. Система поиска по автору дает возможность найти публикации, связанные с данной авторской записью, даже если несколько авторов имеют такую же фамилию. В некоторых случаях персонал MR обращается к авторам, для того чтобы убедиться, что БД корректно ссылается на их статьи. С другой стороны, общее меню поиска использует совпадение строк во всех полях включая поле автора. В настоящее время MathSciNet содержит информацию почти о 2 млн статей из 1900 математических журналов [Wikipedia], большая часть которой извлечена от корки до корки.

7.2.6. Математический портал *Math-Net.Ru* (*http://www.mathnet.ru*) предоставляет российским и зарубежным математикам услуги по поиску информации о математической жизни в России. По данным на март 2011 г. портал содержал 72 журнала,

90500 публикаций, 3032 доклада и лекции, 627 видеозаписей, а также информацию о 168 конференциях и 50 семинарах. Портал создан в 2001 г., поддерживается Математическим институтом им. В. А. Стеклова РАН совместно с Отделением математических наук РАН.

7.2.7. Современной фактографической БД является *arXiv.org* (*http://arxiv.org*) − бесплатный архив электронных публикаций научных статей по физике, математике, астрономии, информатике и биологии. Архив создан в 1991 г. в Лос-Аламосской национальной лабораторией (*Los Alamos National Laboratory*, *LANL, USA*) и первоначально предназначался для статей по физике, однако постепенно возникли разделы, посвященные другим наукам. К середине 2008 г. в нем содержалось более 485 000 публикаций, и каждый месяц добавлялось 3000−4000 статей. Существует несколько десятков зеркал архива, в том числе в России (*http://ru.arxiv.org*). В настоящее время архив спонсируется и обслуживается Корнелльским университетом США [Wikipedia]. При добавлении в архив публикация автоматически попадает в базу цитирования *Citebase*, что позволяет вычислить индекс цитирования на материалах архива. *Citebase* разработана в 2005 г. T. Brody [Brody, 2006].

7.2.8. БД *RePEc* (сокращение от Research Papers in Economics, *http://repec.org*) представляет собой результат работы волонтеров из 73 стран мира по содействию распространению исследований по экономике [Wikipedia]. Главная особенность проекта состоит в децентрализации БД рабочих документов, препринтов, журнальных статей и компонентов программного обеспечения. Проект запущен в 1997 г. на платформе IDEAS. По данным на февраль 2011 г., БД RePEc содержит 390 000 рабочих статей, 615 000 журнальных статей, 2100 программных компонентов, 22 000 книг и глав. Большинство из них можно загружать бесплатно, авторские права остаются за автором или держателем авторских прав. Ве-

дущие издатели “Elsevier” и “Springer” приводят перечень своих материалов по экономике в RePEc. Информация из этой БД используется для ежегодного определения ранга почти 20 000 зарегистрированных экономистов.

7.2.9. БД *JSTOR* (сокращение от слов “Хранилище журналов”, *http://www.jstor.org*) представляет собой сетевую систему архивирования академических журналов, основанную в 1995 г. в США. JSTOR развивается под лозунгом: “…в помощь академическому сообществу для получения всех преимуществ от быстро развивающихся информационных и сетевых технологий”. В сентябре 2009 г. в БД JSTOR хранилось 1 079 наименований журналов, представляющих 51 научную дисциплину, что в сумме составило более 33,7 млн страниц текста [Wikipedia]. Доступность почти всех журналов в JSTOR контролируется “подвижной перегородкой”, которая представляет собой согласованную задержку между текущим томом журнала и последним томом, доступным в JSTOR. Этот срок устанавливается соглашением между JSTOR и издателем и обычно составляет 3−5 лет.

7.2.10. *ProQuest* (*http://www.proquest.co.uk/en-UK/*) − полнотекстовая мультидисциплинарная БД (содержит более 200 БД). Доступна с 1985 г., в настоящее время с ней сотрудничают более 700 ведущих высших учебных заведений мира.

7.2.11. *ACM*, ассоциация по вычислительной технике, основана в 1947 г. как научное и образовательное общество. ACM осуществляет свою деятельность через 170 местных отделений, порядка 500 отделений в колледжах и университетах, а также 35 “специальных групп по интересам” (SIG). В 2009 г. в ACM насчитывалось более 92 000 членов. Издательство ACM выпускает престижные академические журналы “Journal of the ACM”, “Communications of the ACM” и серию журналов “ACM Transactions”. Многие SIG (например, SIGCOMM) организуют регулярные кон-

ференции и получили мировое научное признание. Группы также публикуют труды конференций, специализированные журналы и бюллетени. Полный архив продукции издательства хранится в цифровой библиотеке, которая доступна подписчикам через ACM-портал (*http://portal.acm.org*).

Исторически основным конкурентом ACM было и остается "Общество вычислительной техники IEEE", которое является наиболее крупной подгруппой "Института инженеров по электротехнике и электронике". IEEE бóльшее внимание уделяет аппаратному обеспечению и вопросам стандартизации, чем теоретической вычислительной технике, но отмечается существенное перекрытие с программой ACM.

7.2.12. *IEEE*, или "Институт инженеров по электротехнике и электронике", представляет собой международную некоммерческую организацию по развитию электронных технологий. IEEE состоит из 39 обществ, организованных по специальным техническим областям, с более чем 300 региональными организациями, которые созывают регулярные совещания. В IEEE входит более 365 000 членов приблизительно из 150 стран. Устав IEEE определяет цели организации как "…научные и образовательные, направленные на развитие теории и практики электротехники, электроники, связи и вычислительной техники, а также науки о компьютерах…". Для достижения этих целей IEEE издает научные журналы (более 100 наименований), организует международные конференции. Он является также ведущим разработчиком промышленных стандартов в широком спектре дисциплин. Считается, что IEEE выпускает 30 % мировой литературы в области электронной и вычислительной техники, издавая более 100 рецензируемых журналов с высокими импакт-факторами. Авторы могут по своему выбору выкладывать собственные статьи в открытый доступ. Приблизительно 1/3 авторов IEEE пользуются

этой возможностью. Содержание журналов, наряду с докладами нескольких сотен ежегодных конференций, представлено в сетевой БД *IEEE Xplore* (*http://ieeexplore.ieee.org*), которая в настоящее время содержит свыше двух миллионов записей [Wikipedia].

7.2.13. Технический указатель *Compendex* (*http://www.engineeringvillage2.org*) представляет собой библиографическую БД. Compendex является указателем технических материалов с 1884 г., составленным вручную под исходным названием *Engineering Index*, в качестве *Compendex* он был опубликован издательством "Elsevier". В настоящее время Compendex содержит более девяти миллионов записей и ссылок более чем из 5000 международных источников включая журналы, конференции и отраслевые издания. Ежегодно приблизительно 500 000 новых записей более чем из 190 дисциплин и основных специальностей в инженерной области добавляется к базе данных. Указатель охватывает данные с 1969 г. и обновляется еженедельно.

7.2.14. Сетевая система анализа и поиска медицинской литературы MEDLINE (*http://www.ncbi.nlm.nih.gov/pubmed*) представляет собой БД с информацией по наукам о жизни и биомедицине. MEDLINE включает библиографические данные статей из академических журналов по медицине, уходу за больными, фармацевтике, стоматологии и здравоохранению, а также литературу по биологии и биохимии. В указатель MEDLINE включено около 5 000 журналов этого профиля, свободный доступ к которым возможен через БД PubMed [Wikipedia]. Новые журналы вводятся в состав БД на основе рекомендаций специалистов из Комитета технического анализа литературы с учетом научного охвата и качества журнала.

7.2.15. Существует несколько крупных информационных систем, обеспечивающих доступ к реферативно-библиографическим

и полнотекстовым БД: *STN International* (*http://www.stn-international.de*), *Lexis-Nexis* (*http://www.lexisnexis.com*), *Dialog* (*http://www.dialog.com*). По сути, это системы-интеграторы, которые объединяют большое количество различных БД в рамках единого поискового интерфейса и языка информационных записей. Например, STN International существует с середины 80-х гг., объединяя более 200 БД по всем отраслям науки и технологий. Ряд БД являются ретроспективными, представляют копии публикаций XVII-XIX вв. (!). БД различного профиля для системы STN создаются более ста научно-исследовательскими и информационно-аналитическими учреждениями, организациями и издательствами. В БД STN International представлены публикации приблизительно 50 000 периодических изданий, а также патентные и научно-технические архивы ряда крупнейших национальных и международных организаций. Доступ к этой БД ранее осуществлялся по сетям передачи данных, работавшим на основе исключительно надежного протокола X.25, а ныне по сети Интернет**.**

7.3. Инструменты

7.3.1. *Scopus* (*http://www.scopus.com*) является мультидисциплинарной БД рефератов и цитирований издательства “Elsevier”. В ней проиндексировано более 18 000 наименований научно-технических и медицинских журналов включая 1200 журналов Open Access и около 300 российских журналов, а также 3,7 млн докладов из трудов конференций. Охват − 16500 рецензируемых журналов в области науки, техники, медицины и общественных наук (в том числе искусство и гуманитарные науки). С помощью Scopus Author Preview можно узнать некоторые сведения об авторе, например место его работы. Scopus TopCited дает общее представление о 20 наиболее цитируемых статьях по каждой тематике за последние три, четыре и пять лет. Эти данные доступны через программный интерфейс API, который входит в Scopus.

7.3.2. *Scirus* (*http://www.scirus.com*) является современной поисковой машиной, разработанной в “Elsevier” и специализирующейся на поиске научной информации. В настоящее время поле поиска состоит из 410 млн интернет-страниц по научной тематике с сайтов .edu, .org, ac.uk, .com, .gov и сайтов ведущих университетов мира, а также порядка 17 млн записей из БД ScienceDirect, MEDLINE on BioMedNet, Beilstein on ChemWeb, BioMed Central, SIAM, US Patent Office, E-Print ArXiv, Chemistry Preprint Server, Computer Science Preprint Server, Mathematics Preprint Server, CogPrints и NASA. Перечень доступных БД можно найти на странице *http://scirus.com/srsapp/aboutus*.

7.3.3. *HistCite* (*http://thomsonreuters.com*) − это программное обеспечение (разработчик Ю. Гарфилд), используемое для библиометрического анализа и визуализации его результатов. Ниже приведены некоторые типовые вопросы, на которые можно получить ответ с помощью HistCite. Как много литературы опубликовано в этой области? Когда и какие страны ее публиковали? Какие страны внесли наибольший вклад в эту область? Какие языки чаще всего используются для материалов, публикуемых в этой области? Какие журналы содержат статьи по данной тематике? Какие из них наиболее важные? Кто, вероятно, является ключевыми авторами в этой области? Какие организации представляют эти авторы? Какие статьи самые важные? Как различные исследователи в этой области влияют друг на друга?

Блок визуализации в HistCite преобразует библиографические данные в диаграммы, называемые историографами, − временные связи, установленные на основе цитирования. С помощью историографов легче увидеть и выделить ключевые публикации по изучаемой теме, их хронологию и взаимное влияние.

7.3.4. *Google Scholar* (далее GS, *http://scholar.google.com*) является бесплатной поисковой машиной, которая индексирует пол-

ные тексты литературы по различным научным дисциплинам. По своим функциям GS аналогичен бесплатным Scirus (“Elsevier”), CiteSeerX и getCITED. С другой стороны, GS выполняет функции платных инструментов Scopus (Elsevier) и WoS (“Thomson Reuters”). Его рекламный девиз − “Стоять на плечах гигантов” − является комплиментом в сторону инженеров и ученых, которые сделали вклад в свои области исследований, обеспечив фундамент для новых интеллектуальных достижений [Wikipedia]. Работы по проектированию и реализации GS в 2006 г. начали A. Verstak и A. Acharya (ранее оба работали над основным поисковиком Google) в ответ на выпуск компанией Microsoft поисковой системы Windows Live Academic. Большинство поисковых инструментов дают возможность ранжирования результатов поиска только по одному параметру, например релевантности, числу цитирований или дате публикации. В отличие от них GS может ранжировать результаты поиска по набору параметров. Поэтому первые места занимают часто цитируемые статьи, что подтверждает “эффект Матфея”.

Серьезной проблемой GS является скрытность области его охвата. GS не публикует перечень используемых научных журналов. Вследствие этого невозможно выяснить, насколько результаты его поиска являются современными и исчерпывающими. GS ошибается при идентификации публикаций на сервере препринтов arXiv. Знаки препинания в названиях приводят к неправильным результатам поиска, авторам приписываются чужие статьи, некоторые результаты выдаются без какой-либо понятной причины. Этого достаточно для заключения: GS следует использовать крайне осторожно, особенно для расчета известных метрик, таких как *h*-индекс Хирша (Hirsch) или *JIF*.

7.3.5. *CiteSeerX* (*http://citeseerx.ist.psu.edu*) представляет собой инструмент поиска и цифровую библиотеку, которая содержит 1,5 млн документов и 30 млн цитирований в области вычисли-

тельной техники и информатики. CiteSeer создан в 1997 г. с целью сбора и накопления научных документов, находящихся в Сети в свободном доступе, и автоматического вычисления независимого индекса цитирования. Поскольку эта поисковая машина имеет доступ только к открытым материалам, то с большой вероятностью их авторы будут выглядеть предпочтительнее других.

7.3.6. *CiteULike* (*http://www.citeulike.org*) представляет собой бесплатный сетевой сервис для организации библиографических ссылок. Работает с октября 2004 г. С его помощью можно организовать хранилище копий статей, обладающее социальными функциями.

7.3.7. 2*collab* (*http://www.*2*collab.com*) представляет собой сетевой инструмент, разработанный "Elsevier" для проведения совместных исследовательских работ. Он позволяет делиться закладками, ссылками и любыми присоединенными материалами с коллегами, создавать локальные научные сообщества.

7.3.8. *GetCITED* (*http://www.getcited.org*) представляет собой сетевую базу данных, в создании которой принимают участие сами авторы. Сегодня она содержит более 3 млн публикаций 300 000 тыс. авторов.

7.3.9. Программа *Publish or Perish* [Harzing, 2010] по параметрам запроса проводит поиск и анализирует научные ссылки. Для добычи данных она использует GS, затем анализирует их и представляет статистические результаты: общее количество статей; общее количество цитирований; среднее количество цитирований на одну статью; среднее количество цитирований на одного автора; среднее количество цитирований в год; *h-index* и соответствующие параметры; *g-index* Эгге (Egghe); современный *h-index*; показатель цитирования, оцененный по возрасту; два варианта индиивидуальных *h-index*; анализ количества авторов на одну статью.

Справка. Добыча данных (*Data Mining*). Это процесс обнаружения в "сырых" данных ранее неизвестных, нетривиальных, полезных на практике и доступных в интерпретации знаний, необходимых для принятия решений в различных сферах человеческой деятельности. "Добыча данных" разделяется на задачи классификации, моделирования, прогнозирования и другие. В свою очередь, методы "добычи данных" разделяются на статистические и кибернетические. К статистическим методам относятся дескриптивный анализ, корреляционный и регрессионный анализ, факторный анализ, дисперсионный анализ, компонентный анализ, анализ временных рядов. Кибернетические методы − это нейронные сети, генетические алгоритмы, нечеткая логика, деревья решений, системы обработки экспертных знаний.

7.3.10. *BIBXCEL* (*http://www*8.*umu.se/inforsk/Bibexcel*) – программа для библиометрического анализа данных, представленных в текстовом формате. Ее главная функция состоит в генерации файла, пригодного для загрузки в Excel.

7.3.11. Программный пакет *Bibliometrics MATLAB* [Cardillo, 2010] разработан для проведения индивидуального библиометрического анализа массива своих публикаций. Как правило, для этой цели используют Google Scholar или программу Publish or Perish, которая осуществляет добычу данных с помощью GS. Однако эти программы требуют дополнительных усилий, что обусловлено необходимостью предварительной обработки исходных данных, например решения проблемы "запятой" в названии статьи или преодоления "урезания" списка авторов инструментарием GS. В силу этого некоторые результаты неточны. Вычисление индивидуальных библиометрических параметров с помощью Bibliometrics MATLAB не представляется сложным при наличии доступа к информации о цитировании своих статей. В результате будут получены индивидуальные, проверяемые результаты с высоким уровнем доверия к ним.

Для проведения вычислений в полном объеме необходимо подготовить три массива данных – C,Y и A. Массив C содержит ин-

формацию о цитировании статей. Эти данные можно получить из нескольких БД, например WoS, Researcher ID, или с помощью инструментов GS, Scopus, CiteSeer. Массив Y – год публикаций статей. Наконец, A – массив соавторов. В результате вычислений будут получены: а) параметры описательной статистики публикационной деятельности ученого; б) библиометрические индексы, состоящие, в свою очередь, из нескольких подгрупп.

Параметры описательной статистики публикационной деятельности ученого − это общее количество статей и среднее за каждый год; общее количество цитирований и распределение по годам; коэффициент Джини (*Gini coefficient*) и кривая Лоренца (*Lorenz curve*); статистические характеристики публикаций, цитирований и соавторов. Статистические характеристики – минимум, максимум, среднее, медиана и мода − вычисляются с 95 %-м доверительным интервалом.

Индексы цитирований: *h-index* Хирша, *g-index* Эгге, *A-index* Джин, *h*(2)-*index* Космулского, *e-index* Жанга и *нормализованный h-index* Сидиропулоса.

Индексы, зависящие от времени: *AWCR-index* (взвешенная по возрасту публикации скорость цитирования), *AR-index* Джина – Харзинга и *hc-index* Сидиропулоса.

Индексы, учитывающие количество авторов: h_I*-index* − индивидуальный индекс Батисты, h_I*, norm-index* − индивидуальный, нормированный индекс Харзинга, h_m*-index* − индекс Шрайбера для коллектива авторов.

Пакет Bibliometrics MATLAB оснащен графикой.

7.3.12. *Pajek* (“Паук”, *http://pajek.imfm.si*) – программа для анализа и визуализации больших сетевых структур. *NetDraw* (*http://www.analytictech.com/netdraw/netdraw.htm*) – программа визуализации сетевых структур, например библиографических или социальных сетей. Совместима с Pajek по формату графических данных.

В заключение приведем сведения о разработках библиометрических программных инструментов [Karpagam, et al., 2011] (табл. 7.1).

Таблица 7.1

Библиометрические программные инструменты

Название	**Автор**	**Назначение**	**Ссылка**
Authormap	Howard White	Построение графа цитирования и его визуализация	[White, soft]
Bibcouple	Loet Leydesdorff	Визуализация библиографического coupling авторов на основе информации базы данных WoS	[Leydesdorff, soft]
Citespace	Chaomei Chen	Визуализация структуры связей и тенденций в научной литературе	[Chen, soft]
CleanPoP	Audrey Baneyx	Инструмент для систематизации результатов Publish or Perish	[Baneyx, sof]
Co-auth	Loet Leydesdorff	Визуализация соавторства на основе информации из *WoS*	[Leydesdorff, soft]
Fulltext	Loet Leydesdorff	Составление карты совместного использования слов в полных документах	[Leydesdorff, soft]

HistCite	Eugene Garfield	Коммерческий продукт для библиографического анализа и визуализации	[Garfield, soft]
IntColl	Loet Leydesdorff	Визуализация международного сотрудничества	[Leydesdorff, soft]
ISI	Loet Leydesdorff	Преобразование данных из *WoS* для последующего анализа	[Leydesdorff, soft]
Patent Pictures		Коммерческий продукт “Новости о патентах”	[Patent Pictures, soft]
Publish or Perish	Anne-Wil Harzing	Извлечение и анализ академических цитирований из Google Scholar	[Harzing, soft]
RefViz	Thomson Reuters Research Software	Анализ и визуализация библиографических ссылок, полученных от EndNote, ProCite и Reference Manager	[TRRS, soft]
TI	Loet Leydesdorff	Составление карты совместного использования слов в текстах	[Leydesdorff, soft]

Глава 8. Результативность научной деятельности

Результатом научных исследований и разработок являются “знания”, реализованные в виде описаний открытий, гипотез, теорий и методов и зафиксированные в научных публикациях включая статьи в научных журналах. Однако знания являются объектами, неудобными для измерения научной деятельности. Поэтому для проведения измерений, как правило, используют описания журнальных статей, представленные в виде метаданных, содержащих библиографическую и библиометрическую информацию.

Библиографией называют деятельность по составлению, распространению и использованию библиографической информации. Эта деятельность регламентируется ГОСТ 7.0-99 “Информационно-библиотечная деятельность, библиография”. Например, библиографическая информация о книгах до недавнего времени содержалась на картонных “карточках”, которые хранились в картотеках. Она содержала следующие сведения: имя автора, название, год и место издания, место хранения. Для автоматизации процесса обработки библиографической информации в Библиотеке Конгресса США в начале 80-х гг. XX в. был разработан стандарт Z39.50 [Жижимов, Мазов, 2004].

Сегодня библиографическое описание статьи, опубликованной в научном журнале, помимо названия и списка авторов содержит информацию о месте работы авторов, их адреса, телефоны, аннотацию, ключевые слова и список литературы. Используя эти данные, можно ответить, например, на следующие вопросы: каково общее число публикаций в указанной научной отрасли за интересуемое время и как выглядит распределение публикаций по отраслям, в скольких статьях автор выступает в качестве первого автора, в каких отраслях науки имеются работы с его участием и т. п.

На цитировании построена большая часть метрик количественной оценки научной деятельности. Информация о цитировании зависит от времени, поэтому в точных формулировках следует указывать: количество цитирований данной статьи за фиксированный промежуток времени. Эта информация наряду с библиографическим описанием статьи позволяет получить ответы на “тонкие” вопросы количественного характера, например какое количество статей в выбранной отрасли науки не имеет цитирования, а также строить оценки и ранжирование научных статей по результатам цитирования. Следует отметить, во-первых, что арифметические манипуляции с числом цитирований в настоящее время широко применяются для обоснованных и необоснованных целей. Во-вторых, поскольку по своей природе механизм цитирования является манипулируемым, то делать какие-либо выводы на основе учета результатов цитирования следует весьма осмотрительно и осторожно. Таким образом, использование только этой меры для ранжирования авторов и статей представляется весьма спорным.

8.1. Показатели результативности научной деятельности

Согласно определению, данному в работе [Гохберг, 2003, с. 99], национальная научная система представляет собой совокупность организаций страны, выполняющих исследования и разработки, и взаимосвязей между ними в процессе производства и распространения (передачи) научных знаний. Основные научные области сложились исторически: естественные, технические, медицинские, сельскохозяйственные, общественные и гуманитарные науки. Л. М. Гохберг предлагает иерархическую трехуровневую классификацию отраслей науки применительно к номенклатурам научных специальностей, “гармонизированную” с международными стандартами [Гохберг, 2003, с. 110, табл. 1.15]: первый уровень – области науки, второй – отрасли науки, третий – научные специальности. Используя табл. 8.1., перечень журналов ВАК и

базу данных WoS, можно построить распределение научных журналов по областям и отраслям науки (например, в терминах “область / число журналов”), в которых публикуется большинство российских ученых. Полученное на основе данных WoS распределение отражает сильные и слабые стороны научной системы РФ в сравнении с другими странами, представленными в этой базе.

Таблица 8.1

Классификация науки по областям и отраслям (пример)

Области науки	**Отрасли науки**
1. Естественные науки	1.1. Математика (01.01.00), Механика (01.02.00) 1.2. Астрономия (01.03.00), Физика (01.04.00) 1.3. Химия (02.00.00) 1.4. Биология (03.00.00) 1.5. Науки о Земле (25.00.00)
2. Технические науки	2.1. Технические науки (05.00.00)
3. Медицинские науки	3.1. Медицинские науки (14.00.00), Фармацевтические науки (15.00.00)
4. Сельскохозяйственные науки	4.1. Сельскохозяйственные науки (06.00.00) Ветеринарные науки (16.00.00)
5. Общественные науки	5.1. Экономические науки (08.00.00) 5.2. Юридические науки (12.00.00) 5.3. Педагогические науки (13.00.00) 5.4. Психологические науки (19.00.00) 5.5. Социологические науки (22.00.00)
6. Гуманитарные науки	6.1. Исторические науки (06.00.00) 6.2. Философские науки (09.00.00) 6.3. Филологические науки (10.00.00) 6.4. Искусствоведение (17.00.00) 6.5. Культурология (24.00.00)

8.2. Индекс научной специализации страны

Для оценки места страны в мировой науке в работе [Гохберг, 2003, с. 229] индекс научной специализации страны предлагается вычислять по формуле

$$ISS_{it} = \left(A_{it} \Big/ \sum_i A_{it} \right) \Big/ \left(WA_{it} \Big/ \sum_i WA_{it} \right),$$

где ISS_{it} − индекс научной специализации страны в области i в году t; A_{it} – число статей в области i, принадлежащих национальным авторам, в научных журналах, реферируемых в базах данных WoS, в году t; WA_{it} – число статей в области i в научных журналах, реферируемых в базах данных WoS, в году t. По расчетам Л. М. Гохберга, на протяжении последнего десятилетия научная специализация России практически не менялась.

Получив определенное признание со стороны аналитиков, библиометрические данные, тем не менее, обладают известными недостатками, обусловливающими необходимость их корректной трактовки и ограниченность получаемых выводов. Такие факторы, как физическая невозможность охвата всех мировых изданий, доминирование в отдельных журналах определенных научных парадигм, часто препятствующих публикации нетрадиционных взглядов, недостаточная репрезентативность прикладных исследований, отсутствие качественной оценки содержания статей, существенно снижают надежность библиометрических данных. Например, особенностью базы данных SCI является непропорциональная представительность биомедицинских наук (45 % представленных журналов), что на самом деле отражает ее первоначальное назначение, а также англоязычных публикаций, хотя в известной мере и оправданная ролью английского языка в качестве международного средства общения ученых. Поэтому неслучайно Европейская Комиссия приняла решение о создании аль-

тернативной базы данных по европейским научным публикациям, в которой были бы представлены издания на немецком, французском, итальянском, испанском и других европейских языках [Гохберг, 2003, с. 231].

8.3. Метрика эффективности публикаций страны

Рассмотрим метрику *PEI* (Publication Efficiency Index) [Guan, Ma, 2004]), с помощью которой можно оценить, как соотносится уровень публикаций страны с усилиями ученых страны в данной области исследований. Выберем интересующую область и будем рассматривать только те журналы, которые относятся к этой области. Основная формула для расчета *PEI* имеет вид

$$PEI = (TNC_i / TNC_t) / (TNP_i / TNP_t),$$

где TNC_i – общее количество ссылок, сделанных на журналы страны i; TNC_t – общее количество ссылок, сделанное на публикации всех стран; TNP_i – общее количество статей, опубликованных страной i; TNP_t – общее количество статей, опубликованное всеми странами.

Если $PEI \geq 1$, то воздействие публикаций в данной стране и конкретной области исследований выше (равно) предпринятых усилий исследователей. Анализ основан на сравнении количества ссылок на статью, опубликованную страной, с данным соотношением для всех стран, включенных в анализ. В противоположность этому при $PEI < 1$ воздействие полученных результатов незначительное (или слабое), несмотря на предпринятые усилия.

8.4. Эффект Матфея для стран

Известен эффект Матфея для авторов публикаций, заключающийся в том, что работы известных ученых имеют потенциальное преимущество в цитировании перед работами их менее известных коллег. Группа немецких ученых во главе с М. Боницем [Bonitz,

et al., 1997] провела исследование этого эффекта для стран (Matthew effect for countries).

В работе [Писляков, Дьяченко, 2009] отмечается: «То, что принято называть “научным уровнем журнала”, его “репутацией”, “влиятельностью”, находит выражение в средней цитируемости его статей. …разброс цитируемости отдельных статей вокруг этого среднего неизбежен, но в том случае, если он коррелирует со страной создания публикации, можно говорить об эффекте Матфея». Это означает, что публикация в престижном журнале, скорее всего, получит больше цитирований, чем среднее значение цитируемости этого журнала.

Интерес представляет не одно издание, а набор изданий, агрегированных по научному направлению. Рассмотрим это на примере. Выберем n авторитетных научных журналов по исследуемой области науки и обозначим это множество $J=\{J_1,J_2,\ldots J_n\}$. Из каждого журнала выберем все статьи, написанные учеными стран C_1 и C_2. Обозначим эти множества $AC_1(J_i)$ и $AC_2(J_i)$ соответственно. Пусть z_i – средняя цитируемость статей журнала J_i. Вычислим “ожидаемое число цитирований”. Для этого умножим число статей $|AC_1(J_i)|$, опубликованных в журнале J_i авторами страны C_1, на z_i и получим значение средней цитируемости страны C_1 в журнале J_i. Теперь просуммируем полученные произведения для C_1 по всем журналам множества J и обозначим полученную сумму через

$$ECR(C_1) = \sum_i z_i \times \left|AC_1(J_i)\right| .$$

Мы получили ожидаемое число цитирований в исследуемой области науки для страны C_1 на множестве журналов J – “…такое число цитирований, которое имели бы все публикации страны, если бы каждая из них получала число цитирований, в точности

равное средней цитируемости статьи журнала, в котором она опубликована" [Писляков, Дьяченко, 2009]. Повторив проделанные вычисления для страны C_2, получим значение $ECR(C_2)$. Реальное число цитирований, полученных публикациями авторов из стран C_1 и C_2, обозначим через $OCR(C_1)$ и $OCR(C_2)$ соответственно.

Индекс Матфея (Matthew index, *MI*) для страны *C* вычисляется по формуле

$$MI = (OCR - ECR) / ECR.$$

В результате, если $MI > 0$, то страна *C* получает больше цитирований, чем можно предположить; если $MI < 0$, то статьи авторов из страны *C* цитируются ниже среднего уровня, случай $MI = 0$ не требует комментариев. Кроме того, можно сравнить значения C_1 и C_2. Интерпретация результатов сравнения не входит в нашу компетенцию.

Заметим, что индекс *MI* отличается от простого сравнения среднего числа цитирований на статью для страны с общемировым уровнем, так как последний может быть высоким вследствие того, что ученые публикуются в престижных журналах, а *MI* характеризует уровень цитируемости публикаций страны на фоне других работ в одних и тех же журналах. В работе [Писляков, Дьяченко, 2009] можно найти результаты исследования эффекта Матфея для российских ученых, печатающихся в зарубежных изданиях.

8.5. Euro-Factor

В 2002 г. был запущен проект "Euro-Factor" для библиометрического анализа европейских журналов в области биомедицины [Euro-Factor, 2002] и создана БД EUROFACTOR. Для расчета ко-

личественной оценки журнала j принята метрика $EF(j)$, которая вычисляется по формуле

$$EF(j) = \frac{A(j)}{EFC \times \sqrt{A(j) + B(j)}}.$$

Здесь $A(j)$ – общее число цитирований всех статей журнала j; $B(j)$ – число всех статей; EFC – коэффициент и $\sqrt{\ }$ – неотрицательное значение квадратного корня от суммы $A(j) + B(j)$.

8.6. Оценка результативности РАН

В Институте системного анализа Российской академии наук разработана и вводится в строй автоматизированная система учета результатов интеллектуальной деятельности (АСУ РИД РАН). Подобные научные информационные системы уже функционируют в Великобритании, Норвегии, Бельгии, Германии, Финляндии, Чехии, где они получили название Current Research Information Systems (CRIS).

Президиум Российской академии наук Постановлением № 201 от 12.10.2010 г. утвердил согласованные с Министерством образования и науки РФ Положение о Комиссии по оценке результативности деятельности научных организаций Российской академии наук и Методику оценки [ПОСТ РАН № 201]. Для этого ученые должны представить данные по восьми направлениям оценки, в частности информацию об актуальности, перспективности и ресурсной обеспеченности исследований, вовлеченности в национальное и мировое научное сообщество, экспертной деятельности, коммерциализации результатов исследований, научном и кадровом потенциале, инфраструктуре и состоянии финансовой деятельности научной организации. Согласно предложенной методике число публикаций и цитируемость работников научной организации отнесены (!) к общему числу исследователей этой организации. Для этих целей будет использована информация БД РИНЦ,

WoS, Scopus, Medline, Metadex, Compendex, Pascal, Biosis и др. Импакт-фактор публикаций работников научной организации будет определяться по WoS.

8.7. Три объекта измерений

Наука является особым видом познавательной деятельности, направленной на получение, уточнение и производство объективных, системно организованных и обоснованных знаний о природе, обществе и мышлении и включающей сбор научных фактов, их постоянное обновление и систематизацию, критический анализ и на этой основе синтез новых научных знаний или обобщений, которые не только описывают наблюдаемые природные или общественные явления, но и позволяют строить причинно-следственные связи и как следствие − прогнозировать. Естественно-научные теории и гипотезы, которые подтверждаются фактами или опытами, формулируются в виде законов природы или общества. Термины “наука” и “ученый” введены [Википедия] У. Уэвеллом (1794−1866 гг.).

8.7.1. Автор. Будем полагать, что автором научной статьи является специалист в какой-либо научной области – ученый. В широком смысле понятие ученый относится к любому человеку, который систематически расширяет знания человечества либо участвует в деятельности и поддержании традиций тех или иных научных и философских школ. В более узком смысле учеными называют только тех людей, которые применяют научный метод − совокупность основных способов получения новых знаний и методов решения задач в рамках любой науки.

Научный метод включает способы исследования феноменов, систематизацию, корректировку новых и полученных ранее знаний. Умозаключения и выводы делаются с помощью правил и принципов рассуждения на основе эмпирических (наблюдаемых и измеряемых) данных об объекте. Базой получения данных явля-

ются наблюдения и эксперименты. Для объяснения наблюдаемых фактов выдвигаются гипотезы и строятся теории, на основе которых формулируются выводы и предположения. Полученные прогнозы проверяются экспериментально или путем сбора новых фактов.

Важной стороной научного метода, его неотъемлемой частью для любой науки является требование объективности, исключающее субъективное толкование результатов. Не должны приниматься на веру какие-либо утверждения, даже если они исходят от авторитетных ученых. Для обеспечения независимой проверки проводится документирование наблюдений, обеспечивается доступность для других ученых всех исходных данных, методик и результатов исследований. Это позволяет не только получить дополнительное подтверждение путем воспроизведения экспериментов, но и критически оценить степень адекватности (валидности) экспериментов и результатов по отношению к проверяемой теории.

8.7.2. Научная публикация. По результатам исследований, выполненных с использованием научного метода, ученый готовит публикации. Для этого он единолично или в составе коллектива авторов пишет научную работу (например, статью), после выхода журнала в свет становясь ее автором (или соавтором).

Выбирая формат научной статьи, многие авторы следуют общей схеме. Такие статьи начинаются с резюме, введения и краткого обзора, включая обсуждение подобных исследований. Далее, как правило, следует теоретическая часть, которая состоит из формальной постановки задачи (модели) и описания метода решения. В экспериментальной части приводятся конкретные детали проведения исследования. Затем следуют разделы “Результаты” и “Применение”. В заключении определяют место исследования в изучаемой области и намечают пути дальнейших изысканий.

Кроме научных существуют и другие типы журнальных статей. Письма (называют также сообщениями) представляют собой крат-

кое описание важных текущих исследований, которые обычно быстро обрабатывают для немедленной публикации, поскольку считают их срочными и (или) важными. Заметки − то же, что и письма, только менее срочные или важные. Обзорные статьи не содержат оригинальных исследований, а собирают результаты из множества отдельных статей по определенной тематике в согласованное описание последних достижений в данной области. Обзорные статьи представляют информацию по тематике и дают ссылки на журналы с оригинальными исследованиями.

Некоторые журналы имеют раздел редактора и раздел писем к редактору. Журналы типа “Science” имеют раздел “Новости”, в котором публикуются статьи, написанные научными журналистами, а не учеными. Эти публикации не считаются научными, поскольку не рецензируются коллегами.

8.7.3. Научный журнал. Научный журнал представляет собой периодическое издание, направленное на дальнейший прогресс науки, обычно с помощью сообщения о новых исследованиях. Большинство журналов высокоспециализированны, однако некоторые из старых журналов типа “Nature” публикуют статьи и научные работы в широком научном спектре.

Научные журналы принимают к печати рецензированные статьи, стремясь обеспечить соответствие статей стандартам качества журнала. Несмотря на то что внешне научные журналы подобны обычным журналам, в действительности они совершенно иные. Выпуски научных журналов редко читают мимоходом, как обычно читают “глянцевый” журнал. Публикация результатов исследований служит важнейшей частью научного метода. Результаты должны быть проверяемыми. Современные статьи представляют последние теоретические исследования и экспериментальные результаты в научной области, которой посвящен журнал. Они часто непонятны неподготовленному читателю.

Престиж научных журналов создается со временем и может быть отражением многих факторов, некоторые из них, но не все можно выразить количественно. Например, в XX в. высокий престиж имел журнал “Доклады Академии наук СССР”.

Во многих областях существует информационная иерархия научных журналов, и наиболее престижные в своей области журналы стремятся избирательно подходить к статьям, отбираемым для публикации. Стандарты, используемые журналом для определения возможности публикации, могут меняться в широких пределах. Обычно редакторы вводят правила научного стиля, однако эти правила меняются от журнала к журналу, особенно для журналов разных издателей.

История научных журналов начинается в 1665 г., когда французский “Journal des sçavans” и английский “Philosophical Transactions of the Royal Society” впервые начали систематически публиковать результаты исследований. В XVII в. мнение, что наука может двигаться вперед только за счет прозрачного и открытого обмена идеями, подтвержденными опытными данными, было крайне непопулярным. Сам факт публикации научного исследования был противоречивым и широко осмеивался. О новом открытии было принято объявлять в виде анаграммы, которая охраняла приоритет открывателя и не поддавалась расшифровке. И. Ньютон и Г. Лейбниц пользовались этим методом, однако выяснилось, что он работал недостаточно хорошо. Социолог Р. Мертон обнаружил, что в XVII в. в 92 % случаев одновременные открытия заканчивались спорами. Число споров снизилось до 72 % в XVIII в., до 59 % − во второй половине XIX в. и до 33 % в первой половине XX в. [Wikipedia]. Снижение числа ученых, оспаривающих приоритет научных открытий, может быть обусловлено увеличением числа обращений к публикациям статей в современных научных журналах.

Глава 9. Метрики авторов научных публикаций

На “Science Online”* под солнцем жгучим
“ученый” размышлял, чей индекс круче.

Подражание Вл. Вишневскому

Термин “метрика” имеет много значений. Например, в математике метрикой называют функцию, определяющую расстояние в метрическом пространстве [Википедия]. Здесь и далее под “метрикой” будем понимать метод проведения библиометрического измерения, основанный на анализе цитирования. В простейшем случае результаты научного труда оцениваются с помощью двух параметров: числа статей, опубликованных в реферируемых научных журналах, и числа ссылок на эти статьи.

Частота проведения и масштаб библиометрических измерений зависят от доступности для исследователей библиографических БД – WoS, Scopus, Compendex, PubMed и др. Начиная с 2005 г. наблюдается лавинообразное увеличение числа разработок метрик, характеризующих (оценивающих) активность ученых по публикации результатов исследований (табл. 9.4). Особое место здесь занимает индекс Хирша, поскольку значительную часть других разработок представляют его модификации. В табл. 9.4 приведены далеко не все разработки, достаточно полный их перечень представлен на сайте [SCI2S].

9.1. Метрика *h*-индекс

Индекс Хирша, или *h*-индекс, предложен в 2005 г. американским физиком Х. Хиршем из университета Сан-Диего (Калифор-

* Международная научно-практическая конференция “Science Online: Электронные информационные ресурсы для науки и образования”. http://elibrary.ru/.

ния, США) [Hirsch, 2005]. *h*-индекс представляет собой количественную характеристику продуктивности ученого, основанную на распределении цитирований его работ.

Для “автора” индекс Хирша вычисляется следующим образом. Пусть N ($N \geq 1$) − общее число публикаций некоторого автора. Будем полагать, что этот автор имеет индекс Хирша, равный h, если h из его N научных трудов цитируются как минимум h раз каждая, в то время как оставшиеся ($N - h$) трудов цитируются не более чем h раз каждая. Иными словами, автор имеет индекс Хирша, равный h, если он опубликовал h трудов, на каждый из которых сослались как минимум h раз.

Формальное определение *h*-индекса дано в работе [Glanzel, 2006b]. Предположим, что некоторый автор опубликовал n трудов и каждая i-я публикация имеет x_i цитирований (i=1, 2, ..., n). Упорядочим публикации по “рангу”, т. е. по убыванию значений x_i ($x_1^* \geq x_2^* \geq \ldots \geq x_n^*$), где x_1^* − число цитирований, полученное наиболее цитируемой публикацией, а x_n^* − число цитирований, полученное наименее цитируемой публикацией. Будем считать, что значение *h*-индекса равно h, где

$$h = \max\{j : x_j^* \geq j\}.$$

Таким образом, *h*-индекс автора является результатом баланса между числом его публикаций и числом цитирований, которое получила каждая публикация.

Поясним это определение графически. Упорядочим все N публикаций автора по невозрастанию их ранга, $R(p_i)$ – ранг публикации p_i. На рис. 9.1 номера упорядоченных публикаций будем откладывать по оси абсцисс, а число цитирований $Ct(.)$ − по оси ординат (нуль, если цитирований нет). Таким образом, построим распределение публикаций по рангу цитирования, где каждой

публикации соответствует точка с координатами (x, y). Теперь выберем точку с минимальным значением x, такую что $x = y$. Будем говорить, что $x = h$ – индекс Хирша. Заметим, что такое h существует, если общее число цитирований всех N публикаций этого автора $T \geq 1$.

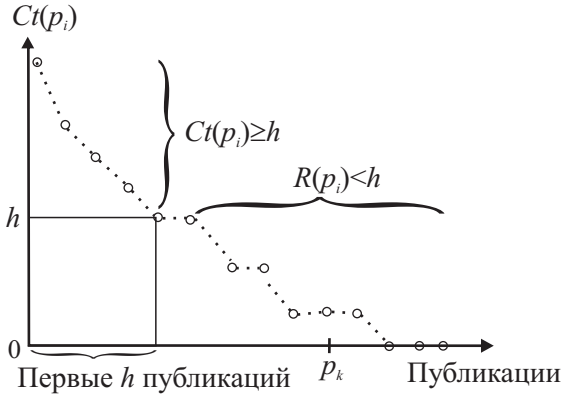


Рис. 9.1. Метрика h-индекс

9.1.1. Ядро Хирша. Все публикации ранга от 1 до h образуют ядро Хирша (h-ядро). В случае если несколько публикаций имеют одинаковое количество цитирований, к определению ядра Хирша существует два подхода: либо в ядро включаются все такие публикации (т. е. ядро будет содержать более h элементов), либо используется дополнительный критерий ранжирования. Например, упорядочим публикации, имеющие одинаковое количество цитирований, в обратном хронологическом порядке, так чтобы наиболее ранняя (по времени опубликования) публикация имела преимущество попасть в ядро.

9.1.2. Вычисление *h*-индекса. Вычислить h-индекс можно вручную на основе информации о цитировании, если таковая имеется. Можно обратиться к автоматическим средствам, например

Scopus или WoK. Можно использовать программный инструмент “Publish or Perish” [Harzing, soft], который вычисляет *h*-индекс на основе информации, предоставленной поисковой машиной Google Scholar. В результате на один и тот же запрос мы получим разные ответы, поскольку в базах данных проиндексированы различные документы.

Детальное изучение состава некоторых БД выполнено в работе [Meho, Yang, 2006]. Установлено, что в WoK достаточно полно представлены журнальные публикации и слабо − труды конференций. Scopus имеет хорошее покрытие материалов конференций и недостаточное покрытие журнальных публикаций до 1996 г. В Google Scholar полно представлены труды конференций и большинство журналов, однако глубина представления ограничена 1990 г. Кроме того, Google Scholar подвергался критике за то, что учитывает цитирования “фантомных публикаций” и “серой литературы” (имеются в виду документы, которые нельзя найти по традиционным каналам). Автор работы [Sanderson, 2008] предлагает в случае, если значения *h*-индекса варьируются для одного и того же ученого, брать наибольшее значение.

9.1.3. Свойства *h*-индекса. Перечислим некоторые свойства индекса Хирша.

9.1.3.1. Значение *h*-индекса дает представление о результативности исследователя в терминах публикационной продуктивности. Это позволяет сравнивать ученых, работающих в одной области, и судить об их научном вкладе.

9.1.3.2. *h*-индекс имеет простое математическое определение.

9.1.3.3. При вычислении *h*-индекса не учитывается отношение общего числа цитирований к общему количеству публикаций, поскольку такая процедура вознаграждает авторов с небольшим количеством высокоцитируемых публикаций.

9.1.3.4. Количество работ, опубликованных автором, имеет прямое влияние на максимальное значение, которое может получить его *h*-индекс.

9.1.3.5. Одинаково хорошие работы скорее повлияют на *h*-индекс молодого ученого, чем на *h*-индекс маститого.

9.1.3.6. *h*-индекс является "устойчивым", поскольку появление высокоцитируемых публикаций (единичных пиков) не дает немедленного повышения значения *h*-индекса, а малоцитируемые публикации вообще не оказывают влияния на его значение, таким образом, публикация "незначительных" работ не поощряется.

9.1.3.7. *h*-индекс может применяться к любому уровню агрегации (автор, группа авторов, научная организация и т. д.).

9.1.3.8. *h*-индекс не пригоден для сравнения научных результатов исследователей, работающих в различных научных областях. Однако в работе [Batista, et al., 2006] утверждается, что значение *h*-индекса можно нормировать с целью учета научной специальности. Например, $h = 3$ для математики соответствует $h = 9$ для биологии.

9.1.4. Критика *h*-индекса. *h*-индекс страдает от проблем, которые присущи всем метрикам, базирующимся на цитировании. Во-первых, это сложность классификации самоцитирований автора или группы авторов. Во-вторых, трудность сбора данных о цитировании, а следовательно, тотальная зависимость значения *h*-индекса от состава метаданных библиографической БД, по данным которой проводится расчет. В-третьих, отсутствие классификации цитирований, в результате чего не учитывается контекст цитирования, а именно ссылки в отрицательном смысле, ссылки на введение, ссылки на недобросовестную работу и т. д. Совокупность перечисленных факторов, в конечном счете, влияет на точность измерений.

Остановимся на особенностях *h*-индекса, мешающих точной оценке продуктивности авторов. Эти особенности отмечены, например, в работах [Jin, et al., 2007; Zhang, 2009; Wikipedia].

9.1.4.1. Длительность публикационного периода автора оказывает влияние на *h*-индекс, он не пригоден для сравнения ученых с "малой" и "большой" научными карьерами. Например, молодые исследователи, имеющие короткую научную карьеру, будут иметь невыгодное положение по сравнению с маститыми учеными, так как у первых заведомо меньше работ и цитирований, чем у последних. Так, *h*-индекс Э. Галуа равен двум и останется таким навсегда [Википедия]. Однако в основополагающей работе Х. Хирша имеются предложения по учету этого эффекта.

9.1.4.2. *h*-индекс является целым числом, что снижает точность измерения. В работе [Ruane, Tol, 2008] определяется понятие рационального *h*-индекса.

9.1.4.3. *h*-индекс не способен убывать, что позволяет ученым "отдыхать на лаврах", так как количество цитирований может только увеличиваться.

9.1.4.4. *h*-индекс не учитывает число авторов цитируемой работы и имеет тенденцию к увеличению в научных областях с большими группами авторов, например в экспериментальной области по сравнению с теоретической. В основополагающей работе Х. Хирша предлагается решать этот вопрос путем деления цитирований между соавторами. При отсутствии информации о вкладе это наиболее простой подход, однако в некоторых областях науки на первое место принято ставить особо значимых авторов.

9.1.4.5. *h*-индекс не учитывает исключительно успешные публикации, т. е. публикации, имеющие высокое число цитирований. Например, два ученых могут иметь одинаковый показатель *h,* скажем $h = 30$, но у одного может быть 20 работ, которые цитировались более 1000 раз, а у другого нет ни одной подобной работы.

9.1.4.6. *h*-индекс в основном предназначен для сравнения выдающихся в своей области ученых.

9.1.4.7. *h*-индекс не учитывает "эффект Матфея" [Wikipedia], заключающийся в том, что известные ученые получают большее число цитирований.

9.1.4.8. *h*-индекс не учитывает тип документа, на который приводится ссылка, тем не менее обзорные работы имеют тенденцию получать большее количество цитирований, чем статьи.

9.1.5. Метрики *a*-индекс и *m*-индекс. В работе [Hirsch, 2005] помимо определения *h*-индекса даны определения еще двух индексов: *a*-индекса и *m*-индекса.

a-индекс определяется следующим образом. Пусть $N_{c,tot}$ – количество всех цитирований одного автора. Тогда в общем случае $N_{c,tot} > h^2$. Для того чтобы сравнить значения $N_{c,tot}$ и h^2, вводится коэффициент пропорциональности a, такой что $N_{c,tot} = a \times h^2$, т. е.

$$a = N_{c,tot} / h^2 .$$

Таким образом, *a*-индекс учитывает "выбросы", которые дают публикации с большим количеством цитирований, а также публикации с меньшим числом цитирований.

Для определения *m*-индекса будем полагать, что n – "длительность карьеры ученого", т. е. количество лет от начала публикаций (при этом автор подчеркивает, что это не обязательно самая первая публикация). Предположим также, что автор публикует работы в течение карьеры стабильно и одинакового качества (практически с одним и тем же количеством цитирований), тогда индекс h можно определить как $h \approx m \times n$, т. е.

$$m \approx h / n .$$

Подчеркивается, что индекс m свой для каждого ученого. В рамках линейной модели индекс m предлагается в качестве ме-

ры, позволяющей сравнить ученых с различной длительностью карьеры.

9.1.6. Сравнение *h*-индекса со стандартными показателями. В работе [Van Raan, 2006] приводятся характеристики статистической корреляции между *h*-индексом и несколькими стандартными библиометрическими показателями. Вопрос изучается на уровне исследовательских групп (147 университетских химических исследовательских групп в Нидерландах за 1991-2000 гг.), которые автор считает наиболее интересным объектом для исследования, особенно в естественных науках. Более того, вместо всего временного периода используется трехгодичное окно, для того чтобы фокусироваться на влиянии более поздних работ, т. е. получать текущую производительность. Рассматриваются следующие показатели:

а) Количество публикаций *P* в журналах, охваченных WoS.

б) Количество цитирований *C*, полученных *P*.

в) Среднее количество цитирований на публикацию *CPP*.

г) Средняя важность (англ. impact), основанная на престиже журнала: если журнал один, *JCS* (journal citation score) − это *CPP* на уровне журнала; если количество журналов больше одного, то используется среднее, *JCSm*.

д) Средняя важность относительно области/подобласти (англ. field/subfield) деятельности, *FCS* (в случае одной области) – это *CPP* на уровне области; *FCSm* (в случае нескольких областей).

е) Сравнение влияния группы со всеобщим средним на основе *JCSm*, *CPP/JCSm*.

ж) Сравнение влияния группы со всеобщим средним на основе *FCSm*, *CPP/FCSm*.

При вычислении средних разного уровня рассматриваются одни и те же типы документов и окна цитирования и публикаций, а самоцитирования не учитываются. Для документов различного

типа средние подсчитываются отдельно, затем берется общая сумма. В качестве “хорошего” индикатора предлагается *CPP/FCSm.* Проведенные исследования показали, что корреляция с общим количеством цитирований C и общим количеством публикаций P у h-индекса значительно выше, чем у индикатора *CPP/FCSm*. Степень корреляции между h-индексом и *CPP/FCSm* низкая.

Поскольку h-индекс указывает на “грубую оценку цитирования”, можно предположить, что он имеет высокую степень корреляции с оценкой коллег. По мнению автора, это можно обнаружить, просто используя количество всех цитирований − C; между C и h-индексом обнаруживается высокая степень корреляции. Однако ситуация меняется для небольших групп в областях науки с “менее мощным потоком цитирования”. Автор пришел к выводу, что в этом случае “хороший” индикатор научной производительности является более приемлемым.

9.1.7. Проект “Кто есть кто в российской науке”. В качестве примера использования h-индекса как одного из параметров ранжирования отечественных ученых приведем одну из таблиц проекта “Кто есть кто в российской науке” [Штерн, 2001]. В рамках этого проекта начиная с 2001 г. публикуются списки российских научных работников, составленные на основе информации БД WoS. Ежегодно составляется два основных списка авторов, работы которых были процитированы не менее 1000 раз начиная с 1986 г., а также тех авторов, у кого суммарное цитирование работ за последние семь лет превысило 100. Также составляются дополнительные списки по географическим признакам (городам РФ, странам вне РФ, институтам в РФ) и областям знаний (физика, астрономия, биология, химия, науки о Земле, математика). Автор проекта Б. Е. Штерн отмечает: “Этот проект имел разнообразные и долгосрочные последствия: многие российские научные работники впервые узнали о существовании индекса цитирования; не-

которые задумались о том, что система профессиональных ценностей может быть выстроена вне степеней и званий; иные усомнились в том, что эта “международная система ценностей” универсальна − уж очень разные бывают науки, даже среди естественных, и не может быть единых “порогов” для определения высокого цитирования; кто-то утверждал, что система обесценивается из-за слишком активного цитирования заведомо спекулятивных работ”.

Таблица 9.1

Первые позиции списка CI_{tot} по состоянию на 10.04.2011 г.

Позиция, параметр	**1**	**2**	**3**	**4**
Name	Polyakov A. M.	Geim A. K.	Novoselov K. S.	Zakharov V. E.
CI_{tot}	23947	21159	20179	17623
$<N>$	1,7	6,1	7,0	0,0
$\underline{h}$	47	46	38	60
CI_{max}	3609	3585	3968	1766
FA	15135, 1914	4717, 2466	8897, 3968	15999, 1766
Field	hep	cond-mat	cond-mat	math-phys plasma
Affiliation	Landau ITP		Microel Techn	LPI
Residence	Princeton, NJ	Manchester	Manchester	Tucson, Arizona
First pub.	1963	1981	1998	1962
Updated	2011-03-22	2011-01-16	2011-03-22	2009-05-07

В табл. 9.1 приведены первые четыре строки списка CI_{tot}. По состоянию на 10.04.2011 г. этот список содержит информацию о

1413 российских ученых, работы которых процитированы не менее 1000 раз.

В табл. 9.1 используются следующие обозначения: CI_{tot} − суммарное цитирование по ISI (WoS), глубина до 1986 г.; $CI_{\max}$ − максимальное цитирование одной работы; CI_7 − суммарное цитирование работ, опубликованных за последние 7 лет; FA, $FA_{\max}$ − суммарное цитирование работ, в которых человек является первым автором, и максимальное цитирование одной из этих работ; FA_7, $FA_{7,\max}$ − то же для публикаций за последние 7 лет; h − индекс Хирша; $<N>$ − среднее число авторов в статье; *First pub* − год первой публикации, упомянутой в WoS.

9. 2. Метрики, подобные *h*-индексу

Нет единого мнения, что какой-либо одиночный
библиометрический индекс явно предпочтительнее *h*.
Х. Хирш [*Hirsch,* 2010]

В 2010 г. автором работы [Hirsh, 2010] сформулированы четыре пожелания о том, какими свойствами должна обладать любая метрика (m), пригодная для измерения результатов научной деятельности. Во-первых, m должна быть значимой в статистическом смысле и в идеале иметь предсказательную силу. Во-вторых, m не должна побуждать к действиям, наносящим ущерб науке. В-третьих, метрика m должна быть нечувствительной к “небольшим” вариациям в списке цитирований. Последнее, но немаловажное формальное определение метрики m не должно создавать трудностей для вычисления значения m на основе существующих библиографических БД.

В свете высказанных критических замечаний относительно *h*-индекса неудивительно появление серии работ, посвященных

простым вариациям идеи, заложенной в определении индекса Хирша, и ставящих цель устранить хотя бы один из указанных недостатков. Представим некоторые из них.

9.2.1. Метрика *g*-индекс. Расчет индекса Хирша ведется по всем высокоцитируемым публикациям. Если определено, что статья имеет число цитирований больше h, то величина превышения не имеет значения и далее не исследуется.

Из определения h-индекса следует, что каждая публикация ранга $1,\ldots,h$ имеет, по крайней мере, h цитирований, так что все публикации имеют не менее h^2 цитирований. На основе подтверждающих примеров L. Egghe [Egghe, 2006a] предполагает, что первые $(h+1)$ публикаций в сумме будут иметь $(h+1)^2$ или более цитирований и т. д. Исходя из этого предположения в работе вводится определение g-индекса: множество публикаций имеет g-индекс, равный g, если g является наивысшим рангом, таким что первые g публикаций вместе имеют, по крайней мере, g^2 цитирований. Это также означает, что первые g публикаций имеют меньше, чем $(g+1)^2$ цитирований. Отсюда следует утверждение: “Во всех случаях верно неравенство $g \geq h$”. Таким образом, g-индекс учитывает фактическое количество цитирований наиболее продуктивных публикаций. В работе приводятся примеры более явной дифференциации ученых в случае сравнения с помощью g-индекса.

Данное исследование получило развитие в работе [Egghe, 2006b], в котрой изложена теория g-индекса, основанная на законе Лотки. Приведена практика применения g-индекса, пригодная для оценки, основанной на любом множестве публикаций, будь то статьи ученого или годовая продукция журнала.

9.2.2. Метрика *hg*-индекс. Индекс hg базируется на комбинации индексов h и g [Alonso, et al., 2010] и вычисляется как геомет-

рическое среднее индексов h и g, $hg=\sqrt{(h\times g)}$. Одним из недостатков h-индекса считается недооценка цитирований в h-ядре. В то же время на значение g-индекса, учитывающего этот недостаток, существенное влияние может оказать одна много цитируемая работа. hg-индекс введен для того, чтобы минимизировать недостатки и сохранить достоинства. К основным достоинствам hg-индекса можно отнести то, что он, во-первых, обеспечивает бóльшую степень детализации, чем h и g, позволяя эффективнее оценивать ученых, например в случае с одинаковым значением h; во-вторых, смягчает влияние часто цитируемых работ и обеспечивает лучший баланс между влиянием большинства лучших работ и небольшого количества очень цитируемых работ. Достаточно показать, что

$$h \le hg \le g \text{ и } hg - h \le g - hg,$$

т. е. hg ближе к h, чем к g.

9.2.3 Метрика $h^{(2)}$-индекс. В работе [Kosmulski, 2006b] вводится понятие $h^{(2)}$-индекса, который определяется как *наибольшее натуральное число, такое что каждая из $h^{(2)}$ наиболее цитируемых публикаций получила не менее $(h^{(2)})^2$ цитирований*. Из определения следует, что

$$g \ge h \ge h^{(2)}.$$

т. е. для вычисления $h^{(2)}$-индекса требуется рассмотрение меньшего числа работ. Автор аргументирует введение индекса тем, что он больше подходит для сравнения авторов, работающих в областях, где принято много цитирований, таких как химия, медицина, биология.

9.2.4. Индексы *A* и *R*. *A*-индекс (от англ. “average”), определяющий среднее количество цитирований на публикацию в

h-ядре, предложен в работе [Jin, 2006] (доступна [Jin, et al., 2007]) и вычисляется по формуле

$$A = (1/h) \sum_{j=1}^{h} cit_j.$$

В данной формуле цитирования (cit_j) упорядочены по убыванию. Поскольку *h*-ядро содержит точно h элементов, то *A*-индекс определяется однозначно. Очевидно, что $h \leq A$, тем не менее для вычисления *A*-индекса используется тот же набор данных, что и для вычисления *h*-индекса.

A-индекс также не лишен недостатков. Проиллюстрируем это на вымышленном примере. Предположим, что ученый X_1 опубликовал 20 работ, одна из которых процитирована 10 раз, другие – по одному разу. Ученый X_2 опубликовал 30 работ, одна из которых процитирована 10 раз, все остальные – по два раза. Очевидно, что ученый X_2 успешнее ученого X_1. Это проявляется и в значении *h*-индекса, который равен единице для X_1 и двум для X_2. Однако *A*-индекс равен десяти для X_1 и равен шести для X_2. Ученый X_2 "наказан" за то, что он имеет больший *h*-индекс, который в формуле для вычисления *A*-индекса стоит в знаменателе дроби. Для устранения этой "несправедливости" в работе [Jin, et al., 2007] вводится *R*-индекс, который определяется как квадратный корень из общего количества цитирований на публикации в *h*-ядре:

$$R = \sqrt{\sum_{j=1}^{h} cit_j}.$$

R-индекс измеряет интенсивность цитирования в рамках *h*-ядра. Очевидно, что индекс R можно представить в виде $R = \sqrt{A \times h}$. Поскольку количество цитирований каждой статьи не меньше *h*, то $h \leq R$. В случае если количество цитирований каждой статьи

точно равно h, имеем $R = h$. Этот результат еще раз показывает, что при конструировании метрик использовать корень из суммы предпочтительнее, чем саму сумму.

9.2.5 Метрика *AR*-индекс. В целях преодоления проблемы, заключающейся в том, что h-индекс не может уменьшаться и ученые "могут отдыхать на лаврах", в работе [Jin, 2007] определяется зависящий от возраста публикаций (age-dependent) R-индекс – AR-индекс.

Обозначив через a_j возраст публикации j, определим AR в виде

$$AR = \sqrt{\sum_{j=1}^{h} \left(cit_j / a_j \right)}.$$

В случае если существует несколько публикаций, имеющих точно h цитирований, в h-ядро включается более поздняя публикация.

Преимущество AR-индекса очевидно: он позволяет учитывать не только все цитирования, но и возраст публикации. В этом случае h-индекс дополняется индексом, способным убывать. По мнению автора, это является необходимым условием хорошего индикатора. Заметим, что AR-индекс базируется на h-индексе, так как использует h-ядро. Для AR-индекса уже не обязательно верно неравенство $h \leq AR$, в отличие от соотношения h с R-индексом (см. 9.2.4). Вычисление AR-индекса по сложности не превосходит вычисления h, так как дополнительно к данным, требующимся для вычислений h-индекса, нужен только возраст публикаций. Положительные черты R-индекса здесь сохраняются, так что пара (h, AR) предлагается автором как хороший индикатор научной производительности. При этом подчеркивается, что при использовании пары индикаторов целесообразно рассматривать временно́е окно, а не принимать во внимание все достижения ученого.

Anne-Wil Harzig [Harzing, soft] указывает, что в программном продукте "Publish or Perish" вместо метрики AR вводятся три мет-

рики: *AWCR* (Age-weighted citation rate), *AW*, *AWCRpA* (per author), в которых сумма берется по цитированиям всех публикаций (а не только публикаций, входящих в *h*-ядро):

$$AWCR = \sum_{j \in P} \left(cit_j / a_j \right).$$

Здесь P – количество всех публикаций за рассматриваемый период

$$AW = \sqrt{AWCR}.$$

Индикатор *AWCRpA* подобен *AWCR,* но цитирование для публикации делится на число авторов публикации.

9.2.6. Метрика *e*-индекс. Характерным свойством *h*-индекса является то, что на его значение не влияют мало- и высокоцитируемые публикации. Это является причиной как минимум двух недостатков, которые ограничивают аккуратное и беспристрастное сравнение личной производительности ученых. Первый недостаток – не рассматривается информация о цитированиях, кроме h^2 цитирований, соответствующих *h*-ядру. Вследствие этого возможны недоразумения, так как ученые, имеющие более низкий *h*-индекс, могут получить большее число цитирований, чем те, кто имеет высокий *h*-индекс. Второй недостаток – *h*-индекс имеет низкое разрешение, поскольку выражается в натуральных числах, а не в действительных. В работе [Zhang, 2009] указанные недостатки предлагается устранить путем введения *e*-индекса, учитывающего цитирования, проигнорированные при вычислении *h*-индекса.

Дадим определение *e*-индекса. Дополнительные цитирования в *h*-ядре, обозначаемые как e^2, определяются следующим образом:

$$e^2 = \sum_{j=1}^{h} \left(cit_j - h \right) = \sum_{j=1}^{h} cit_j - h^2$$

(cit_j – количество цитирований, полученных *j*-й публикацией).

Таким образом, *e*-индекс определяет вклад дополнительных цитирований, относящихся к высокоцитируемым работам.

Пусть

$$d^2 = \sum_{j=1}^{h} cit_j .$$

Тогда

$$d^2 = h^2 + e^2 , \qquad (9.2.1)$$

или

$$e = \sqrt{\left(d^2 - h^2 \right)} .$$

Заметим, что $e \geq 0$ и является вещественным числом. Соответственно

$$0 \leq e < \infty.$$

Без потери общности будем считать, что значения cit_j, j=1, …, N могут быть представлены интегрируемой функцией $C(t)$, $t \in [0, N]$, где $C(0)=0$, $C(j)= cit_j$, j=1, …,N. Тогда

$$e^2 = \int_0^h \left(C(t) - h \right) dt = \int_0^h C(t)dt - h^2 .$$

Опираясь на $C(t)$, дадим геометрическое объяснение функции e^2 (рис. 9.2). Напомним, что e не зависит от h, и e^2 представляет собой остаток множества цитирований в h-ядре, в дополнение к h^2. Заметим, что чем больше e, тем больше остаток и тем значительнее потеря информации о цитированиях, если использовать только h-индекс. Иными словами, если h-индекс используется для оценки производительности ученого, то чем меньше e, тем более достоверен h-индекс. Если e=0, что неправдоподобно, то h-индекс полностью описывает информацию, касающуюся публикаций в h-ядре. В противном случае имеет место потеря информации о цитировании.

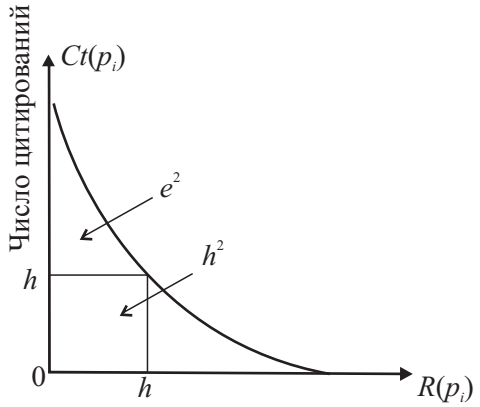


Рис. 9.2. Метрика *e*-индекс

Теперь покажем, как индексы A и R, представленные в подп. 9.2.4, выводятся из индексов h и e.

В десятичной системе координат построим прямоугольник со сторонами h по оси ординат и e по оси абсцисс (рис. 9.3). Очевидно, что он содержит полную информацию о цитированиях, полученных всеми публикациями h-ядра. Эвклидово расстояние от начала координат до точки $P(e, h)$ равно

$$R = \sqrt{\left(h^2 + e^2 \right)} = d. \qquad (9.2.2)$$

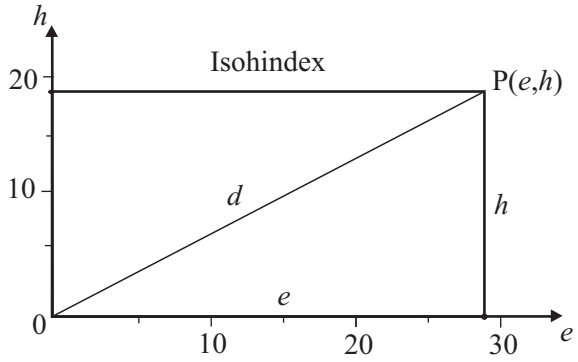


Рис. 9.3. Метрика *R*-индекс

Из равенств (9.2.1), (9.2.2) получаем

$$A = h + e^2 / h .$$

9.2.7. Соотношения метрик *h, e, A, R.* В работе [Zhang, 2009] проведено сравнение рассмотренных способов оценки продуктивности авторов. Индексы *h, e, A, R* можно разделить на два типа: фундаментальные и производные. Фундаментальные индексы удовлетворяют двум условиям: они не зависят друг от друга и могут быть использованы для получения новых индексов. Здесь индексы h и e являются фундаментальными, поскольку не зависят друг от друга и могут использоваться для получения A и R. Индексы A и R, наоборот, являются производными. Пусть

$$f = (e/h)^2, \quad h \neq 0,$$

где f обозначает отличие проигнорированных цитирований, полученных в рамках h-ядра, от h^2.

Как видно на рис. 9.2, общее число цитирований в рамках h-ядра равно $h^2 + e^2$. Следовательно, комбинация (h, e) обеспечивает полное представление о цитированиях в рамках h-ядра. В то же время индексы A и R избыточны по отношению к h. В случае если A или R используется совместно с h, он маскирует значения f, что проиллюстрировано в п. 9.2.8 при сравнении цитирований ряда ученых.

Сравним производительность ученых, принадлежащих одной isohindex-группе. Пусть имеется группа ученых и H – максимальный h-индекс. Точки $P(e,h)$ могут быть расположены только на горизонтальных линиях h-плоскостей. На одной и той же горизонтальной линии все точки имеют одинаковый h-индекс. Эту линию будем называть isohindex (одна из таких линий показана на рис. 9.3). Соответственно группа ученых, имеющих один и тот же

h-индекс, называется isohindex-группой. В такой группе адекватно оценить производительность можно только с привлечением *e*-индекса.

Преимущество использования *e*-индекса заключается в том, что e^2 указывает на все проигнорированные цитирования в *h*-ядре, при этом *A* и *R*, являясь производными от *h,* включают вклад h^2 цитирований и дополнительных e^2 цитирований и зависят от *h* и *e*, в то время как *e* не зависит от *h*. Следовательно, использование индексов *A* и *R* совместно с *h в* рамках одной isohindex-группы может исказить результаты.

Необходимо напомнить о *g*-индексе, который представлен как "чувствительный к уровню высокоцитируемых статей" и определен как "наибольшее число *g* статей, такое что все вместе они получили g^2 или больше цитирований" [Egghe, 2006a]. Несмотря на ряд достоинств, *g*-индекс имеет и недостатки. Например, с его помощью нельзя ранжировать ученых, имеющих высокий уровень цитируемости. Например, если для любого *k*

$$\sum_{j=1}^{k} cit_j > N^2, \quad k = 1,\ldots,N, \tag{9.2.3}$$

то *g*-индекс не определен.

Фактически для любых *N* условий в неравенстве (9.2.3) *g*-индекс может быть неопределенным. Среди *N* условий наиболее сильным условием является

$$cit_1 > N^2, \tag{9.2.4}$$

а наиболее слабым –

$$\sum_{j=1}^{N} cit_j > N^2. \tag{9.2.5}$$

Неравенства (9.2.3)–(9.2.5) ассоциируются с реальными ситуациями. Соответствующие примеры приведены в работе [Zhang, 009].

В заключение следует отметить, что *h*-индекс уже используется некоторыми БД для оценки личной производительности ученых. Вследствие потери информации о цитировании, сравнение, основанное только на *h*-индексе, может искажать действительную картину. Общее количество цитирований некоего ученого может быть значительно больше, чем количество цитирований многих других, имеющих больший *h*-индекс; количество проигнорированных цитирований (e^2) может быть в пять раз больше, чем h^2. Таким образом, для получения адекватной картины необходимо совместное использование *h*-индекса и *e*-индекса. Другие индексы *h*-типа, такие как *A* и *R*, являются *h*-зависимыми и также страдают от потери информации, а при использовании совместно с *h*-индексом маскируют различие между учеными. Несмотря на свою простоту, *e*-индекс необходим как дополнение к *h*-индексу, особенно при ранжировании высокоцитируемых ученых или для более точного сравнения производительности в группе ученых, имеющих одинаковый *h*-индекс.

9.2.8. Индивидуальный *h*-индекс. Для вычисления индивидуальной средней продуктивности в работе [Batista, et al., 2006] предложена метрика h_I. Пусть $Na^{(T)}$ – общее количество авторов в статьях, которые были учтены при вычислении h (если у каждой статьи только один автор, то $Na^{(T)} = h$).

Пусть $< Na > = Na^{(T)}/ h,$ т. е. $< Na >$ − среднее количество авторов в изучаемых *h* публикациях. Определим h_I как

$$h_I = h \,/ < Na >= h^2 / Na^{(T)} .$$

Метрика h_I определяет количество публикаций, имеющих не менее h_I цитирований, которое опубликовал бы исследователь в течение своей карьеры, если бы работал один. Эта метрика используется в библиографических расчетах [Harzing, soft].

В расчетах [Harzing, soft] используется также альтернативный подход учета индивидуального *h*-индекса. Рассматривается нормализованное количество цитирований: количество цитирований совместной работы делится на количество соавторов до вычисления *h.* Такой подход считается лучшей аппроксимацией влияния каждого автора, так как несколько работ с большим количеством соавторов не оказывают такого сильного влияния на значение *h*-индекса, как в первом случае.

Иной подход используется в работе [Schreiber, 2008], в которой для учета индивидуальной продуктивности предложена метрика h_m (multi-authored). Количество соавторов у публикаций учитывается здесь при определении ранга статьи. Пусть публикации упорядочены по невозрастанию количества цитирований, ранг определен по Хиршу нумерацией публикаций, начиная от 1. Обозначим через $c(r)$ количество цитирований работы ранга r, а через $a(r)$ − количество соавторов у статьи ранга r. Новые ранги у публикаций определим следующим образом:

$$r_{eff}(r) = \sum_{r'}^{r} 1/a(r').$$

Например, $r_{eff}(2)=1/a(1)+1/a(2)$. Аналогично тому, как определяется *h*-индекс, определим h_m-индекс:

$$h_m = \max_r \left(r_{eff}(r) \le c(r) \right).$$

В работе [Schreiber, 2008] подчеркивается, что с учетом соавторства *h*-индекс меняется значительно, поэтому, во-первых, требуются дальнейшие поиски адекватного решения, во-вторых, это показывает, насколько опасно пользоваться единственной метрикой при сравнении деятельности ученых. Метрика h_m также рассчитывается [Harzing, soft].

9.2.9. Метрика h_w-индекс. В работе [Egghe, Rousseau, 2008] вводится понятие взвешенного относительно цитирований h-индекса (citation-weighted h-index) h_w, который определен для непрерывного и дискретного случаев. Рассмотрим вычисление h_w для дискретного случая. Построим таблицу взвешенных рангов (табл. 9.2), и на ее основе определим h_w:

$$h_w = \sqrt{\sum_{i=1}^{r0} cit_i}.$$

Здесь r_0 – наибольший индекс в табл. 9.2, для которого $r_w(j) \le j$. Заметим, что $r_w(j)$ – не обязательно натуральное число, так же как и cit_j, которое может быть взвешено относительно количества со-

Таблица 9.2

Таблица для определения h_w в дискретном случае

Взвешенный ранг r_w	**Количество цитирований публикаций в h-ядре**
$r_w(1)=cit_1/h \ge 1$	$cit_1 \ge h$
$r_w(2)=(cit_1/h + cit_2/h) \ge 2$	$cit_2 \ge h$
…	…
$r_w(j) = \left(\sum_{i=1}^{j} cit_i \right) \Big/ h \ge j$	$cit_j \ge h$
…	…
$r_w(h) = \left(\sum_{i=1}^{h} cit_i \right) \Big/ h \ge h$	$cit_h \ge h$
$r_w(h+1) = \left(\sum_{i=1}^{h+1} cit_i \right) \Big/ h$	$cit_{h+1} < h$

авторов. Если h=0, то h_w устанавливается равным нулю. Кроме того, первые h или h+1 рядов необходимы для вычисления h-индекса.

Перечислим свойства h_w-индекса:

а) Если все первые h статей имеют одинаковое количество цитирований, например $k \geq h$, то $h_w = \sqrt{k \times h}$.

б) Ранг $r_0 \leq h$.

в) $h_w > \sqrt{h \times (h-1)}$.

9.2.10. Метрика *p*-индекс. В работе [Prathap, 2010] введена метрика p-индекс (performance index), а в работе [Prathap, 2011] дано ее расширение с учетом соавторства. Метрика предлагается как дополнение к h-индексу для преодоления неэффективности в тех случаях, когда h имеет тенденцию иметь небольшое значение, т. е. распределение цитирований имеет длинный хвост малоцитируемых работ. Автор считает, что формула C^2/P, где P – общее количество работ, C − общее количество цитирований, хорошо отражает сущность научной деятельности, сочетая количество и качество (C – количественный показатель, C/P − качественный). На основе анализа размерностей h и C^2/P автор делает вывод, что C^2/P имеет размерность h^3, а $(C^2/P)^{1/3}$ имеет ту же размерность, что и h.

Для учета соавторства вводятся два индекса: p_f (fractional), когда соавторы считаются равноправными, и p_h (harmonic), когда авторы публикации упорядочены по степени значимости. При расчете индекса p_f публикации рассматриваются в любом порядке; в случае если публикация i имеет $a(i)$ авторов, то ее дробное влияние r_i=1/$a(i)$ распространяется и на количество цитирований, и на количество публикаций: P_f=Σr_i, C_f=$\Sigma(r_i \times c_i)$. Соответственно $p_f = (C_f^2/ P_f)^{1/3}$. При расчете индекса p_h публикации рассматрива-

ются в любом порядке, но влияние статьи *i,* имеющей $a(i)$ авторов, рассчитывается для автора номер j по формуле $r_i=(1/j) / (1+1/2+...+1/a(i))$:

$$P_h = \sum r_i,\ C_{h,} = \sum\left(r_i \times c_i \right),\ p_h = \left(C_h^{\,2} / P_h \right)^{1/3}.$$

Данный подход универсален по сути и применим практически к любой схеме определения степени влияния публикации.

9.2.11. Метрика *ħ*-индекс. Новый индекс *ħ* (hbar) введен в работе Х. Хирша [Hirsch, 2010] для учета соавторства, не оказывающего влияния на оригинальный *h*-индекс. В отличие от авторов работ, приведенных в пп. 9.2.8, 9.2.10, Х. Хирш пошел другим путем − не взвешивая публикации или цитирования согласно количеству соавторов. В первоначальном варианте определение имело следующий вид. Ученый имеет индекс *ħ,* если *ħ* из его публикаций принадлежит его *ħ*-ядру. Публикация принадлежит *ħ*-ядру ученого, если имеет $\geq ħ$ цитирований и принадлежит *ħ*-ядру каждого из соавторов.

Поскольку одним из требований, которыми должна обладать метрика, Х. Хирш считает возможность прозрачного расчета на основе имеющихся баз данных, им был введен второй вариант, более простой для счета и “несамосогласованный” (англ. “non-self-consistent”): ученый имеет индекс *ħ*, если *ħ* из его публикаций принадлежит его *ħ*-ядру; публикация принадлежит *ħ*-ядру ученого, если имеет не менее *ħ* цитирований и принадлежит *h*-ядру каждого из соавторов.

В работе [Hirsch, 2010] приводится итерационная процедура получения индивидуального индекса *ħ* на основе *h*-индексов всех соавторов.

Индекс *ħ* может убывать, так как могут увеличиваться *h*-индексы соавторов. Так, совместная работа ученых с практически равными *h*-индексами незначительно уменьшит значение *ħ* для

каждого из них. Также, значение $\hbar$-индекса маститого ученого не сильно пострадает от соавторства с начинающими исследователями, поскольку его базовое значение было достигнуто ранее. Х. Хирш считает, что этот индекс можно использовать совместно с h-индексом, особенно для молодых ученых, но если желательна единственная метрика, то следует использовать среднее между h и $\hbar$.

9.2.12. Иерархический *h*-индекс. Рассмотренный выше h-индекс используется для оценки индивидуальной продуктивности автора научных статей. Подобные показатели уровней продуктивности также могут быть полезными для оценки других объектов измерений, например коллектива авторов, научных журналов, организаций и стран. В этом контексте важно отметить работу [Schubert, 2007], в которой вводится понятие "successive h-index". Это h-индекс, полученный на основе h-индексов более низкого уровня. В табл. 9.3 приведены результаты анализа четырех уровней объектов измерения, выполненного в работе [Egghe, 2008].

Таблица 9.3

Уровни объектов измерения

Уровень	**Объект измерений**	**Результат**
1	Статьи	Индивидуальная продуктивность автора
2	Авторы	Продуктивность организации
3	Организации	Продуктивность страны
4	Страны	Глобальная продуктивность

На первом уровне значение h-индекса определяет индивидуальную научную продуктивность автора по его публикациям. На втором уровне рассматривается продуктивность организации, при этом учитываются упорядоченные по убыванию h-индексы работающих в них авторов. Полученная последовательность позволяет говорить об h-индексе данной организации. Одним из вариантов

такого подхода является вычисление *I*-индекса [Prathap, 2006]. На третьем уровне рассматриваются все научные организации страны. Анализируя их *h*-индексы, можно получить *h*-индекс этой страны. В заключение, на четвертом уровне, *h*-индексы стран могут быть упорядочены, чтобы получить "глобальную картину" – *h*-индекс группы стран. Математическая основа для моделирования *h*-индексов разработана автором [Egghe, 2008], предложившим концепцию "глобального *h*-индекса" некоего "метаавтора", который объединяет все статьи и их цитирования различными авторами в одну группу.

9.2.13. Аннотации метрик

9.2.13.1. В работе [Vaidya, 2005] определяется *v*-индекс как уточнение *m*-индекса. При этом учитывается не только продолжительность карьеры, но и время, которое ученый тратит на исследование. Например, ученые, одновременно работающие в научной лаборатории и клинике, затрачивают на исследования лишь 40−50 % времени.

9.2.13.2. В работах [Kosmulski, 2006a; Prathap, 2006] независимо друг от друга авторы определяют индекс для измерения продуктивности организации. Организация имеет *I*-index, равный *i*, если, по меньшей мере, *i* исследователей имеют *h*-индекс, по меньшей мере равный *i.* В работе [Prathap, 2006] определяемый аналогичным образом индекс назван *h*2. В этой же работе приведен альтернативный способ определения продуктивности организации, рассматриваются все работы всех сотрудников и все цитирования, полученные ими. К этим данным применяется процедура вычисления *h*-индекса, полученный результат называют *h*1-индексом.

9.2.13.3. В работе [Sidiropoulos, et al., 2007] определяются следующие метрики: contemporary *h*-index, учитывающий взвешенное количество цитирований согласно возрасту публикации; trend

h-index, учитывающий возраст цитирований. Также дается определение версии *h*-индекса − индекса h^n, нормализованного по общему количеству публикаций (normalized *h*-index): исследователь имеет индекс $h^n=h / N_p$, если h из его N_p работ получили не менее h цитирований каждая, а остальные $(N_p - h)$ работ получили не более h цитирований.

9.2.13.4. В работе [Anderson, et al., 2008] определяется индекс h_T (tapered *h*-index), учитывающий все публикации автора. Рассматриваются "диаграммы Ферре" (англ. Ferrers diagram [Wikipedia]), являющиеся графическим представлением разбиения целого числа (в данном случае − разбиения общего числа цитирований по публикациям). Показано, что *h*-индекс ассоциируется со стороной "квадрата Дюрфи" (англ. Durfee square [Wikipedia]), наибольшего квадрата диаграммы Ферре. Приводятся рассуждения о том, как следует расставлять веса точек диаграммы и вычислять ее общий вес, который определяет индекс h_T.

9.2.13.5. В работе [Gągolewski, Grzegorzewski, 2009] представлена обобщенная версия *h*-индекса, предоставляющая дополнительную информацию о форме функции цитирований автора (с тяжелым "хвостом", гладкая, с "пиками" и т. д.)

9.2.13.6. В работе [Bras-Amorys, et al., 2011] определяется *c*-индекс, который учитывает качество цитирования в терминах степени сотрудничества между цитирующим и цитируемым авторами. Ученый имеет *c*-index, равный n, если n из N цитирований исходят от авторов, которые имеют коллаборационную дистанцию, по меньшей мере, равную n, а остальные $(N - n)$ цитирований исходят от авторов, имеющих дистанцию не больше n. В результате новый индекс учитывает только цитирования, которые достаточно важны и в которых важность пропорциональна дистанции.

Для того чтобы помочь читателю ориентироваться в изложенном материале, приведем краткую таблицу метрик и их определений (табл. 9.4).

Таблица 9.4

h-индекс и другие меры продуктивности научной деятельности

Метрика	**Определение / формула**
h [Hirsch, 2005]	Ученый имеет индекс h, если h из его Np публикаций имеет, по меньшей мере, h цитирований каждая, а другие $(Np\text{-}h)$ публикаций имеют не более h цитирований каждая
g [Egghe, 2006a]	Наибольшее число g публикаций такое, что все вместе они получили g^2 или более цитирований
A [Jin, et al., 2007]	$A = 1/h\sum_j cit_j$
$\boldsymbol{h^{(2)}}$ [Kosmulski, 2006b]	Индекс $h^{(2)}$ ученого определяется как наибольшее натуральное число, такое что каждая из $h^{(2)}$ наиболее цитируемых публикаций получила не менее $(h^{(2)})^2$ цитирований
$\boldsymbol{h_I}$ Индивидуальный *h*-индекс [Batista, et al., 2006]	$Na^{(T)}$ – общее количество авторов в h публикациях; $< Na > = Na^{(T)} / h$ – среднее количество авторов на публикацию; $h_I = h / < Na > = h^2 / Na^{(T)}$
$\boldsymbol{h^n}$ Нормализованный *h*-index [Sidiropoulos, et al., 2007]	$h^n = h/Np$

R [Jin, et al., 2007]	$R = \sqrt{\sum_j cit_j}$
AR [Jin, 2007]	$AR = \sqrt{\sum_j \left(cit_j / a_j \right)}$
successive h-index [Schubert, 2007]	Последовательность упорядоченных по убыванию *h*-индексов членов группы определяет *h*-индекс группы
$\boldsymbol{h_w}$ [Egghe, Rousseau, 2008]	Индекс, взвешенный относительно цитирований в *h*-ядре, $h_w = \sqrt{\sum_{i=1}^{r_0} cit_i}$
$\boldsymbol{h_m}$ (multi-authored) [Schreiber, 2008]	Индивидуальный *h*-индекс. До вычисления индекса ранги публикаций изменяются с учетом количества соавторов
e [Zhang, 2009]	$e = \sqrt{\sum_j \left(cit_j - h^2 \right)}$
***ħ* (*hbar*)** [Hirsch, 2010]	Ученый имеет индекс *ħ*, если *ħ* из его публикаций принадлежит его *ħ*-ядру. Публикация принадлежит ħ-ядру ученого, если имеет $\geq \hbar$ цитирований, и принадлежит *h*-ядру каждого из соавторов
hg [Alonso, et al., 2010]	$hg = \sqrt{h \times g}$
p [Prathap, 2010]	$p = (C^2/P)^{1/3}$

Глава 10. Метрики научных журналов

А в попугаях–то я гораздо длиннее!
Г. Остер. 38 попугаев*, 1977*

В данной главе рассматриваются метрики научных периодических изданий, основанные на анализе цитирования и помогающие ранжировать научные журналы. При выполнении процедуры ранжирования следует учитывать также цену журнала, эта информация доступна, например, на сайте [Journal Price].

Необходимо сделать важную оговорку. “Продуктивностью” журнала принято считать число статей, опубликованных журналом в данной области исследований за фиксированный период времени [Tsay, 2009]. Если принять, что все опубликованные статьи имеют одну и ту же вероятность цитирования, то из этого должно вытекать, что чем больше статей публикует журнал, тем выше частота, с которой данный журнал будет цитироваться. Следовательно, частота цитирования журнала не только является функцией научной значимости опубликованного материала, но и зависит от количества статей, которое он публикует. По сути, продуктивность журнала является фактором, оказывающим влияние на подсчет цитирований. В работе [Magyar, 1974] проведено сравнение функции продуктивности с законом Брэдфорда.

Очевидно, что число цитирований журнала влияет на сам журнал, поскольку журнал, содержащий часто цитируемые статьи, имеет более широкое распространение. В свою очередь, статьи из журнала, имеющего большое распространение, будут цитироваться с большей вероятностью. Следовательно, частота цитирования журнала отражает его ценность [Garfield, 1972]. Однако практически в каждой области исследований существуют журналы, которые цитируются менее часто, несмотря на их ценность. Таким об-

разом, необходимо очень аккуратно делать выводы на основе анализа цитирований.

10.1. Импакт-фактор научного журнала

Импакт-фактор разработан для идентификации наиболее значимых журналов в данной области исследований. Первыми важность создания перечня значимых журналов осознали авторы работы [Gross, Gross, 1927], которые рассмотрели метрику, в настоящее время называемую импакт-фактором. Они были заинтересованы идентификацией того, какие периодические научные издания необходимы в библиотеке их колледжа. В 1955 г. вышла работа [Garfield, 1955], в которой изложен подход к вычислению метрики "импакт-фактор научных журналов" (*JIF*), а в 1963 г. – работа [Garfield, Sher, 1963], в которой представлен индекс научного цитирования. Идеи и результаты этих ключевых работ используются и в настоящее время. *JIF* ежегодно рассчитывается по БД журналов, индексированных "Thomson Reuters". Результаты публикуются на сайте WoS и в отчетах по цитированию журналов – "Journal Citation Reports" (JCR).

В общем случае значение *JIF* для журнала получается путем деления числа ссылок, сделанных на документы, опубликованные в журнале J за время T, на число документов, опубликованных за время T. Таким образом, для определения точного значения *JIF* важно знать, какие ссылки (точнее, ссылки на какие документы) включены в числитель *JIF* и какие документы (статьи, письма и т. п.) учитываются в знаменателе этой дроби.

В первой формулировке *JIF* [Gross, Gross, 1927] самоцитирование исключено из анализа, для того чтобы избежать переоценки опубликованных материалов. В настоящее время для вычисления значения *JIF* применяется метод, предложенный в работе [Garfield, Sher, 1963]. Этот метод учитывает любые цитирования,

т. е. количество ссылок в числителе является суммой всех цитирований и самоцитирований, которые получает журнал. В то же время в знаменателе дроби просуммированы не все документы, опубликованные в журнале. Отсутствуют, например, письма, колонка редактора, интервью, информация о конференциях, конкурсах и награждениях – эти документы в расчете *JIF* не участвуют. Данный метод различного подсчета значений числителя и знаменателя дроби *JIF* был и остается объектом критики. Таким образом, значение *JIF* во многом (если не полностью) зависит от исходных документов в БД, по метаданным которых ведется расчет.

10.1.1. Синхронный и диахронный *JIF*. Говорят, что публикация A цитирует публикацию B, если в тексте A имеется ссылка на B, которая содержится в списке литературы или в постраничной сноске публикации A. Говорят, что журнал J цитирует журнал I, если существует хотя бы одна публикация из J, которая цитирует хотя бы одну публикацию из I. Оба определения предполагают наличие некоторого временного интервала, в котором выполняется цитирование. Пусть $Cit_J(Y_1, Y_2)$ − суммарное число цитирований за год Y_1, полученное всеми публикациями журнала J, вышедшими в году Y_2. Рассматриваются только журналы, входящие в БД, на массивах которой вычисляется *JIF*. Пусть $Pub_J(Y)$ − суммарное число публикаций, вышедших в журнале J в году Y. Заметим, что термин “публикация” требует уточнения, которое можно найти в работе [Писляков, 2005].

Классический (синхронный) двухлетний импакт-фактор 2*JIF* (рис. 10.1) журнала J в году Y определяется дробью:

$$[Cit_J(Y, Y\text{-}1) + Cit_J(Y, Y\text{-}2)] / [Pub_J(Y\text{-}1) + Pub_J(Y\text{-}2)],$$

где числитель – сумма ссылок, появившихся в массиве журналов за год Y, на публикации журнала J, вышедшие в годах $Y-1$ и

$Y-2$; знаменатель − суммарное число публикаций, вышедших в журнале J за годы $Y-1$ и $Y-2$.

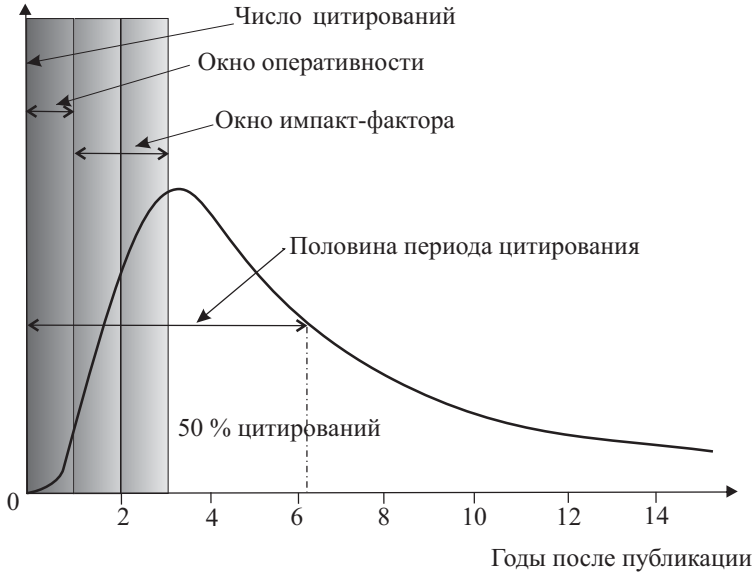


Рис. 10.1. Двухлетний импакт-фактор журнала

Окном публикации будем называть период времени $T=[Y_1, Y_2, ..., Y_k]$ (целое число лет), за который собираются цитирования на публикации журнала J. Окном цитируемости будем называть период времени $T=[Y_1, Y_2, ..., Y_n]$, к которому относятся цитируемые публикации журнала J. Для метрики 2*JIF* размеры окон составляют k=1, n=2.

Пятилетний импакт-фактор журнала (5*JIF*) вычисляется по аналогичной схеме. Для метрики 5*JIF* размеры окон составляют k=1, n=5.

Диахронный импакт-фактор (*DJIF*) журнала J в году Y определяется дробью

$$[Cit_j(Y+1, Y) + Cit_j(Y+2, Y)] / Pub_j(Y),$$

где числитель – сумма ссылок на вышедшие в году Y публикации журнала J, появившихся в массиве журналов за годы Y+2 и Y+1; знаменатель – число публикаций, вышедших в журнале J в году Y. В отличие от синхронного диахронный импакт-фактор учитывает цитирования, которые получают публикации журнала, вышедшие в фиксированном году. Для метрики *DJIF* размеры окон составляют k=2, n=1.

10.1.2. Обсуждение *JIF*. Значения *JIF* для журналов могут значительно различаться в зависимости от областей исследований. Например, журналы в области биомедицины имеют высокие уровни цитирования, а следовательно, высокие значения *JIF*. В противоположность этому наиболее признанные математические журналы имеют низкие *JIF* вследствие малой склонности к цитированию в этой научной дисциплине [Archamblaut, Lariviere, 2009]. Таким образом, прямое сравнение *JIF* может поставить в невыгодное положение сравниваемые дисциплины. Однако с помощью *JIF* можно идентифицировать журналы, оказывавшие сильное воздействие на область исследований в течение фиксированного периода времени.

Современная критика *JIF* направлена на определение размеров окон цитируемости и публикации. В работе [Moed, et al., 1998] высказано мнение, что двухгодичное окно цитируемости искажает значение *JIF* даже при сравнении журналов, принадлежащих одной области исследований [Glanzel, Schoepfin, 1995]. Авторы данной работы аргументированно утверждают, что возраст, при котором *JIF* достигает своего наивысшего значения, не обязательно равен двум годам после публикации.

В настоящее время на практике *JIF* применяется как минимум в двух случаях. Во-первых, библиотеки используют эту метрику при формировании подписки на научную периодику. Во-вторых, на значение *JIF* ориентируются авторы, принимая решения о том,

в какой журнал представить свои работы. Действует общее правило: журналы с высоким значением импакт-фактора рассматриваются как наиболее престижные.

Автор работы [Kutateladze, 2009] высказывается против замены механизма экспертных суждений системой числовых индикаторов. Он приводит общую схему для вычисления двухлетних и пятилетних *JIF* и сравнивает формальные показатели математических журналов. Далее цитируется эта работа.

Для российской математики имеют значения следующие метрики. *MCQ* – показатель цитирования "Американского математического общества", основанный на БД журнала "Mathematical Reviews"; *JIF* – синхронный импакт-фактор корпорации "Thomson Reuters"; РИНЦ – Российский индекс научного цитирования на БД "Научной электронной библиотеки"; *MNRU* – импакт-фактор "Общероссийского математического портала Math-Net.Ru" на собственной базе.

Пусть $Q_{N,k}$ − число ссылок в году N на работы, опубликованные в рассматриваемом журнале в году N-k, P_N − число работ, опубликованных этим журналом за весь год N. В этих обозначениях величина MCQ_N за год N вычисляется по формуле

$$MCQ_N = [Q_{N,1} + Q_{N,2} + \ldots + Q_{N,5}] / [P_{N-1} + P_{N-2} + \ldots + P_{N-5}].$$

Импакт-факторы *JIF* и *MNRU* в году N обозначим через JIF_N и $MNRU_N$ соответственно. Тогда по определению имеем

$$JIF_N = MNRU_N = [Q_{N,1} + Q_{N,2}] / [P_{N-1} + P_{N-2}].$$

Итак, показатели *MCQ*, *JIF* и *MNRU* вычисляются по одной схеме, но при разной глубине учета данных. Важно отметить, что три показателя вычисляются на разных, хотя и пересекающихся, базах данных. Процедура вычисления индекса *РИНЦ* неизвестна.

Далее автор проводит сравнение *MCQ*, *JIF* и *MNRU* для групп отечественных и зарубежных математических журналов и делает

неутешительный вывод: "…вышеприведенные показатели, взятые за конкретный год, характеризуют, прежде всего, сами БД и лишь в небольшой части некоторые феномены реального функционирования науки".

10.1.3. Индекс оперативности *IMI.* *IMI* (immediacy index) журнала J в году Y определяется дробью

$$Cit_J(Y, Y) / Pub_J(Y).$$

Здесь числитель – число ссылок, полученных в году Y на публикации журнала J, вышедшие в этом же году; знаменатель – суммарное число публикаций, вышедших в журнале J в году Y.

Значение *IMI* указывает на то, как быстро содержание данного журнала было отмечено (т. е. как быстро на статьи, опубликованные в журнале, начали ссылаться другие авторы) и как высоко было оценено специалистами в данной области исследований [Davarpanah, Aslekia, 2008]. Следует отметить, что этот показатель может быть подвержен искажениям. Например, более вероятно, что статья, опубликованная в начале года, будет цитироваться большее число раз в течение этого года, нежели статья, опубликованная в конце того же года. Очевидно, что это будет влиять на показатель оперативности журнала, в котором опубликованы статьи. Данное рассуждение применимо к журналам, выходящим в свет ежемесячно и ежеквартально. Другим фактором, который должен учитываться, является "размер" журнала в смысле количества статей, публикуемых в течение года. Небольшие журналы в этом отношении имеют преимущество, так как знаменатель в формуле (общее количество статей) будет меньше, что может повлиять на показатель. С другой стороны, крупные журналы (которые имеют неблагоприятные условия по знаменателю в формуле) публикуют большое количество статей и поэтому имеют большую вероятность цитирования. Несмотря на то что индекс оперативно-

сти используется достаточно широко для оценки непосредственного воздействия журнала, он зависит от области научных исследований, к которой принадлежит журнал, уровня самоцитирования журнала и языка, на котором написаны статьи.

10.1.4. Половина периода цитирования и самоцитирование. Рассматриваются метрики, отражающие годы публикации ссылок. Метрика *CitIn* (*cited half-life*) определяет долговременную ценность статей, опубликованных журналом. Она вычисляется по медиане "возраста" ссылок (год, в котором была сделана ссылка), полученных на статьи журнала, опубликованного в изучаемом году, и означает количество лет (возможно, неполных), по прошествии которых журналом было получено 50 % общего количества ссылок. Метрика *CitOut* (*citing half-life*) определяет ценность статей, на которые ссылаются авторы журнала. Эта метрика также вычисляется по медиане "возраста" ссылок, сделанных в статьях журнала, опубликованного в изучаемом году. Обе метрики вычисляются по БД WoK, как правило, не ранее, чем через 10 лет после выхода журнала в свет.

Половина периода цитирования характеризует процесс старения публикаций. Как объясняет автор работы [Moed, et al., 1998], этот процесс можно рассматривать как комбинацию двух процессов: созревания и спада. В годы после публикации статьи количество сделанных на нее ссылок будет возрастать до достижения максимального уровня цитирования. За этой фазой созревания затем последует уменьшение числа сделанных ссылок, которое представляет собой спад.

Практическая значимость метрик *CitIn* и *CitOut* обсуждалась в работе [Ladwig, Sommese, 2005], авторы которой считают, что данные показатели могут быть полезны в процессе управления коллекциями журналов и создания архива. Они разработали статистическую модель, основанную на полупериоде цитирования

журналов для управления подпиской на онлайновую периодику в университетских библиотеках.

Самоцитирование часто составляет значительную долю ссылок на журнал. По оценкам “Thomson Reuters” (1994), приблизительно 13 % общего количества ссылок на журнал представляют собой ссылки журнала на себя. Тот факт, что журнал цитирует самого себя, оказывает прямое влияние на значение его *JIF*, поэтому активно обсуждается, включать ли самоцитирование в расчет *JIF* журнала.

Поведение самоцитирования было и остается предметом исследований. В работе [Biglu, 2008] проведен анализ самоцитирований 500 журналов, выбранных случайным образом за период с 2000 по 2005 гг. Установлено, что в течение этого периода количество самоцитирований, как и общее количество ссылок, возрастало. Также было обнаружено различие между журналами высокого или низкого ранга (в терминах количества ссылок на них). В данной выборке журналы высокого ранга имели низкий уровень самоцитирования (~2 %), тогда как у журналов низкого ранга уровень самоцитирования был высоким (~17 %).

Справка. Квантиль в математической статистике − такое число, что заданная случайная величина β не превышает его с фиксированной вероятностью. У одномерного распределения медианой называется квантиль уровня 0,5. Таким образом, медианой m является такое число, что $\mathrm{P}\{\beta < m\} = 0{,}5$ или $\mathrm{P}\{\beta \le m\} = 0{,}5$. Медиана определяет возможное значение, которое делит ранжированную совокупность данных (например, результатов измерений) на две равные части: 50 % “нижних” единиц ряда данных будут иметь значение признака не более, чем медиана, а “верхние” 50 % − значения признака не менее, чем медиана. Медиана согласуется с интуитивным пониманием “среднего” и может быть использована для центрирования распределения. Медиана робастна, поэтому ее использование предпочтительно для распределений с так называемыми “тяжелыми хвостами”.

10.2. Метрика Eigenfactor

Метрика Eigenfactor (*EF*), введенная в работе [Bergstrom, 2007], основана на вычислении "взвешенных" цитирований, которые определяют влияние журнала J на другие журналы. Новый метод подсчета цитирований был предложен для исправления недочета *JIF*, связанного с невозможностью учета престижа цитирующих журналов.

Метрика *EF* ранжирует журналы для определения наиболее влиятельных, т. е. журналов, на которые делается значительное число ссылок. Она основана на предпосылке, что одна цитата, поступившая из высокопрофессионального журнала, может быть более ценной, нежели несколько цитат из второстепенных журналов. Этот факт является очень важным, поскольку другие методы анализа, основанные на цитированиях, не учитывают то, откуда цитирования получены. Цитата из обзорной статьи, содержащей большое количество ссылок, будет оценена ниже, чем цитата из исследовательской статьи, которая цитирует только те работы, которые относятся к ее содержанию [Bergstrom, 2007]. Цитата цитате рознь!

Метрика *EF* основана на простой модели чтения научной литературы, в которой читатели следуют по цепочке ссылок, перемещаясь от одной статьи к другой и как следствие от одного журнала к другому. Так образуется граф статей и соответственно журналов. Алгоритм вычисления значения *EF* использует структуру графа для оценки важности каждого журнала. Учитывается пятилетний период активности цитирования, самоцитирования исключаются.

Метрика *EF* отражает влияние рассматриваемого журнала на научную периодику путем определения частоты его использования. Грубо говоря, *EF* журнала J в году X определяется в виде процентного отношения суммы "взвешенных" цитирований, по-

лученных всеми статьями журнала за рассматриваемый год, к сумме "взвешенных" цитирований, полученных за предыдущие пять лет статьями всех журналов, проиндексированных в данной БД. Очевидно, что объем журнала и число статей влияют на эту меру, поскольку крупные журналы будут иметь больше цитирований. Этот эффект учитывается путем масштабирования.

Метрика *EF* включена в отчеты "Thomson Reuters" по цитированию журналов и доступна начиная с 2007 г. Значение *EF* для конкретного журнала можно получить на сайте http://www.eigenfactor.org.

10.2.1. Вычисление *EF*. Ниже приведена общая схема вычисления *EF*, которая используется "Thompson Reuters" приблизительно для 8000 основных журналов. Главным блоком в этой схеме является вычисление весовых коэффициентов (см. [Eigenfactor]). Исходные данные о цитировании извлекаются из *JCR*. Изложение ведется согласно работе [Franceschet, 2010].

Зафиксируем год проведения измерения y. Пусть $C = (c_{i,j})$ – матрица взаимного цитирования журналов, в которой c_{ij} – число цитирований из статьи, опубликованной в журнале i за год y, на статьи, опубликованные в журнале j в течение фиксированного интервала времени Y, состоящего из пяти предыдущих лет, т. е. $Y = \{(y-1), (y-2),\ldots, (y-5)\}$. Таким образом, строка i матрицы C представляет цитирования журналом i остальных журналов, а столбец j содержит цитирования журнала j остальными журналами. Самоцитирование журналов исключается, чтобы не повышать рейтинг тех журналов, где этот факт был замечен. Для этого положим $c_{ij} = 0$ для всех $j = i$.

Матрицу C можно представить в виде сети цитирования, в которой узлы представляют собой журналы и имеется ребро от узла i к узлу j с весом c_{ij} тогда и только тогда, когда $c_{ij} > 0$. Очевидно,

что в этой сети может существовать узел (журнал) i, который не цитирует никакие другие журналы. Следовательно, если i – "повисший" узел, то у i нет исходящих ребер, и i-я строка матрицы цитирования имеет все элементы, равные 0.

Далее матрица цитирования C преобразуется в нормированную матрицу $H = (h_{ij})$, в которой все строки, не отвечающие "повисшим" узлам, нормированы на сумму строки:

$$h_{ij} = c_{ij} \Big/ \sum_j c_{ij} .$$

Пусть вектор-строка a такой, что $a_i > 0$ представляет собой число статей, опубликованных в журнале i за интервал Y, деленное на общее число статей, опубликованных за период Y всеми журналами. Следует отметить, что $\sum_i a_i = 1$. Преобразуем матрицу H в матрицу H', в которой все строки, отвечающие "повисшим" узлам, заменены вектором a. По построению матрица H' является стохастической.

Определим новую стохастическую матрицу P следующим образом:

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

Здесь матрица A состоит из одинаковых строк a, а значение параметра α в соответствии с алгоритмом Google PageRank [Brin, Page, 1998] положено равным 0,85. Пусть π – левый собственный вектор P, такой что $\pi = \pi P$. Поскольку матрица P является неприводимой, то по теореме Перрона-Фробениуса [Гантмахер, 1967, с. 354] такой вектор существует и является единственным.

Справка. Для вычисления собственного вектора матрицы, как правило, требуется многократно повторять умножение плотных матриц, что приводит к значительным вычислительным затратам. Вычислить собственный вектор матрицы P можно, например, с помо-

щью степенного метода [Писсанецки, 1988, с. 245], который дает достаточно точные результаты при решении задач с разряженными матрицами большого размера. Шаг итерации имеет вид

$$f^{(k+1)} = \alpha H f^{(k)} + \left[\alpha d f^{(k)} + (1-\alpha) \right] a,$$

где d – строковый вектор, отражающий "повисшие" столбцы: $d_i = 1$, если журнал никого не цитирует, $d_i = 0$, если имеются цитирования. В качестве $f^{(0)}$ берется столбец, каждый элемент которого равен $1/n$, где n – размерность матрицы. Итерационный процесс вычисления сходится к максимальному собственному вектору P. После каждой операции вычисляется $\tau = f^{(k+1)} - f^{(k)}$. Если $\tau < e$, то $f \approx f^{(k+1)}$ считается собственным вектором π. Обычно требуется не более 100 итераций для $e = 0{,}00001$. Рассуждения приведены для правого собственного вектора.

Вектор π, называемый "вектором влияния", содержит оценки, используемые для определения весов цитированиям, содержащимся в матрице H. Вектор r метрики EF вычисляется по формуле $r = \pi H$, т. е. взвешенные цитирования для журнала j определяются в виде

$$r_j = \sum_i \pi_i h_{ij}.$$

Значения нормируются, так что EF_j – это процент от всех взвешенных цитирований, т. е. EF_j определяется следующим образом:

$$EF_j = 100 \times r_j \Big/ \sum_j r_j .$$

Метрика EF является метрикой, зависимой от размера журнала: при равных условиях "большие" журналы имеют более высокие значения EF, поскольку в них больше статей, и поэтому можно ожидать, что они цитируются чаще.

10.2.2. Метрика *AI*. Метрика Article Influence (далее *AI*) определена в работе [Bergstrom, West, 2008] на основе метрики *EF*. Она оценивает важность статьи в соответствии с журналом, в ко-

тором та опубликована. Заметим, что в данном случае размер журнала не имеет значения. *AI* использует структуру всего графа цитирования для оценки важности каждого журнала. *AI* журнала *j* вычисляется по формуле

$$AI_j = 0,01 \times EF_j / a_j .$$

10.2.3. Аргументы в пользу *EF*. Приведем ключевые рассуждения о метриках *EF* и *AI,* изложенные в работе [Bergstrom, West, 2008]. Авторы делают акцент на том, что *EF* и *AI* являются агрегированными мерами уровня цитирования, т. е. учитывают все статьи всех журналов этой БД. Следовательно, оценка одной статьи должна выполняться с осторожностью, поскольку эти меры не сконцентрированы на индивидуальных статьях. Кроме того, следует помнить, что данные о цитировании не являются единственным методом оценки журналов, и для анализа цитирования могут использоваться другие методы, например прямые измерения круга читателей и использования журналов и статей.

В работе [Franceschet, 2010] приводятся аргументы, побуждающие к использованию метрики *EF*. Обращая внимание читателя на важность этой метрики, приводим краткое содержание аргументов.

10.2.3.1. *EF* придает веса цитированиям в соответствии с важностью цитирующих журналов. В отличие от нее, *JIF* просто суммирует цитирования, не придавая им веса. В результате метрика *JIF* отражает меру “популярности”, тогда как метрика *EF* включает понятия “престижность” и “авторитет”. Журнал, поддержанный престижными и авторитетными источниками, с большей вероятностью является престижным и авторитетным.

10.2.3.2. *EF* учитывает интенсивность ссылок цитирующих журналов. Интенсивность ссылок журнала представляет собой длину списка ссылок статей, опубликованных в журнале. Цитирование журналами с короткими библиографиями считается более

важным, чем цитирование журналами с высокой интенсивностью цитирования.

10.2.3.3. Для сбора информации о цитировании *EF* использует 5-летний временной интервал. Этот период превышает заданный интервал в два года, который обычно применяется для вычисления *JIF*, что позволяет собрать больше информации о цитировании журнала, особенно в областях исследований с длительным периодом цитирования опубликованных работ.

10.2.3.4. *EF* использует всю сеть цитирования. Значение *EF* вычисляется рекурсивно по оценкам цитирующих журналов, распространенных по всему графу цитирования. В отличие от этого для расчета *JIF* используется информация о цитировании, базирующаяся на данных из журналов-предшественников в графе цитирования.

10.2.3.5. *EF* не учитывает самоцитирование журнала.

10.2.3.6. *EF* имеет математическое обоснование. Помимо сказанного выше, стохастическую матрицу P можно представить как матрицу перехода в конечное число состояний (журналов) марковской цепи. Поскольку P примитивная, то на нее распространяется теорема Маркова, и вектор влияния π отвечает единственному стационарному распределению. Авторы работы [Bergstrom, et al., 2008] предлагают интерпретировать стохастический марковский процесс переходов от одного журнала к другому в процессе чтения научных журналов в терминах случайного блуждания по сети цитирования.

10.2.3.7. Сайт [Eigenfactor], являющийся открытым ресурсом, содержит информацию о цитировании приблизительно для 8 000 включенных в JCR журналов с 1995 г. “Thomson Reuters” с 2007 г. включила также показатель *EF* в *JCR* по разделам Science и Social Science. Сайт обновляется один раз в шесть месяцев.

10.3. Метрика *SNIP*

Метрика “Source-Normalized Impact per Paper” (*SNIP*) разработана с целью исправления недостатка метрики *JIF*, не учитывающей специфику научной области, к которой относится цитируемый журнал, а именно принятые нормы цитирования, скорость отклика на публикацию и степень покрытия области БД, по метаданным которой проводится вычисление значения этой метрики [Moed, 2010a]. Исходной БД для *SNIP* является Scopus. Документы в этой БД делятся на несколько типов, однако для вычисления *SNIP* рассматриваются только статьи, труды конференций и обзоры, остальные документы игнорируются.

Ключевым понятием является потенциал цитирования (R_j), который указывает, насколько часто публикации, относящиеся к данной тематической области, цитируют другие публикации. R_j определяется как среднее количество ссылок в статьях, относящихся к данной тематической области. В отличие от этого, *JIF* учитывает, насколько часто журналы, относящиеся к данной области, цитируются.

Пусть ${}^1a_j,..,\ {}^m a_j$ – документы, в которых имеются ссылки на журнал j. Этот набор документов составляет тематическую область журнала. Пусть ${}^i r_j$ – количество ссылок, содержащихся в документе ${}^i a_j$. Тогда потенциал цитирования журнала j в тематической области равен

$$R_j = \frac{\sum_{i=1}^{m} {}^i r_j}{m}.$$

Здесь тематическая область журнала определяется как набор публикаций, цитирующих этот журнал. Следует отметить, что цитирования самого журнала могут составлять только 1 % цитирований из публикаций этой тематической области. Потенциал ци-

тирования журнала в данной тематической области зависит от того, насколько полно БД покрывает эту область. Вводится понятие потенциала цитирования базы данных (R_j^{db}), который учитывает только количество ссылок на документы, опубликованные в журналах и проиндексированные в БД.

Пусть теперь ${}^{i}r_j^{db}$ – количество ссылок журнала i на документы, содержащиеся в db. Тогда потенциал цитирования базы для журналов в тематической области j вычисляется следующим образом:

$$R_j^{db} = \frac{\sum_{i=1}^{m} {}^{i}r_j^{db}}{m}.$$

Введем коэффициент f_j , такой что $R_j^{db} = f_j \times R_j$. Здесь f_j – это доля цитируемых статей, которые проиндексированы в db, т. е. область покрытия. Для определения значения *SNIP* необходимо знать среднее количество цитирований на публикацию для выбранного журнала. Этот параметр обозначим *RIP* (*raw impact per paper*). Пусть C_j – количество всех цитирований рассматриваемого года на статьи в журнале j за три предыдущих года, A_j – количество всех документов, опубликованных за эти три года в журнале j. Тогда

$$RIP_j = C_j / A_j.$$

Пусть N – количество журналов в БД, M^{db} – медиана всех потенциалов цитирования базы для тематических областей журналов:

$$M^{db} = \text{Med}\ \{R_j^{db}\},\ j=1,\ldots N.$$

Относительный потенциал цитирования базы для тематической области j получаем в результате деления потенциала цитирования базы для тематической области на медиану (Med):

$$R_j^{db} / M^{db}.$$

Значение метрики *SNIP* для журнала *j* определяется дробью

$$SNIP_j = (RIP_j) / (R_j^{db} / M^{db}).$$

Из определения следует, что если $R_j^{db} / M^{db} = 1$, то значения метрик *SNIP* и *RIP* совпадают. Для журналов из тематической области, в которой принято большое количество цитирований, *SNIP* < *RIP*, а для областей с небольшим количеством цитирований *SNIP* > *RIP*.

Сильные стороны *SNIP*:

а) Границы тематической области не определены заранее, а полностью базируются на взаимосвязи цитирований. Они вытекают из взаимоотношений цитирования статья − статья, а не журнал − журнал.

б) Тематическое поле может быть точно определено, даже если в нем важную роль играют универсальные или мультидисциплинарные, покрывающие несколько научных областей журналы.

в) Метрика корректирует различие в практике цитирования, принятое в разных тематических областях, особенно частоту цитирования и скорость отклика.

г) Метрика корректирует различие в покрытии базой данных тематических областей; малое покрытие к малому потенциалу цитирования базы и как следствие ведет к более высоким значениям *SNIP* по сравнению с индикатором цитирования на статью *RIP*.

д) *SNIP* корректирует не только различие между тематическими областями журналов, но и различие между журналами, охватывающими разные предметы и объекты исследования внутри тематических областей.

е) *SNIP* базируется на понятии цитирования от проверенной рецензентами публикации к проверенной рецензентами публикации. Это делает его менее зависимым от манипуляций, особен-

но со стороны редакторов, “свободные” цитирования и редакторские самоцитирования не включаются.

ж) Позволяет оценивать междисциплинарные журналы, такие как “*Nature*” или “*Science*”.

з) Индекс умеренно зависит от изменения временных окон цитируемости и публикации.

В заключение приведем цитату из работы [Писляков, 2011]: “*SNIP* появился как попытка преодолеть зависимость импакт-фактора от активности цитирования в той или иной научной области. *SNIP* – это нормализованный по источникам уровень цитируемости статьи. Каждое цитирование дает “единицу”, но ее вес пересчитан относительно потенциала цитирования в каждой дисциплине и даже в каждом “персональном окружении” конкретного журнала. *SNIP* показывает долю цитирований, которые ухватил себе конкретный журнал из всего массива доступных ссылок…”.

10.4. Метрика *SJR*

Метрика важности журналов *SJR* (SCImago Journal Rank, [González-Pereira, et al., 2009]) относится к новому типу метрик, которые учитывают не только количество полученных цитирований, но и научное влияние этих источников, т. е. является функцией, комбинирующей количество и качество. Основная идея, позволяющая достигнуть такого результата, заключается во взвешивании цитирований на основе “важности” журналов, их производящих. Эта важность вычисляется рекурсивно, т. е. важными журналами будут те, которые, в свою очередь, получают много цитирований из других важных журналов.

Авторы метрики *SJR* пользуются подходом и вычислительной моделью, предложенной в работе [Brin, Page, 1998], т. е. рассматривают сеть цитирований как связанный граф, где узлы (англ. node) соответствуют журналам, а связь между узлами отражает вероятность того, что исследователь переходит от журнала к жур-

налу в соответствии со свободно выбранной ссылкой. Весь процесс представляет собой “свободное блуждание”, начинающееся с произвольно выбранного журнала.

Метод определяет и итеративную процедуру, которая начинается с первоначальных, предварительно установленных величин важности журналов, пересчитывает эти значения до достижения устойчивого состояния. Важность журналов перераспределяется с каждой итерацией. Процесс можно описать с помощью обобщенной формулы

$$P\left(Node_i,\ it_k \right) = (1-\lambda)/N +$$

$$+\lambda \times \sum_{j=1}^{N} \left(Connection_{(i,j)} \right) \times P\left(Node_j,\ it_{k-1} \right).$$

Таким образом, важность узла i во время итерации k – это сумма относительных важностей, передаваемых всеми связанными с i узлами. Важность, передаваемая от j к i, взвешивается относительно силы связи между ними (*Connection*). Сила связи определяется долей количества цитирований от j к i в течение рассматриваемого года. Фактор случайного перехода (англ. jump factor) $(1-\lambda)/N$ вводится для обеспечения сходимости и обозначает вероятность выбора журнала, в случае если исследователь не следует ссылкам в рассматриваемых публикациях, а выбирает журнал случайным образом.

В качестве источника данных рассматривается БД Scopus, выбранная по следующим показателям:

а) Охват журналов.

б) Соотношение между первичной (статьи, обзоры и труды конференций) и вторичной продукцией для каждого журнала.

в) Установленные критерии для определения типа документов.

г) Аккуратность переходов от ссылок к цитируемым единицам.

Документы классифицируются по научным областям и категориям. Рассматриваются 295 областей, которые группируются в 26 более крупных областей. Имеется и обобщенная область, содержащая мультидисциплинарные журналы, такие как “Nature” и “Science”. В свою очередь области группируются по четырем категориям: науки о жизни, физика, социология и здравоохранение.

Окно цитируемости устанавливается равным трем годам, так что рассматриваются ссылки, встречающиеся в журналах года расчета и относящиеся к публикациям за три предыдущих года. Для того чтобы предотвратить излишнее влияние самоцитирований, количество ссылок, которые направляются из журнала на него же, лимитировано 33 % общего числа ссылок.

Вычисление проводится по приведенной выше схеме. Первоначально всем журналам присваивается равный престиж. Далее, согласно итеративной схеме престиж перераспределяется до тех пор, пока разность значений между последовательными итерациями не превысит некоторого предварительно заданного порога.

Индикатор *SJR* вычисляется в два этапа. Сначала вычисляется мера *Prestige SJR* (*PSJR*), которая отражает престиж всего журнала и зависит от его размеров. Независимая метрика *SJR* получается путем нормализации на втором этапе.

Первый этап. Сначала все журналы получают престиж, равный $1/N$, где N – количество журналов в базе данных. Затем начинается итеративная процедура. Каждая итерация определяет новые значения престижа каждому журналу в соответствии с тремя критериями: 1) минимальный престиж, определяемый тем, что журнал просто включен в базу; 2) публикационный престиж, зависящий от количества публикаций, включенных в базу; 3) престиж от цитирований, зависящий от “важности” журналов, от которых получены цитирования.

Таким образом, формулу для определения ранга журнала i можно представить в виде

$$PSJR_i = P1+P2+P3,$$

где

$$P1 = \left(1-d-e\right)/N, \quad P2 = e\times\left(\ Art_i \Big/ \sum_{j=1}^{N} Art_j\ \right),$$

N − количество журналов в базе; e, d – константы, представляющие собой веса для количества престижа, получаемого от публикаций и цитирований соответственно (d =0,9, e=0,0999); Art_j – количество первичных единиц (статьи, обзоры и труды конференций) журнала j; компоненты $P1$, $P2$ являются константами относительно итераций;

$$P3 = d\times\left[\ \sum_{j=1}^{N} C_{ji}\times\left(\ PSJR_j/C_j\ \right)\times CF + \right.$$
$$\left. +\left(\ Art_i \Big/ \sum_{j=1}^{N} Art_j\ \right)\times \sum_{k\in DN} PSJR_k\ \right], \qquad (10.4.1)$$

C_j – общее количество ссылок в журнале j; C_{ji} – количество ссылок из журнала j на журнал i; CF – фактор корреляции, определяемый далее.

В соотношении (10.4.1) первое слагаемое в квадратных скобках $\sum_{j=1}^{N} C_{ji}\times\left(\ PSJR_j/C_j\ \right)\times CF$ представляет собой престиж журнала i, полученный от цитирований его публикаций из других журналов. Каждое цитирование взвешивается престижем, полученным цитирующим журналом за время предыдущих операций и деленным на количество ссылок, найденных в журнале (на публи-

кации любой степени давности). Поскольку при перераспределении престижа используются только публикации в трехгодичном окне, вводится фактор корреляции, для того чтобы восполнить потерю престижа, соответствующего оставшимся цитированиям. Фактор корреляции находим по формуле

$$CF = \left(1 - \sum_{k \in DN} PSJR_k \right) \Big/ \sum_{h=1}^{N} \sum_{k=1}^{N} C_{kh} \times \left(PSJR_k / C_k \right).$$

Здесь DN – множество "отвисших узлов" (англ. dangling nodes), ассоциирующихся с журналами, которые никого не цитируют. Знаменатель соответствует количеству престижа, распределенного между цитированиями, попадающими в трехгодичное окно, а числитель соответствует тому, сколько могло бы быть престижа, т. е. единица минус престиж, сосредоточенный в отвисших журналах.

Второе слагаемое в квадратных скобках выражения (10.4.1)

$$\left(Art_i \Big/ \sum_{j=1}^{N} Art_j \right) \times \sum_{k \in DN} PSJR_k$$

перераспределяет престиж, аккумулирующийся в отвисших журналах из множества DN пропорционально количеству первичных единиц журнала i. После каждой итерации сумма престижей нормализуется до единицы. Процесс заканчивается, когда разность между значениями престижа для всех журналов, полученными за две следующие друг за другом итерации, становится незначительной.

Второй этап. Метрика *PSJR*, вычисляемая во время первого этапа, определяет престиж всего журнала. Эта метрика не пригодна для сравнения журналов различного размера, так как более толстые журналы имеют тенденцию иметь больше цитирований, поэтому она нормализуется по количеству первичных единиц:

$$SJR_i = c \times (PSJR_i / Art_i).$$

Метод вычисления *SJR* аналогичен методу, применяемому при вычислении метрики Article Influence, *AI* (см. п. 10.2). Основное отличие заключается в том, что, во-первых, при вычислении *SJR* самоцитирования журналов ограничиваются 33 %, тогда как при вычислении *AI* такие цитирования опускаются; во-вторых, при вычислении *SJR* нормализация проводится по всем ссылкам в цитирующем журнале, независимо от окна цитирования, а при вычислении *AI* − только по идентифицированным ссылкам. Кроме того, индикатор *SJR* вычисляется на основе данных из базы Scopus, являющейся наиболее полной в библиографическом смысле, с использованием трехгодичного окна цитируемости, а *AI* – на основе данных из Web of Science с использованием окна цитируемости, равного пяти годам.

В работе приводятся результаты сравнения новой метрики с метрикой “импакт-фактор” для журналов. Делается вывод о том, что, несмотря на сильную корреляцию этих метрик, *SJR* более точен, он сокращает ранг журналов, цитирование которых больше их научного влияния.

10.5. Индекс Хирша для журнала (*hJ*)

Данная метрика, предназначенная для оценки влияния того или иного журнала на определенную область исследований за фиксированное время, разработана как дополнение к *JIF* [Braun, et al., 2006]. Ее преимущества перед *JIF* состоят в следующем. Во-первых, метрика *hJ* менее “чувствительна” к большому числу недостаточно высоко цитируемых статей, а также к чрезмерно цитируемым статьям. Во-вторых, авторы трактуют как положительный тот факт, что *hJ* “балансирует” количество публикаций и уровень цитирования, что снижает переоценку некоторых журналов. Также отметим, что для вычисления *h*-индекса автора рассматривается вся карьера ученого, тогда как при вычислении *hJ* учитываются публикации за фиксированный период времени, например за один год.

Метрика *hJ* вычисляется следующим образом. Выберем журнал *J*, имеющий *N* публикаций за период времени, так чтобы число цитирований было “достаточно представительным”. Публикации A_i упорядочим в порядке убывания цитирований. Из этой последовательности выберем *h* статей, имеющих, по крайней мере, *h* цитирований, так, чтобы остальные (*N*–*h*) статей имели каждая меньше или равное *h* число цитирований. Определенное таким образом число *h* будет значением метрики *hJ*. В этом случае говорят, что журнал *J* за период *T* имеет индекс Хирша журнала *hJ*, который вычислен на основе информации базы данных “имя БД”, и сообщают дату проведения вычислений − “число, месяц, год”.

Анализируя значения метрик *hJ* журналов, относящихся к одной и той же области исследований, можно ранжировать эти журналы. Важно понимать, что для сравнения исходную информацию необходимо брать из одного источника, например WoS. Подчеркнем, что часто журналы, популярные в некоторой области и имеющие высокое значение *hJ,* могут иметь низкое значение *JIF*.

10.6. Метрика связности *R*(*A, B, T*)

Взаимосвязи между научными журналами изучались в работе [Pudovkin, Garfield, 2002]. Метод, предложенный авторами данной работы, основан на подсчете количества ссылок за данный период времени. Вводится “метрика связности” $R(A,B,T)$ для журналов *A* и *B* за период времени *T*, которая определяется равенством

$$R(A,B,T)=Cit(A,B,T) / (P(B,T)\times Cit(A,T)).$$

Здесь $Cit(A,B,T)$ означает, сколько раз *A* цитирует *B* за время *T*; $P(B,T)$ – количество статей, опубликованных в журнале *B* за время *T*; $Cit(A,T)$ − общее число ссылок в журнале *A* за время *T*.

С помощью метрики $R(A,B,T)$ можно определить, какие журналы и насколько тесно связаны с данным (целевым) журналом на

основе ссылок, которые он делает и которые делаются на него. Следует отметить, что существуют другие методы изучения меры связи между журналами, например метод анализа совместного цитирования, также применяются методы кластерного и фрактального анализа.

10.7. Эволюция журнальных метрик

> Одна метрика не может отразить все области научной деятельности и обеспечить единственную совершенную систему ранжирования.
>
> *H. Moed* [*Moed, 2010b*]

10.7.1. *JIF* как оценка заслуг. В большинстве сфер научной деятельности анонсирование новых результатов происходит через статьи, труды конференций, книги и патенты. В этих публикациях имеются ссылки на более ранние статьи, которые оказались полезны или на которые имеется ответ в данной статье, что привлекает ответное цитирование. Таким образом, можно измерять количество цитирований, для того чтобы оценить важность публикации. В 1955 г. Гарфилд высказал идею оценки важности журналов с помощью подсчета количества его цитирований, а в 1961 г. были опубликованы Science Citation Index и Journal Citation Reports, содержащие импакт-факторы журналов (*JIF*). С этого времени редакторы стали рассматривать метрику *JIF* как показатель производительности журналов.

Матрица *JIF* имеет беспрецедентный уровень влияния на редакторов, которые фокусируют свое внимание на повышении именно количества цитирований как на основном показателе качества. Однако эта метрика является манипулируемой. Например, известно, что обзоры цитируются чаще других работ, а обзорные журналы имеют более высокие *JIF*, чем другие журналы этой же области деятельности. Журналы, стремящиеся повысить *JIF*, могут начать

печатать больше обзоров. Открытым остается вопрос, полезно ли это для общества и (или) науки. Кроме того, при вычислении *JIF* в качестве числителя рассматриваются все публикации, в том числе письма и редакторские заметки, а в знаменателе подсчитывается сумма только статей и обзоров.

В ряде случаев *JIF* считают мерой скорости реакции на новые публикации и их количество. Однако это утверждение не всегда верно, поскольку в различных областях науки приняты различные нормы цитирования и скорость реакции, в результате, например, журналы в области науки о живой природе имеют более высокий *JIF*, чем журналы математические или общественных наук.

В докладе "Статистики цитирования" [Игра в цыфирь, 2011] говорится: "Двухгодичный срок при подсчете *JIF* был взят, чтобы сделать статистику своевременной. Для ряда областей, например для биомедицины, это подходит, так как большинство публикаций быстро получают отклик. Но для таких, как математика, этого срока недостаточно. Изучение цитирований в математических журналах (Math Reviews Citation database) показывает, что около 90 % цитирований выпадает из двухгодичного окна. Поэтому *JIF* в этой области науки учитывает только 10 % цитирований и пропускает основную массу". "Индексы Thomson Scientific не покрывают и половины математических журналов, упомянутых в двух крупнейших математических реферативных журналах, "Mathematical Reviews" и "Zentralblatt". Там же далее: "Если уже ясно, что не стоит оценивать статьи на основании *JIF* журналов, то отсюда следует, что не имеет смысла оценивать на основании *JIF* и авторов этих статей, программ, согласно которым они работают, а особенно дисциплин, которые они представляют. *JIF* без дополнительной информации – слишком грубая оценка для такой деятельности. И, разумеется, оценка людей – это не то же самое, что оценка их публикаций. Но, если все же хочется ранжировать каче-

ство конкретной публикации на основе цитирования, следует считать цитирования именно к этой публикации".

10.7.2. Потребность в новых метриках. Оценка производительности и ранжирование касаются большинства людей, относящихся к научному миру. Любой уровень академической жизни – от индивидуальных исследователей и групп до учреждений и даже целых стран или регионов – все чаще оценивается и ранжируется в целях определения выгодности инвестирования. Несмотря на то что никто не сомневается в необходимости оценки производительности, в академических кругах не утихают дебаты о том, как эти оценки производятся и для чего они используются [Journal Metrics, 2010].

Финансирование исследований поступает из различных источников, однако каждый хочет, чтобы вложения окупились. Государственные органы с возрастающим вниманием следят за научными исследованиями и разработками, рассматривая их в качестве двигателей экономического прогресса. Инвесторы и акционеры используют различные индикаторы производительности, для того чтобы определить, где и кем именно осуществлены лучшие исследования, они также нуждаются во все более подробных отчетах о том, как используются фонды. Руководители, ответственные за принятие решения, ищут прозрачные методы измерения социальных и экономических результатов различных сторон исследований. Чиновники, занимающиеся финансированием науки, в большей мере начинают относиться к научным институтам как к коммерческим организациям. Они нуждаются в метриках, которые дают информацию о различных аспектах производительности научного труда, для сравнения своих показателей с показателями конкурентов.

На индивидуальном уровне многие ученые также используют метрики, чтобы оценить собственную производительность и карь-

ерный рост. В области управления журналами редакторы используют метрики, чтобы определить рейтинг журнала, а исследователи изучают информацию о ранжировании, чтобы оценить выгоды от публикаций. Ранжирование на уровне учреждений используется для многих целей. Политики используют его для измерения экономического потенциала нации, научные учреждения − для распределения ресурсов и финансирования, ученые − для поддержания репутации и статуса, студенты − для определения потенциального места обучения, а государственные и частные акционеры − для принятия решений по размещению инвестиций. Таким образом, измерение производительности и результатов научного труда является востребованным и актуальным. Это стимулирует исследователей к разработкам новых, более точных метрик и устранению замеченных недостатков в уже апробированных методах измерений.

Одной из наиболее известных является метрика *h-индекс*, первоначально разработанная для оценки ученых. Ниже приведен пример, который демонстрирует один из недостатков этой метрики и предостерегает от неправильного ее использования.

Рассмотрим публикационную активность трех ученых.

- Автор A1 опубликовал семь статей, пять из них цитируются не менее пяти раз. *h*-индекс автора A1 равен пяти.
- Автор A2 также опубликовал семь статей с такой же частотой цитирования, но у него есть еще много статей, цитируемых четыре или менее раз. Таким образом, автор A2 проделал больший объем работы, чем автор A1, однако имеет такой же *h*-индекс, что и A1.
- Автор A3 опубликовал семь работ, две из них имеют “высокое” (значительно больше h) количество цитирований, но у него такой же *h*-индекс, как у автора A1.

Видно, что различные распределения цитирований могут порождать один и тот же *h*-индекс, однако сомнительно, что они от-

ражают одинаковую производительность. Таким образом, метрика *h*-индекс не пригодна для оценки ученых с длинной карьерой и работающих в области, где принята высокая активность цитирования. Если *h*-индекс используется в качестве единственной метрики, то реальная картина будет искажена.

В последнее время появилось множество "новых" метрик, альтернативных *JIF*, для ранжирования журналов. В основном они пытаются преодолеть недостатки *JIF*, связанные с перекосом в сторону много цитирующих областей науки, характерными качествами которых являются обильное цитирование и высокая скорость ответной реакции на публикацию. Среди них Source-Normalized Impact per Paper (*SNIP*), Eigenfactor (*EF*) & Article Influence (*AI*) и SCImago Journal Rank (*SJR*).

Метрика *SNIP* разработана в целях корректировки недостатка *JIF,* связанного со скоростью ответной реакции на публикацию. *SNIP* – это результат деления среднего количества цитирований на статью на "потенциал цитирования" в рассматриваемой тематической области (англ. subject field). Потенциал цитирования − это оценка среднего количества цитирований на статью, которое ожидается получить соответственно тематической области. (Здесь тематическая область журнала – это множество документов, его цитирующих.)

Метрика *SJR* устанавливает взвешенную оценку цитированиям, получаемым от материалов, опубликованных в данном журнале. Согласно этой метрике цитирования из журнала с большим значением *SJR* "весят больше".

Метрика *AI* является результатом деления *EF* журнала на нормализованное количество статей, в нем печатающихся. Эту метрику можно интерпретировать как процент времени, которое будет потрачено на чтение журнала, если использовать модель "свободного блуждания" по сети цитирования.

Все “новые” метрики делают поправку на контекст, в частности *SJR* и *AI* достигают этого, опираясь на базу данных, что позволяет проводить сравнения между областями исследований. Кроме того, при вычислении *SJR* и *SNIP* размер окна увеличен до трех лет, а при вычислении *AI* − до пяти. Наконец, *SNIP, SJR* и *EF & AI* являются открытыми метриками, соответствующие методы их вычисления и применения опубликованы и общедоступны.

10.7.3. Сравнение некоторых журнальных метрик. Погоня за количеством цитирований побуждает авторов обращать внимание на то, где печататься, а для выбора журнала нужна информация о его ранге. В библиометрических измерениях, например при ранжировании журналов, недостаточно одной метрики. Хорошим примером являются метрики *SNIP* и *SJR*, которые используются как взаимодополняющие и тем самым подчеркивают нежелательность использования одной из них.

По мнению авторов работы [Journal Metrics, 2010], для получения и последующего обсуждения оценки производительности научных исследований необходимо использовать максимальное число доступных метрик. Научные исследования слишком многогранны и важны, чтобы пренебрегать такой возможностью.

В табл. 10.1 использованы следующие обозначения и сокращения:

– A (article) – статья; L (letter) – письмо; R (report) – доклад на конференции; S (survey) – обзор.

– JCR (Journal Citation Reports) – отчет, содержащий *JIF* ведущих журналов.

– Scopus (версия официального названия: SciVerse Scopus) − библиографическая и реферативная БД, индексирующая научные журналы, материалы конференций и серийные книжные издания. Разработчиком и владельцем Scopus является издательская корпорация “Elsevier”.

– a8 и c8 − увеличение числа обзоров в журнале увеличивает значение метрики.

– a10 − не вводит поправку на различия в степени покрытия базой данных предметных областей.

– b8 − сдерживается значением *EF* журналов, из которых цитируются обзоры.

– b10 и d10 − влияние распределяется согласно областям, наиболее полно представленным в БД.

Таблица 10.1

Параметры некоторых журнальных метрик

№ п/п	Параметры	Метрики журнала			
		JIF (a)	***EF&AI*** (b)	***SNIP*** (c)	***SJR*** (d)
1	Окно цитируемости	2 года	5 лет	3 года	3 года
2	Окно публикации	1 год	1 год	1 год	1 год
3	Учет самоцитирования	Есть	Нет	Есть	Есть, см. d3
4	Нормализация по предметной области	Нет	Есть	Есть	Есть
5	Типы публикаций, используемых в числителе	Любые публикации	Любые публикации	A, R, S	A, R, S
6	Типы публикаций, используемых в знаменателе	A, S	A, L, S	A, R, S	A, R, S
7	Статус источника цитирования	Не учитывается	Учитывается	Не учитывается	Учитывается
8	Эффект от учета цитирований обзоров	См. a8	См. b8	См. c8	См. d8
9	Основная БД	JCR	JCR	Scopus	Scopus
10	Эффект от покрытия БД предметной области	См. a10	См. b10	См. c10	См. d10

– c10 − вводит поправку на различия в степени покрытия базой данных предметных областей.

– d3 − процент самоцитирований для журнала ограничивается 33 %.

– d8 − Цитирования взвешиваются на основе престижа журнала, от которого они поступают.

11.1. Бенфорд, Франк (*Benford, Frank*) – американский инженер, физик [Wikipedia].

Достижения. В 1938 г. Ф. Бенфорд открыл закон, получивший его имя – закон Бенфорда (Benford's Law), или, другое название, “феномен первой цифры”. Согласно этому закону в статистических данных первая цифра d ($d \in \{1,\ldots 9\}$) встречается с вероятностью $P(d) = \log (d+1) - \log (d) = \log (1 + 1/d)$. Впервые проявление этого закона заметил американский астроном С. Ньюкомба в 1881 г. Он обнаружил, что справочники, содержащие логарифмические таблицы, истрепаны там, где содержатся логарифмы чисел, начинающихся с единицы, и целы для чисел, начинающихся на цифру девять. Закон оставался незамеченным до тех пор, пока Ф. Бенфорд, незнакомый, очевидно, с работой С. Ньюкомба, не вывел этот же закон и не опубликовал его в 1938 г., обосновав огромным объемом данных. Он проанализировал информацию о площадях бассейнов 335 рек, удельной теплоемкости и молекулярном весе сотен химических соединений, номерах домов первых 342 лиц, указанных в справочнике. Анализ чисел показал, что единица является первой значащей цифрой с вероятностью не 1/9, как следовало бы ожидать, а приблизительно 1/3. Закон Бенфорда применим, например, к банковским счетам, остаткам товаров на складах, ценам на акции, численности населения, смертности, площадям стран, высотам высотных сооружений и т.п.

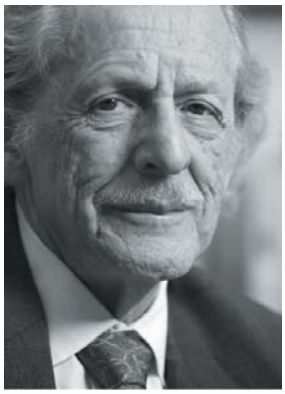
Биографическая справка. Франк Бенфорд родился 29 мая 1883 г. в Джонстауне, Пенсильвания. В 1910 г. закончил Мичиган-

ский университет. Работал в “General Electric”. Создал инструмент для измерения индекса рефракции стекла. Ф. Бенфорд умер 4 декабря 1948 г. в Нью-Йорке.

11.2. Брэдфорд, Самюэль (*Bradford, Samuel*) – английский математик и библиотекарь [Wikipedia].

Достижения. В 1934 г. установил эмпирический закон, названный его именем (Bradford’s Law). Рассмотрим множество журналов, в которых опубликованы статьи по выбранной тематике. Упорядочим это множество по возрастанию числа статей. Закон гласит, что ранжированное таким образом множество журналов можно разбить на три группы так, что в каждой группе будет одинаковое количество статей по заданной теме. При этом количество журналов в каждой группе будет соответствовать отношению $1{:}n{:}n^2$, где n − некоторое число, большее единицы. Несомненной заслугой С. Брэдфорда является успешное внедрение в документооборот крупнейшей европейской научной библиотеки десятичной системы классификации.

Биографическая справка. Самюэль Брэдфорд родился в Лондоне 10 января 1878 г. С 1899 г. работал в библиотеке Музея науки (South Kensington). С 1925 по 1937 гг. руководил библиотекой музея. С. Брэдфорд умер 13 ноября 1948 г.



11.3. Гарфилд, Юджин (*Garfield, Eugene*) − американский ученый, один из основателей библиометрии [Wikipedia].

Достижения. Ю.Гарфилд, следуя идеям изобретателя гипертекста В. Буша, изложенным в 1945 г. в работе “Как мы можем думать” (русский перевод см. [Буш, Ванневар]), разработал Science Citation Index (SCI, индекс научного цитирования). Ю. Гарфилду принадлежит

открытие: небольшое количество журналов типа “Nature” и “Science” служит ядром всех естественных наук. Следует отметить, что такое явление не наблюдается для гуманитарных и общественных наук.

Биографическая справка. Ю. Гарфилд родился 16 сентября 1925 г. в Нью-Йорке. В 1960 г. Ю. Гарфилд основал Институт научной информации (ISI) − коммерческую организацию, занимающуюся вопросами разработки библиографических БД научных публикаций и их индексированием, а также определением индекса цитируемости, импакт-фактора и других статистических показателей выполнения научных работ. В 1961 г. Ю. Гарфилд защитил диссертацию по структурной лингвистике в Пенсильванском университете. В 2007 г. Ю. Гарфилд создал программный продукт HistCite для библиометрического анализа. Он является редактором и издателем журнала “The Scientist”.

11.4. Лотка, Альфред (*Lotka, Alfred*) − американский математик, статистик [Lotka, CV].

Достижения. Альфред Лотка эмпирическим путем установил, что число авторов, написавших n статей, составляет $1/n^a$ от числа авторов, написавших одну статью, где значение a близко к двум. А. Лотка получил известность за работы в области динамики популяций. Он изучал процесс смены поколений, дал современное аналитическое выражение возраста поколения, анализировал процесс демографического развития семьи. А. Лотка ввел интегральное уравнение воспроизводства населения. Цикл этих работ принес Лотке известность как основателю современного демографического анализа и автору теории стабильного населения, которую он впоследствии распространил на процессы развития самообновляющихся совокупностей. Исследовал экономические и демографические аспек-

ты здравоохранения и эволюции продолжительности жизни, заложив основы экономической демографии.

Биографическая справка. Альфред Лотка родился 2 марта 1880 г. в городе Лемберг, Австро-Венгрия (сейчас Львов, Украина). В 1901 г. окончил Бирмингемский университет в Великобритании. В 1912 г. защитил диссертацию. С 1924 по 1947 гг. руководил математическими исследованиями в американской страховой компании "Метрополитен лайф иншуренс". С 1938 по 1939 гг. являлся президентом "Американской ассоциации населения". С 1942 г. руководил "Американской статистической ассоциацией", являлся членом "Международного союза по научному изучению населения". Альфред Лотка умер 5 декабря 1949 г.

11.5. Мандельброт, Бенуа (Mandelbrot, Benoît) − французский и американский математик, создатель фрактальной геометрии [Wikipedia].

Достижения. В 1952 г. Бенуа Мандельброт в Сорбонне защитил диссертацию, которая положила начало его междисциплинарным исследованиям. В результате соединения элементов теории информации, лингвистики и теории вероятностей в диссертации выводились законы статистической структуры языка. В основу были положены работы гарвардского филолога Дж. К. Ципфа, который в 1949 г. эмпирически установил, что в любом достаточно объемном содержательном тексте частоты употребления слов описываются степенным законом. Мандельброт уточнил закон Ципфа и показал, что он следует из принципа наименьших усилий, если предположить, что при передаче имеющегося количества информации (в смысле Шеннона) и говорящий, и слушающий стремятся затратить как можно меньше усилий, минимизировав среднюю стоимость слова. В области экономики Б. Мандельброт изучил статистику цен на хлопок за большой период времени (бо-

лее 100 лет) и смог выявить тенденцию их изменения. Он проследил симметрию в длительных и кратковременных колебаниях цены. По сути, для решения этой проблемы Б. Мандельброт применил зачатки своего рекурсивного (фрактального) метода. Б. Мандельброт ввел в оборот термин “фрактал” (от слова “подобный”).

Биографическая справка. Бенуа Мандельброт родился в Варшаве в 1924 г.

В 1936 г. эмигрировал во Францию и попал под влияние своего дяди Шолема Мандельбройта, члена группы “Николя Бурбаки”. Уже в школе Б. Мандельброт обладал великолепным пространственным воображением, позволявшим ему даже алгебраические задачи решать геометрическим способом. В 1945 г. Б. Мандельброт поступает в Сорбонну и защищает диссертацию. Затем он переезжает в США, где заканчивает Калифорнийский институт технологии. С 1958 г. Б. Мандельброт работает в научно-исследовательском центре IBM. Занимался лингвистикой, теорией игр, экономикой и астрономией. Бенуа Мандельброт умер 14 октября 2010 г. в Кембридже (Массачусетс, США).

11.6. Маршакова-Шайкевич Ирина Владимировна (*Marshakova-Shaikevich, Irina*) [Маршакова-Шайкевич, CV].

Достижения. И. В. Маршакова-Шайкевич является автором метода коцитирования (co-citation − публикация 1973 г.), который используется при построении карт науки в Институте научной информации ISI США , а также в базах Web of Knowledge. Автор метода расчета стандартного показателя воздействия для научного журнала и области знания на основе данных, представленных в базе Journal Citation Reports (публикации 1988 г. – на русском языке, 1996 г. – на английском языке). Этот метод позволяет оценивать журнал незави-

симо от области знания и, в частности, используется для оценки национального корпуса научной периодики.

Биографическая справка. В 1973 г. опубликована основополагающая работа [Маршакова, 1973] “Система связей между документами, построенная на основе ссылок”. В 1975 г. в ВИНИТИ РАН защищена диссертация на соискание ученой степени кандидата технических наук “Алгоритмические классификации динамических документальных массивов информации”. В 1988 г. опубликована монография [Маршакова, 1988] “Система цитирования научной литературы как средство слежения за развитием науки”. В 1993 г. в Институте философии РАН защищена диссертация на соискание ученой степени доктора философских наук “Методы количественного анализа научного знания”. В 2002 г. опубликована работа [Маршакова-Шайкевич, 2002] “Вклад России в развитие мировой науки”. В 2004 г. вышла работа [Маршакова-Шайкевич, 2004] “Классификация научных журналов методом коцитирования”. В 2008 г. опубликована монография [Маршакова-Шайкевич, 2008] “Россия в мировой науке. Библиометрический анализ”. В настоящее время И. В. Маршакова-Шайкевич работает ведущим научным сотрудником в Институте философии РАН.

11.7. Налимов Василий Васильевич (*Nalimov, Vasili*) − видный советский и российский ученый [Wikipedia].

Достижения. В. В. Налимов является одним из создателей научного направления наукометрия. В соавторстве с З. М. Мульченко им написана работа [Налимов, Мульченко, 1969]. В. В. Налимов владел английским, немецким, французским, польским, арабским языками. Он подготовил 18 кандидатов и трех докторов наук. Действительный член РАЕН с 1996 г., член редколлегий международных научных журналов. Награжден медалью Дерека де Солла Прайса (1985 г.), почетным знаком РАЕН “За заслуги в развитии науки и экономики” (1996 г.), удостоен почетного звания “Классик цитирования”

(по оценкам ISI, 1990 г.). Творческое наследие Налимова В. В. включает более 30 книг и 250 статей по различным отраслям знания.

Биографическая справка. В. В. Налимов родился 4 ноября 1910 г. В 1929 г. поступил на математическое отделение физико-математического факультета МГУ. В 1930 г. ушел из МГУ в знак протеста против травли интеллигенции и вскоре (1930 г.) поступил на работу лаборантом, затем инженером-лаборантом во Всесоюзный электротехнический институт (ВЭИ). Служил в армии в научно-техническом центре ВВС. После демобилизации работал в Институте контрольно-измерительных приборов, где прошел аттестационную комиссию, давшую право на защиту кандидатской диссертации без окончания вуза.

1936 г. − первый арест. Второй арест − 18 июня 1937 г., приговор − 5 лет по ст. 58 пп. 10−11. Провел пять лет (после формального освобождения в 1942 г., не дававшего права на возвращение в Москву) в Магадане, в основном на Оротуканском заводе горного оборудования. В 1947 г. − возвращение в Москву. С 1955 г. работает младшим научным сотрудником в ВИНИТИ АН СССР (редактор в отделе "Оптика"). Кандидат технических наук, тема диссертации: "Дифференциальное изучение ошибок спектрального и химического анализа с применением методов математической статистики".

В 1959 г. переводится в Государственный институт редких металлов (ГИРЕДМЕТ). В 1964 г. защищает диссертацию на соискание ученой степени доктора технических наук "Методологические аспекты химической кибернетики".

В 1970 г. занимает должность профессора на кафедре теории вероятностей и математической статистики МГУ. Создал (вместе с Б. В. Гнеденко) секцию "Математические методы исследования" журнала "Заводская лаборатория" и руководил ею с 1962 по 1997 гг.

1965−1975 гг. − первый заместитель заведующего межфакультетской Лабораторией статистических методов МГУ (заведующий лабораторией − акад. А. Н. Колмогоров). После расформирования Лаборатории − заведующий лабораторией (1975−1988 гг. − главный научный сотрудник) математической теории эксперимента Биологического факультета МГУ. С 1993 г. − главный научный сотрудник лаборатории системной экологии биологического факультета МГУ. В. В. Налимов скончался 19 января 1997 г.

11.8. Парето, Вильфредо (Pareto, Vilfredo) − итальянский инженер, экономист и социолог [Wikipedia].

Достижения. Вильфредо Парето разработал ряд теорий, названных его именем: распределение Парето, кривая Парето, закон Парето, эффективность по Парето. Он является одним из основоположников теории элит, согласно которой общество имеет пирамидальную структуру: на вершине находится элита − социальный слой, руководящий жизнью всего общества; залог успешного развития состоит в своевременной ротации (обновлении) элиты.

Биографическая справка. В. Парето родился 15 июля 1848 г. в Париже. В 1858 г. семья Парето переехала в Италию, где Вильфредо получил одновременно классическое гуманитарное и техническое образование. В 1869 г. после окончания Политехнической школы в Турине В. Парето защитил диссертацию “Фундаментальные принципы равновесия в твердых телах”. В течение ряда лет он занимал важные должности в железнодорожном ведомстве и в металлургической компании. В первой половине 1890-х гг. В. Парето публикует ряд исследований в области экономической теории и математической экономики. С 1893 г. и до конца жизни он был профессором политической экономии Лозаннского университета в Швейцарии, сменив в этой должности известного экономиста Леона Вальраса. Умер Парето 20 августа 1923 г. в Селиньи (Швейцария).

11.9. Прайс, Дерек Джон де Солла (Price, Derek John de Solla) − британский и американский историк науки, ученый в области информатики [Wikipedia].

Достижения. Научный вклад Д. Прайса включает, во-первых, исследования экспоненциального роста науки и полупериода актуальности научной литературы (1963 г.); во-вторых, исследования цепи цитирования между научными работами включая открытие, что цепи цитирования имеют распределение по степенному закону (1965 г.); в-третьих, создание математической теории цепей цитирования на основе “Matthew effect” (1976 г.). Прайсу принадлежит первая подробная работа об антикитерском механизме (см. фото) − статья “Древнегреческий компьютер” (англ. *An ancient Greek computer*) в журнале *Scientific American* [Price, 1959]. Среди наиболее значительных работ Прайса − монография 1963 г. “Малая наука, большая наука” (англ. *Little Science, Big Science*), заложившая основу новой отрасли знания − наукометрии.

Биографическая справка. Д. Прайс родился 22 января 1922 г. в Лейтоне (Англия), изучал физику и математику в Лондонском университете, который окончил в 1942 г. Там же в 1946 г. Д. Прайс защитил диссертацию по экспериментальной физике. Идея экспоненциального роста науки пришла к нему, когда он заметил такой рост по журналу “Philosophical Transactions of the Royal Society” в период 1665−1850 гг., полная подшивка которого находилась у него дома. В 1049 г. Прайс вернулся в Англию, чтобы работать над диссертацией доктора философии по истории науки, на этот раз в Кембриджском университете. В дальнейшем Д. Прайс работал в США в качестве консультанта Смитсоновского института и сотрудника Института перспективных исследований, а затем много лет, вплоть до своей кончины, был профессором истории науки в Йельском университете. Д. Прайс умер 3 сентября 1983 г. В 1984 г. он посмертно удостоен

премии “Американской ассоциации информатики и технологии” за выдающийся вклад в исследования в области информатики.

11.10. Хирш, Хорхе (Hirsch, Jorge) – американский профессор физики [Wikipedia].

Достижения. Х. Хирш является автором *h*-индекса, предназначенного для оценки публикационной производительности ученых. Этот индекс базируется на множестве наиболее цитируемых публикаций и множестве цитирований, которые они получили от других публикаций. Индекс может применяться также для измерения продуктивности и важности групп ученых, таких как отделение, организация или страна. В 2010 г. Х. Хирш предлагает *hbar* (*ħ*-индекс), который полезен для определения публикационной производительности исследователей с учетом эффекта соавторства.

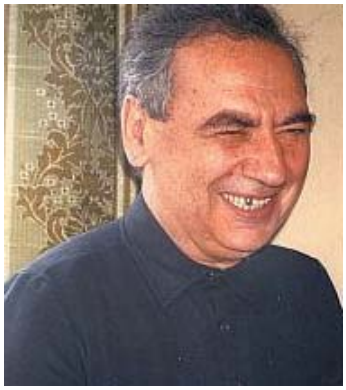
Биографическая справка. Хорхе Хирш родился в 1953 г. Работает в Калифорнийском университете Сан-Диего (США).

11.11. Ципф, Джордж (Zipf, George) – американский статистик и лингвист [Wikipedia].

Достижения. Исследуя распределения частоты встречаемости слов естественного языка в достаточно большом тексте, в 1949 г. Ципф эмпирически установил, что в любом достаточно объемном содержательном тексте частоты употребления слов описываются степенным законом. Эта закономерность получила название закона Ципфа. Формально это можно записать в виде $P_n \sim 1/n^a$, где P_n – частота использова-

ния элемента (слова) с порядковым номером *n,* параметр *a* близок к единице.

Биографическая справка. Джордж Ципф родился 7 января 1902 г. Учился в Гарварде, изучал китайский язык и демографию. Профессор Гарвардского университета. Умер 25 сентября 1950 г. в Ньютоне (Массачусетс, США).



11.12. Шрейдер Юлий Анатольевич (Schreider, Juli) − советский и российский ученый и философ [Шрейдер, CV].

Достижения. Серия работ в области математической лингвистики, теории и практики ранговых распределений. Ю. А. Шрейдер является специалистом по информатике, методологии науки, философии, религии. Область научных исследований: математика (функциональный анализ и вычислительные методы), вычислительная техника, информатика, семиотика, логика, философия и методология науки, проблемы сознания и философия религии. В трудах Ю. А. Шрейдера представлены новые результаты по теории бинарных отношений типа сходства; подведены итоги исследований по теории классификации, логике принятия решений “большинством”, исследуется отличие информации от личностного знания. В последние годы Ю. А. Шрейдер занимался исследованием моделей рефлексивных структур и вопросами ценностного выбора, связанными с возникновением и разрешением конфликтов; сформулировал основные принципы инженерии знаний. В 1960-е гг. Ю. А. Шрейдер заинтересовался религиозными проблемами, и этот интерес положил начало его философским исследованиям.

Биографическая справка. Юлий Анатольевич Шрейдер родился в Днепропетровске 28 октября 1927 г. В 1946 г. окончил механико-математический факультет МГУ; 1949 г. – аспирант МГУ; в 1950 г. защитил кандидатскую диссертацию по функциональному анализу. В 1949–1950 гг. и в 1956–1961 гг. работал в Институте электронных машин. 1951–1956 гг. − доцент Московского института стали; в 1961–1989 гг. работал в отделе семиотики ВИНИТИ АН СССР; 1962 г. – редактор и соавтор монографий "Основы метода Монте − Карло" (М., 1962). В 1981 г. защитил докторскую диссертацию по философии "Гносеологические особенности современной науки в свете системного подхода"; 1984 г. – профессор информатики; 1989 г. – главный научный сотрудник Института проблем передачи информации РАН. Ю. А. Шрейдер умер 24 августа 1998 г.

Ученый и философ Ю. А. Шрейдер преподавал в МГУ на механико-математическом факультете и на отделении структурной и прикладной лингвистики филологического факультета. Академик РАЕН (по отделению "Наука и теология"), профессор. Опубликовал около 800 трудов.

11.13. Яблонский Анатолий Иванович (Yablonsky Anatoly) – видный российский специалист по философии, методологии науки и проблемам науковедения [Яблонский, CV].

Достижения. Главный объект исследований А. И. Яблонского − законы строения, функционирования и развития науки как особой социальной системы. А. И. Яблонскому удалось значительно усовершенствовать методы анализа таких систем. Он является автором монографии "Математические модели в исследовании науки" [Яблонский, 1986], в которой изучаются математические методы моделирования динамики и структуры науки, а также закономерности развития науки на основе законов Ципфа − Парето. А. И. Яблонский внес большой вклад в исследование "негауссо-

вых” вероятностных распределений и их использование для анализа социальных систем.

Биографическая справка. А. И. Яблонский родился в Уссурийске 20 апреля 1936 г. Переехал в Москву и окончил Московский физико-технический институт. В 1959−1970 гг. работал в различных отраслевых НИИ, занимаясь анализом функционирования больших технических систем. 1971−1978 гг. − старший научный сотрудник Института истории естествознания и техники им. С. И. Вавилова (ИИЕТ АН СССР). С 1979 г. − старший научный сотрудник Всесоюзного научно-исследовательского института системных исследований ВНИИСИ (ныне − ИСА РАН). А. И. Яблонский умер 14 февраля 1986 г.

[Adams, Yellen, 1976] Adams W. J., Yellen J. L. Commodity bundling and the burden of monopoly // Quaterly J. Economics. 1976. V. 90. P. 475–498.

[Almind, Ingwersen, 1997] Almind T. C., Ingwersen P. Informetric analyses on the World Wide Web: Methodological approaches to webometrics // J. Document., 1997. N 53, V 4. P. 404–426.

[Alonso, et al., 2010] Alonso S., Cabrerizo F., Herrera-Viedma E., Herrera F. hg-index: a new index to characterize the scientific output of researchers based on the h- and g-indices // Scientometrics. 2010. V. 82, N 2. P. 391−400.

[Anderson, et al., 2008] Anderson T. R., Hankin R. K. S., Killworth P. D. Beyond the Durfee square: Enhancing the h-index to score total publication output // Scientometrics. 2008. V 76, N. 3. P. 577–588.

[Andres, 2009] Andres A. Measuring Academic Research. Oxford: Chandos Publ., 2009. P. 169.

[Archamblaut, Lariviere, 2009] Archamblaut E., Lariviere V. History of the journal impact factor: contingencies and consequences // Scientometrics. 2009. V. 79, N 3. P. 1−15.

[Bakos, Brynjolfsson, 1999] Bakos Y., Brynjolfsson E. Bundling information goods: pricing, profits and efficiency // Management Sci. 1999. V. 45, iss. 12. P. 1613–1630.

[Bakos, Brynjolfsson, 2000a] Bakos Y., Brynjolfsson E. Aggregation and disaggregation of information goods: implications for bundling, site licensing, and micropayment systems // Internet publishing and beyond: The economics of digital and intellectual property / Ed. by D. Hurley, B. Kahin, H. Varian. Cambridge: MIT Press, 2000.

[Bakos, Brynjolfsson, 2000b] Bakos Y., Brynjolfsson E. Bundling and competition on the Internet // Marketing Sci. 2000. V. 19. P. 63–82.

[Bakos, et al., 1999] Bakos Y., Brynjolfsson E., Lichtman D. Shared information goods // J. of Law and Economics. 1999. V. 42. P. 117–155.

[Baneyx, sof]. [Electron. resource]. http://cleanpop.ifris.net/.

[Bannerman, 1998] Bannerman J. Pricing on-line journals // J. Serials Community. 1998. V. 11, N 1. P. 23–26.

[Batista, et al., 2006] Batista P. D., Campiteli M. G., Kinouchi O., Martinez A. S. Is it possible to compare researchers with different scientific interests? // Scientometrics. 2006. V. 68, N 1. P. 179−189.

[Bergstrom, 2007] Bergstrom C. T. Eigenfactor: Measuring the value and prestige of scholarly journals // College Res. Libraries News. 2007. V. 68, N 5. P. 314−316.

[Bergstrom, West, 2008] Bergstrom C. T., West C. J. D. Assessing citations with the Eigenfactor Metrics // Neurology. 2008. N 71. P. 1850−1851.

[Bergstrom, et. al., 2008] Bergstrom C. T., West C. J. D., Wiseman M. A. The eigenfactor metrics // J. Neurosci., 2008. V 28, N . 45 P. 11433−11434.

[Bibliometrics, Def01] [Electron. resource]. http://en.wikipedia.org /wiki/Bibliometrics.

[Bibliometrics, Def02] [Electron. resource]. http://informetrics.ru /pedia/.

[Bibliometrics, Def03] [Electron. resource]. http:// polygraphicbook.narod.ru/text/statiy/2/0/051.html.

[Biglu, 2008] Biglu M. H. The influence of references per paper in the SCI to impact factors and the Matthew effect // Scientometrics. 2008. V. 74, N 3. P. 453−470.

[Bonitz, et al., 1997] Bonitz M., Bruckner E., Scharnhorst A. Characteristics and impact of the Matthew effect for countries // Scientometrics. 1997. V. 40, N 3. P. 407−422.

[Bot, et al., 1998] Bot M., Burgemeester J., Roes H. Cost of publishing an electronic journal // D-Lib Mag. November 1998. V. 4, N 11.

[Bras-Amorys, et al., 2011] Bras-Amorys M., Domingo-Ferrer J., Torra V. A. bibliometric index based on the collaboration distance // Proc. 7th Intern. conf. MDAI 2010, Perpignan (France), Oct. 27–29, 2010. P. 5−6 // Lecture Notes Computer Sci. Springer. 2010. V. 6408. P. 5−6.

[Braun, et al., 2006] [BGS, 2006] Braun T., Glanzel W., Schubert A. A Hirsch-type index for journals // Scientometrics. 2006. V. 69, N 1. P. 169−173.

[Brin, Page, 1998] Brin S., Page L. The anatomy of a large-scale hypertextual Web search engine // Computer Networks. 1998. V. 30, N 1–7. P. 107−117.

[Brody, 2006] Brody T. Evaluating research impact through open access to scholarly communication: Thesis. S. l., 2006. [Electron. resource]. http://eprints.ecs.soton.ac.uk/13313/.

[Brookes, 1969] Brookes B. C. Bradford's law and the bibliography of science // Nature, 1969, 224. P. 953–956.

[Brooks, 1986] Brooks T. Evidence of complex citer motivations // J. Amer. Soc. Inform. Sci. 1986. V. 37, iss. 1. P. 34–36.

[Brooks, 1990] Brooks B. C. Biblio-, sciento-, infor-metrics? What are we talking about. Informetrics 89/90 / Ed. by L. Egghe, R. Rousseau. Elsevier Sci. Publ., 1990. P. 31−43.

[Burrell, 1991] Burrell Q. L. The Bradford Distribution and the Gini index // Scientometrics, 1991. V. 21. P. 117–130.

[Burrell, 2005] Burrell Q. L. Symmetry and other transformation features of Lorenz / Leimkuhler representations of informetric data // Information Processing & Management. 2005. V. 41, iss. 6. P. 1317–1329.

[Butler, 1999] Butler D. The writing is on the Web for science journals in print // Nature. 1999. V. 397. P. 195−200.

[Cardillo, 2010] Cardillo G. Bibliometrics: the art of citations indices. 2010. [Electron. resource]. http://www.mathworks.com/matlabcentral /fileexchange/28161-bibliometrics-the-art-of-citations-indices.

[Chen, soft] [Electron. resource]. http://cluster.cis.drexel.edu/.

[Cole, Cole, 1972] Cole J. R., Cole S. The Ortega Hypothesis: Citation analysis suggests that only a few scientists contribute to scientific progress // Science. 1972. V. 178, N 4059. P. 368−375.

[Condon, 1928] Condon E. V. Statistics of vocabulary // Science. 1928. V. 67. P. 300.

[Cozzens, 1989] Cozzens S. E. What do citations count? The rhetoric-first model // Scientometrics. 1989. V. 15, N 5−6. P. 437−447.

[Davarpanah, Aslekia, 2008]. Davarpanah M. R., Aslekia S. Scientometric analysis of international LIS Journals: productive and characteristics // Scientometrics. 2008. V. 77, N 1. P. 21−39.

[Egghe, 1985] Egghe L. Consequences of Lotka's law for the law of Bradford // J. Documentation. 1985. V 41, iss. 3. P. 173–189.

[Egghe, 1986] Egghe L. The dual of Bradford's law // J. American Society for Information Science. 1986, V. 37, iss. 4. P. 246–255.

[Egghe 1990] Egghe L. Application of the Theory of Bradford's Law to the Calculation of Leimkuhler's Law and to the Completion of Bibliographies // J. American Society for Information Science. 1990. V 41, №. 7. P. 469–492.

[Egghe, 2006a] Egghe L. An improvement of the h-index: the g-index // ISSI Newslett. 2006. V. 2, N 1. P. 8−9.

[Egghe, 2006b] Egghe L. Theory and practice of the g-index // Scientometrics. 2006. V. 69, N 1. P. 131−152.

[Egghe, 2008] Egghe L. Modelling successive h-indices // Scientometrics. 2008. V. 77, N 3. P. 377−387.

[Egghe, Rousseau, 2008] Egghe L., Rousseau R. An h-index weighted by citation impact // Inform. Proc. Management. 2008. V. 44, N 2. P. 770–780.

[Eigenfactor] Метрика Egenfactor. [Electron. resource]. http://www.eigenfactor.org/.

[Euro-Factor, 2002] Метрика Euro-Factor. [Electron. resource]. http://lea.univ-lille1.fr/Menu_du_Site/Publications/Acrobat/VICER-EUROFACTOR.pdf.

[Fairthorne, 1969] Fairthorne R. A. Empirical hyperbolic distributions (Bradford-Zipf-Mandelbrot) for bibliometric description and prediction // J. Document. 1969. V. 25, N 4. P. 319−343.

[Fano, 1956] Fano R. M. Information theory and the retrieval of recorded information // Documentation in Action. N. Y.: Reinhold Publishing Corp., 1956. P. 241.

[Fishburn, et al., 1997] Fishburn P. C., Odlyzko A. M., Siders R. C. Fixed fee versus unit pricing for information goods: competition, equilibria, and price wars // First Monday. 1997. V. 2, N 7.

[Fishwick, et al., 1998] Fishwick F., Edwards L., Blagden J. Economic implications of different models of publishing scholarly journals for professional societies and other small or specialist publishers: Rep. to the joint information systems committee, electronic libraries program, 1998. [Electron. resource]. http:// www.ukoln.ac.uk/services/elib /papers/supporting/.

[Franceschet, 2010] Franceschet M. Ten good reasons to use the eigenfactor metrics // Inform. Proc. Management. 2010. V. 46, N 5. P. 555−558.

[Gągolewski, Grzegorzewski, 2009] Gągolewski M., Grzegorzewski P. A geometric approach to the construction of scientific impact indices // Scientometrics. 2009. V. 81, N 3. P. 617–634.

[Garfield, 1955] Garfield E. Citation indexes for science: a new dimension in documentation through association of ideas // Science. 1955. V. 122 (3159). P. 108−111.

[Garfield, 1972] Garfield E. Citation analysis as a tool in Journal evaluation // Science. 1972. V. 178. P. 471−479.

[Garfield, Sher, 1963] Garfield E., Sher I. H. New factors in the evaluation of scientific literature through citation indexing // Amer. Document. 1963. V. 14, N 3. P. 195−201.

[Garfield, soft] [Electron. resource]. http://thomsonreuters.com/.

[Gaston, 1973] Gaston J. Originality and competition in science: A study of the British high energy physics community. Chicago: Univ. Chicago Press, 1973.

[Getz, 1992] Getz M. Electronic publishing in academia: an economic perspective / Serials Review. [Electron. resource]. 1992. V. 18. P. 25–31.

[Ginsparg, 1996] Ginsparg P. Winners and losers in the global research village // Serials Librarian. 1996. V. 30, iss. 3/4. P. 83–95.

[Glanzel, 2003] Glanzel W. Bibliometrics as a research field: A course on theory and application of bibliometric indicators. 2003. [Electron. resource]. http://nsdl.niscair.res.in/bitstream/123456789/968/1/Bib_Module_KUL.pdf.

[Glanzel, 2004] Glanzel W. A bibliometric approach to the role of autor self-citations in scientific communication // Scientometrics. 2004. V. 59, N 1. P. 63−77.

[Glanzel, 2006] Glanzel W. On the H-index. A mathematical approach to a new measure of publication activity and citation impact // Scientometrics. 2006. V. 67, N 2. P. 315−321.

[Glanzel, Schoepfin, 1995] Glanzel W., Schoepfin U. A bibliomertic study on ageing and reception processes of scientific literature // J. Inform. Sci., 1995. V. 21. P. 37−53.

[Goffman, Newill, 1964] Goffman W., Newill V. Generalization of epidemic theory: an application to the transmission of ideas // Nature. 1964. V. 204. P. 225–228.

[González-Pereira, et al., 2009] González-Pereira B., Guerrero-Bote V., De Moya F. The SJR indicator: A new indicator of Journals' scientific prestige // 2009. arXiv:0912.4141.

[Gross, Gross, 1927] Gross P. L. K., Gross E. M. College libraries and chemical education // Science. 1927. V. 66, N 1713. P. 385−389.

[Guan, Ma, 2004] Guan J., Ma N. A comparative study of research performance in computer science // Scientometrics. 2004. V. 61, N 3. P. 339−359.

[Hagstrom, 1965] Hagstrom W. The scientific community. N. Y.: Basic Books. 1965. 304 p.

[Harzing, 2010] Harzing A.-W. The publish or perish book. Melbourne: Tarma Software Res. Pty Ltd., 2010. 250 p.

[Harzing, soft] [Electron. resource]. http://www.harzing.com /pop.htm/.

[Hirsch, 2005] Hirsch J. E. An index to quantify an individual's scientific research output // Proc. Nation. Academy of Sciences of the USA, 2005. V. 102, N 46. P. 16569–16572.

[Hirsch, 2010] Hirsch J. E. An index to quantify an individual's scientific research output that takes into account the effect of multiple coauthorship // Scientometrics. 2010. V. 85, N 3. P. 741–754. См. также: 2010. arXiv:0911.3144v2.

[Ingwersen, 2006] Ingwersen P. Webometrics – ten years of expansion // Proc. of the Intern. workshop on webometrics, informetrics and scientometrics & 7th COLLNET Meeting, Nancy (France),10–12 May, 2006.

[Irvine, Martin, 1976] Irvine J., Martin B. R. Evaluating big science: CERN's past performance and future prospects // Scientometrics. V. 7, N 3−6. P. 281−308.

[ISSI Society] International society for scientometrics and infometrics. [Electron. resource]. http://www.issi-society.info/past.html.

[Jin, 2006] Jin B. H. *h*-index: An evaluation indicator proposed by scientist // Science Focus. 2006. V. 1, N 1. P. 8−9.

[Jin, 2007] Jin B. H. The AR-index: complementing the *h*-index // ISSI Newsletter. 2007. V. 3, N 1. P. 6.

[Jin, et al., 2007] Jin B. H., LiMing L., Rousseau R., Egghe L. The *R*- and *AR*-indices: Complementing the *h*-index // Chinese Science Bulletin. 2007. V. 52, N 6. P. 855−863.

[Journal Metrics, 2010] Booklet the evolution of journal assessment. 2010. [Electron. resource]. http://www.J.metrics.com.

[Journal Price] J. Price Guide [Electron. resource]. http://mulibraries.missouri.edu/guides/rankings/J.prices.htm.

[Karpagam, et al., 2011] Karpagam R., Gopalakrishnan S., Natarajan M. Scientific measures ahd tools for research literature output // Indian J. Scien. and Technol. 2011. V. 4, N 7. P 828–833.

[Kessler, 1963a] Kessler M. M. Bibliographic coupling between scientific papers // Amer. Documentation. 1963. V. 14, N 1. P. 10−25.

[Kessler, 1963b]. Kessler M. M. An experimental study of bibliographic coupling between technical papers // IEEE Transaction on Information Theory. 1963. V. 9, N 1. P. 49.

[Kosmulski, 2006a] Kosmulski M. I - a bibliometric index // Forum Akademickie. 2006. N 11. P. 31.

[Kosmulski, 2006b] Kosmulski M. A new Hirsch-type index saves time and works equally well as the original h-index // ISSI Newslett., 2006. V 2, N 3. P. 4–6.

[Kutateladze, 2009] Kutateladze S. S. The Game of Cipher Beads // J. Appl. Indust. Math. 2009. V. 3, N 3. P. 364–366.

[Ladwig, Sommese, 2005] Ladwig J. P., Sommese A. J. Using cited half-life to adjust download statistics // College and Res. Libraries. 2005. V. 66, N 6. P. 527−542.

[Leimkuhler, 1967] Leimkuhler F. F. The Bradford distribution // J. of Documentation. V 23, iss. 3. P. 197–207.

[Leydesdorff, soft] [Electron. resource]. http://home.medewerker.uva.nl/l.a.leydesdorff/.

[Liebowitz, 1985] Liebowitz S. J. Copying and indirect appropriability: Photocopying of journals // J. Polit. Economy. 1985. V. 72. N 4. P. 816–824.

[LINDO.soft] [Electron. resource]. http://www.lindo.com/.

[Lotka, 1926] Lotka A. The frequency distribution of scientific productivity // J. Washington Acad. Sci. 1926. V. 16. N 12. P. 317–324.

[Lotka, CV] [Electron. resource]. http://users.telenet.be /ronald.rousseau/html/lotka.html.

[Magyar, 1974] Magyar G. Bibliometric analysis of a new research sub-field // J. Document. 1974. V. 30, N 1. P. 32−40.

[McAfee, et al., 1989] McAfee R. P., McMillan J., Whinston M. D. Multiproduct monopoly, commodity bundling, and correlation of values // The Quarterly J. Economics. 1989. V. 104. N 2. P. 371–383.

[McCabe, 2000] McCabe M. Academic journal pricing and market power: A portfolio approach // Proc. of the Amer. econ. assoc. conf., Boston (MA), 2000. [Electron. resource]. http://www.si.umich.edu/ ~mccabe /JournPub.PDF.

[McCabe, 2002] McCabe M. Journal pricing and mergers: a portfolio approach // Amer. Econ. Rev. 2002. V. 92. N 1. P. 259−269.

[McCabe, Snyder, 2007] McCabe M. J., Snyder C. M. Academic journal prices in a digital age: A two-sided market model // B. E. J. Econ. Analysis Policy. 2007. V. 7, iss. 1 (Contributions). Article 2.

[Merton, 1988] Merton R. K. The Matthew effect in science. 2. Cumulative advantage and the symbolism of intellectual property // ISIS. 1988. V. 79. P. 606−623.

[Moed, 2010a] Moed H. Measuring contextual citation impact of scientific journals // 2010. arXiv:0911.2632.

[Moed, 2010b] [Electron. resource]. http://www.researchtrends.com/issue15-january-2010/expert-opinion-5/.

[Moed, et al., 1998] Moed H. F., van Leeuwen T. N., Reedijk J. A new classification system to describe the ageing of scientific journals and their impact factors // J. Document. 1998. V. 54. P. 387−419.

[Mas-Collel, et al., 1995] Mas-Collel A., Whinston M. D., Green J. R. Microeconomic Theory. N. Y.: Oxford Univ. Press, 1995.

[Meho, Yang, 2006] Meho L. I., Yang K. A new era in citation and bibliometric analyses: Web of Science, Scopus, and Google Scholar // 2006. arXiv:cs/0612132.

[Moravcsik, Murugesan, 1975] Moravcsik M. J., Murugesan P. Some results on the function and quality of citations // Social Studies of Sci. 1975. V. 5, iss. 1. P. 86−92.

[NEICON] [Electron. resource]. http://www.neicon.ru/.

[Odlyzko, 1997] Odlyzko A. M. The economics of electronic journals // First Monday. 1997. V. 2, N 8.

[Odlyzko, 1999] Odlyzko A. M. Competition and cooperation: libraries and publishers in the transition to electronic scholarly journals // J. Electronic Publ. 1999. V. 30, N 4. P. 163–185.

[Ortega, 1932] Ortega y Gasset J. The revolt of the masses. N. Y.: Norton, 1932. P. 84−85.

[Patent Pictures, soft] [Electron. resource]. http://www.researchinformation.info/rijanfeb04patents.html/.

[Price Medal, 2012] http://www.issi-society.info/price.html.

[Prathap, 2006] Prathap G. Hirsch-type indices for ranking institutions' scientific research output // Current Sci. 2006. V. 91, N 11. P. 1439.

[Prathap, 2010] Prathap G. Is there a place for a mock h-index? // Scientometrics. 2010. V. 84, N 1. P. 153−165.

[Prathap, 2011] Prathap G. The fractional and harmonic p-indices for multiple authorship // Scientometrics. 2011. V. 86. P. 239–244.

[Price, 1959] Price D. An ancient Greek // Scientific American. 1959. V. 200. N 6. P. 60–67.

[Price, 1961] Price D. Science since Babylon. New Haven: Yale Univ. Press., 1961.

[Price, 1963] Price D. Little science, big science. N. Y.: Columbia Univ. Press, 1963. 119 p.

[Price, 1965] Price D. Networks of scientific papers // Science. 1965. V. 149 (3683). P. 510−515.

[Price, 1976] Price D. A general theory of bibliometric and other cumulative advantage processes // J. Amer. Soc. Inform. Sci. 1976. N 27. P. 292–306.

[Pritchard, 1969] Pritchard, A. Statistical Bibliography or Bibliometrics? // J. Document., 1969. V. 25, N 4. P. 348–349.

[Pudovkin, Garfield, 2002] Pudovkin A. I., Garfield E. Algorithmic procedure for finding semantically related journals // J. Amer. Soc. Inform. Sci. Technol. 2002. N 53. P. 1113−1119.

[Raisig, 1960] Raisig M. Mathematical evaluation of the scientific serial // Science. 1960. N 131. P. 1417−1419.

[Rochet, Tirole, 2003] Rochet J. C., Tirole J. Platform competition in two-sided markets // J. Eur. Econ. Ass. 2003. N 1. P. 990−1029.

[Rousseau, 1992] Rousseau. R. Homepage. [Electron. resource]. http: //users.telenet.be/ronald.rousseau/html/timeline_of_bibliometrics.html.

[Rousseau, 1994] Rousseau R. Bradford curves // Inf. Proc. and Manag. 1994, № 30. P. 267–277.

[Rousseau, Leimkuhler, 1987] Rousseau R., Leimkuhler F. F. The nuclear zone of a Leimkuhler curve // J. of Documentation, 1987, 43(4). P. 322–333.

[Ruane, Tol, 2008] Ruane F., Tol R. Rational (successive) h-indices: An application to economics in the Republic of Ireland// Scientometrics. 2008. V. 75. N 2. P. 395–405.

[Sanderson, 2008] Sanderson M. Revisiting h measured on UK LIS and IR academics // J. Amer. Soc. Inform. Sci. Technol. 2008. V. 59, N 7. P. 1184–1190.

[Sarabia, 2008] Sarabia J. M. A general definition of the Leimkuhler curve // J. Informetrics. 2008. V. 2, iss. 2. P. 156−163.

[Schreiber, 2008] Schreiber M. To share the fame in a fair way, h_m modifies h for multi-authored manuscripts // New J. Physics. 2008. V. 10. P. 1−8.

[Schubert, 2007] Schubert A. Successive h-indices // Scientometrics. V. 70, N 1. P. 201−205.

[SCI2S] [Electron. resource]. http://sci2s.ugr.es/hindex/.

[Scientometrics] [Electron. resource]. http://www.springerlink.com/content/101080/.

[Shannon, 1948] Shannon C. E. A mathematical theory of communication // Bell System Techn. J. 1948. T. 27. P. 379−423, 623–656.

[Sidiropoulos, et al., 2007] Sidiropoulos A., Katsaros D., Manolopoulos Y. Generalized Hirsch h-index for disclosing latent facts in citation networks // Scientometrics. 2007. V. 72, N 2. P. 253–280.

[Small, 1972] Small H. G. Co-citation in the scientific literature: a new measure of the relationship between two documents // J. Amer. Soc. Inform. Sci. 1973. V. 24. P. 265−269.

[Sudhier, 2010] Sudhier K. G. Application of Bradford's law of Scattering to the Physics Literature: A Study of Doctoral Theses Citations at the Indian Institute of Science // J. Library and Information Technology. 2010. V. 30, № 2. P. 3–14.

[TRRS, soft] [Electron. resource]. http://www.refviz.com/.

[Tsay, 2009] Tsay M. Y. An analysis and comparison of scientometric data between journals of physics, chemistry and engineering // Scientometrics. 2009. V. 78, N 2. P. 279−293.

[Vaidya, 2005] Vaidya J. *V*-index: A fairer index to quantify an individual's research output capacity // British Medical Journal. 2005. V. 331, N 7528. P. 1339.

[Van Raan, 1988] Handbook of quantitative studies of science and technology / Ed. by A. F. J. Van Raan. Amsterdam: North-Holland, 1988.

[Van Raan, 2006] Van Raan A. F. J. Comparison of the Hirsch-index with standard bibliometric indicators and with peerjudgment for 147 chemistry research groups // Scientometrics. 2006. V. 67, N 3. P. 491−502.

[Van Raan, 2008] Van Raan A. F. J. Self-citations as an impact-reinforcing mechanism in the science system // J. Amer. Soc. Inform. Sci. Technology. 2008. V. 59, N 10. P. 1631−1643.

[Varian, 1995] Varian H. R. Pricing information goods // Proc. scholarship in the new inform. environment symp. Harvard Law School, May 2–5, 1995. Cambrig: MA.

[Varian, 1996] Varian H. R. Pricing electronic journals // D-Lib Mag. 1996. V. 2, N 6.

[Varian, 2000] Varian H. R. Buying, sharing and renting information goods // J. Industrial Economics. 2000. V. 48, N 4. P. 473–488.

[White, soft] [Electron. resource]. http://project.cis.drexel.edu/.

[Wikipedia] Wikipedia, the free encyclopedia. [Electron. resource]. http://en.wikipedia.org/wiki /Main_Page.

[WolframMW, Gini] [Электрон. ресурс]. http://mathworld.wolfram.com/GiniCoefficiennt.html.

[WolframMW, Lorenz] [Электрон. ресурс]. http://mathworld.wolfram.com/LorenzCurve.html.

[Yancey, 2005]. Yancey R. Fifty years of citation indexing and analysis // Thomson Routers, Sept., 2005.

[Zhang, 2009] Zhang C. The e-index, complementing the h-index for excess citations // PLoS ONE. 2009. V. 5, iss. 5. e5429. P. 1–4.

[Академик, словари] [Электрон. ресурс]. http://dic.academic.ru/dic.nsf/dic_wingwords/.

[Арапов и др., 1975] Арапов М. В., Ефимова Е. Н., Шрейдер Ю. А. О смысле ранговых распределений // НТИ. Сер. 2. 1975. № 1. С. 9–20.

[Арнольд, 1998] Арнольд В. И. Статистика первых цифр степеней двойки и передел мира // Квант. 1998. № 1. С 2–4.

[Бешенов, 2007] Бешенов А. Единицы метаданных DCMI. [Электрон. ресурс]. http://beshenov.ru/dcmi-terms.html.

[Бредихин и др., 2008] Бредихин С. В., Кузнецов А. Ю., Хуторецкий А. Б. Оптимизация подписки на электронные журналы // Сиб. журн. индустр. математики. 2008. Т. 11. № 2. С. 21–28.

[Буш, Ванневар] [Электрон. ресурс]. http://www.uic.unn.ru /pustyn/lib /vbush.ru.html.

[Википедия] Википедия, свободная энциклопедия. [Электрон. ресурс]. http://ru.wikipedia.org/wiki/ Main_Page.

[ВИНИТИ, 2011] [Электрон. ресурс]. http://www2.viniti.ru/.

[Гантмахер, 1967] Гантмахер Ф. Р. Теория матриц. М.: Наука, 1967. 576 с.

[Гельфанд, 2010] Гельфанд М. Проведите поиск в РИНЦ самостоятельно! // Троицкий вариант. 2010. 20 июля. (№ 58). С. 4–5,7.

[Горькова, 1988] Горькова В. И. Инфометрия: (количественные методы в НТИ). М.: ВИНИТИ, 1988. Сер. Информатика. Итоги науки и техники. Т. 10. С. 3–326.

[Гохберг, 2003] Гохберг Л. М. Статистика науки. М.: ТЕИС, 2003. 478 с.

[Гражданский кодекс РФ] [Электрон. ресурс] http://www. grazkodeks.ru.

[Гэри, Джонсон, 1982] Гэри М., Джонсон Д. Вычислительные машины и труднорешаемые задачи. М.: Мир, 1982.

[Еременко, 2010] Еременко Г. Через два месяца РИНЦ станет достаточно объективным. [Электрон. ресурс]. http://www.strf.ru /organization.aspx? CatalogId=221&d_no=33357.

[Жижимов, Мазов, 2004] Жижимов О. Л., Мазов Н. А. Принципы построения распределенных информационных систем на основе протокола Z39.50. ОИГГМ СО РАН, Новосибирск: Изд-во ИВТ СО РАН, 2004. 361 с.

[Игра в цыфирь, 2011] Игра в цыфирь, или как теперь оценивают труд ученого / Сб. ст. о библиометрике. М.: Изд-во МЦНМО, 2011. 72 с.

[Каленов, 2011]. Каленов Н. Е. Еще раз о РИНЦ. [Электрон. ресурс]. http://trv-science.ru/2011/02/01/eshhe-raz-o-rinc/#more-10080.

[Когаловский, 2002] Когаловский М. Р. Энциклопедия технологий баз данных. М.: Финансы и статистика, 2002. 799 с.

[Кристева, 1967] Словарь литературоведческих терминов. [Электрон. ресурс]. http://slovar.lib.ru/dictionary/intertextualnost.htm.

[Лука, 19:26] Евангелие от Луки. Гл. 19. [Электрон. ресурс]. http://biblia.org.ua/bibliya/lk.html.

[Манцивода, 2005] Манцивода А. В. Система метаописаний Dublin Core. [Электрон. ресурс]. http://teacode.com/concept/eor/dc.html.

[Марк 4:25] Евангелие от Марка. Гл. 4. [Электрон. ресурс]. http://biblia.org.ua/bibliya/mk.html.

[Маршакова, 1973] Маршакова И. В. Система связей между документами, построенная на основе ссылок (по указателю “Science Citation Index”) // НТИ. Сер. 2. Информ. системы. 1973. № 6. С. 3–8.

[Маршакова, 1988] Маршакова И. В. Система цитирования научной литературы как средство слежения за развитием науки. М.: Наука, 1988. 287 с.

[Маршакова-Шайкевич, 2002] Маршакова-Шайкевич И. В. Вклад России в развитие мировой науки // Отеч. записки. Блеск и нищета российской науки. 2002. № 7. С. 314–345.

[Маршакова-Шайкевич, 2004] Маршакова-Шайкевич И. В. Классификация научных журналов методом коцитирования // Науч.-техн. информация. Сер. 1. 2004. № 8. С. 31–35.

[Маршакова-Шайкевич, 2008] Маршакова-Шайкевич И. В. Россия в мировой науке. Библиометрический анализ. М.: ИФ РАН 2008. 227 с.

[Маршакова-Шайкевич, CV] Curriculum vitae. [Электрон. ресурс]. http://iph.ras.ru/shaykevich.htm.

[Матфей 25:29] Евангелие от Матфея. Гл. 25. [Электрон. ресурс]. http://biblia.org.ua/bibliya/mf.html.

[Мертон, 1968] Мертон Р. К. Эффект Матфея в науке, II: накопление преимуществ и символизм интеллектуальной собственности. [Электрон. ресурс]. http://igiti.hse.ru/data/033/314/1234/3_6_1Merto.pdf.

[Налимов, Мульченко, 1969] Налимов В. В. Наукометрия. Изучение науки как информационного процесса / В. В. Налимов, З. М. Мульченко. М.: Наука, 1969. 192 с.

[Пападимитриу, Стайглиц, 1985] Пападимитриу Х. Комбинаторная оптимизация. Алгоритмы и сложность / Х. Пападимитриу, К. Стайглиц. М.: Мир, 1985.

[Пенькова, Тютюнник, 2001] Пенькова О. В., Тютюнник В. М. Информетрия, наукометрия и библиометрия: наукометрический анализ современного состояния // Вестн. Тамбов. гос. ун-та. Сер. “Естеств. науки”. 2001. Т. 6, вып. 1. С. 6–8.

[Петров, Яблонский, 1980] Петров В. М. Математика и социальные процессы // В. М. Петров, А. И. Яблонский. М.: Знание, 1980. С. 64.

[Писляков, 2005] Писляков В. В. Наукометрические методы и практики, рекомендуемые к применению в работе с российским индексом научного цитирования: Отчет о НИР по теме “Разработка системы статистического анализа российской науки на основе данных российского индекса цитирования” / М.: 2005. [Электрон. ресурс]. http://www.elibrary.ru/projects/citation/docs/scientometrics.pdf.

[Писляков, 2007a] Писляков В. В. Зачем создавать национальные индексы цитирования? // Науч. и техн. библиотеки. 2007. № 2. С. 65–71.

[Писляков, 2007b] Писляков В. В. Моделирование процесса обращения к электронным информационным источникам на основе информетрического закона Брэдфорда // Ученые записки Казан. гос. ун-та. Сер. Физ.-мат. науки. 2007. Т. 149, кн. 2. С. 116–127.

[Писляков, 2011] Писляков В. В. Не импактом единым // Наука и жизнь. 2011. № 1. С. 13–15.

[Писляков, Дьяченко, 2009] Писляков В. В., Дьяченко Е. Л. Эффект Матфея в цитировании статей российских ученых, опубликованных за рубежом // НТИ. Сер. 2. Информ. процессы и системы. 2009. № 3. С. 19–24. [Электрон. ресурс]. http://www.hse.ru/data/694/515/1239/Pislyakov2.pdf.

[Писсанецки, 1988] Писсанецки С. Технология разряженных матриц. М.: Мир, 1988. 412 с.

[ПОСТ РАН № 201] Постановление Президиума Российской академии наук № 201 от 12.10.2010. Об утверждении Положения о Комиссии по оценке результативности деятельности научных организаций Российской академии наук и методики оценки результативности деятельности научных организаций Российской академии наук [Электрон. ресурс]. http://www.ras.ru/presidium/documents/directions.aspx.

[Прайс, 1966] Прайс Д. Малая наука, большая наука // Наука о науке. М.: Прогресс, 1966. С. 281–384.

[Редькина, 2005] Редькина Н. С. Формализованные методы анализа документальных информационных потоков // Библиосфера. 2005. № 2. С. 51–59.

[Редькина, 2006] Редькина Н. С. Модель многоуровневого изучения результативности научных исследований // Тр. VII Всерос. конф. молодых ученых по математическому моделированию и

информационным технологиям (с участием иностранных ученых). Красноярск, 1–3 ноября 2006 г. [Электрон. ресурс]. http://www.nsc.ru/ws/YM2006/10617/Redkina.pdf.

[Сериал 38 попугаев, 1977] Цитата из сериала советских мультфильмов 38 попугаев, 1977 год. Режиссер И. Уфимцев, художник Л. Шварцман, сценарий Г. Остер. См. http://www.youtube.com/watch?v=8tzmTQIAv28.

[Сонин, 2010] Сонин К. И. Оценка научной статьи: взгляд рецензента. [Электрон. ресурс]. http://www.youtube.com/user/NewEconomicSchool.

[Сторчевой, 2010] Сторчевой М. А. История теории потребления и спроса. [Электрон. ресурс]. http://www.ecsocman.edu.ru/db/msg/184377.html.

[Тироль, 2000] Тироль Ж. Рынки и рыночная власть: теория организации промышленности. СПб.: Эконом. шк., 2000.

[Хайтун, 1983] Хайтун С. Д. Наукометрия: состояние и перспективы. М.: Наука, 1983. 344 с.

[Хайтун, 1989] Хайтун С. Д. Проблемы количественного анализа науки. М.: Наука, 1989. 280 с.

[Шрейдер, CV] Curriculum vitae. [Электрон. ресурс]. http://www.biografija.ru /show_bio.aspx?id=137863

[Штерн, 2001] Кто есть кто в российской науке. [Электрон. ресурс]. http://www.expertcorps.ru/.

[Яблонский, 1986] Яблонский А. И. Математические модели в исследовании науки. М.: Наука, 1986. 352 с.

[Яблонский, CV] Curriculum vitae. [Электрон. ресурс]. http://print.biografija.ru /?id=140137.

Алфавитный указатель терминов и сокращений

hg-индекс	*hg*-index (Alonso)	161
h_I-индекс	h_I-index (Batista)	170
hJ-индекс Хирша для журнала	Journal *hJ*-index (Hirsch)	203
h_m-индекс	h_m-index (Schreiber)	171
h^n-индекс	h^n-index, normalized *h*-index (Sidiropoulos)	177
h_T -индекс	h_T –index, tapered *h*-index (Anderson)	177
h_w-индекс	h_w-index, citation-weighted *h*-index (Egghe)	172
hc-индекс	*hc*-index, contemporary *h*-index (Sidiropoulos)	176
I-индекс	*I*-index (Kosmulski, Prathap)	176
IMI, индекс оперативности	*IMI*, Immediacy index	186
ISS, индекс научной специализации страны	*ISS*-index	141
m-индекс Хирша	*m*-index (Hirsch)	156
MI-индекс Матфея для стран	*MI*-Matthew Index	144
p, p_f, p_h-индексы	p, p_f, p_h-indexes (Prathap)	173
R-индекс	*R*-index (Jin)	162
v-индекс	*v*-index (Vaidya)	176
ђ-индекс Хирша	*ђ*-index (Hirsch)	174
Иерархический *h*-индекс	Successive *h*-index (Schubert)	175
Импакт-фактор журнала (*JIF*)	Journal Impact Factor, *JIF* (Garfield)	181
Индекс индивидуальный, нормализованный	*norm*-index (Harzing)	170
Институт научной информации (ISI)	Institute of Scientific Information, ISI	116
Инфометрия	Infometrics	17
Квантиль	Qantile	188

Коцитирование	Co-citation	112
Коэффициент Джини	Gini Ratio	40
Кривая Леймкулера	Leimkuhler Curve	38
Кривая Лоренца	Lorenz Curve	39
Математическое ожидание	Mean value	33
Медиана	Median	33
Метрика *AI*	Article Influence, *AI* (Bergstrom)	192
Метрика *CitIn*	Cited half-life, *CitIn*	187
Метрика *CitOut*	Citing half-life, *CitOut*	187
Метрика *EF*	Eigenfactor, *EF* (Bergstrom)	189
Метрика *EF*(*j*)	Journal Euro-Factor, *EF*(*j*)	145
Метрика *SJR*	SCImago Journal Rank, *SJR* (Gonzalez-Pereira)	198
Метрика *SNIP*	Source-Normalized Impact per Journal, *SNIP* (Moed)	195
Метрика связности журналов *R(A,B,T)*	*R*(*A,B,T*) (Pudovkin)	204
Метрика эффективности публикаций страны (PEI)	Publication Efficiency Index, PEI (Guan)	142
Мода	Mode	33
Наукометрия (НМ)	Scientometrics	16
Накопленное преимущество (НП)	Accumulated advantage	107
Открытый доступ (ОД)	Open Access	79
Поисковая машина Google Scholar (GS)	Google Scholar, GS	131
Распределение Парето	Pareto Distribution	31
Распределение ранговое	Ranking	28
Распределение Ципфа	Zipf distribution	41

Российский индекс научного цитирования (РИНЦ)	RINC-Index	124
Самоцитирование (СЦ)	Selfcitation	107
Указатель научного цитирования (SCI)	Science Citation Index, SCI. Thomson Reuters	116
Указатель цитирования (WoS)	Web of Science, WoS	123
Указатель цитирования “Американского математического общества” (MCQ)	MCQ-index	185
Указатели цитирования “Искусство и гуманитарные науки” (A&HCI) и “Общественные науки” (SSCI)	Art&Humanities Citation Index (A&HCI) and Social Science Citation Index (SSCI), Thomson Reuters	116
Функция плотности вероятности (PDF)	Probability Density Function, PDF	33
Функция полезности	Utility Function	85
Функция распределения вероятности (CDF)	Cumulative Distribution Function, CDF	32
Функция распределения масс (PMF)	Probability Mass Function, PMF	42
Цитата, Цитирование	Quotation, Citation	100
Эффект Матфея	Matthew Effect	108
Ядро Хирша	Hirsch Core	152

ОГЛАВЛЕНИЕ

Сергей Всеволодович Бредихин,
Александр Юрьевич Кузнецов

МЕТОДЫ БИБЛИОМЕТРИИ И РЫНОК ЭЛЕКТРОННОЙ НАУЧНОЙ ПЕРИОДИКИ

Редактор *Л. Н. Ковалева*

Верстка *О. Г. Заварзина*

Подписано в печать 10.02.12. Формат 60×84/16. Печать офсетная.
Усл. печ. л. 14,4. Уч.-изд. л. 9,4. Тираж 100 экз. Заказ № .

Отпечатано в типографии ООО "Омега-принт". 630090, Новосибирск,
просп. Акад. Лаврентьева, д. 6. Тел. (383) 335-65-23