Казанский государственный университет
НИИ математики и механики им. Н. Г. Чеботарева

На правах рукописи

Писляков Владимир Владимирович

ИНФОРМЕТРИЧЕСКОЕ МОДЕЛИРОВАНИЕ ПРОЦЕССА ОБРАЩЕНИЯ К ЭЛЕКТРОННЫМ ИНФОРМАЦИОННЫМ РЕСУРСАМ

Специальность 05.13.18 – Математическое моделирование,
численные методы и комплексы программ
Диссертация на соискание ученой степени кандидата
физико-математических наук

Научный руководитель
Доктор физ.-мат. наук, профессор,
заслуженный деятель науки РТ
А. М. Елизаров

Казань – 2008

СОДЕРЖАНИЕ

Введение

Статистический анализ научной и образовательной деятельности получает всё большее распространение как в мировой, так и в отечественной практике [99; 105; 2; 5; 6; 4; 110]. Одним из наиболее важных направлений такого анализа является выявление и изучение статистических закономерностей производства, поиска и использования информации — направление, выделенное в отдельную молодую отрасль, *информетрию* (см., в частности, [66; 1])[1].

Широко распространенным методом изучения информационных потоков является *информетрическое моделирование* — математическое моделирование информационных процессов с использованием информетрических законов. Такое моделирование заключается в выявлении эмпирических закономерностей, наблюдаемых в значительном числе информационных процессов, облечении их в строгую математическую форму и распространении данных моделей на остальные процессы, однотипные исследованным. Одним из упрощенных примеров использования готовой модели, основанной на информетрическом законе Лотки, служит оценка суммарного количества I статей, написанных группой из S произвольно выбранных пишущих авторов (т. е. ученых, опубликовавших хотя бы одну статью):

[1] К информетрии плотно прилегают еще две области знания — библиометрия и наукометрия. Разграничение их «полномочий» зависит от конкретных определений этих наук, которые подчас отличаются у разных авторов. Мы придерживаемся представления, что библиометрия занимается только *опубликованной* информацией, *документами*, являясь, таким образом, подмножеством информетрии. Наукометрия значительным образом пересекается с информетрией, хотя есть и несовпадающие области: например, вопросы использования ненаучной информации или вопросы чистого спроса на информацию, без ее воспроизводства, не являются предметом собственно наукометрического исследования, в то же время аспекты занятости населения в науке или потоков средств, вкладываемых в R&D, находятся вне рамок информетрии.

$$I \sim \frac{6S}{\pi^2} \sum_{i=1}^{\frac{1}{\pi}\sqrt{6S}} \frac{1}{i} \sim 0{,}3\, S\left(0{,}7 + \ln S\right).$$

Подробнее закон Лотки исследован в параграфе 4.2, а пока заметим, что использование информетрического моделирования дает нам в данном случае возможность без сбора подробных сведений о публикациях каждого ученого, без вникания в особенности научной дисциплины, в которой работает эта группа, исходя только из ее размера, оценить по порядку величины ее суммарную продуктивность.

Отличительное свойство применения математических моделей к социальным процессам (которыми являются процессы производства и использования научной информации) — возможность не интересоваться глубинными причинами наблюдаемых закономерностей и приводящими к ним явлениями, происходящими на микроуровне. Несмотря на то что были предложены объяснения возникновения ряда информетрических законов («успех рождает успех», «принцип наименьших усилий», «принцип максимума энтропии»), собственно информетрическое моделирование строится без оглядки на микроуровень, с использованием закономерностей, обнаруженных в эмпирике, как данного.

Одна из основных задач информетрического моделирования, проводимого на конкретном информационном процессе, — это определение, насколько применим к рассматриваемому процессу, насколько корректно и насколько точно выполняется тот или иной информетрический закон, каковы особенности его применения в данном случае и каковы основные коэффициенты, получаемые в процессе приближения экспериментальных данных используемой моделью. Обобщение подобного рода исследований на целом ряде однотипных информационных процессов позволяет говорить о той или иной степени применимости изучаемых законов к произвольному аналогичному процессу.

Типичная сложность, с которой сталкивается ученый в процессе информетрического моделирования, — это переход от одних условий, в которых тестируется модель, к качественно или количественно другим. Это может быть вызвано изменением масштаба исследования (пример: переход от изучения рассеяния по журналам статей, написанных исследовательской лабораторией, к рассеянию публикаций целой страны), сменой дисциплинарной области, в которой происходит информационный процесс (анализ цитирований статей по математике или по биологии), или заменой самого процесса на аналогичный (переход от изучения статистики чтения какого-либо журнала к статистике его цитирования). Последний случай нам наиболее интересен.

С одной стороны, любой из процессов, подлежащих информетрическому моделированию, сходен с другим, являясь собственно «информационным процессом», или точнее, в терминологии Л. Эгге, «информационным процессом производства» (Information Production Process, IPP [59; 66, pp. 292, 313; 63, p. 101–102]). В нем можно выделить пары «источники — продукты». Например, при написании статей «источником» является автор, «продуктом» — публикация; при изучении рассеяния статей по журналам «источник» — это журнал, а «продукт» — вновь статья; в ходе анализа цитируемости продуктом будет цитирование, а источником — цитировавший или цитируемый автор, статья или журнал и т. д.

С другой стороны, нельзя априорно, без специальных исследований утверждать применимость информетрического закона для всех типов «информационных процессов производства» на основании его подтверждения для одного из таких процессов. Эти вопросы составляют основную проблематику настоящей диссертационной работы. Все законы информетрии, с использованием которых мы будем проводить информетрическое моделирование, исходно были открыты на информационных процессах иной

природы, чем те, применимость к которым мы исследуем. Мы моделируем обращение читателей к информационным источникам, т. е. процесс использования информации, в то время как законы Брэдфорда и Леймкулера были открыты для рассеяния статей по журналам, закон Лотки — для распределения статей по авторам, закон Ципфа сформулирован в лингвистических исследованиях частоты употребления слов, а правило Парето — вообще при изучении распределения доходов среди населения.

Итак, фокусировка нашего исследования на том, насколько корректно и гладко можно совершить перенос информетрического моделирования от «исходных» информационных процессов, давших рождение изучаемым моделям, к процессу обращения к документам, информационным ресурсам. Отметим, что сам процесс использования информации иногда выделяется из остальных «информационных процессов производства», так как существенным образом является вторичным информационным процессом: прежде чем информацию использовать, ее необходимо создать. Это позволяет Л. Эгге относить спрос на информационные источники даже к особому виду информетрии: «линейной трехмерной информетрии» (linear three-dimensional informetrics [63, p. 161–163]) или «type/token-taken informetrics» [61; 43; 63, p. 168–172] (в лингвистике пара «type/token» [84] идентична паре «источники — продукты» в информетрии, а прибавка «taken» указывает на изучение востребованности, «взятия» информации). «Трехмерность» здесь подразумевает наличие в информетрическом процессе сразу трех узлов, которые назовем: первичные источники, промежуточные продукты-источники и финальные продукты. Например, первичными источниками могут быть авторы статей (или журналы, опубликовавшие эти статьи); промежуточными продуктами-источниками — собственно опубликован-

ные статьи; финальными продуктами — цитирования, полученные данными статьями[2].

В нашем моделировании обращения к электронным информационным ресурсам в качестве источников будут фигурировать журналы, в качестве финальных продуктов — загрузки читателями полнотекстовых статей из этих журналов. Мы опускаем промежуточное звено «продуктов-источников» — собственно статьи, публикуемые журналами. Для простоты так регулярно делается в информетрических исследованиях. Например, при изучении числа публикаций различных стран опускается промежуточное звено — ученые этих стран, которые создавали данные публикации; при изучении цитируемости целых журналов нередко не интересуются отдельными статьями, которые и собирают цитируемость издания. В настоящей работе нам не понадобится промежуточный уровень, но, тем не менее, надо отдавать себе отчет, что мы моделируем информационный процесс более сложный, «составной», в отличие от простых процессов, на которых открывались используемые законы. Это, безусловно, добавляет важности вопросу об успешности планируемого перенесения информетрических моделей на исследуемый нами процесс.

Несмотря на изложенные соображения, рассматриваемые в настоящей работе информетрические законы уже неоднократно применялись к интересующему нас процессу — процессу использования информации. Подобных экспериментов проводилось достаточно много для законов Брэдфорда и Парето, где остро стоит вопрос о выделении «ядра» наиболее спрашиваемой и необходимой литературы, меньше — для законов Леймкулера,

[2] Вообще говоря процесс цитирования во многом близок к процессу обращения к информационным источникам: и в том и другом случае можно говорить об «использовании» информации — или для написания собственной работы или для простого ознакомления с публикацией. Изучению информетрического процесса цитирования и инструментов для его исследования посвящен ряд работ автора настоящей диссертационной работы [11; 12; 13; 15; 14; 109].

Лотки и Ципфа. Нередко такие опыты были достаточно успешными (обзоры библиографии приведены в соответствующих параграфах).

Отличительной чертой нашего исследования является попытка переноса целого комплекса информетрических моделей из традиционной «печатной» среды в среду электронную. Мы ограничиваемся чтением, обращением к *электронным* источникам, а именно базам данных периодики с онлайновым доступом, размещенным в сети Интернет (электронные информационные ресурсы JSTOR, EBSCO, ProQuest, ScienceDirect). Таким образом, «перенос» классических информетрических законов осуществляется нами одновременно в двух плоскостях: с тех информационных процессов, где они были исходно сформулированы, на процесс использования информации (при этом мы также переходим от «двумерной» информетрии к «трехмерной») и из печатного окружения — в электронное.

Актуальность проблемы заключается, прежде всего, в том, что исследование соединяет, во-первых, набирающие силу и авторитет в исследованиях науки и образования информетрические и библиометрические методы, во-вторых — завоевывающие всё большую популярность и получающие всё более широкое распространение в наше время электронные информационные ресурсы. Таким образом, в настоящей диссертационной работе объединяются актуальные методы исследования и современный объект, к которым эти методы применяются. В ней также затрагиваются такие насущные вопросы, как трактовка статистических показателей чтения онлайновых ресурсов в вузе, выделение наиболее важных, «ядерных» изданий из многотысячной их совокупности, построение наиболее оптимального фонда электронных документов при минимизации затрат.

Цели и задачи исследования

Целью исследования является применение информетрического моделирования к процессу обращения к электронным информационным ресурсам.

Задачами исследования являются:

— информетрическое моделирование при помощи законов Брэдфорда, Парето, Леймкулера, Лотки, Ципфа на основе экспериментальных данных об обращении к электронным информационным ресурсам, полученных в Государственном университете – Высшей школе экономики;
— получение и обсуждение основных коэффициентов, найденных в процессе приближения модели к экспериментальным данным;
— анализ применимости информетрических законов для моделирования спроса на электронные документы и исследование особенностей, характерных для такого рода спроса с точки зрения информетрии;
— выделение при помощи законов Брэдфорда, Парето и индекса Джини «ядра», т. е. наиболее спрашиваемых научных журналов в электронной подписке вуза.

Методы исследования

В диссертационной работе используются статистические методы, методы математического моделирования, методы интегрального и дифференциального исчисления, методы мониторинга обращений к онлайновым базам данных.

Современные средства анализа процесса обращения к документальным онлайновым базам данных позволяют с высокой степенью достоверности получить статистические показатели чтения пользователями электронных изданий, для того чтобы в дальнейшем применять их в процессе моделирования.

В качестве таких данных в настоящей работе выступает статистика использования электронных онлайновых источников в Государственном Университете – Высшей школе экономики. Сбор и анализ этой статистики производился на протяжении всего 2004 г. (январь-декабрь). Таким образом, имеется «живой» экспериментальный набор данных информетрического процесса. Обладая достаточным объемом для репрезентативности (всего зафиксировано 41959 обращений к статьям из 2590 различных изданий, размещенных в базах данных), этот набор может служить базой для информетрического моделирования.

Далее для каждого информетрического закона осуществляется приближение его коэффициентов для наилучшего совпадения с экспериментальными данными. Регрессионный анализ при определении оптимальных коэффициентов информетрической модели проводился при помощи статистического пакета SPSS версии 12.0 и 14.0. Степень совпадения модели и экспериментальных данных и, таким образом, степень применимости моделирования при помощи того или иного информетрического закона оценивается, там где это возможно, при помощи коэффициента детерминации (R-квадрат).

Научная новизна настоящего исследования содержит несколько аспектов. Прежде всего, в случае большинства информетрических законов (Брэдфорда, Леймкулера, Лотки, Ципфа) нам неизвестны труды предшественников, которые изучали бы применимость этих законов для моделирования процесса использования информации в электронной, онлайновой среде. Кроме того, столь объемный экспериментальный массив (41959 обращений к статьям из 2590 различных изданий) практически невозможно получить в «традиционном» печатном окружении, а потому вообще существует крайне мало исследований, которые осуществляли бы информетрическое моделирование на таком солидном материале. Наконец, использо-

вание для эксперимента онлайновых библиотек и систем учета статистики обращений к последним является более точным и достоверным.

К научной новизне данной диссертационной работы также относится введение в отечественный научный оборот ряда современных зарубежных источников по информетрии, которые прежде либо цитировались крайне редко, либо не рассматривались в российских исследованиях вообще.

Достоверность результатов работы в значительной степени определяется достоверностью исходного экспериментального массива, собранного методом анализа лог-файлов системами учета использования электронных ресурсов. Этот метод свободен от многих технических и методологических погрешностей других способов оценки обращений к периодическим изданиям (опросы, статистика возвратов журналов на полки, статистика выдач периодики на абонемент), а потому дает авторитетную экспериментальную базу для исследования. Кроме того, достоверность полученных результатов обеспечивают точные математические методы, применяемые в исследовании, а также профессиональные программные пакеты обработки и анализа статистической информации, использованные нами при информетрическом моделировании.

Практическая значимость диссертационного исследования заключается в возможности применить его методику и основные выводы к созданию и регулированию оптимального фонда онлайновой периодики учебной или научной организации, выделению информационного «ядра» и тех источников, которые являются ключевыми для обеспечения научной и образовательной деятельности.

Апробация результатов проводилась на ряде международных конференций: «SCIENCE ONLINE: электронные информационные ресурсы для науки и образования» (на II, IV, V, X конференциях в 2003, 2004, 2005 и 2007 гг.); «Международная конференция Крым: Библиотеки и информаци-

онные ресурсы в современном мире науки, культуры, образования и бизнеса» (на XI и XII конференциях в 2004 и 2005 гг.). Также результаты излагались на семинаре Отделения математического моделирования НИИ математики и механики им. Н. Г. Чеботарева Казанского государственного университета (2007 и 2008 гг., руководитель проф. А. М. Елизаров) и на семинаре «Математические методы анализа решений в экономике, бизнесе, политике» (2007 г., ГУ-ВШЭ, руководители семинара проф. Ф. Т. Алескеров и проф. В. В. Подиновский).

Публикации: результаты исследования опубликованы в 5 работах, в том числе в четырех статьях [3; 8; 9; 13] в журналах из Перечня, рекомендованного ВАК для публикации результатов диссертационных работ (одна из публикаций в соавторстве). Одна из этих статей [9] также была перепечатана в виде главы в коллективной монографии [10].

Структура и объем диссертации

Работа состоит из введения, четырех глав, заключения, списка литературы и трех приложений.

Содержание работы

Во **введении** производится постановка проблемы, определяется понятие информетрического моделирования и излагается краткая история вопроса.

Первая глава содержит обзор методов, применяемых в работе, и характеристик объектов настоящего исследования. Приведены описания и информетрические параметры электронных баз данных, включенных в исследование; изложено представление о процессе обращения к электронным изданиям с точки зрения информетрического моделирования; описаны инструментарий и процесс получения статистики использования электронных информационных ресурсов; дана характеристика полученного статистического массива и указаны операции, подготавливающие его к использова-

нию в качестве опорных экспериментальных данных при информетрическом моделировании.

Вторая глава посвящена моделированию процесса обращения к электронным информационным ресурсам при помощи закона Брэдфорда и состоит из трех параграфов. В параграфе 2.1 дается определение информетрического закона Брэдфорда в двух формах («вербальной», т. е. математической, и графической), излагаются основные методы определения оптимальных коэффициентов в модели Брэдфорда, а также приводится обзор литературы с фокусом на применении закона Брэдфорда к различным процессам использования информации. Далее в параграфе 2.2 при помощи закона Брэдфорда строится модель обращения к электронным источникам — реальные данные разбиваются на «зоны Брэдфорда» согласно математической формулировке закона, проводится подбор эмпирических коэффициентов закона различными методами, а также строится кривая Брэдфорда («библиограф»). Устанавливается, что закон Брэдфорда не выполняется в своей классической «вербальной» формулировке, однако график обращения к источникам хорошо соответствует предсказанному виду кривой Брэдфорда. В параграфе 2.3 излагаются подходы к выделению «информационного ядра» (наиболее спрашиваемых информационных источников) на основе применения модели Брэдфорда. Далее из имеющегося экспериментального массива выделяется ядро, двумя методами — кубической аппроксимацией кривой Брэдфорда с нахождением точки перегиба и привлечением математической формулировки закона для определения границы квазипрямолинейного участка библиографа.

Глава 3 касается информетрического моделирования с использованием закона (принципа, правила) Парето и состоит из четырех параграфов. В параграфе 3.1 дается представление о принципе Парето и приводится ряд его проявлений в различных областях знания, с обзором соответствующей

литературы. Параграф 3.2 содержит определение кривой Лоренца и индекса Джини, а также изложение одной из современных методик поиска ядра информационного спроса путем наблюдения за изменением индекса Джини в усеченной выборке журналов. Далее в параграфе 3.3 реальные данные по обращению к электронным информационным ресурсам проверяются на соответствие эмпирическому принципу Парето. Выясняется, что принцип Парето выполняется даже в более сильной форме, чем его классическое определение. Обсуждается различие между соответствием модели Парето экспериментальным данным для отдельных составляющих информационного массива (отдельных баз данных) и для всего массива в совокупности. Строится кривая Лоренца, численными методами определяется площадь под кривой и вычисляется индекс Джини. Делается вывод о высокой концентрации распределения количества обращений по информационным источникам, после чего в параграфе 3.4 определяется ядро изданий из электронных коллекций ГУ-ВШЭ при помощи закона Парето и индекса Джини. Это ядро сопоставляется с результатом, полученным в параграфе 2.3 с использованием закона Брэдфорда.

Глава 4 охватывает еще три информетрических закона, при помощи которых осуществляется моделирование процессов обращения к электронным информационным ресурсам. Это законы Леймкулера (параграф 4.1; здесь площадь под прямой Леймкулера - Лоренца вычисляется в аналитическом виде и сравнивается со значением, полученным в параграфе 3.3 численными методами), Лотки (параграф 4.2) и Ципфа (параграф 4.3). Делается вывод о принципиальной применимости информетрических моделей, базирующихся на данных законах, к процессам обращения к электронным источникам, и обсуждается наблюдаемое отклонение от «классических» вариантов данных закономерностей.

В **заключении** диссертационной работы подводятся итоги информетрического моделирования процессов обращения к электронным информационным ресурсам при помощи различных законов, делается вывод о пригодности большинства из них к такого рода исследованиям и обсуждаются перспективы дальнейшего применения информетрии к моделированию процессов производства и использования научной информации.

Положения диссертации, выносимые на защиту:

— Методология применения основных информетрических законов (Брэдфорда, Парето, Леймкулера, Лотки, Ципфа) для математическому моделированию процесса обращения читателей к электронным информационным источникам.

— Доказательство применимости основных информетрических моделей к экспериментальным данным по обращению читателей к электронным информационным источникам.

— Сформулированный подход к применению основных информетрических моделей для выявления «ядра литературы» (наиболее важных информационных источников).

— Построенная кривая Леймкулера – Лоренца, вычисленный индекс Джини и найденное «ядро литературы» для конкретных экспериментальных данных по обращению читателей к электронным информационным источникам.

1. Электронные издания и процесс их использования

1.1. Базы данных научных информационных источников: структура и классификация

Последние два десятилетия в информационном обслуживании научных исследований и образовательного процесса наметился существенный сдвиг, связанный с переводом всё большей доли литературы в электронный вид. При этом понятие «электронный вид» является достаточно широким, оно может подразумевать целый ряд принципов организации обращения к информации (локальный доступ, удаленный доступ, гибридный доступ, мобильный доступ) и различных информационных носителей (от магнитной ленты до DVD). В настоящем исследовании нас будет интересовать наиболее технологичная на данный момент форма предоставления научной информации — размещение ее в электронном виде в сети Интернет (размещение «в онлайне»).

Типы научных документов, присутствующих в Интернете, разнообразны. Это монографии, диссертации, препринты, труды конференций, энциклопедии и справочники, статистические таблицы и банки данных проведенных экспериментов и др. Однако со времени первого проникновения академической информации в сеть самым распространенным типом научного Интернет-документа является научная статья. Журналы не только наиболее полно представлены в сети, но и именно к ним идет максимальное число обращений пользователей, читателей.

Можно встретить в Интернете журнал, размещенный отдельно, на собственном специальном сайте, однако нередко академическая периодика

объединяется в журнальные коллекции — так выглядят наиболее серьезные ресурсы научных периодических изданий. Размещенные на единой технологической платформе журналы принято называть «журнальными базами данных». Как правило, в рамках базы данных предоставляются два основных способа доступа к информации — Browse и Search. В режиме Browse доступ осуществляется пользователем по иерархии: наименования журналов → список томов конкретного журнала → список выпусков тома журнала → список статей в выпуске → статья периодического издания. В режиме Search используется стандартный поиск по ключевым словам в тех или иных библиографических полях: автор, заглавие статьи, заглавие журнала, аннотация, полный текст, тематическая рубрика и др.

В журнальных базах данных уровень полноты представления журнала может быть разным: либо только выходные данные о статьях (чисто библиографическая информация), либо, в дополнение, аннотации статей («реферативные базы данных»), либо также списки цитируемой литературы (особые ресурсы, носящие название индексов цитирования), либо — наиболее полное представление журнала — полные тексты статей. Существуют базы, в которых часть изданий представлена в полном тексте, часть — аннотациями или библиографией.

В случае наличия полнотекстовых статей, последние могут быть представлены различными типами файлов. Наиболее распространены html- и pdf-файлы, однако существуют и более специализированные форматы: txt, gif, jpeg, tiff, djvu, а также специальные комбинации, когда, например, в html-текст графики, таблицы, иллюстрации и формулы внедряются в виде графических объектов.

Наконец, говоря о разнообразии представления научной информации в сети, следует уделить внимание и режиму доступа к ней. Немало научных журналов может быть найдено сегодня в открытом доступе, особенно

ввиду активизировавшегося в последние годы движения за открытый доступ к информации, Open Access Movement. Тем не менее большинство академических изданий (особенно из наиболее влиятельных) по-прежнему выпускается коммерческими издательствами, а значит как правило размещаются в тех или иных коммерческих ресурсах. В настоящей диссертационной работе будут фигурировать четыре именно таких ресурса.

Производителей коммерческих научных информационных ресурсов принято обобщенно называть «вендорами». Существует два выделенных типа вендоров — издательства и «агрегаторы». Первые (издательства) самостоятельно производят знание (вернее, участвуют в его производстве) — издают научные журналы — и размещают их на своих Интернет-платформах, предоставляя доступ по подписке различным организациям или частным лицам. Обычна ситуация, когда в электронном ресурсе, созданном и поддерживаемом издательством, размещаются журналы только этого издательства и его дочерних структур. Что касается агрегаторов, то они не производят журналы сами, однако заключают договоры, нередко с сотнями различных издательств, и размещают их издания на своей технологической платформе (созданной и поддерживаемой агрегатором). Базы данных издательств обычно содержат несколько сот журналов (для самых крупных издательств — может быть более тысячи), базы данных агрегаторов — несколько тысяч. У агрегаторов может быть смешанное представление журналов — часть в полных текстах, часть в аннотациях — у издательств же все журналы размещаются в полноценном, полнотекстовом варианте.

Издательству, не имеющему собственной технологической платформы, разумно разместить полные тексты своих журналов у какого-либо агрегатора, предоставив тому работу по поддержке (а может быть и созданию) электронных версий собственных изданий, зарабатывая при этом еще

некоторые отчисления из средств, полученных агрегатором от своих подписчиков. Однако даже те издательства, которые размещают изданные журналы в собственных электронных ресурсах, нередко предпочитают параллельно заключить договор с одним или несколькими агрегаторами и диверсифицировать таким образом потоки прибыли. Один из способов не создать в таком случае конкуренции своей платформе — отдавать издания в базы агрегаторов с некоторой задержкой относительно выхода печатной версии. Эта искусственная задержка называется в индустрии электронных ресурсов «эмбарго» и составляет чаще всего от 6 до 24 месяцев. В собственную базу данных издательство помещает журнал немедленно, эмбарго в таких случаях не бывает.

В настоящем диссертационном исследовании мы будем обращаться к статистике использования электронных ресурсов трех агрегаторов и одного издательства.

1.2. Процесс использования электронных изданий как объект информетрического моделирования

Выдающийся бельгийский специалист по информетрии, главный редактор выходящего с 2007 г. журнала «Journal of Informetrics» Л. Эгге вводит понятие «обобщенной библиографии» или «информационного процесса производства» (Information Production Process, IPP [59; 66, pp. 292, 313; 63, pp. 8, 101–102]). Простейший пример такого процесса — собственно библиография, набор статей по некоторой тематике. Статьи опубликованы в различных журналах, и в этом смысле журналы «производят» статьи. Также в качестве информационного процесса производства могут быть рассмотрены ссылки, размещенные в различных статьях (статьи «произво-

дят» ссылки), авторы, публикующиеся в разных дисциплинах (дисциплины в данном случае «производят» авторов) и др. Ключевой момент в информационном процессе производства — наличие этой пары «источники — продукты», когда «источники» производят «продукты». Например, в случае классической библиографии источником является журнал, продуктом — статья; в процессе цитирования продуктом будет цитирование, а источником — цитировавший или цитируемый автор, статья или журнал и т. д.

Формальное описание IPP, которое приводит Л. Эгге, заключается в наличии множества источников $\mathrm{S} = [0, T]$, множества «продуктов» $\mathrm{I} = [0, A]$ и строго возрастающей дифференцируемой функции $V : \mathrm{S} \to \mathrm{I}$, причем $V(r)$ равна кумулятивному числу продуктов, содержащихся в r наименее продуктивных источниках. Отметим, что это «непрерывный» IPP, т. к. r может принимать произвольные значения от 0 до T, в отличие от реального дискретного информационного процесса производства (такая замена нередко проводится, чтобы применить хорошо разработанный математический аппарат интегрального и дифференциального исчисления).

Л. Эгге также приводит «более абстрактный, но совершенно естественный» [63, p. 8] пример IPP — книговыдачу. В данном случае книга является источником, ее выдача читателю — продуктом. В этом смысле книга «производит» свою выдачу.

Аналогичные представления справедливы и в случае использования электронных изданий. Источник — электронный журнал, продукт — открытая статья. Заметим, что, когда мы говорим об использовании электронного издания, в стороне остаются вопросы о том, «насколько полезны» оказываются полученные пользователем материалы, насколько внимательно он ознакомился с ними, насколько в полной мере прочел их. Даже единица измерения статистических показателей, как мы увидим дальше, достаточно крупна — открытая статья. Мы не сможем зафиксировать откры-

тие и тем более чтение пользователем отдельных страниц статьи[3]. Можно сказать, что для учета «полноты» использования информации, заключенной в статье, логичнее обращаться к статистике цитирования документов — как правило, для цитирования публикации в собственном научном труде необходимо достаточно глубокое знакомство с ее содержанием. В этом случае мы до какой-то степени переходим к исследованию «полноты» использования, однако сразу теряем целый пласт читателей, — тех, которые не становятся «писателями», изучая научные статьи не для того, чтобы публиковать собственные исследования в соответствующих областях. Впрочем, ничто не мешает объединять данные количественные исследования чтения и цитирования, и одной из тенденций последнего времени является совместный анализ этих показателей [102; 32; 119; 15].

Можно провести деление информационных процессов производства на «простые» и «составные», для которых в терминологии Л. Эгге применима соответственно «двумерная» и «трехмерная» информетрия (см., например, [59]). Простой процесс имеет одно множество источников, одно множество продуктов и процесс, непосредственно производящий продукты из источников. Такого рода IPP возникают, когда исследуются, например:

— рассеяние статей по журналам;

— рассеяние статей по авторам;

— рассеяние цитирований по статьям;

— распределение выдач по книгам;

Существуют также ситуации, когда имеется одно множество «продуктов», которое производится двумя наборами источников (статьи «производят» журналы, но их же производят авторы), или, наоборот, одно множест-

[3] Лишь в одной из рассматриваемых здесь баз данных корректно считается показатель загрузки отдельных страниц — в ресурсе JSTOR. На основании этой уникальной статистики можно сделать вывод, что в среднем пользователи ГУ-ВШЭ открывают около двух страниц в каждой статье.

во источников, которое производит два множества продуктов (журналы производят статьи и они же производят выдачи своих номеров на абонемент). Эти ситуации являются «наложением» простых процессов, Л. Эгге поэтому называет описывающую их информетрию «трехмерной».

Однако есть особый случай — как мы говорим, «составной» информационный процесс производства. В этом случае имеется множество «первичных источников», производящих промежуточные продукты, которые одновременно являются, в свою очередь, источниками, производящими «окончательные», финальные продукты. Составной IPP при этом снова распадается на два простых, однако теперь они как бы «следуют друг за другом», один является продолжением и завершением другого, продукт в одном из них служит источником во втором. Например, журнал производит статьи, статьи получают (производят) цитирования. Тем самым итоговый составной процесс заключается в том, что журнал получает цитирования: исследуется цитируемость всего журнала.

Процесс использования электронных периодических изданий также является составным информационным процессом производства — ввиду того что пользователь обращается не к журналу в целом, а к отдельным его статьям. Первичными источниками являются журналы, производящие статьи, которые, в свою очередь, производят свои загрузки — акты открытия их пользователем базы данных. Можно не рассматривать отдельно этот промежуточный узел (статьи), однако стоит иметь в виду, что процесс использования онлайновой периодики является составным и его допустимо рассматривать через призму «линейной трехмерной информетрии» или, применяя вслед за Л. Эгге термины из математической лингвистики, «type/token-taken informetrics» [61; 43]. Некоторые математические последствия этого см., например, в [62; 63, p. 163–172].

Наконец, рассматривая процесс обращения к электронным изданиям как объект информетрического моделирования, необходимо отметить еще один крайне важный аспект: методы сбора статистики использования онлайновых источников лишены обыкновенных технических и методологических погрешностей, свойственных сбору статистики спроса на печатные документы (подробнее см. в следующем разделе). Это позволило провести в настоящем диссертационном исследовании корректное и достоверное информетрическое моделирование.

1.3. Системы сбора и обработки статистики использования электронных изданий

Остановимся теперь на процессе получения статистических данных по использованию электронных источников и изложим особенности методологии и технологии их сбора и обработки.

Прежде всего, напомним, какие существуют методы получения статистических данных по востребованности читателями фонда в печатной среде. Идеальной здесь является статистика выдач на абонемент или в читальный зал, зафиксированная автоматизированной информационно-библиотечной системой — компьютерной программой, при помощи которой книги выдаются читателям. Она легкодоступна (может быть автоматически получена из программы) и точна. Та же статистика, зафиксированная в бумажных формулярах (при отсутствии полной автоматизации процесса выдачи), хотя и остается точной, но на ее обработку будут потрачены значительные усилия.

Однако проблема заключается в том, что нередко — а в настоящее время все чаще и чаще — значительная часть фонда библиотеки находится

в «открытом доступе», где читателю необязательно обращаться за помощью к библиотекарю, чтобы взять издание в читальный зал. Следовательно, статистика спроса на литературу, не выдаваемую на абонемент, либо вообще отсутствует, либо для ее оценки применяются крайне трудоемкие и неточные методы (например, подсчет возвратов библиотекарями изданий на полки с читательских столов; опрос читателей; просьба к ним помечать на вкладыше в издании факт просмотра). При этом проблема состоит в том, что именно периодика (особенно текущая) как правило помещается в открытом доступе.

Но даже если бы в печатной среде удалось перечисленными методами получить точную статистику по обращению к тому или иному журналу, это все равно осталось бы принципиально недостижимым на уровне отдельных статей — только на уровне номеров периодического издания. После того как печатный номер попал в руки читателю, мы не можем сказать, сколько публикаций оттуда было открыто хотя бы на одной странице.

Перечисленные сложности технического и методологического характера успешно разрешаются в электронной среде. Прежде всего, в отличие от выдач печатной литературы, фиксирование статистических данных по востребованности онлайновых ресурсов происходит не на стороне библиотеки, а на стороне самого производителя ресурса, «вендора»[4]. Это делает сбор статистики минимально трудоемким для библиотеки: собрав статистику активности пользователей на основе анализа лог-файлов обращений

[4] Некоторые библиотеки все-таки предпочитают собирать статистику самостоятельно: это возможно, если доступ ко всей электронной подписке организован строго через аутентифицирующий сервер, «библиотечный прокси-сервер» (он управляет удаленным доступом, позволяя обращаться к базам данных не только из зданий организации-подписчика, но и из дома, общежития и т. д.). Можно также поручить фиксирование обращений к ресурсам третьей стороне, если доступ к подписке осуществляется через специальные аутентифицирующие системы, например Athens. В России на текущий момент такого рода схемы доступа практически отсутствуют: одна из первых установок библиотечного прокси проведена в ГУ-ВШЭ в 2008 г., а первый и пока тестовый доступ к Athens организован для российских организаций также в середине 2008 г.

к своим веб-серверам, вендор сам предоставляет ее в виде, удобном для анализа и интерпретации.

Формат представления данных может быть разнообразным, на выбор пользователя — веб-страницы или различные типы файлов данных, готовых для импорта в Excel или СУБД. Что касается фиксируемых показателей, то это, как правило, число пользовательских сессий (заходов в базу данных), совершенных поисков, просмотров аннотаций документов и их полных текстов и др. Для просмотра аннотаций/полных текстов возможно разбиение по отдельным наименованиям периодических изданий, к которым были обращения. Именно такое распределение по наименованиям потребуется нам в настоящем исследовании.

Все данные отфильтрованы для конкретной организации и предоставляются ее представителю («администратору» подписки на месте) на специальном служебном сайте по логину и паролю или, в редких случаях, доступны всем пользователям, вошедшим в ресурс с компьютеров соответствующей организации.

Особенность, специфическая для статистических систем большинства онлайновых информационных ресурсов (автору известно только одно исключение), — это подсчет не *уникальных* документов, к которым были обращения, а суммарного числа обращений к документам ресурса. При этом многократное обращение к одной и той же электронной публикации считается каждый раз — по числу отдельных обращений. Иными словами, для такой системы нет различия, было ли запрошено 100 различных документов по одному разу — или один определенный документ был открыт за исследуемый период времени 100 раз (разными пользователями или даже одним и тем же пользователем). И в том и в другом случае будет зафиксировано 100 обращений к цифровому объекту. Существуют лишь некоторые правила отсева случайных повторных открытий: при открытии дважды од-

ним и тем же пользователем одного и того же документа в html-формате с интервалом менее 10 секунд или при открытии дважды одним и тем же пользователем одного и того же документа в pdf-формате с интервалом менее 30 секунд, загрузка соответствующей статьи считается единственной. Надо отметить, что на момент проведения данного исследования даже это правило носило рекомендательный характер и нам неизвестно, насколько следовали ему рассматриваемые вендоры.

1.4. Методы и набор статистических данных, используемые в моделировании при помощи информетрических законов

В настоящем исследовании математические модели процесса обращения к электронным источникам будут применены к экспериментальным данным, полученным в библиотеке Государственного университета — Высшей школы экономики (ГУ-ВШЭ). Библиотека создана в 1994 г. и является небольшой по меркам российских университетских библиотек: в 2004 г. (время исследования) печатный фонд составлял ок. 70 тыс. наименований и ок. 350 тыс. экземпляров. За три года до исследуемого периода началось развитие электронной подписки на базы данных зарубежной периодики: в период с апреля по ноябрь 2001 г. была осуществлена подписка на ресурсы компаний-агрегаторов JSTOR, ProQuest и EBSCO, которые до сих пор являются ядром электронного фонда библиотеки. С 2004 г. также была оформлена подписка на электронные журналы издательства Elsevier (платформа ScienceDirect), которые активно использовались в течение всего года и даже превысили по востребованности одну из хорошо освоенных баз данных (EBSCO). Именно эти четыре ресурса (EBSCO, JSTOR, Pro-

Quest, ScienceDirect) пользовались наибольшим спросом в ГУ-ВШЭ и поэтому включены в настоящее исследование. Следует отметить, что, хотя все 4 компании предоставляют доступ к электронным изданиям, их ресурсы отличаются по своим задачам, структуре контента, объему доступной информации, равно как по другим параметрам. Так, база данных JSTOR является «архивным» проектом, не содержит текущих номеров журналов (за последние несколько лет) и включает в себя несколько сот изданий. В то же время EBSCO и ProQuest, напротив, делают особый акцент на текущей периодике, а счет журналов в этих ресурсах идет на тысячи. ScienceDirect, в отличие от остальных баз данных-агрегаторов, предоставляет доступ к журналам только одного издательства Elsevier и его дочерних структур.

Для удобства клиентов каждый из перечисленных вендоров позволяет подписаться не на весь имеющийся у него информационный массив, а на некоторый набор узко дисциплинарных или же политематических баз данных, образующих подмножество контента ресурса. Укажем те составляющие каждого ресурса, которые включены в анализ. В 2004 г. ГУ-ВШЭ имел доступ к четырем «коллекциям журналов» электронной библиотеки JSTOR: «Arts & Sciences I», «Arts & Sciences II», «Business», «Language & Literature» (всего 416 наименований). Ресурс ProQuest в течение 2004 г. был представлен базами данных «ABI/INFORM Global», «Academic Research Library», «ProQuest Education Journals», «Social Sciences PlusText», «ProQuest Psychology Journals» и «ProQuest Computing», что в сумме составило около 3600 полнотекстовых периодических изданий. Полнотекстовые ресурсы компании EBSCO, находящиеся в подписке ГУ-ВШЭ, включали в себя базы данных «Academic Search Premier», «Business Source Premier», «MasterFILE Premier», «Newspaper Source», «Regional Business News» и «Health Source», в сумме ок. 11000 полнотекстовых источников. Наконец, подписка на ScienceDirect охватывала коллекции «Business, Management

and Accounting», «Economics, Econometrics and Finance» и «Social Sciences» (всего 310 наименований периодики). Данные по составу и объему баз данных, которые были доступны для пользователей ГУ-ВШЭ в указанных ресурсах, суммированы в таб. 1.4.1. Введено ограничение на полнотекстовый материал, т. к. именно его использование будет учитываться в настоящем исследовании (см. далее).

платформа	доступные полнотекстовые базы данных	всего наименований источников
EBSCO	Academic Search Premier, Business Source Premier, MasterFILE Premier, Newspaper Source, Regional Business News, Health Source	ок. 11000
JSTOR	Collections: Arts & Sciences I, Arts & Sciences II, Business, Language & Literature	416
ProQuest	ABI/INFORM Global, Academic Research Library, ProQuest Education Journals, Social Sciences PlusText, ProQuest Psychology Journals, ProQuest Computing	ок. 3600
ScienceDirect	Business, Management and Accounting + Backfiles; Economics, Econometrics and Finance + Backfiles; Social Sciences + Backfiles	310

Таб. 1.4.1. Полнотекстовые ресурсы, участвующие в информетрическом моделировании

Как говорилось в параграфе 1.1, между контентом различных ресурсов существуют значительные пересечения (их структура была частично исследована в [7]). Поэтому в общей сложности в данных электронных ресурсах полнотекстовый доступ для библиотеки ГУ-ВШЭ открыт, за вычетом дубликатов, к около 13400 изданий. В ГУ-ВШЭ неоднократно проводились тренинги по использованию всех перечисленных ресурсов, актив-

ность обращений к ним достаточно высока: в течение периода наблюдения (2004 г.) в совокупности было отправлено более 57000 поисковых запросов и открыто почти 42000 полнотекстовых статей. Это позволяет проводить обобщения и обеспечивает статистическую значимость результатов исследования.

В трех из четырех рассматриваемых электронных ресурсов (за исключением JSTOR) содержатся не только полнотекстовые журналы, но также и издания, доступ к которым предоставлен читателям ГУ-ВШЭ лишь на уровне выходных данных статей или их аннотаций. Тем не менее, в настоящей работе нас будут интересовать только обращения к полным текстам, т. к. именно такое использование ресурсов можно считать «полноценным», эта статистика легче поддается интерпретации и более свободна от случайных заходов пользователей.

Для единообразного представления результатов анализа необходимо выбрать показатель использования, подсчитываемый статистическими модулями всех четырех ресурсов. Такой характеристикой является число открытых полнотекстовых статей. Этот показатель не только присутствует в сведениях, предоставляемых каждой электронной библиотекой, но и является наиболее значимым и поддающимся корректной интерпретации. В частности, именно на него предписывают опираться эксперты «Project COUNTER», выработавшие рекомендации для производителей баз данных по представлению статистики использования их ресурсов [89].

Статистические модули систем EBSCO, JSTOR, ProQuest и ScienceDirect подсчитывают не уникальные открытые документы, а суммарное число обращений к полным текстам ресурса: многократное обращение к одной и той же статье считается отдельно каждый раз — как это пояснялось в параграфе 1.3.

Данные по использованию объединены по всем четырем ресурсам, с суммированием показателей по наименованиям источников, представленных более чем в одном из них. Журналы идентифицировались по ISSN, при отсутствии ISSN идентификация проводилась по заглавию. Издания с различными ISSN и одинаковым заглавием считались разными. Не проводилось объединение данных по одному и тому же журналу, если он в течение своей истории изменял ISSN или наименование.

В качестве временного промежутка, на котором собиралась статистика обращений к базам данных, выбран весь 2004 г., с января по декабрь включительно. За данный период было зафиксировано 41959 обращений к полным текстам статей из 2590 уникальных наименований.

Существенную оговорку следует сделать относительно обработки статистики использования баз данных EBSCO и JSTOR. Во время периода наблюдения на одном из веб-серверов, расположенных в ГУ-ВШЭ, работал специальный робот «link checker» — программа, которая проверяла работу ссылок, размещенных на страницах данного веб-сервера. Ряд этих ссылок вел на ресурсы EBSCO и JSTOR, в связи с чем робот неоднократно за день обращался к данным ресурсам, чтобы проверить, что ссылки работают. Таким образом, активность робота во много раз превысила обращения к ресурсам читателей, и необходимо было провести дополнительную работу по вычленению заходов обычных пользователей. Было установлено, что роботом затронуты 28 журналов в JSTOR и 4 журнала в EBSCO, робот повлиял на статистику за август (частично), сентябрь, октябрь (частично), декабрь (частично). Был проведен детальный анализ показателей по этим журналам за указанные месяцы, и, исходя из предположения, что в период действия робота доля каждого затронутого им журнала в суммарной статистике была такой же, как и в течение всего остального года, были оценены показатели «чистого» спроса на журнал, без учета влияния link checker’a. Именно

эти исправленные цифры далее использовались в исследовании. Отметим, что суммарный вклад журналов из JSTOR, затронутых роботом, по восстановленным данным составляет за 4 указанных месяца менее 9,5% от всей годовой статистики ресурса. В случае журналов EBSCO этот показатель еще меньше — 0,8%.

В таб. 1.4.2 собраны абсолютные, а также процентные значения числа журналов, из которых была открыта хотя бы одна статья (для каждого ресурса). В столбце «число полнотекстовых изданий» приводятся лишь приблизительные цифры для EBSCO и ProQuest: специфика данных ресурсов, крупных агрегаторов, не позволяет установить с точностью до единиц число доступных журналов. Последняя строка отражает всю подписку как единый массив, показывая число уникальных изданий и убирая дублеты.

	число полнотекстовых изданий	изданий, к полным текстам которых были обращения	доля востребованных изданий
EBSCO	11000	1105	10%
ProQuest	3600	1452	40%
JSTOR	416	306	74%
ScienceDirect	310	236	76%
всего	**13400**	**2590**	**19%**

Таб. 1.4.2. Количество полнотекстовых журналов и число изданий, которые были хотя бы один раз востребованы за период исследования

Как видно из таб. 1.4.2, далеко не все издания из электронных библиотек используются. При этом для крупных агрегаторов (EBSCO и ProQuest) характерен меньший процент востребованных журналов. Кроме того, наблюдается яркая обратная зависимость процента востребованных изданий от суммарного их количества в ресурсе, коэффициент корреляции составляет $-0{,}97$. Это закономерно: все ресурсы обеспечивают «пакетную» под-

писку, когда журналы не выбираются заказчиком один за другим, а берутся сразу в сформированных вендором коллекциях. Таким образом, вместе с остро необходимыми подписчику изданиями в электронный фонд практически неизбежно поступает и «информационный шум» — источники, не представляющие интереса для читателей. И чем больше размер таких коллекций, чем больше размер информационного массива, предоставляемого вендором, тем большая будет в нем доля изданий, не отвечающих запросам пользователей и поэтому не востребованных ими. Для сравнения, в недавно появившемся исследовании [24] по данным 2001 г. сообщается об использовании за полугодие учеными Лос-Аламосской национальной лаборатории лишь 1892-х электронных журналов из 20000 доступных (9,5%).

Во всем процессе информетрического моделирования нами учитываются только источники (журналы), к которым были обращения. Не востребованные ни разу журналы полностью исключаются из рассмотрения. Аналогичный подход использует, например, К. Бёррел [40], аргументируя это тем, что истинное число неиспользуемых источников определяется с большой неуверенностью. Он также ссылается на свои работы [38; 39; 44], где тоже использует такой подход. В [41] К. Бёррел подробно останавливается на этой проблеме: невозможно знать, сколько ученых *не* написали статьи по той или иной тематике за некоторый промежуток времени или сколько журналов не опубликовали статьи по определенной дисциплине. Даже в том случае, когда «непродуктивные» источники более-менее ясны (например, при исследовании выдач книжного фонда), неопределенность сохраняется: книга может не выдаваться по той причине, что она, например, украдена. В нашем случае применима сходная логика: не всегда списки источников, предоставляемые производителями электронных ресурсов, на 100% оперативно отражают фактический контент баз данных, который, к тому же, еще может несколько изменяться за время исследования (осо-

бенно у агрегаторов). Отметим, однако, что в работах [19; 85; 92] К. Бёррелу оппонирует по данному вопросу ряд других ученых.

В заключение настоящего параграфа диссертационной работы, носящего методический характер, укажем, что все статистические вычисления, приближения, регрессии выполнены нами при помощи статистического пакета SPSS версий 12.0 и 14.0.

2. Моделирование процесса обращения к электронным информационным источникам при помощи информетрического закона Брэдфорда

2.1. Формулировка закона Брэдфорда

Закон рассеяния информации Брэдфорда (Bradford’s law of scattering) был впервые сформулирован Самюэлем Брэдфордом в 1934 г. [29] и в более поздней работе [30]. В отечественной традиции закон также называется «законом концентрации и рассеяния информации». Важность закона Брэдфорда позволяет некоторым исследователям особенно подчеркивать его роль в плеяде информетрических закономерностей — см. заглавие обзорной статьи Ю. Гарфилда «Закон Брэдфорда и связанные с ним статистические закономерности» [75], где речь идет о законах Лотки, Ципфа, Лейм-кулера и др.

Закон Брэдфорда получил два типа формулировки — «вербальную» и «графическую».

2.1.1. Вербальная формулировка закона Брэдфорда

«Вербальный» вариант закона Брэдфорда формулируется следующим образом. Выделим множество журналов, в которых появляются статьи, посвященные некоторой выбранной научной тематике, и упорядочим их в порядке убывания количества этих статей — от изданий, в которых имеется наибольшее количество публикаций данной тематики, до тех, в которых встретилась лишь одна статья, относящаяся к выбранной дисциплине. Тогда это упорядоченное множество журналов можно разбить на три зоны так, чтобы в каждой было одинаковое количество статей по заданной теме.

При этом, утверждает закон Брэдфорда, числа журналов в соответствующих зонах будут относиться друг к другу как $1 : q : q^2$, где q — некоторое число большее единицы.

Иными словами, в наборе журналов существует некоторое «ядро» изданий, содержащее треть всех статей по заданной тематике. Для того чтобы найти столько же статей в других журналах, необходимо взять их в q раз большее число. Наконец, оставшиеся публикации будут рассеяны по изданиям, которых в q^2 раз больше, чем в «ядре». Закон при этом не регулирует величину коэффициента q (кроме очевидного неравенства $q > 1$), он лишь утверждает, что отношение мощности второй «зоны» к мощности первой равняется отношению мощностей третьей и второй зон (естественно, как и в случае любого информетрического закона, речь идет не о точном равенстве, а о приблизительном совпадении).

Если исходить из оригинальной вербальной формулировки закона, можно связать коэффициент q с относительной величиной «ядра» журналов следующим образом. Пусть всего источников (журналов) имеется S, при этом в ядро (первую зону Брэдфорда) входит S_1 из них. Тогда во второй зоне имеется qS_1 источников, в третьей — q^2S_1. Следовательно, справедливо равенство

$$S_1 + qS_1 + q^2S_1 = S, \qquad (2.1.1)$$

откуда

$$1 + q + q^2 = \frac{S}{S_1}.$$

Решив уравнение относительно q, получим для положительного корня:

$$q = -\frac{1}{2} + \sqrt{\frac{S}{S_1} - \frac{3}{4}}. \qquad (2.1.2)$$

На практике ядро, как правило, составляет небольшая доля от общего числа изданий и $\frac{S}{S_1} >> 1$, а квадратный корень из этого отношения также ощутимо превышает единицу. Так, у Брэдфорда в [29] S/S_1 для прикладной геофизики равно 36, для литературы по смазочным технологиям — 20. Поэтому справедлива упрощенная оценка:

$$q \approx \sqrt{S/S_1}\,, \tag{2.1.3}$$

т. е. коэффициент пропорциональности для мощностей последовательных зон Брэдфорда приближенно равен квадратному корню из отношения суммарного числа всех рассматриваемых источников (журналов) к величине первой, «ядерной» зоны. Заметим, что равенство (2.1.3) эквивалентно пренебрежению в (2.1.1) всеми членами кроме q^2.

Хотя сам С. Брэдфорд исследует три «зоны» источников, формулировка закона в [29] позволяет предположить, что его можно трактовать в виде, обобщенном относительно количества зон. Множество журналов может быть разбито не на три, а на произвольное число зон. При этом число изданий в каждой последующей зоне будет в q раз больше, чем в предыдущей, а число статей, приходящихся на журналы каждой зоны, будет одинаковым и равным, очевидно, отношению общего числа статей к количеству зон.

В случае n зон Брэдфорда справедливо равенство (аналогичное (2.1.1) для трех зон):

$$S_1 + qS_1 + \ldots + q^{n-1}S_1 = S,$$

следовательно,

$$\frac{S}{S_1} = 1 + \ldots + q^{n-1} = \frac{q^n - 1}{q - 1}. \tag{2.1.4}$$

Эта формула верна для идеального брэдфордского распределения. Для него важно правильно установить размер «ядра» S_1, далее по количеству статей, приходящихся на ядерные журналы, устанавливается число n зон Брэдфорда (оно очевидно равно отношению общего числа публикаций I к числу «ядерных» статей), а по формуле (2.1.4) определяется коэффициент в распределении Брэдфорда q.

Однако на практике закон Брэдфорда, как и любая информетрическая модель, редко выполняется в совершенстве. Поэтому можно выбирать различные величины для S_1, n, q и получать более или менее хорошие разбиения на зоны Брэдфорда — в большей или меньшей степени удовлетворяющие закону. Отсюда возникает задача определения таких параметров распределения Брэдфорда, чтобы оно максимально хорошо приближало исходные экспериментальные данные. При этом необходимо, чтобы искомые параметры могли быть найдены из «макроскопических» характеристик набора данных. Таковыми можно считать четыре величины:

— общее число «источников» (журналов) S;

— общее число «продуктов» (статей) I;

— число источников с минимальной продуктивностью s_1 (т. е. количество журналов, которые опубликовали лишь одну статью по рассматриваемой тематике);

— количество продуктов, произведенных источником с максимальной продуктивностью i_{max} (т. е. число статей, опубликованных в самом продуктивном по рассматриваемой тематике журнале).

Если мы имеем перед собой набор экспериментальных данных, все перечисленные характеристики могут быть немедленно получены из него. Необходимо определить следующие параметры модели Брэдфорда:

— число источников в «ядре» (первой зоне Брэдфорда) S_1;

— число продуктов в каждой зоне Брэдфорда I_B;

— коэффициент распределения Брэдфорда q;

— число зон Брэдфорда n.

Решить данную задачу по моделированию невозможно, опираясь только на модель Брэдфорда. Потребуются дополнительные ограничения, накладываемые на набор данных, а именно предположение их соответствия некоторым иным информетрическим законам. Впрочем, информетрические модели тесно связаны друг с другом и такого рода предположения, по меньшей мере, обоснованны.

К решению данной задачи имеется несколько подходов, например изложенный в работах [133; 17] (с некоторыми вариациями) метод отечественного ученого А. И. Яблонского. Приведем его выкладки, более детально и с некоторым обобщением.

А. И. Яблонский исходит из предположения, что данные удовлетворяют закону Ципфа-Мандельброта (см. далее, раздел 4.3, формула (4.3.3)) и, соответственно, подчиняются распределению:

$$i_r = \frac{s_1}{r + a}, \qquad (2.1.5)$$

где i_r — число статей в источнике, занимающем r-е место в списке журналов, упорядоченном в порядке убывания продуктивности (т. е. от наиболее к наименее продуктивному), а коэффициент $a \approx s_1 / i_{\max}$.

Далее для выполнения закона Брэдфорда приравниваются вычисленные при помощи (2.1.5) количества статей в журналах первой и последующих зон Брэдфорда (число этих журналов, по обозначению, равно $S_1, S_2, S_3 \ldots$):

$$\sum_{1}^{S_1} \frac{s_1}{r + a} = \sum_{S_1+1}^{S_1+S_2} \frac{s_1}{r + a} = \ldots = \sum_{S_1+\ldots+S_{n-1}+1}^{S_1+\ldots+S_{n-1}+S_n} \frac{s_1}{r + a}.$$

Суммирование можно заменить интегрированием — метод приближения, часто используемый при математическом анализе информетрических моделей:

$$\int_{1}^{S_1} \frac{s_1}{r+a}\,dr = \int_{S_1+1}^{S_1+S_2} \frac{s_1}{r+a}\,dr = \ldots = \int_{S_1+\ldots+S_{n-1}+1}^{S_1+\ldots+S_{n-1}+S_n} \frac{s_1}{r+a}\,dr .$$

Следовательно, после сокращения на s_1 и проведения интегрирования:

$$\ln\frac{S_1+a}{a+1} = \ln\frac{S_1+S_2+a}{S_1+a+1} = \ldots = \ln\frac{S_1+\ldots+S_{n-1}+S_n+a}{S_1+\ldots+S_{n-1}+a+1}. \qquad (2.1.6)$$

Из первого равенства этой цепочки следует, что S_2 так выражается через S_1:

$$S_2 = S_1\frac{S_1+a}{a+1}.$$

Данное выражение позволяет предположить, что коэффициент Брэдфорда q равен $\frac{S_1+a}{a+1}$. Докажем это по индукции. Пусть для решений уравнений (2.1.6) верно равенство:

$$S_j = S_1 q^{j-1}, \quad j = 1\ldots m, \qquad (2.1.7)$$

где $q = \frac{S_1+a}{a+1}$. Покажем, что для j=m+1 равенство также справедливо. Из m-го уравнения цепочки (2.1.6) получаем

$$\frac{S_1+a}{a+1} = \frac{S_1+\ldots+S_m+S_{m+1}+a}{S_1+\ldots+S_m+a+1}$$

или

$$q = 1+\frac{S_{m+1}-1}{(S_1+\ldots+S_m)+a+1}.$$

Заменим, использовав (2.1.7), $\sum_{j=1}^{m} S_j$ на $S_1\frac{q^m-1}{q-1}$ и запишем выражение для S_{m+1}:

$$S_{m+1} = (q-1)\left(S_1 \frac{q^m - 1}{q-1} + a + 1 \right) + 1 = S_1 q^m + (a+1)q - (S_1 + a).$$

Так как $q = \frac{S_1 + a}{a+1}$, в правой части остается только первый член и мы получим искомое равенство $S_{m+1} = S_1 q^m$. Таким образом, мы доказали, что коэффициент Брэдфорда q связывается с мощностью «ядра» журналов выражением

$$q = \frac{S_1 + a}{a+1}. \qquad (2.1.8)$$

Далее, исходя из того факта, что распределение удовлетворяет закону Ципфа-Мандельброта и подчиняется закону (2.1.5), можно записать зависимость кумулятивного числа статей от числа рассматриваемых журналов (упорядоченных, как обычно, в порядке убывания в них статей):

$$I(S) = \int_1^S \frac{s_1}{r+a} dr = s_1 \ln \frac{S+a}{a+1} \qquad (2.1.9)$$

(здесь, напомним, S — это кумулятивное число журналов, а s_1 — число журналов с минимальной продуктивностью, т. е. производящих только одну статью по рассматриваемой тематике).

С другой стороны, согласно (2.1.4), число журналов в первых m зонах Брэдфорда равно

$$S(m) = S_1 \frac{q^m - 1}{q-1} \qquad (2.1.10)$$

(S_1 — число источников в первой зоне Брэдфорда). Выразим здесь m через количество статей, относящихся к этому числу зон Брэдфорда. Очевидно, что $m = I / I_B$ (I_B — число статей в первой, следовательно, и в каждой, зоне). Сделав эту подстановку в (2.1.10), можно снова выразить I через S:

$$I(S) = \frac{I_B}{\ln q} \ln\left(\frac{q-1}{S_1} S + 1\right). \qquad (2.1.11)$$

Теперь сопоставим (2.1.9) и (2.1.11), приравняем коэффициенты при логарифмической функции и получим еще одну зависимость между параметрами распределения Брэдфорда:

$$I_B = s_1 \ln q. \qquad (2.1.12)$$

Для получения заключительного равенства, которого не хватает для определения параметров Брэдфорда, А. И. Яблонский в [17] вводит предположение, что последняя зона Брэдфорда состоит только из источников с минимальной продуктивностью, т. е. из журналов с одной статьей. В этом случае число продуктов в последней зоне Брэдфорда I_B совпадает с числом источников в этой зоне, которое равняется $S_1 q^{n-1}$ (n — общее число зон Брэдфорда). При этом суммарное число источников выводится из формулы (2.1.4), поэтому мы имеем два равенства:

$$I_B = S_1 q^{n-1},$$

$$S = S_1 \frac{q^n - 1}{q - 1},$$

откуда, поделив второе на первое, получим

$$\frac{S}{I_B} = \frac{q^n - 1}{q^n} \frac{q}{q-1}. \qquad (2.1.13)$$

Так как q^{n-1} равно отношению числа источников в последней и первой зонах Брэдфорда, $q^n > q^{n-1} >> 1$. Следовательно, в последнем равенстве можно пренебречь членом $\frac{q^n - 1}{q^n}$ по сравнению с $\frac{q}{q-1}$. Учитывая это и подставив в (2.1.13) выражение для I_B из (2.1.12), получим:

$$\frac{S}{s_1 \ln q} = \frac{q}{q-1}.$$

В итоге найдем уравнение, из которого может быть найден коэффициент распределения Брэдфорда q:

$$\ln q = \frac{\mathrm{S}}{s_1}\frac{q-1}{q}. \qquad (2.1.14)$$

Точное решение уравнения не может быть получено в аналитическом виде. Если же предположить, что q близко к единице (как делает А. И. Яблонский), то $\ln q \approx q-1$ и (2.1.14) дает решение

$$q = \frac{\mathrm{S}}{s_1}. \qquad (2.1.15)$$

Далее, из (2.1.13), вновь пренебрегая членом $\frac{q^n-1}{q^n}$ и подставив (2.1.15), получим выражение для числа статей в каждой зоне Брэдфорда:

$$\mathrm{I_B} = \mathrm{S}\frac{q-1}{q} = \mathrm{S} - s_1. \qquad (2.1.16)$$

Это размер каждой зоны Брэдфорда для *продуктов* (статей). Формула для размера первой зоны Брэдфорда для *источников* (т. е. размер ядра журналов) может быть получена из (2.1.8):

$$\mathrm{S_1} = a(q-1) + q = s_1(q-1)/i_{\max} + q \qquad (2.1.17)$$

(здесь А. И. Яблонский отбрасывает второй член, пренебрегая q по сравнению с $a >> 1$).

Совокупность уравнений (2.1.15) (т. е. приближенного (2.1.14)), (2.1.16) и (2.1.17) определяет параметры распределения Брэдфорда, исходя из «макрохарактеристик» экспериментальных данных. Суммарное число зон Брэдфорда получается делением общего числа продуктов I (до сих пор, заметим, не фигурирующего в полученной системе решений) на $\mathrm{I_B}$, вычисляемое по формуле (2.1.16). А. И. Яблонский проверил выведенные параметры на экспериментальных данных из [56] и получил хорошее совпадение эмпирики с математической моделью.

А. Букстейн (по сообщению Л. Эгге [58]) указал на следующий недостаток формул Яблонского. В принципе выбор «ядра» журналов может иметь некоторую свободу и при хорошем следовании распределения закону Брэдфорда возможно получить различные коэффициенты q. Этой свободы в строгом результате Яблонского не заложено. Таким образом, более общее решение должно иметь дополнительный параметр.

Кроме того, мы можем отметить как недостаток то, что вычисление q опирается на предположение, что в последней зоне Брэдфорда присутствуют только журналы с минимальной продуктивностью. Это также накладывает некоторые ограничения на результаты Яблонского. Поэтому Л. Эгге в [58] предложил усовершенствование метода оценки параметров закона Брэдфорда исходя из того, что распределение удовлетворяет закону Лотки (см. далее параграф 4.2, ср. формулу (4.2.6)):

$$s_i = \frac{s_1}{i^2},$$

где s_i — число источников (журналов) с продуктивностью i (с i статьями). Рассмотрим тогда зону Брэдфорда с номером p от конца (т. е. первой считается самая большая зона, этот «обратный» отсчет удобен в последующих выкладках). Пусть в нее входят журналы с m_{p-1}, $m_{p-1}+1$, …, m_p статьями в каждом. Тогда в эту зону Брэдфорда входит следующее число статей (применяем закон Лотки):

$$\begin{aligned} \mathrm{I_B} = \left(1-\alpha_{p-1}\right)\left(\frac{s_1}{m_{p-1}^2}\right)m_{p-1} + \left(\frac{s_1}{\left(m_{p-1}+1\right)^2}\right)\left(m_{p-1}+1\right) + \ldots + \\ + \left(\frac{s_1}{\left(m_p-1\right)^2}\right)\left(m_p-1\right) + \alpha_p\left(\frac{s_1}{m_p^2}\right)m_p = \\ = \left(1-\alpha_{p-1}\right)\frac{s_1}{m_{p-1}} + \frac{s_1}{m_{p-1}+1} + \ldots + \frac{s_1}{m_p-1} + \alpha_p\frac{s_1}{m_p}, \end{aligned} \qquad (2.1.18)$$

где α_{p-1} — доля журналов с m_{p-1} статьями, которая принадлежат (p–1)-й зоне Брэдфорда (остальная доля $(1-\alpha_{p-1})$ принадлежит уже p-й зоне), α_p — доля журналов с m_p статьями, которая принадлежит p-й зоне. В левую часть данного равенства можно подставить выражение (2.1.12) для I_B (которое, заметим, было получено без предположения о том, что последняя зона Брэдфорда состоит из источников с минимальной продуктивностью), в результате получим

$$\begin{aligned}\ln q &= \left(1-\alpha_{p-1}\right)\frac{1}{m_{p-1}} + \frac{1}{m_{p-1}+1} + \ldots + \frac{1}{m_p - 1} + \alpha_p \frac{1}{m_p} = \\ &= \sum_{j=m_{p-1}}^{m_p-1} \frac{1}{j} - \frac{\alpha_{p-1}}{m_{p-1}} + \frac{\alpha_p}{m_p}.\end{aligned} \tag{2.1.19}$$

Отсюда выразим α_p через α_{p-1}:

$$\alpha_p = m_p\left(\ln q + \frac{\alpha_{p-1}}{m_{p-1}} - \sum_{j=m_{p-1}}^{m_p-1}\frac{1}{j}\right). \tag{2.1.20}$$

Докажем по индукции, что

$$\alpha_p = m_p\left(p\ln q - \sum_{j=1}^{m_p-1}\frac{1}{j}\right). \tag{2.1.21}$$

Выражение для α_1 получим непосредственно из (2.1.18), которое в случае первой зоны Брэдфорда будет выглядеть так:

$$I_B = \left(\frac{s_1}{1^2}\right)\cdot 1 + \left(\frac{s_1}{2^2}\right)\cdot 2 + \ldots + \left(\frac{s_1}{\left(m_1-1\right)^2}\right)\left(m_1-1\right) + \alpha_1\left(\frac{s_1}{m_1^2}\right)m_1.$$

Подставив (2.1.12), найдем

$$\alpha_1 = m_1\left(\ln q - \sum_{j=1}^{m_1-1}\frac{1}{j}\right).$$

Таким образом, база индукции проверена, α_1 удовлетворяет (2.1.21). Далее предположим, что равенство (2.1.21) выполняется для p, проверим, что тогда оно верно и для $p+1$. Из (2.1.20) найдем выражение для α_{p+1} и подставим в него (2.1.21):

$$\alpha_{p+1} = m_{p+1}\left(\ln q + \frac{\alpha_p}{m_p} - \sum_{j=m_p}^{m_{p+1}-1}\frac{1}{j}\right) = m_{p+1}\left(\ln q + p\ln q - \sum_{j=1}^{m_p-1}\frac{1}{j} - \sum_{j=m_p}^{m_{p+1}-1}\frac{1}{j}\right) =$$

$$= m_{p+1}\left((p+1)\ln q - \sum_{j=1}^{m_{p+1}-1}\frac{1}{j}\right)$$

Поэтому шаг индукции также доказан и (2.1.21) верно.

По определению m_p в p-й зоне Брэдфорда есть как минимум один журнал с m_p статьями и нет ни одного с m_p+1 статьями. Это означает, что коэффициенты α_p должны быть строго больше нуля и не больше единицы. Записав это двойное ограничение и использовав (2.1.21), получим:

$$0 < \alpha_p \le 1 \quad \Rightarrow \quad 0 < m_p\left(p\ln q - \sum_{j=1}^{m_p-1}\frac{1}{j}\right) \le 1 \quad \Rightarrow$$

$$\Rightarrow \quad \frac{1}{p}\sum_{j=1}^{m_p-1}\frac{1}{j} < \ln q \le \frac{1}{p}\sum_{j=1}^{m_p}\frac{1}{j}. \qquad (2.1.22)$$

На самом деле неравенства (2.1.22) могут рассматриваться как ограничение на m_p. Очевидно, что уже при относительно небольших m_p это двойное неравенство можно приблизить равенством и, учитывая, что по определению постоянная Эйлера-Маскерони $\gamma = \lim_{n\to\infty}\left(\sum_{j=1}^{n}\frac{1}{j} - \ln n\right)$, получить

$$\ln q \approx \frac{1}{p}\sum_{j=1}^{m_p}\frac{1}{j} \approx \frac{1}{p}\left(\ln m_p + \gamma\right),$$

откуда

$$m_p \approx q^p \Big/ e^{\gamma} \quad . \qquad (2.1.23)$$

При помощи этой формулы можно вывести теоретическое значение для q, использовав то, что максимальное число статей в журнале *из последней зоны* (если вести подсчет, как в данных выкладках, от зоны с журналами наименьшей продуктивности, т. е. из первой зоны при «обычной» нумерации) просто равняется максимальной продуктивности журнала во всей выборке, $i_{\max}$. Таким образом, если n — суммарное количество зон Брэдфорда, то $m_n = i_{\max}$ и из (2.1.23) выразим коэффициент Брэдфорда q:

$$q = \left(e^{\gamma}\, i_{\max}\right)^{1/n}. \qquad (2.1.24)$$

На основании изложенных расчетов, Л. Эгге в работе [60] (см. также [66, p. 343–345]) сформулировал методику «подгонки» параметров модели Брэдфорда к конкретным данным, которая в разделе 2.2 будет реализована нами на эмпирическом массиве показателей обращений к электронной периодике:

— выбрать n — число групп Брэдфорда; это произвольное число, обычно в промежутке от 4 до 10, однако для больших библиографий оно может быть и больше; выбор его может быть регламентирован тем, чтобы в итоге получить более удобное, т. е. более близкое к целому числу, значение S_1, величины первой зоны Брэдфорда (см. формулу далее);

— по формуле (2.1.24) вычислить, коэффициент Брэдфорда q;

— применить (2.1.4) для вычисления размера первой зоны Брэдфорда S_1, т. е. использовать формулу:

$$S_1 = S\frac{q-1}{q^n - 1}; \qquad (2.1.25)$$

— если S_1 получается не целым, то надо использовать округление до ближайшего меньшего целого $[S_1]$; число статей, пришедшихся на $[S_1]$ журналов, будет числом статей в первой зоне Брэдфорда;

— проводить дальнейшее разбиение статей, находя по библиографии, сколько их приходится на следующие группы Брэдфорда, содержащие $\{q\mathrm{S}_1\}$, $\{q^2\mathrm{S}_1\}$ и т. д. журналов (причем берется именно $\{q\mathrm{S}_1\}$, а не $\{q[\mathrm{S}_1]\}$); здесь и далее фигурными скобками будем обозначать округление до ближайшего целого, т. е. $\{x\} = [x + 0{,}5]$.

Обратим внимание на то, что округление для величины первой группы Л. Эгге предлагает делать в меньшую сторону, $[\mathrm{S}_1]$, оговариваясь, что можно округлять и до ближайшего целого, но тогда, в случае если дробная часть S_1 (т. е. $\mathrm{S}_1 - [\mathrm{S}_1]$) немногим больше 0,5, то S_1 округляется в большую сторону, а последняя группа Брэдфорда не будет заполнена до конца. Тем не менее это не так: в связи с тем что для вычисления размера последующих групп используется точное, а не округленное значение S_1, ошибки округления могут потребовать лишь незначительных корректив величины последней группы (в случае если сумма округлений $\{\mathrm{S}_1\} + \{q\mathrm{S}_1\} + \{q^2\mathrm{S}_1\} + \ldots + \{q^{n-1}\mathrm{S}_1\}$ окажется больше или меньше суммарного числа источников S). Так как последняя группа включает в себя источники с меньшей продуктивностью, эта коррекция практически не повлияет на выраженные в «продуктах» величины групп Брэдфорда, зато «выровняет» отношение величин групп, выраженных в числе источников. Сам Л. Эгге также использовал округление в большую сторону, но только в том случае когда вычисленное значение S_1 оказывалось очень близким к большему целому: в [60] он округляет 8,93 до 9 и 2,95 до 3. Но, например, для значения $\mathrm{S}_1 = 2{,}7$ он в явном виде рекомендует использовать округление до двух.

Описанный способ определения параметров модели Брэдфорда (как с введенными нами поправками правил округления, так и без них) будет применен в разделе 2.2 к набору статистических данных по использованию электронных источников в библиотеке ГУ-ВШЭ.

2.1.2. Графическая формулировка закона Брэдфорда

Графическая формулировка закона Брэдфорда также ведет свою историю от основополагающей работы С. Брэдфорда [29]. Вновь рассматривается множество журналов, в которых появляются статьи, посвященные некоторой научной тематике, причем журналы упорядочены в порядке убывания количества этих статей. Графическую формулировку закона Брэдфорда иллюстрирует рис. 2.1.1. Отложим по оси абсцисс натуральный логарифм числа первых *s* наиболее продуктивных журналов, а по оси ординат — кумулятивное (суммарное) число релевантных статей *i*, содержащихся в этих *s* изданиях. Тогда, согласно графической формулировке закона Брэдфорда, получившаяся кривая в полулогарифмических координатах (ее называют «библиограф») будет вогнутой до некоторого значения $s = S_1$, после которого перейдет в прямую (рис. 2.1.1). При этом, согласно С. Брэдфорду, именно S_1 соответствует границе первой зоны Брэдфорда.

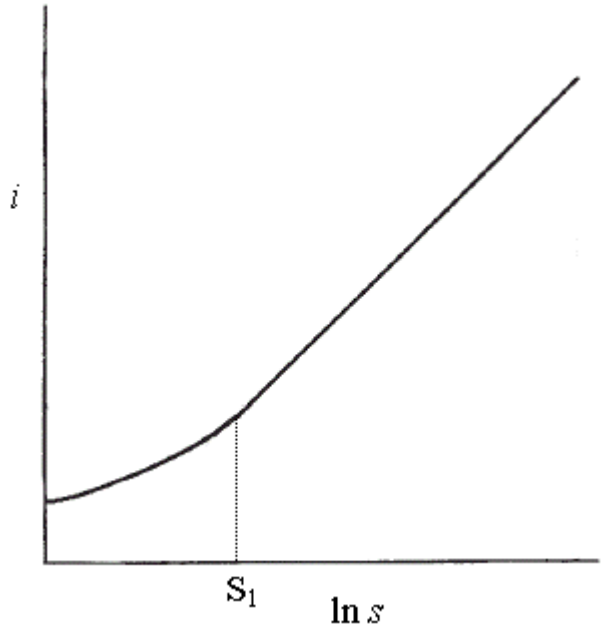

Рис. 2.1.1. Библиограф: график зависимости кумулятивного числа статей в *s* первых журналах от ln *s*

Впервые внимание на математическую нетождественность графической и вербальной формулировок закона Брэдфорда обратил Б. К. Викери [127].

Б. Брукс [33; 34] предложил конкретные формулы для фрагментов библиографа. Выпуклый фрагмент он предложил задавать выражением типа

$$i(s) = \alpha s^{\beta}, \tag{2.1.26}$$

а линейный фрагмент — зависимостью

$$i(s) = k \ln\left(s/a\right), \tag{2.1.27}$$

где α, β, k, a — некоторые константы. Оценку наиболее важной константы k, характеризующей угол наклона прямолинейной части библиографа, Б. Брукс проводит следующим образом. Все библиографии конечны, и вклад последнего журнала в библиографию не менее одной статьи. Поэтому если S — это суммарное число журналов (источников), инкремент в числе статей при переходе от $S-1$ к S источникам должен быть равен единице: $i(S) - i(S-1) = 1$. Подставив сюда зависимость для $i(s)$ из (2.1.27), получим

$$k \ln\left(S/a\right) - k \ln\left((S-1)/a\right) = k \ln\left(\frac{S}{S-1}\right) = -k \ln\left(1 - \frac{1}{S}\right) = 1,$$

откуда, приблизив $\ln\left(1 - \frac{1}{S}\right) \approx -\frac{1}{S}$, найдем простую оценку наклона библиографа

$$k \approx S \tag{2.1.28}$$

и можем переписать (2.1.27) в виде

$$i(s) = S \ln\left(s/a\right). \tag{2.1.29}$$

Во многих исследованиях наблюдался переход в верхней части библиографа (т. е. в «области высоких рангов») прямолинейного отрезка в вы-

пуклую кривую, т. е. отклонение библиографа от прямой вниз. Данный эффект впервые отмечен О. Грусом [82] и получил поэтому название Groos Droop. На рис. 2.1.2 приведен библиограф из [82], его отклонение от прямой начинается в районе 16000 статей.

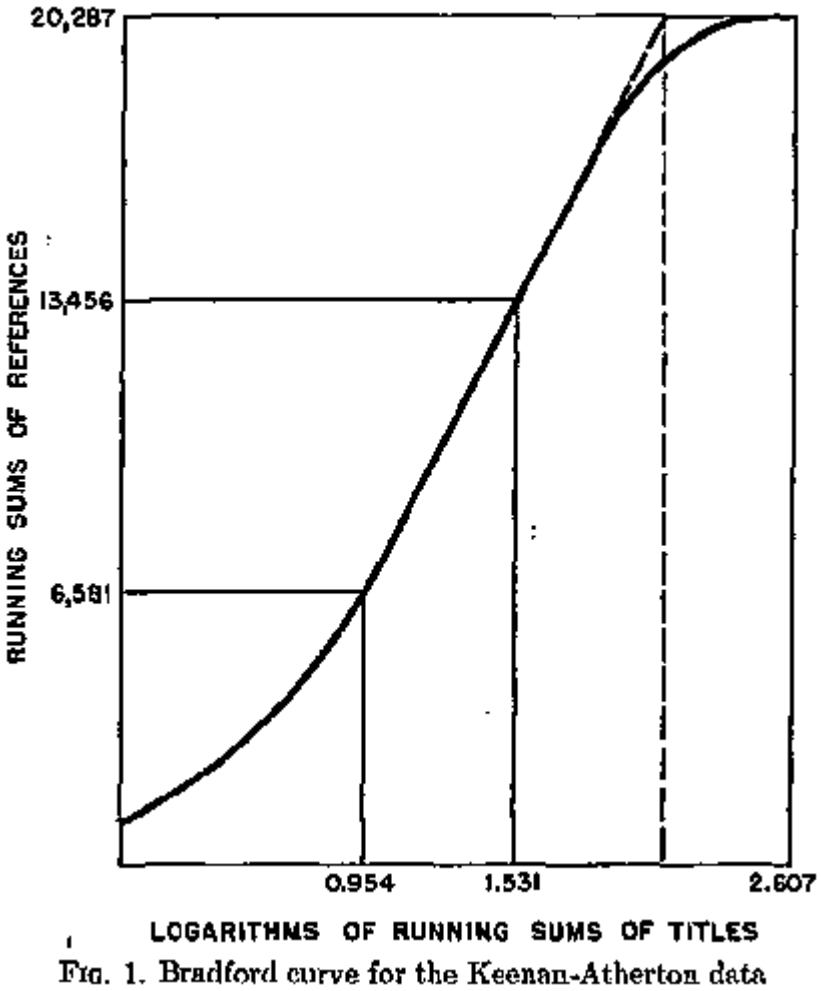

Fig. 1. Bradford curve for the Keenan-Atherton data

Рис. 2.1.2. Иллюстрация Groos Droop из [82].

Данному эффекту предлагалось несколько объяснений, из которых самое частое — влияние неполноты охватываемой библиографии, т. е. наличие журналов, публикующих по нескольку (1–5) релевантных статей и не учтенных по тем или иным причинам при построении библиографа (см., например, [33; 34]). Действительно: если бы такие «дополнительные» жур-

налы существовали, то они внесли бы свой вклад в последние (правые) точки библиографа вида рис. 2.1.2 и, вероятно, «приподняли» бы его конец в сторону прямой линии.

Заметим, однако, что данной попытке объяснения Groos Droop противоречат результаты эксперимента по искусственному моделированию «выпадения» из библиографии низкопродуктивных источников [113]. Здесь авторы при помощи компьютерной модели воспроизвели эффект невключения в библиографию части журналов. Выраженный Groos Droop не был обнаружен ни при «равномерном» (случайном) исключении журналов, ни при «взвешенном» исключении, когда вероятность попадания журнала в неполную библиографию была пропорциональна числу релевантных статей в журнале или даже квадрату этого числа — последние методики пытаются восстановить эффект «незамечания» именно журналов с малым числом статей по исследуемой тематике. Анализируя результаты [113], логично предположить, что неполнота библиографии может быть причиной Groos Droop только в том случае, если вероятность попадания журнала в библиографию зависит от числа релевантных статей в нем еще более сильным образом — например, экспоненциально.

Еще одно из устоявшихся объяснений отклонения от прямой в верхней части кривой Брэдфорда предложено в работе [65], опирающейся на экспериментальные данные из [25]. Если соединить несколько библиографий, относящихся к разным дисциплинам и при этом строго подчиняющихся закону Брэдфорда, то совокупная библиография может не подчиняться закону в области высоких рангов, что приведет к Groos Droop. Исходно мультидисциплинарность в этой модели требует непересечения источников (журналов), именно этот случай получил подробное теоретическое исследование в [65]. Однако авторы резонно замечают, что аналогичный эффект (быть может, в меньшей мере) должен отмечаться и при не-

полной изоляции библиографий, частичном их пересечении. Действительно, именно в области высоких рангов (правой части библиографа) источники скорее всего будут пересекаться мало в случае библиографии по широкой тематической области, которая может быть рассмотрена как композиция узких малопересекающихся библиографий. Ведь если источник (журнал) малопривлекателен для одной подобласти, то он, скорее всего, случаен и маловероятно, что он будет присутствовать и в другой. Эти соображения будут полезны для нашего анализа результатов моделирования спроса на электронные журналы в ГУ-ВШЭ, и мы вернемся к ним вновь в параграфе 2.2.

Так или иначе, наличие отклонения от прямой в области журналов с высокими рангами не считается невыполнением графической формулировки закона Брэдфорда: некоторые исследователи даже приходят к выводу, что такого рода отклонения (в виде Groos Droop или, наоборот, с превращением прямой в вогнутую кривую) внутренне присущи данной модели [111].

2.1.3. Охват законом Брэдфорда различных информационных процессов

В своей основополагающей работе [29] С. Брэдфорд открыл закономерность для распределения по журналам статей, посвященных прикладной геофизике и смазочным материалам. Позднее закон с большим или меньшим успехом применялся к многочисленным данным библиографии по различным тематическим направлениям: по исследованию операций [91; 60], термофизическим свойствам веществ [94], мышечным волокнам и компьютерной науке [34], тучным клеткам и шистосомозу [78; 60], тропическому сельскому хозяйству [93] и др. Более того, было выдвинуто пред-

положение, что закон Брэдфорда приложим не только к рассеянию статей определенной тематики по научным журналам, но и к информационным процессам иной природы. С различной степенью успеха закон проверялся на распределении статей по авторам [56] или странам публикации [71], распределении сделанных цитирований по журналам [48; 121] или полученных цитирований по авторам [47; 120]. В работе [3] было показано хорошее выполнение закона Брэдфорда в его графической форме на материале цитирования обширного массива публикаций журнала «Успехи химии»: библиограф имеет классический вид, с небольшим Groos Droop на конце.

Библиография по закону Брэдфорда очень обширна (см., например, [95; 128]), и мы не ставим здесь цели охватить ее целиком. Остановимся на более узкой тематике — изучении приложимости данной информетрической модели не к производству информации, а к ее потреблению, причем потреблению не через цитирование (что также может быть вполне корректно названо «использованием»), а через чтение.

Одними из первых, кто коснулся данной темы, еще в 1948 г., были Дж. Бернал [21] и Б. Викери [127], использовавшие массив из 1600 статей, затребованных у библиотеки Исследовательских лабораторий Баттервика другими библиотеками в рамках межбиблиотечного обмена. В целом было установлено принципиальное выполнение закона Брэдфорда, хотя с некоторыми неточностями, которые Б. Викери приписал неоднородности исследуемого дисциплинарного поля (у разных ученых, обращавшихся к литературе, были различные научные интересы). Анализом спроса на литературу в свете закона Брэдфорда занимался также П. Ф. Коул [50], анализируя как данные Дж. Бернала из [21], так и собственное исследование [49], охватывающее статистику запросов в Службу технической и библиотечной информации компании «Бритиш Петролеум». Характерной особенностью последнего было то, что обращения к источникам фиксировались не для

самих «конечных пользователей», а для специалистов справочной службы, выполнявшей их запросы. Как для данных Дж. Бернала [21], так и для своего набора [49] П. Ф. Коул нашел хорошее выполнение графической формулировки закона Брэдфорда (равно как еще для одного банка данных [23], основанного не на информации об использовании литературы, а на сборе статистики росписи журналов в «Chemical Abstracts», подобном исходному исследованию Брэдфорда).

Упоминание закона Брэдфорда в применении к процессу использования (чтения) литературы встречается в статье 1964 г. [73]. Однако здесь авторы, хотя и пишут о том, что их данные «подтверждают корректность распространения закона Брэдфорда на использование журналов», но трактуют последний слишком широко, утверждая, что он гласит: «Существует несколько журналов, дающих много статей, больше изданий дают среднее число публикаций, а подавляющее большинство дополнительных журналов содержат малое количество статей». Таким образом, в [73] не проверяется соотношение между числом журналов в каждой зоне Брэдфорда, не строится библиограф и т. д.

Это проводится в явном виде в работе Б. К. Брукса [34] на основании материала из [132] — здесь информетрическая модель Брэдфорда применяется к процессу обращения к реферативным периодическим изданиям по общественным наукам. Б. К. Брукс констатирует, что построенный им библиограф хорошо соответствует модели. Кроме того, выдача книг в библиотеке приводится Б. К. Бруксом в другой его работе [35] в качестве типичного примера пары «источники-продукты» для закона Брэдфорда.

В работе [37] закон Брэдфорда также применяется к процессу использования литературы, а именно к статистике книговыдач в Хиллмановской библиотеке Питтсбургского университета, собранной в течение всего 1972 г. Автор также приходит к выводу, что статистика хорошо соответст-

вует модели Брэдфорда (отмечая, что «вербальная» формулировка подходит в данном конкретном случае лучше, чем графическая), при этом построенный им библиограф действительно имеет классический вид. Этот труд является, по сути, распространением одного из выводов более раннего обширного исследования [36], проведенного в той же библиотеке, в котором констатируется близость статистики книговыдач к логнормальной функции, упоминается распределение Брэдфорда, однако не проводится подробный анализ параметров этой модели и не строится библиограф.

В [103] анализируется статистика межбиблиотечного обмена библиотек поликлиник Массачусетса. Автор, Д. Мортон, исследует вербальную формулировку закона Брэдфорда, разделяя источники/продукты на 20 зон, коэффициент Брэдфорда при этом варьируется от 1,09 до 2,00. Вообще говоря, это свидетельствует о плохом выполнении закона Брэдфорда для такого количества зон. Тем не менее, графическая формулировка прилагается прекрасно, библиограф имеет классический вид и даже без Groos Droop на конце. Последнее обстоятельство автор списывает на то, что из исследования была исключена вся «случайная» периодика, не имеющая непосредственного отношения к медицинской тематике.

Наконец, работа [122] также вплотную подходит к теме настоящего исследования. В ней осуществляется попытка приложить информетрическую модель Брэдфорда к массиву заказов на периодику, поступивших за два года в национальную службу доставки документов Турецкого академического сетевого и информационного центра (TANIC). Авторы пишут, что «частота использования наименований в первых двух зонах (часто и умеренно используемые журналы) удовлетворяет тому, что следует ожидать согласно закону рассеяния Брэдфорда. Но частота обращений к менее используемым журналам оказывается выше, чем предполагалось...». Однако остается непонятным, каким образом можно говорить о выполнении закона

Брэдфорда для двух зон, если для него требуется как минимум три зоны. Соотношение между числом журналов в первой и второй зонах в [122] оказывается равным 1 : 3,0, в то время как между второй и третьей — 1 : 9,6 (разделение на зоны проведено делением на три суммарного количества обращений ко всем журналам). Это скорее свидетельствует о том, что закон Брэдфорда выполняется плохо. Обратное утверждение авторов можно трактовать лишь в том смысле, что наблюдается высокая концентрация востребованности различных журналов, без соблюдения специальной математической закономерности, постулируемой Брэдфордом.

Все перечисленные выше работы показывают, что с тем или иным успехом закон Брэдфорда применим к большому кругу дисциплин и к разнообразным информационным ситуациям. Эта модель имеет по-настоящему широкий спектр приложения в мире научной информации, к процессам различной природы. Отметим, например, любопытный подход М. Бейтс [20], которая связывает зоны Брэдфорда с различными *стратегиями поиска информации*. Информацию, размещенную в первой зоне, следует искать при помощи непосредственного «просмотра» (прежде всего, сплошного изучения оглавления «ядерных» журналов). Информация, размещенная во второй зоне, может быть обнаружена методами «прямого поиска» (по индексированному массиву, например по реферативной базе данных). Рассеянная по большому объему журналов информация из третьей зоны Брэдфорда извлекается поиском по цитируемой и цитирующей литературе (например при помощи научных индексов цитирования).

Несмотря на достаточное количество трудов, применяющих закон Брэдфорда к различным ситуациям и информационным процессам, автору диссертационной работы неизвестны исследования, посвященные приложению данного информетрического распределения к статистике использования онлайновых баз данных научной периодики (библиотек электронных

журналов). Все разработки в области статистики чтения касаются либо выдач печатных изданий, либо работы копировальных служб и служб доставки документов. Это представляется серьезным пробелом в литературе по закону Брэдфорда, т. к., во-первых, современное потребление научной информации всё в большей мере происходит через электронные сетевые источники, во-вторых — как уже говорилось в параграфе 1.2 — онлайновые ресурсы дают возможность собирать статистику обращений к ним на больших временных массивах, будучи избавленными от серьезных технических (и неизбежно следующих за ними методологических) сложностей подсчета обращений к печатным изданиям, особенно на уровне отдельных статей в журнале.

2.2. Закон Брэдфорда как модель процесса обращения к электронным информационным источникам

Теперь обратимся к имеющимся у нас экспериментальным данным по обращению к электронным базам данных документальных источников в библиотеке ГУ-ВШЭ и попробуем применить к ним закон Брэдфорда. Возьмем суммарные данные использования четырех ресурсов (EBSCO, JSTOR, ProQuest, ScienceDirect), т. е. массив, получение которого подробно описано в параграфе 1.4.

Упорядочим журналы в порядке убывания количества открытых из них статей. Таким образом, первым журналом будет American Economic Review с 1306 загруженными статьями, вторым — American Journal of Sociology с 1166 и т. д. В конце упорядоченного списка будет находиться длинный «хвост» из 778 изданий, из которых было открыто лишь по одной статье. Данные приведены в таблице из Приложения 1, устроенной анало-

гично таблицам, фигурировавшим в классической работе [29] и других (отметим, что в нашей таблице 155 строк, в то время как у Брэдфорда в его двух таблицах соответственно 24 и 14, что еще раз подтверждает ценность и представительность нашей выборки для моделирования информационного процесса).

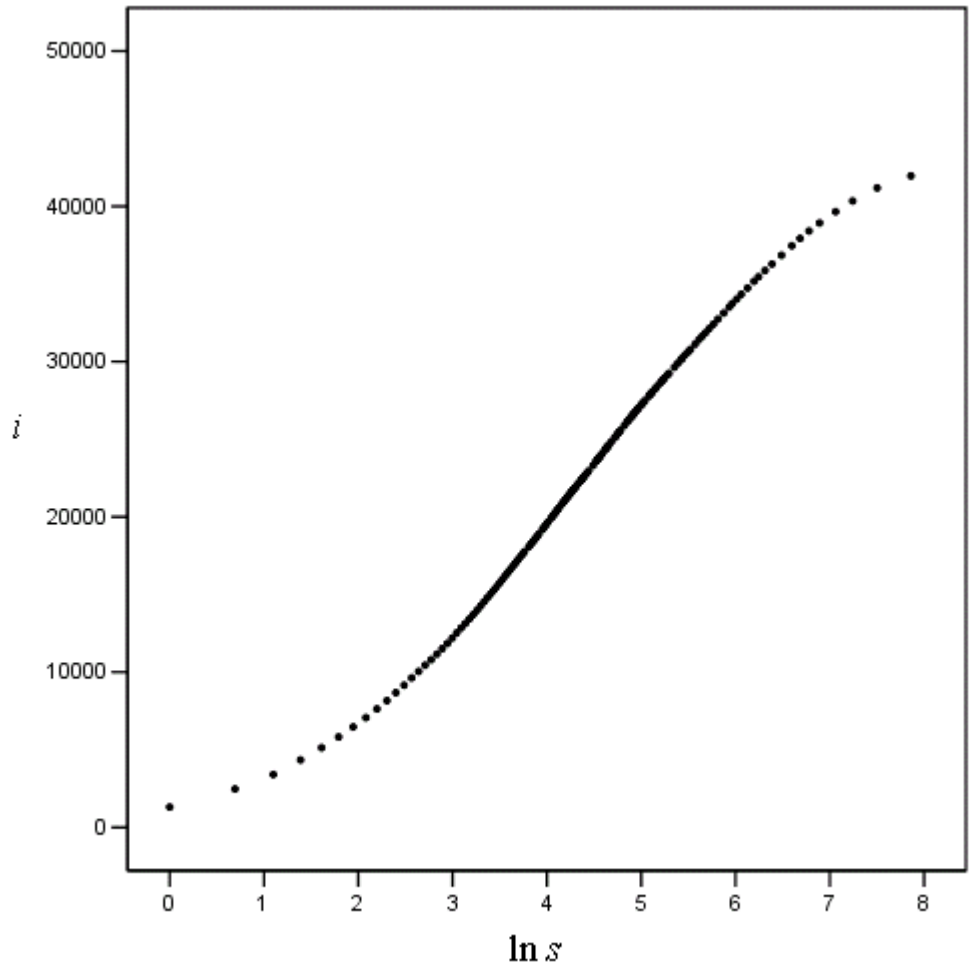

Рис. 2.2.1 Зависимость кумулятивного числа открытых статей в *s* журналах от натурального логарифма *s* (издания упорядочены от наиболее спрашиваемых к наименее спрашиваемым).

Построим кривую Брэдфорда («библиограф») для этого распределения. По оси абсцисс отложим натуральный логарифм количества *s* наиболее продуктивных источников (т. е. данные из столбца Д Приложения 1). По ординате — кумулятивное число статей, пришедшихся на эти *s* журна-

лов (т. е. данные из столбца Г Приложения 1). Итоговый график, построенный по отдельным точкам, показан на рис. 2.2.1, а на рис. 2.2.2 изображен сглаженный вариант данного графика.

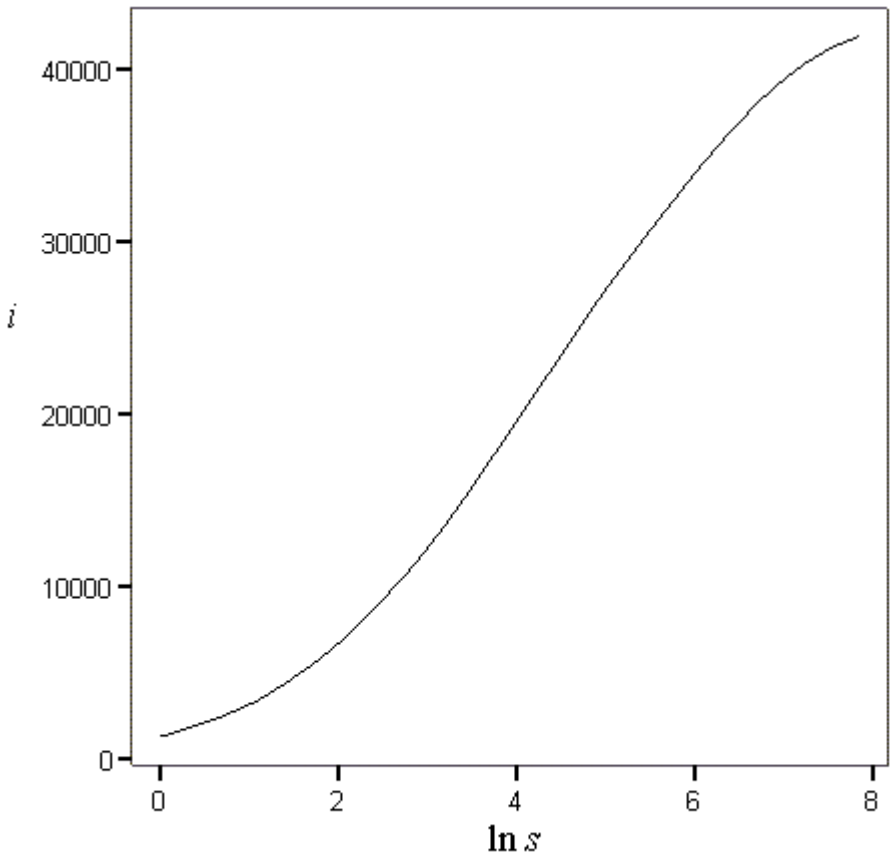

Рис. 2.2.2. Зависимость кумулятивного числа открытых статей в *s* журналах от натурального логарифма *s* (сглаженный график)

Мы видим, что полученный «библиограф», описывающий процесс обращения к четырем электронным ресурсам, имеет вполне классический вид, совпадающий, например, с графиком, приведенным в [93]. Выделяется начальный фрагмент кривой, имеющий положительную вторую производную, далее более или менее прямолинейный участок и в конце — отклонение от прямой в виде выпуклого участка библиографа. Последняя часть — это «Groos Droop», подробно обсуждавшийся в разделе 2.1.2. Из характер-

ных особенностей графика, полученного по экспериментальным данным, можно, пожалуй, отметить недостаточно ярко выраженную прямолинейную часть, служащую как будто переходом от начального вогнутого к финальному выпуклому участку.

На время отвлечемся от графического представления эмпирических данных и обратимся к проверке «вербальной» формулировки закона Брэдфорда. Для этого нам необходимо выделить зоны Брэдфорда и посмотреть на соотношения между числами журналов, попадающих в разные зоны. Прежде всего, воспользуемся элементарным методом, который напрашивается исходя из классической формулировки закона Брэдфорда: разобьем всё множество публикаций на три равные части и посмотрим, на какое число журналов они приходятся (в частности, так поступают авторы в [121; 122]).

зона Брэдфорда	число журналов в зоне	процент журналов	*q*	число статей, открытых из каждого журнала	число статей в зоне
I	26	1,0%	–	1306–275	13987
II	139	5,4%	5,35	272–43	13975
III	2425	93,6%	17,45	42–1	13997
всего	2590	100%	–	–	41959

Таб. 2.2.1. Разбиение на три зоны Брэдфорда методом простого деления множества статей на (примерно) равные части (*q* — коэффицент Брэдфорда, отношение числа журналов в *n*-й зоне к числу в *n*–1-й)

Мы имеем 41959 «продуктов» (статей), т. е. по 13986–13987 в каждой зоне. Следовательно (опираемся на данные Приложения 1), первую зону Брэдфорда формируют 26 журналов, из которых открыто от 1306 до 275 публикаций, всего 13987 статей. Во вторую зону входят 139 изданий, к которым обращались от 272 до 43 раз и в сумме открыли 13975 документов. На оставшиеся 2425 журналов пришлось от 42 до 1 запроса на каждый, в

сумме из них было открыто 13997 статей, и они образуют третью зону Брэдфорда. Результаты отражены в таб. 2.2.1.

Из данной таблицы видно, что концентрация спроса на журналы очень велика: первую зону Брэдфорда, на которую приходится треть всех открытых статей, составляет лишь 1,0% изданий (26 журналов). Для сравнения, в классической работе самого С. Брэдфорда [29] процент журналов, составляющих первую зону, «ядро», равен 2,8% для литературы по прикладной геофизике и 4,9% — по смазке. Тем не менее, в литературе по закону Брэдфорда известна и бóльшая концентрация: например, в исследовании [120], где закон Брэдфорда применяется к распределению цитирований, в первую зону вошли лишь 0,3% всех цитировавшихся изданий.

Также из анализа таб. 2.2.1 можно заключить, что с использованными для ее построения параметрами информетрическая модель Брэдфорда плохо соответствует рассматриваемым эмпирическим данным: числа журналов в трех зонах не образуют геометрической прогрессии, и коэффициент q, который при идеальном выполнении закона должен оставаться постоянным, варьируется более чем в три раза, от 5,4 до 17,5.

Теперь применим более тонкую методику подгонки параметров модели Брэдфорда, предложенную Л. Эгге в [60] и подробно изложенную в параграфе 2.1.1:

— выберем число групп Брэдфорда n равным трем. В методике Эгге предлагается выбирать это число от 4 до 10, но для того чтобы сравнить результаты, полученные «элементарным» способом, с новой методикой, начнем с такого значения n. Опишем процедуру подробно для $n = 3$, в дальнейшем для других n детальные выкладки, аналогичные данным, будут опущены;

— по формуле (2.1.24) вычислим коэффициент Брэдфорда q:

$$q = \left(e^{\gamma}\, i_{\max}\right)^{1/n} \approx \left(1,781 \cdot 1306\right)^{1/3} \approx 13,25;$$

— по (2.1.4) получим размер первой зоны Брэдфорда S_1:

$$S_1 = S\frac{q-1}{q^n - 1} = 2590 \cdot \frac{13,25-1}{13,25^3 - 1} \approx 13,65; \qquad (2.2.1)$$

— т. к. S_1 получается не целым, используем округление до ближайшего целого $\{S_1\} = 14$ (здесь мы используем нашу корректировку метода Эгге, предлагавшего округлять S_1 строго в сторону меньшего целого); на первые 14 наиболее продуктивных журналов приходится, согласно Приложению 1, 10049 статей;

— в двух остальных зонах число источников (журналов) вычисляется как результат округления до ближайшего целого произведений qS_1 и q^2S_1, что дает 181 и 2396 журналов соответственно; отметим, что вычисленные таким образом зоны Брэдфорда дают в сумме 2591 журнал, что превышает общее число журналов S на единицу. Это произошло в результате накопления ошибок округлений, поэтому размер последней группы скорректируем на единицу, приравняв его 2395. По Приложению 1 устанавливаем, что на эти зоны приходится соответственно 19063 и 12847 статей.

В результате получим таб. 2.2.2.

зона Брэдфорда	число журналов	процент журналов	q	число статей
I	14	0,5%	–	10049
II	181	7,0%	12,9	19063
III	2395	92,5%	13,2	12847
всего	2590	100%	–	41959

Таб. 2.2.2. Три зоны Брэдфорда, определенные скорректированным методом Л. Эгге

Сравнивая таб. 2.2.1 и таб. 2.2.2, можно сопоставить результаты выделения трех зон Брэдфорда двумя различными методами — методом самого

Брэдфорда и методом Эгге. Прежде всего, заметим, что в таб. 2.2.1 практически совпадает число статей в каждой зоне, но при этом сильно варьируется коэффициент q. В таб. 2.2.2 наоборот: отношение числа источников в группах друг к другу (параметр q) практически неизменно, однако число статей по зонам серьезно (почти в два раза) отличается друг от друга. Это отличие двух таблиц не является чем-то неожиданным и всего лишь отражает внутреннюю логику, по которой строилось разбиение на зоны двумя различными способами.

Что важно — это то, что, согласно полученным результатам (таб. 2.2.1 и таб. 2.2.2), «вербальная» формулировка закона Брэдфорда для имеющихся экспериментальных данных выполняется малоудовлетворительно, вне зависимости от того, какой способ разбиения применяется. Несмотря на сильную концентрацию и неравномерность распределения «продуктов» по «источникам», которые предсказывает закон Брэдфорда, математическая форма этой концентрации отличается от описываемой вербальным вариантом закона.

Заметим, тем не менее, что введенная нами корректировка метода округления числа источников первой зоны Брэдфорда (округление вычисленного значения $S_1 = S\frac{q-1}{q^n-1}$ в сторону ближайшего целого, а не строго в сторону меньшего целого) привела к улучшению разбиения на группы Брэдфорда. Для демонстрации этого в таб. 2.2.3 приведены результаты разбиения исходным алгоритмом, описанным Л. Эгге. В первой группе здесь 13 журналов вместо 14 (результат 13,65, полученный в (2.2.1), округляется в меньшую сторону), а в последней — 2396 вместо 2395. Сравнив таб. 2.2.2 и таб. 2.2.3, мы видим, что предложенная нами коррекция метода округления величины первой группы сократила разницу между значениями q для двух пар зон (I–II и II–III) более чем в два раза, а отношение максимально-

го и минимального чисел статей в полученных группах Брэдфорда — уменьшила с 2,0 до 1,9 (напомним, что при идеальном выполнении вербальной формулировки закона Брэдфорда q должно быть постоянным, а число статей в группах Брэдфорда — одинаковым). В связи с этим в дальнейшем при использовании метода Эгге мы будем пользоваться нашими правилами округления как доказавшими свою предпочтительность.

зона Брэдфорда	число журналов	процент журналов	q	число статей
I	13	0,5%	–	9629
II	181	7,0%	13,9	19448
III	2396	92,5%	13,2	12882
всего	2590	100%	–	41959

Таб. 2.2.3. Три зоны Брэдфорда, определенные исходным методом Л. Эгге [60] (округление величины первой зоны в меньшую сторону)

Однако Л. Эгге не предлагает ограничиваться тремя зонами Брэдфорда и даже наоборот — рекомендует разбивать множество источников и продуктов на 4–10 зон. Более того, «для больших библиографий может быть целесообразным выбор больше десяти» [60, p. 471]. При этом определение конкретного числа зон n регулируется получающейся в результате расчетов по (2.1.25) величиной первой зоны Брэдфорда S_1: она должна быть по возможности близкой к целому числу.

Заметим, что существует естественное ограничение сверху числа зон Брэдфорда при использовании метода Эгге. Если подставить в формулу (2.1.25) равенство (2.1.24), то легко видеть, что размер первой зоны Брэдфорда S_1 убывает с ростом n:

$$S_1 = S\frac{q-1}{q^n-1} = S\frac{\left(e^{\gamma} i_{\max}\right)^{1/n}-1}{e^{\gamma} i_{\max}-1}.$$

Очевидно, что вычисленная первая зона должна содержать по меньшей мере один источник, отсюда вытекает ограничение на n:

$$S_1 \geq 1 \quad \Rightarrow \quad S\frac{\left(e^{\gamma} i_{\max}\right)^{1/n} - 1}{e^{\gamma} i_{\max} - 1} \geq 1 \quad \Rightarrow \quad n \leq \frac{\ln\left(e^{\gamma} i_{\max}\right)}{\ln\left(1 + \frac{e^{\gamma} i_{\max} - 1}{S}\right)}. \quad (2.2.2)$$

Подставив в (2.2.2) значения параметров нашего набора данных, получим ограничение $n \leq 12{,}1$. Таким образом, разбиение более чем на 12 групп Брэдфорда бессмысленно. В таб. 2.2.4 сведены результаты разбиения на n зон, $n = 4 \ldots 12$. Ввиду того что округление размера первой зоны в сторону ближайшего целого, как показано выше, приводит к лучшим результатам, разбиение на зоны проводится с учетом этой поправки, введенной нами в методику Эгге. В связи с этим размер последней зоны также иногда подвергается коррекции на 1 источник. В случае $n = 5,\ 11,\ 12$ размер последней зоны на 1 журнал больше, чем вычисленный и округленный, в случае $n = 8,\ 9,\ 10$ — на 1 журнал меньше. Для остальных n корректировки не потребовалось.

зона Брэдфорда	число журналов	q	число статей
n = 4; q = 6,94; S_1 = 6,62			
I	7	–	6469
II	46	7,67	12865
III	319	6,93	14088
IV	2218	6,95	8537
n = 5; q = 4,71; S_1 = 4,14			
I	4	–	4336
II	19	4,75	8806
III	92	4,84	12165
IV	433	4,71	10516
V	2042	4,72	6136
n = 6; q = 3,64; S_1 = 2,94			
I	3	–	3411

зона Брэдфорда	число журналов	q	число статей
II	11	3,67	6638
III	39	3,55	9285
IV	142	3,64	9778
V	516	3,63	8174
VI	1879	3,64	4673
n = 7; q = 3,03; S_1 = 2,26			
I	2	–	2472
II	7	3,50	5165
III	21	3,00	7385
IV	63	3,00	8644
V	189	3,00	7937
VI	573	3,03	6658
VII	1735	3,03	3698
n = 8; q = 2,64; S_1 = 1,82			
I	2	–	2472
II	5	2,50	3997
III	13	2,60	5726
IV	33	2,54	7139
V	88	2,67	7510
VI	232	2,64	6596
VII	610	2,63	5487
VIII	1607	2,63	3032
n = 9; q = 2,37; S_1 = 1,52			
I	2	–	2472
II	4	2,00	3353
III	9	2,25	4619
IV	20	2,22	5737
V	48	2,40	6602
VI	113	2,35	6364
VII	267	2,36	5661
VIII	632	2,37	4567
IX	1495	2,37	2584
n = 10; q = 2,17; S_1 = 1,30			
I	1	–	1306
II	3	3,00	3030
III	6	2,00	3828
IV	13	2,17	4978
V	29	2,23	6042

зона Брэдфорда	число журналов	q	число статей
VI	63	2,17	6123
VII	137	2,17	5543
VIII	297	2,17	4984
IX	644	2,17	3906
X	1397	2,17	2219
n = 11; q = 2,02; S_1 = 1,14			
I	1	–	1306
II	2	2,00	2105
III	5	2,50	3659
IV	9	2,25	4094
V	19	2,11	5230
VI	39	2,05	5637
VII	78	2,00	5397
VIII	158	2,03	4819
IX	320	2,03	4373
X	648	2,03	3378
XI	1311	2,02	1961
n = 12; q = 1,91; S_1 = 1,01			
I	1	–	1306
II	2	2,00	2105
III	4	2,00	3058
IV	7	1,75	3580
V	13	1,86	4210
VI	26	2,00	5075
VII	49	1,88	5055
VIII	93	1,90	4723
IX	176	1,89	4292
X	339	1,93	3874
XI	646	1,91	2951
XII	1234	1,91	1730

Таб. 2.2.4. Разбиение на зоны Брэдфорда скорректированным методом Эгге: варьирование числа зон (n = 4 … 12)

Анализируя таб. 2.2.4, прежде всего обратим внимание на динамику коэффициента q. При увеличении числа n рассчитанный (теоретический) коэффициент Брэдфорда q, как и положено, уменьшается с $q \approx 7$ для $n = 4$

до $q \approx 2$ для $n = 12$. В рамках одного значения n реальный (наблюдаемый) коэффициент q колеблется около своего теоретического значения для первых зон и практически совпадает с последним для последних зон Брэдфорда. Колебания вызваны погрешностями округления расчетных размеров зон до целого числа журналов (в применяемой методике в отсутствие этих погрешностей отношение размеров соседних зон было бы константой). Соответственно, рост величины зоны Брэдфорда при переходе от первых зон к последним обеспечивает более точное совпадение расчетного и реального значений коэффициента, так как округление до целого вносит все меньший относительный вклад.

Анализ последнего столбца таб. 2.2.4 (число статей в зонах Брэдфорда) должен показать, насколько хорошо выполняется для нашего массива данных вербальная формулировка закона Брэдфорда: в случае идеального его выполнения количество статей в зонах было бы постоянным. Как можно видеть, это далеко не так: например, отношение максимального числа статей в зоне к минимальному варьируется от 1,90 ($n = 3$, см. таб. 2.2.2) до 4,69 ($n = 10$). Интересна динамика количества статей при переходе от первых зон к последним. Сначала число продуктов в каждой последующей зоне увеличивается, затем достигает своего максимума, вокруг которого (при больших n) имеются 3–4 зоны с близким друг другу числом статей, после чего в последних зонах снова количество статей заметно уменьшается. Область, где размер зон (выраженный в статьях) близок друг к другу — это область приемлемого выполнения вербальной формулировки закона Брэдфорда. Нетрудно заметить, что она совпадает с линейным участком библиографа на рис. 2.2.2.

Например, при $n = 12$ три зоны VI–VIII содержат практически одинаковое число продуктов: отклонение составляет менее 7%. В этих зонах содержатся журналы с номерами от 28 до 195 (естественно, эти номера опре-

деляются по позиции в рейтинге наиболее продуктивных источников). На графике рис. 2.2.2 этому соответствуют точки $\ln 28$ и $\ln 195$, т. е. отрезок оси абсцисс от 3,3 до 5,3, который — если обратиться к виду библиографа — расположен в той части, где график представляет собой практически прямую линию.

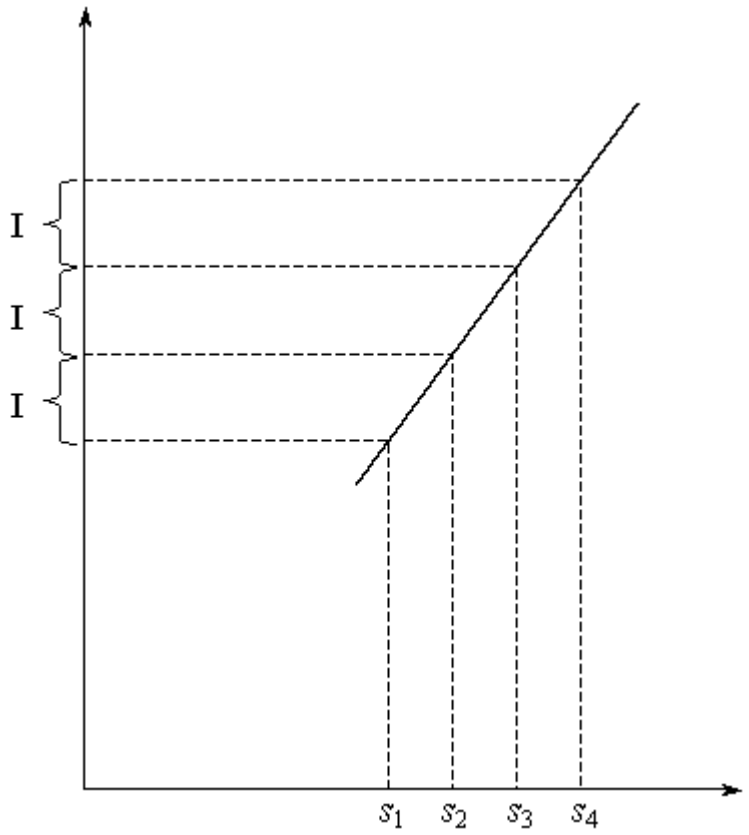

Рис. 2.2.3. Прямолинейный участок библиографа в полулогарифмических координатах (шкала оси абсцисс — логарифмическая)

Этот факт естественным образом отражает взаимосвязь вербальной и графической формулировок закона Брэдфорда. Действительно, предположим, что в полулогарифмических координатах участок библиографа выглядит как прямая линия (рис. 2.2.3). Ординаты границ соседних зон Брэдфорда на библиографе должны отсекать равные отрезки, т. к. продуктивность каждой зоны постоянна и равна I ($I = I_B$ в обозначениях предыдуще-

го раздела). Ввиду линейности данного участка библиографа абсциссы соответствующих границ зон Брэдфорда также отсекают равные отрезки. Если границы зон (выраженных в источниках) обозначить, как на рис. 2.2.3, s_1, s_2, s_3, s_4, то отсюда следует — т. к. шкала абсцисс логарифмическая, — что

$$\ln s_2 - \ln s_1 = \ln s_3 - \ln s_2 = \ln s_4 - \ln s_3,$$

откуда

$$\frac{s_2}{s_1} = \frac{s_3}{s_2} = \frac{s_4}{s_3}. \qquad (2.2.3)$$

Сравним теперь величины отношений между размерами зон Брэдфорда, т. е. пропорции $\frac{s_3 - s_2}{s_2 - s_1}$ и $\frac{s_4 - s_3}{s_3 - s_2}$. Использовав подстановку для s_3 и s_4 из (2.2.3), легко получить равенство этих двух отношений. Следовательно, отношение величин соседних зон Брэдфорда равно константе.

Таким образом, мы получили важный вывод, связывающий графическую и вербальную модели Брэдфорда: если имеется линейный участок библиографа и три соседние зоны Брэдфорда целиком лежат на нем, то «вербальная» формулировка выполняется идеально, т. е. $\frac{S_{k+1}}{S_k} = \frac{S_{k+2}}{S_{k+1}}$, где S_k, S_{k+1}, S_{k+2} — размеры соответствующих зон.

Заметим также, что обратное утверждение, вообще говоря, неверно. В частности, закон Брэдфорда может выполняться идеально в его классическом случае, для трех зон, несмотря на то что стартовый участок библиографа при этом вогнутый, а не прямой. Тем не менее в нашем случае, на исследуемых в настоящей работе данных, получается — как следует из приведенного выше анализа и таб. 2.2.4, — что соответствие вербальной информетрической модели Брэдфорда наблюдается только на прямолиней-

ном участке библиографа. Только здесь одновременно равны отношения зон между собой и постоянно число статей в каждой зоне.

Как отмечалось выше, вид построенного нами на рис. 2.2.2 библиографа привлекает внимание своим достаточно обширным отклонением от прямой в области высоких рангов: Groos Droop ярко выражен. Если продолжить говорить о связи графической и вербальной формулировок модели Брэдфорда в нашем случае, то логично заключить, что именно это свойство графика отвечает за нестрогое — только на определенном интервале и для достаточно «частого» деления на зоны — выполнение математического закона Брэдфорда. Можно предположить, что причины столь ярко выраженного Groos Droop совпадают с описанными в литературе: согласно [34], таковой могут стать недостаточно полный учет, потеря библиографии в области малопродуктивных по исследуемой теме изданий. Согласно [65], мультидисциплинарность массива исследуемых документов также способна привести к Groos Droop. В нашем случае могут иметь место оба эффекта, если перенести их формулировки от процесса появления журнальных публикаций к их использованию: исследуемое множество онлайновых источников, конечно, не исчерпывает все информационные потребности читателей вуза, и многие периферийные издания, публикующие мало статей, релевантных тематике ГУ-ВШЭ, остаются за рамками исследуемых баз данных и настоящего анализа. Это эмулирует «неполноту» исследуемой библиографии, если возвращаться к работе [34]. С другой стороны, учтенные онлайновые библиотеки существенно политематичны, равно как и интересы, спрос на информацию у читателей вуза широкого профиля, специализирующегося на экономике, менеджменте, социологии, праве, политологии и др. Здесь возможен эффект, аналогичный соединению малопересекающихся узкоспециализированных библиографий, описанный в [65].

Именно сочетанием этих причин может объясняться особенно сильная деформация библиографа в области малоспрашиваемых изданий.

Завершая анализ таб. 2.2.4, отметим, что как в первых, так и в последних зонах Брэдфорда число статей меньше, чем в средних зонах. На то, что такой эффект наблюдается в эмпирике, обращают внимание, например, авторы [57]: «При работе с реальными данными были замечены некоторые отклонения от математической формы <используется формула Брукса $R(n) = k \ln n$. — *В. П.*>. А именно, журналы с низким рангом (те, к которым относится максимальное число статей) обыкновенно дают меньше публикаций, чем предсказывает формула. С другой стороны, отмечалось, что журналы с высоким рангом (представляющие мало статей) также дают меньше статей, чем предполагается».

2.3. «Ядро» изданий и закон Брэдфорда

2.3.1. Методы выделения ядра изданий при помощи закона Брэдфорда

Во многих работах, посвященных закону Брэдфорда, начиная с исходной [29], эта информетрическая модель также используется для определения «ядра» («nucleus» у Брэдфорда, далее чаще «core») литературы, совпадающего с первой зоной Брэдфорда и концентрирующего в себе журналы, на которые приходится основная часть документов. В случае распределения опубликованных статей по журналам таким образом определяется состав ядра периодических изданий по соответствующей дисциплине. В случае изучения чтения книг или заказов на доставку документов — выявляются «ядерные» издания, представляющие ключевой интерес для читательской аудитории.

Понятие «ядра» литературы существует не только для журналов. Так, издание «Booklist» регулярно публикует рубрику, в которой помещает подборки наиболее важной книжной литературы определенной тематики (например, по глобальному изменению климата [117], истории и современности Китая [88] и др.), называя его «ядро фонда» (core collection). В своей фундаментальной статье, посвященной спискам «ядерных журналов» [107], Т. Нисонджер приводит различные работы, использующие сочетания «core…», в частности для списка избранных книг современной поэзии [87] или для комиксов, необходимых юношеским отделам публичных библиотек [81]. Также упоминаются подборки «ядерных списков» информационных объектов, входящих в новую развивающуюся концепцию «library 2.0» — блогов и википедий. Тем не менее отмечается, что, если положиться на поиск по базе данных «Library Literature and Information Science Full Text», в последнее время чаще всего ядерные списки составляются именно для журналов.

Автор [107] указывает, что методы, которыми могут определяться ядерные периодические издания, разнообразны: опрос мнения экспертов, исследования статистики использования, учет индексирования реферативными службами или числа библиотек-подписчиков, анализ данных по цитируемости (здесь также упоминается несколько различных методик), анализ сетей цитирования и ко-цитирования, фокусировка на продуктивности журналов, метод, использующий закон Брэдфорда, анализ данных публикационной активности базовой организации.

Среди этих методов Т. Нисонджер выделяет те, которые, на его взгляд, лучше всего подходят для библиотек учебных и научно-исследовательских заведений. Это анализ цитирований, статистика использования, продуктивность журналов и закон Брэдфорда. Отметим, что мы в своей работе опираемся сразу на две из этих предпочтительных четырех

методик: собираем и обрабатываем статистику использования, после чего к ней применяем Брэдфорд-анализ.

Итак, ядро периодических изданий определяется при помощи информетрической модели Брэдфорда следующими двумя способами, с использованием соответственно вербальной и графической формулировок. При использовании вербальной формулировки закона «ядро» считается совпадающим с первой зоной Брэдфорда, определенной математически. Как правило здесь имеется в виду «классическое» разбиение на три зоны Брэдфорда. Иными словами, ядро — это те журналы, которые дают одну треть (33%) всех статей, релевантных исследуемой тематике.

При более распространенном для данной задачи графическом подходе ядро ограничивается точкой перехода вогнутого участка библиографа в прямую. Это соответствует точке S_1 на рис. 2.1.1. Все журналы, лежащие левее данной границы, считаются включенными в ядро.

Нельзя не упомянуть, что работа [106] скептически высказывается о потенциале Брэдфорд-анализа для определения списка ядерных изданий, пытаясь продемонстрировать расплывчатость понятия «тематики» при таком анализе. Авторы находят распределение статей по журналам, полученных при поиске (а) по тематическому слову (например «virology») в названии или аннотации статьи; (б) по тому же тематическому слову в ключевых словах; (в) публикаций, цитируемых в ведущем журнале соответствующей тематики (например, «Journal of Virology»). Полученные при последующем Брэдфорд-анализе списки ядерных журналов существенно отличаются друг от друга — в зависимости от того, каким образом проводился поиск по данной теме. Это критическое замечание, однако, относится к классическому построению тематического «ядра» журналов на основе опубликования в них работ той или иной тематики и едва ли может быть применено к нашему случаю, когда речь идет о рассеянии *востребованных*

статей по многопрофильному набору журналов, входящих в мультидисциплинарную подписку вуза. В настоящем исследовании вопрос о четкости определения «тематики» просто не стоит.

Наконец, стоить отметить, что в перспективе, ввиду нарастающей активности движения открытого доступа и создания «университетских репозитариев» (institutional repositories: см., например, [98]), миссия и функция научного журнала как такового могут понизиться, уступив место тому, что Д. Гудман называет «статейной базой данных», article database [80], «полностью независимой от журналов». При этом, как справедливо отмечает Т. Нисонджер, понятие «ядра» для журналов обессмыслится [107]. Тем не менее сама концепция поиска и определения ядра глубже и базируется не на строгой привязке к той или иной форме представления материалов, а на выстраивании иерархии объектов в соответствии с их информационной значимостью. Такого рода подходы фундаментальны и не зависят от эволюции форматов представления информации.

2.3.2. Ядро изданий при моделировании процесса обращения к электронным информационным источникам с применением закона Брэдфорда

Государственный Университет – Высшая школа экономики обладает одной из самых обширных онлайновых подписок среди научных и образовательных организаций в России, спрос на англоязычную электронную периодику среди его студентов и преподавателей растет из года в год и уже ощутимо превышает востребованность зарубежных журналов, поступающих в печатном виде. Поэтому представляют интерес определение и анализ ядра читательского спроса на научную информацию именно по электронным источникам. Это, помимо решения теоретической задачи инфор-

метрического моделирования, также будет иметь несомненные практические следствия при решении задач формирования фондов электронной периодики.

Как говорилось в предыдущем параграфе, определение ядра с использованием закона Брэдфорда возможно двумя путями: используя вербальную и графическую формулировки закона Брэдфорда. В связи с тем что, как мы установили в параграфе 2.2, для исследуемого статистического массива графическая формулировка в целом выполняется (библиограф имеет классический вид), а вербальная выполняется достаточно плохо, мы сфокусируемся на методе выделения ядра наиболее спрашиваемых изданий при помощи графического варианта закона Брэдфорда (вербальная формулировка, впрочем, еще будет привлечена на одной из стадий работы с библиографом).

Граница ядра, определяемая с помощью графической формулировки модели Брэдфорда, совпадает с переходом построенного нами библиографа (рис. 2.2.2) от вогнутой части к прямолинейной. Проблема, однако, в том, что на настоящий момент нет общепризнанного метода определения точки этого перехода. Так, про приближение библиографа прямой авторы [57] пишут: «В большинстве литературы данные просто отображаются на графике и приближение осуществляется на глаз. Мы не обнаружили такого подхода к приближению подобных данных, который был бы привлекателен с теоретической точки зрения». Сами авторы [57], осуществляя приближение прямой, удаляют из библиографа 4 самых продуктивных журнала (нижняя часть графика) и все журналы с одной статьей (верхняя часть), после чего остальные точки приближают регрессионной прямой. Для нашего случая такой метод, очевидно, бесполезен, т. к. он фактически заранее предполагает, что ядро состоит из четырех журналов.

Для того чтобы не пользоваться подгонкой ядра «на глаз», предложим два различных метода определения искомой границы. Прежде всего, судя по форме графика рис. 2.2.2, логично предположить, что его можно достаточно точно приблизить кубическим многочленом. Такое приближение методом наименьших квадратов показано на рис. 2.3.1.

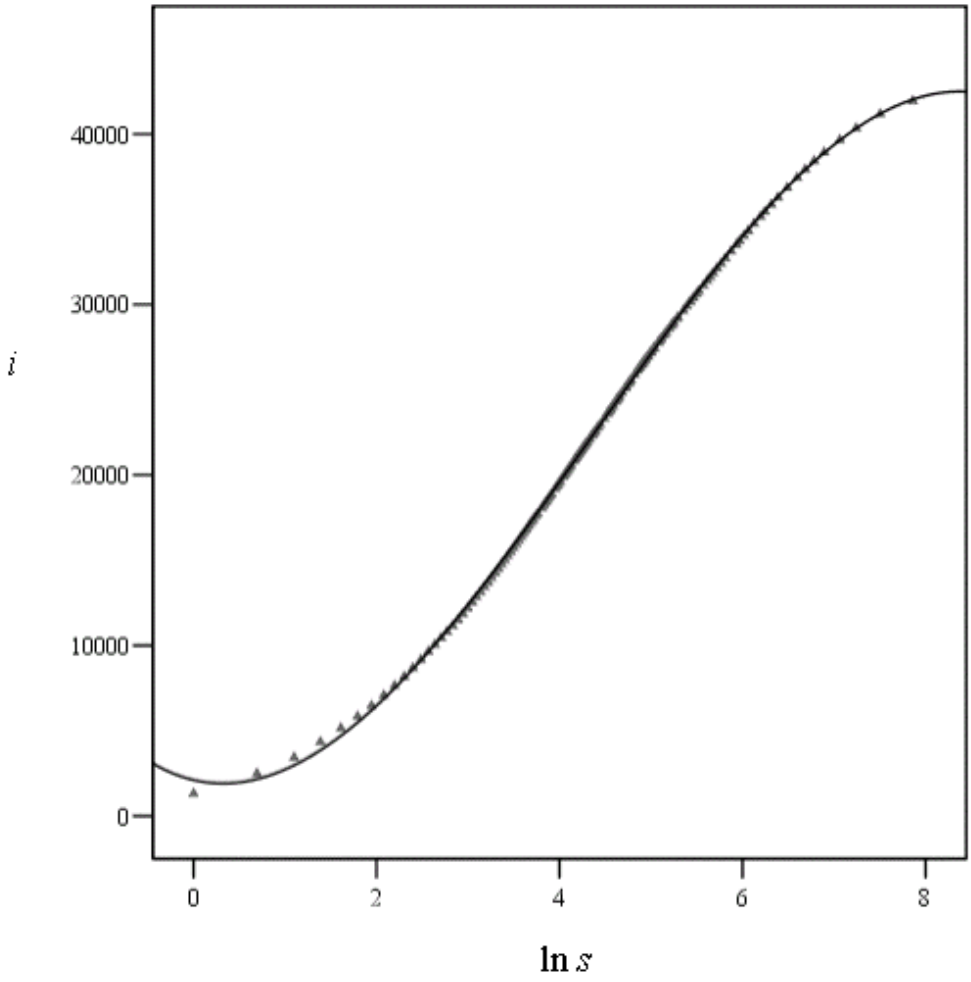

Рис. 2.3.1. Приближение библиографа полиномом третьей степени: точки — исходные данные, кривая — кубический многочлен

В результате аппроксимации значения коэффициентов полинома получились равными:

$$y = a_3 x^3 + a_2 x^2 + a_1 x + a_0;$$
$$a_0 = 2106,7;\ a_1 = -1255,4;\ a_2 = 2030,1;\ a_3 = -155,7. \qquad (2.3.1)$$

Мы видим на рис. 2.3.1, что приближение практически идеальное, это подтверждает коэффициент детерминации (R квадрат): он равен единице с точностью до третьего знака после запятой. Теперь найдем точку перегиба аппроксимирующей функции — вблизи нее график имеет квазипрямолинейный вид. Использовав полученные коэффициенты аппроксимации в (2.3.1), найдем точку перегиба:

$$y'' = 0 \;\Rightarrow\; 6a_3x + 2a_2 = 0 \;\Rightarrow\; x = -\frac{a_2}{3a_3} \approx 4{,}35. \qquad (2.3.2)$$

Так как ось абсцисс на рис. 2.3.1 логарифмическая, то найденной абсциссе точки перегиба соответствует размер ядра:

$$S_1 = e^{4{,}35} \approx 77. \qquad (2.3.3)$$

Значит, ядро, полученное при помощи использованного метода, содержит 77 журналов. Естественно, как следует из самого подхода, таким способом мы нашли «оценку сверху»: дальше этой точки аппроксимирующий полином из вогнутого становится выпуклым, значит, она *уже* находится на прямолинейном участке (скорее всего, где-то в его середине). Итак, ядро спроса литературы, полученное методом Брэдфорда, точно не превышает 77 наименований журналов.

Уточнить искомую границу поможет второй предлагаемый нами метод.

Обратимся к таб. 2.2.4 и вербальной формулировке информетрической модели Брэдфорда. Как мы видели в параграфе 2.2, при определении параметров модели методом Эгге и числе зон, равном $n = 12$, существуют зоны с шестой по восьмую, где закон Брэдфорда практически точно выполняется — количество статей в этих зонах близко к константе. Если предположить, что это сопровождается линейностью библиографа на данном участке (что не является необходимым условием, но в данном случае такое предположение вполне закономерно), то можно сказать, что граница ядра прибли-

женно совпадает с началом шестой зоны при разбиении на 12 зон. Из таб. 2.2.4 следует, что это точка $x = \ln 28$, и ядро, таким образом, включает в себя 28 журналов. Это можно считать более реалистичной оценкой объема «ядерной», самой спрашиваемой литературы. Отметим, что весь участок, на котором хорошо выполняется закон Брэдфорда при делении методом Эгге на 12 зон, охватывает, как было сказано ранее (параграф 2.2), журналы с 28-го по 195-й и отрезок на (логарифмической) оси абсцисс [3,3; 5,3]. Середина этого участка находится в точке 4,3, что совпадает с найденной в (2.3.2) точкой перегиба аппроксимирующего полинома. Наше предположение о размещении этой точки в середине прямолинейного участка подтвердилось, и два способа оценки величины ядра (оценка сверху и уточненный показатель) находятся в согласии друг с другом.

В Приложении 2 в явном виде приведен список из 28 журналов, составляющих выделенное нами ядро. В результате проделанной работы по моделированию и выделению ядра можно приступать к практическим действиям по корректировке фонда: обеспечивать полноту электронного доступа университета к найденным ключевым изданиям (наличие полного архива и доступа к самым свежим номерам), выписывать ядерные журналы также и в печатном виде и т. д. Кроме того, возможен анализ важности баз данных, включенных в исследование. Так, 19 из 28 ядерных изданий представлены в базе данных EBSCO, 19 — в JSTOR, 18 — в ProQuest и лишь 5 — в ScienceDirect. В то же время стоит вопрос об «эксклюзивности» представления журналов в каждом ресурсе. Так, если отказаться от подписки на ScienceDirect, то будут полностью потеряны 5 изданий из ядра спроса, если от EBSCO — то одно, остальные же ядерные журналы дублируются в той или иной форме (за те или иные годы) в двух и более базах данных. Всё это целесообразно принимать во внимание при принятии решений об изменении состава электронной подписки.

3. Моделирование процесса обращения к электронным информационным источникам при помощи информетрического принципа Парето

3.1. Формулировка принципа Парето и примеры его действия

Правило (принцип) Парето, известное также как «правило 80/20», было обнаружено в 1897 г. итальянским экономистом и социологом Вильфредо Парето. Путем эмпирического наблюдения он установил, что 80% земель в Италии принадлежат 20% населения (соответственно, остальные 80% граждан довольствуются 1/5 частью итальянской земли) [108]. Подобно тому как распределение Гаусса описывает целый спектр разнообразных процессов, принцип Парето оказался применим ко многим явлениям совершенно различной природы. Для примера можно привести следующие варианты данного правила:

— 20% клиентов приносят компании 80% дохода;

— за 20% рабочего времени делается 80% работы;

— 80% посещений лекций обеспечивают 20% студентов;

и т. д.

Следует отметить, что сам В. Парето формулировал свой принцип только для экономической сферы. На роль исследователя, распространившего правило на явления иной природы, претендует Дж. Джуран [90].

В самом общем случае принцип Парето формулируется так: *20% причин отвечают за 80% следствий*.

Как и в случае других информетрических законов, данное правило — эмпирическое, оно является попыткой приблизительно описать сложные

неравномерные процессы и должно проверяться в каждом конкретном случае. Это не закон, а, скорее, «закономерность» и, кроме того, сама пропорция может оказаться иной — не 80/20, а, например, 90/10 или 70/30. Правило Парето указывает именно на существенное отклонение от пропорции 50/50 в различных системах, а не на конкретную величину отклонения.

Пионерскими работами в области исследования сильной концентрации спроса на отдельных наименованиях, возникающей при использовании традиционной периодики, являются, например, доклад Д. Дж. Уркхарта [125] (на материале заказов из других библиотек, поступивших в британскую Библиотеку Музея науки), его же статья [126] или работа А. Дж. Медоуза [101] (в последней, правда, в качестве меры «использования» берутся цитирования различных журналов: исследуются пристатейные списки литературы в трехгодичном комплекте журнала «Monthly Notices of the Royal Astronomical Society»).

Впервые явно упоминаемое правило 80/20 в применении к библиотекам было рассмотрено и обобщено в 1969 г. Р. Трюсвэллом [123] (который при этом не упоминал В. Парето или Дж. Джурана, называя закономерность просто «80/20 Rule»). Он, исследовав распределение спроса по единицам хранения для трех библиотек (в том числе по данным [73]), пришел к выводу, что принцип Парето хорошо выполняется для статистики выдачи литературы: примерно 80% требований читателей приходилось на 20% библиотечного фонда. Его исследование касалось только печатных изданий (как журналов, так и монографий).

Вслед за Р. Трюсвэллом статистику выдач монографий изучали авторы [36], которые на данных статистики книговыдач в Хиллмановской библиотеке Питтсбургского университета, собранной в течение более чем семи лет, получили, что 80% запросов читателей пришлось на 42,5% единиц хранения (не более 5% фонда имели несколько экземпляров одной книги,

поэтому корректнее говорить «около 42% наименований»). При этом цифра 42,5% берется от объема использовавшегося фонда, т. е. экземпляров, которые хотя бы раз выдавались. Если бы брался процент от всего фонда, предназначавшегося для выдачи на абонемент, была бы получена пропорция 80/22.

Л. Хардести [83] (цит. по [69]) установил, что 30% наименований обеспечивали 80% читательского спроса в библиотеке Университета Де По.

У. Бриттен [31] провел более детальное исследование восьмилетней статистики обращения фонда монографий, сериальных изданий (исключая периодику), диссертаций, микрофильмов и аудиовидеозаписей в библиотеках Университета Теннеси в Кноксвилле. Выяснилось, что совокупно 20% единиц хранения ответственны чуть больше чем за 80% спроса, т. е. правило Парето в целом выполняется. При этом, если ограничиться только теми документами, которые были хотя бы один раз востребованы, пропорция изменится на 80/44. Однако если исследовать аналогичные соотношения, рассмотрев фонд как совокупность различных тематических рубрик (автор опирается на классификацию Библиотеки Конгресса, LC classes), то окажется, что внутри каждого класса 80% спроса читателей приходится на весьма различные доли фонда — от 40% для класса RG (гинекология) до 1,5% для класса J (акты законодательной и исполнительной власти). Аналогичные различия могут быть найдены и внутри отдельного класса, в подрубриках.

В то же время ряд исследований книговыдач в библиотеках показал неидеальное выполнение принципа Парето. В [69] в результате анализа пятилетней работы абонемента в библиотеке Центра наук о здоровье Университета Нью-Мехико было получено, что 80% всех выдач приходится на 36% наименований (к тому же процент брался от всего фонда, включая

16% книг, не выданных за период исследования ни разу). Аналогичный результат зафиксирован для другой библиотеки наук о здоровье — Университета Иллинойса в г. Чикаго [22]. По анализу выдач книг в течение трех лет после их поступления в фонд была получена пропорция 80/38. Авторы [69] и [22] склонны объяснять значительное отклонение в результатах от правила Парето, зафиксированного в других исследованиях, ссылкой на специфику медицинских библиотек и низким отношением числа доступных в фонде монографий к количеству читателей.

Проводились подобные исследования и в отношении периодических печатных изданий. В одной из ранних работ [73] (на нее также ссылается Трюсвэлл), приводятся данные, из которых можно установить, что 80% суммарного спроса на журналы в медицинских библиотеках Колумбийского и Йельского университетов обеспечивают 28% наименований (от «объединенного фонда»). Однако авторы не указывают этого процентного отношения в явном виде и не связывают результаты с правилом Парето (правда, упоминается вскользь закон Брэдфорда, см. раздел 2.1.3).

Работа [46] изучает спрос на периодику по физике в Библиотеке естественных наук Массачусетского технологического института. Здесь также не упоминается принцип Парето, однако из приведенных таблиц и графиков можно установить, что за 3,5 месяца, когда длилось исследование, 80% обращений пришлось примерно на 23% изданий, спрошенных хотя бы один раз за это время. Интересно, что стоимость подписки на данные журналы при этом составляет около 35% от суммарной стоимости всех периодических изданий, использовавшихся хотя бы раз.

По статистике, опубликованной Д. Мортоном [103], изучавшим запросы на периодику, поступившие в рамках межбиблиотечного обмена в системе библиотек массачусетских клиник, можно получить соотношение примерно 80/40, правило Парето не выполняется. При этом автор, сравни-

вая на разных уровнях процента использования литературы (в зависимости от доступных ему данных) результаты своего исследования с работами [131; 130; 46; 45], предполагает, что сходство полученных им соотношений с наблюдаемыми в двух первых публикациях и резкое отличие от двух последних объясняется тем, что существенно бóльшую концентрацию спроса на отдельных изданиях следует ожидать при анализе выдач внутри библиотеки, а не при исследовании межбиблиотечного обмена: наиболее спрашиваемые журналы, которые и обеспечивают концентрацию, как правило имеются в каждой библиотеке и поэтому не будут затребованы у коллег из другой организации.

Исследование [100] посвящено сравнению распределений по журналам их «суммарного использования» и «плотности использования» (берется отношение спроса к протяженности занимаемой изданием книжной полки, т. е. приближенно к общему объему издания). Для суммарного использования правило Парето выполняется практически идеально — 80% спроса приходится на 21% наименований (при выделении только текущей, не более чем пятилетней давности, периодики это отношение несколько иное, 73/20). Для плотности использования правило существенно отличается (80/37), но, вопреки далеко идущим выводам авторов [100], этот эффект обусловлен самой методикой: журналы пересортировываются по *плотности* использования, а потом берутся *n* верхних из них, которые обеспечивали бы 80% *суммарного* спроса. Очевидно, что при таком подходе число изданий окажется бóльшим, чем при сортировке по суммарному использованию. Фактически, авторами решается задача о нахождении такой части фонда, которая обеспечивает 80% спроса, занимая при этом минимальное пространство на полке. Надо, тем не менее, отметить, что найденное подмножество журналов занимает 46% полочного пространства, т. е. для рас-

пределения «спрос/полочное пространство» правило Парето не выполняется, в отличие от распределения «спрос/число наименований».

В статье [116] сотрудница библиотеки Трентского университета (Канада) М. Щильяно приходит к выводу, что 80% спроса пришлось всего на 15% наименований журналов. Точно такую же пропорцию 80/15 наблюдали авторы из Северной Каролины (США) [118] (в обоих исследованиях учитывались также издания с нулевым использованием, если их отбросить, отношение получилось бы более «мягким»). Правда, в последнем случае стоимость подписки на журналы, обеспечивающие 80% использования, составила 38% от всего бюджета на периодику — однако здесь справедливо замечание, высказанное выше: для проверки правила Парето на распределении «спрос/цена» необходимо упорядочить издания не по их суммарному использованию, а по отношению использования к стоимости подписки.

Сообщалось, что статистика выполнения заказов по доставке документов из научных периодических изданий также подчиняется принципу Парето. Так, в двухгодичном анализе работы службы доставки документов Турецкого академического сетевого и информационного центра (TANIC) было зафиксировано, что 80% запросов на получение статей приходилось на 21% журналов [122]. Трехгодичные данные о функционировании службы фотокопирования французского Национального института научной и технической информации при Национальном центре научных исследований (INIST-CNRS), собранные по запросам на европейские журналы по фармакологии, дают пропорцию 80/28 [18]. В работе [114] про эту же национальную службу сообщается, что «стандартное правило 80/20 (80% запросов относятся лишь к 20% наименований) подтверждалось из раза в раз на протяжении шести лет».

В последнее время появились аналогичные работы по статистике использования электронных изданий. В масштабном исследовании дея-

тельности консорциума OhioLINK (всего за период наблюдения было открыто более полутора миллионов электронных статей) оказалось, что спрос распределен более равномерно, чем говорит «классическое» правило Парето, и 80% загруженных статей относятся к 40% наименований периодики [55]. Сходные результаты получены в проекте PEAK, во время которого предоставлялся доступ к 1100 электронным журналам издательства Elsevier для консорциума из 12 крупных и средних по масштабу библиотек: 80% открытых файлов содержали статьи из 37% представленных наименований. В то же время для одного из участников проекта, Университета Индианы, пропорция оказалась почти «классической»: на 23% наименований периодических изданий пришлось 80% спроса пользователей университета [129]. В работе [72] проверяется принцип Парето для всей совокупности обращений за 2004 г. пользователей онлайновой базы данных периодических изданий Emerald Management Xtra к содержащимся в ней 108 журналам (исключались издания, прекратившие выходить к 2004 г.). Выясняется, что принцип существенным образом не выполняется: 80 процентов загруженных статей пришлись на 47% наименований журналов, представленных в продукте. Также не удовлетворяют классическому правилу Парето данные по трехлетнему использованию комплекса из четырех ресурсов (Academic Press, Kluwer, MCB Emerald, Wiley), полученные в [124] для Консорциума каталонских университетских библиотек (CBUC). Для различных платформ 80% загруженных читателями электронных статей приходится на от 27% до 46% наименований журналов, имеющихся в соответствующем ресурсе, а обобщенные по всем ресурсам показатели дают пропорцию 80/34.

3.2. Кривая Лоренца, индекс Джини и выделение ядра журналов с использованием принципа Парето

Один из наиболее наглядных способов проверки принципа Парето — построение так называемой «кривой Лоренца» (или «Леймкулера-Лоренца», см. параграфы 3.3 и 4.1). Впервые такого рода кривая была построена М. О. Лоренцем в работе [96] для изучения концентрации распределения дохода населения. По одной оси откладывается кумулятивный процент числа населения, по другой — процент общего дохода, который приходится на данную долю населения. При этом представители населения упорядочены согласно их доходам — в методике Лоренца, от самых бедных до самых богатых. Пример трех кривых Лоренца представлен на рис. 3.2.1.

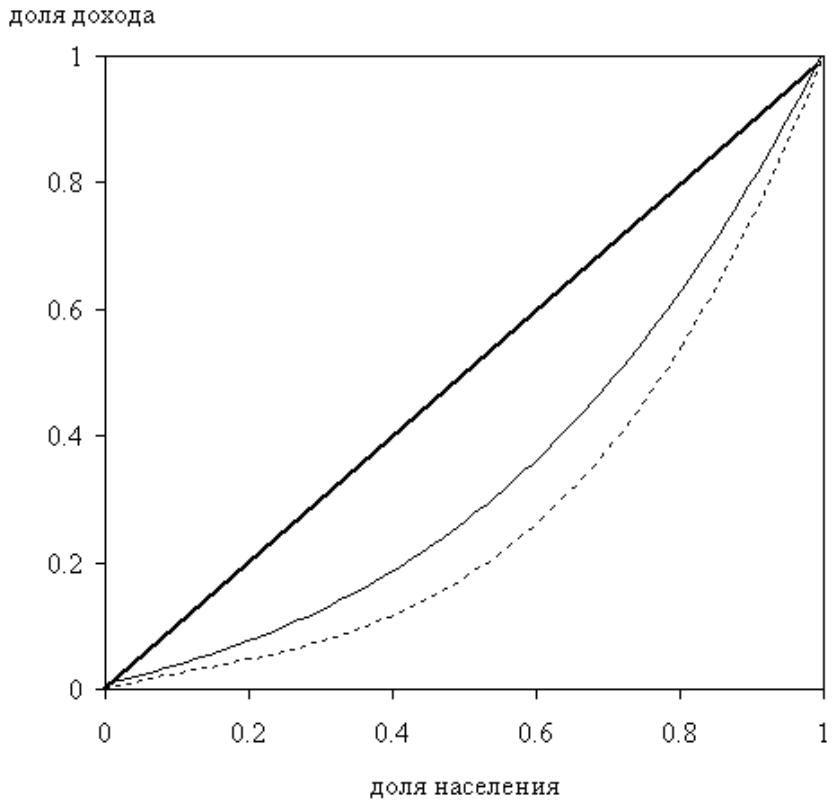

Рис. 3.2.1. Кривые Лоренца

Если бы распределение доходов среди населения было идеально равномерным, то кривая Лоренца имела бы вид прямолинейного отрезка, проведенного из точки (0; 0) в точку (1; 1) (полужирный отрезок прямой на рис. 3.2.1). Чем больше доходы концентрируются на небольшой богатой части населения, тем сильнее выгнута кривая Лоренца, тем больше она отходит от «равномерной» диагонали. Так, очевидно, что на рис. 3.2.1 кривая Лоренца, изображенная пунктиром, характеризует более неравномерно распределенные доходы, чем сплошная вогнутая кривая.

Отметим, что при помощи кривых Лоренца правило Парето проверяется элементарным образом: достаточно посмотреть, какая абсцисса у точки на кривой с ординатой 0,2. Разность между 1 и этой абсциссой покажет, какому проценту населения принадлежат 80% всего дохода страны. Опять же, при абсолютно равномерном распределении дохода, очевидно, 80% дохода придется на 80% населения, что и отражает кривая Лоренца в виде диагонального отрезка прямой. Аналогичный подход можно применить и в нашем случае, заменив долю населения долей журналов, а долю доходов — долей статей, прочитанных из данных журналов.

Для количественной оценки неравномерности описанных распределений применяется показатель, впервые предложенный итальянским статистиком К. Джини [77] и называемый индексом (или коэффициентом) Джини *G*. Он тесно связан с кривой Лоренца и графически может быть выражен как отношение площади между кривой Лоренца и абсолютно равномерным распределением ко всей площади под абсолютно равномерным распределением. В обозначениях рис. 3.2.2 индекс Джини

$$G = \frac{A}{A + B}, \qquad (3.2.1)$$

где A и B — площади соответствующих областей.

Индекс Джини может изменяться от нуля ($A = 0$, кривая Лоренца совпадает с диагональным отрезком, и распределение абсолютно равномерное) до единицы ($B = 0$, кривая Лоренца на полуинтервале [0; 1) совпадает с осью абсцисс, а в точке $x = 1$ принимает значение 1, распределение абсолютно неравномерное). Чем больше *G*, тем больше неравномерность исследуемого распределения. Так как площадь под диагональным отрезком, очевидно, равняется $1/2$, формулу (3.2.1) можно переписать в виде

$$G = 2A = 1 - 2B\,. \qquad (3.2.2)$$

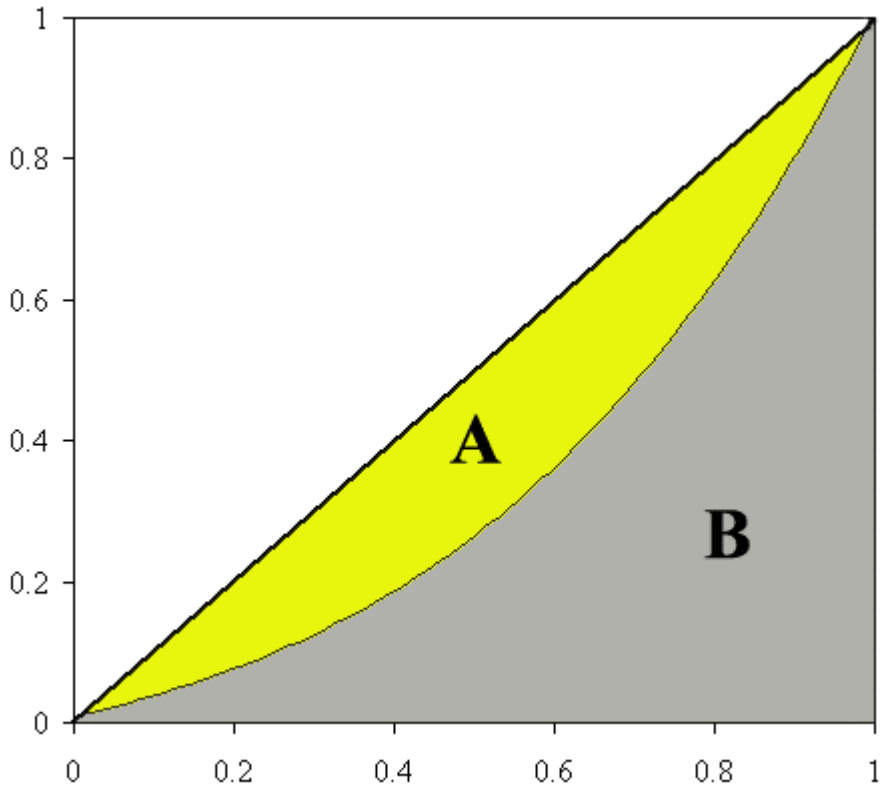

Рис. 3.2.2. Вычисление индекса Джини по кривой Лоренца.

Для нашего случая можно подсчитать площадь под кривой Лоренца, приблизив ее трапециями и просуммировав полученные результаты. В результате получим:

$$B = \left(2SI\right)^{-1} \sum_{k=0}^{S-1} \left(I_{k+1} + I_k\right), \tag{3.2.3}$$

где S — общее число источников (журналов), I — общее число продуктов (открытых статей), а I_k — суммарное число продуктов в первых k наименее продуктивных журналах. Так как мы работаем с кривой Лоренца, то упорядочиваем источники в порядке неубывания их продуктивности. По формуле (3.2.2):

$$G = 1 - \left(SI\right)^{-1} \sum_{k=0}^{S-1} \left(I_{k+1} + I_k\right). \tag{3.2.4}$$

В работе [67] был предложен метод выделения ядра журналов — наиболее важных в той или иной предметной области изданий — на основе использования кривых Лоренца (см. также теоретические выкладки и критику в [42]). Суть метода заключается в следующем. Упорядочим журналы, публикующие статьи по выделенной дисциплине, в порядке убывания в них числа статей данной тематики (пусть всего имеется N журналов). Подсчитаем индекс Джини G для получившейся выборки. Заменим в выборке число статей в $N - s$ менее продуктивных изданиях на нули так, чтобы ненулевые значения остались для s наиболее продуктивных журналов. Снова подсчитаем индекс Джини для новой выборки (включающей в себя не s значений, а все N, однако последние $N - s$ равны нулю). Обозначим такой коэффициент G_s. Очевидно, что $\forall\, s > 0$ справедливо неравенство $G_s > G$ (обнулением $N - s$ точек мы могли только «увеличить неравенство» в распределении, следовательно индекс Джини возрастает). Кроме того, чем меньше s, тем выше G_s.

Для журнала с номером s (в выборке, упорядоченной по убыванию продуктивности) введем специальную характеристику, «коэффициент принадлежности к ядру», вычисляемую по формуле

$$m_s = \frac{G_s - G}{G_1 - G} \qquad (3.2.5)$$

Легко видеть, что $m_s = 1$ для первого, наиболее продуктивного источника (ядро должно состоять как минимум из одного журнала), убывает с ростом s и для максимального $s = N$ обращается в нуль ($G_N = G$). Авторы рассматриваемого метода предлагают выбрать некоторый уровень $m^{\min}$ (в 90 или 95 процентов) и считать ядром те журналы, для которых m_s превышает соответственно 0,9 или 0,95. Несмотря на некоторую произвольность выбора уровня «отсечения», данная методика более прогрессивна, чем, например, простой выбор 10 или 5 процентов наиболее продуктивных журналов для ядра. Она учитывает реальную степень «отличия» журналов, входящих в ядро, от остальных изданий, а не отсекает массив по формальным количественным критериям.

В разделе 3.4 мы применим данный метод к статистике обращения к электронным журналам, что позволит еще одним способом получить ядро читательского спроса, наиболее важные издания, и сравнить полученные результаты с множеством ядерных журналов, полученным в разделе 2.3 при моделировании с помощью закона Брэдфорда.

3.3. Принцип Парето как модель процесса обращения к электронным информационным источникам

Информетрическое моделирование с использованием принципа Парето проведем для статистики обращений к каждой из четырех онлайновых платформ, входящих в наше исследование, а также для статистических показателей, объединенных по всем ресурсам.

Подсчитаем для каждого из четырех ресурсов количество открытых статей в двух наиболее востребованных журналах, в трех наиболее востребованных журналах и т. д., получая в итоге зависимость кумулятивного числа открытых статей от количества журналов, взятых из верхних строк таблицы востребованности изданий. Отразим полученную зависимость в виде четырех кривых, нанесенных на одном графике (рис. 3.3.1).

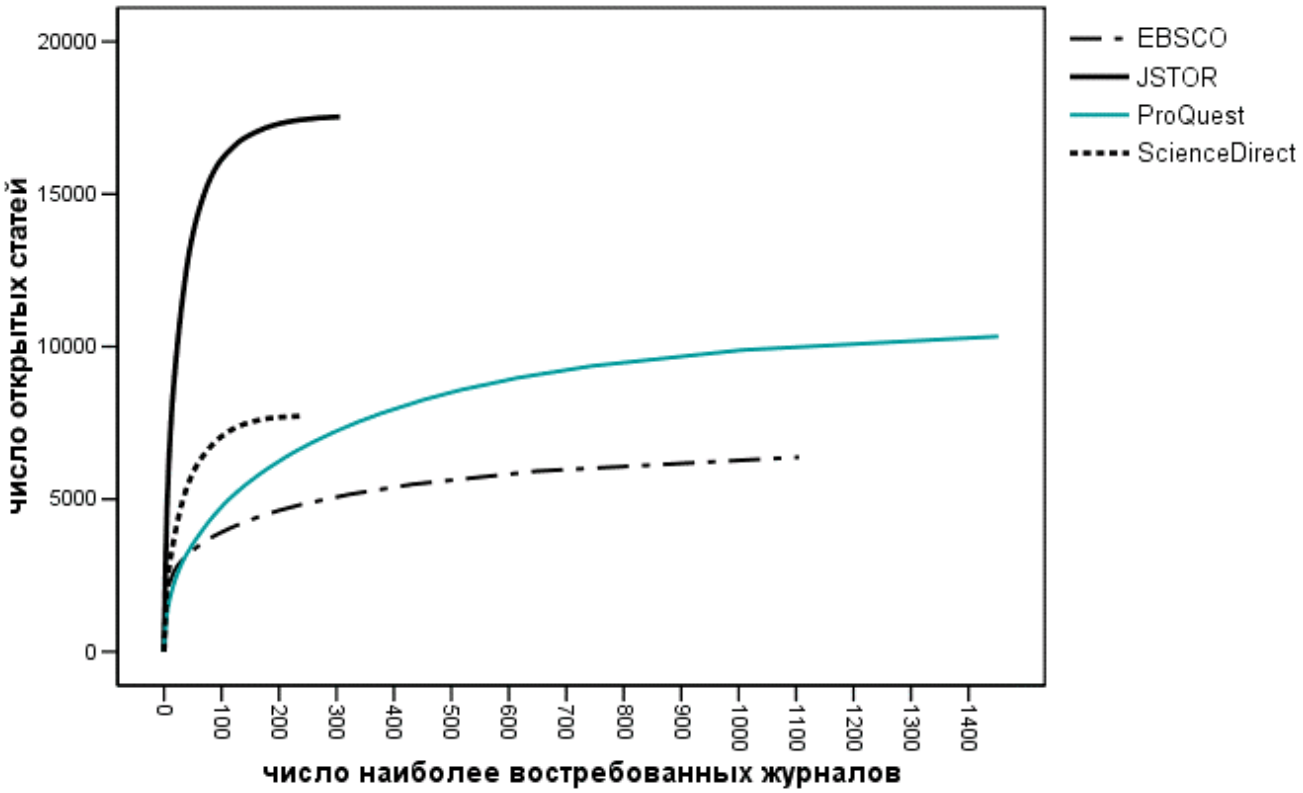

Рис. 3.3.1. Зависимость кумулятивного числа открытых статей от числа наиболее востребованных журналов (для четырех электронных ресурсов)

Далее преобразуем полученные графики, нормировав их по обеим осям: выберем в качестве 100% по оси абсцисс число журналов из соответствующего ресурса, к которым было хотя бы одно обращение пользователей, а 100% по оси ординат — суммарное количество статей, востребованных из базы данных. Заметим, что полученные кривые абсолютно анало-

гичны кривым Лоренца, несут ту же смысловую нагрузку и отличаются лишь способом упорядочивания «источников»: не от наименее продуктивных к наиболее продуктивным, а наоборот. Такие кривые называют кривыми Леймкулера, см. также дальше раздел 4.1. Результаты построения отражены на рис. 3.3.2.

Рис. 3.3.2. Зависимость кумулятивного процента открытых статей от процента наиболее востребованных журналов («кривые Леймкулера»).

На основании рис. 3.3.2 можно наглядно оценить степень «неравномерности» спроса на различные наименования журналов для каждой базы данных. Чем больше площадь под кривой Леймкулера, тем меньше площадь под классической кривой Лоренца (одна является отражением дру-

гой). Следовательно, тем больше коэффициент Джини, являющийся индексом концентрации и, в итоге, тем больше неравномерность спроса. Поэтому можно заключить, например, что «на старте кривой», т. е. для наиболее активно используемых журналов, максимальную концентрацию спроса демонстрирует EBSCO, однако в дальнейшем востребованность различных наименований в этом ресурсе становится более равномерной. В целом наибольшее «неравенство спроса» демонстрирует не EBSCO, а JSTOR — именно под этим графиком площадь наибольшая. Напротив, наиболее равномерный спрос на различные издания имеем в случае базы данных ProQuest.

При помощи построенных графиков можно наглядным образом проверить правило Парето. Те точки, в которых кривые пересекают отметку по оси y в 80%, показывают пропорцию Парето для соответствующей базы данных. Видно, что для JSTOR правило Парето выполняется практически точно, а для трех остальных ресурсов концентрация спроса меньше и, соответственно, пропорция Парето более «мягкая», около 80/30 (80% востребованных статей опубликованы примерно в 30% журналов). Точные цифры приведены в Таб. 3.3.1, там же приведены индексы Джини, рассчитанные по формуле (3.2.4).

Теперь объединим данные по всем четырем ресурсам (просуммировав показатели для тех журналов, которые представлены более чем в одной базе данных) и построим кривую Леймкулера для совокупной электронной подписки (рис. 3.3.3). 80% востребованных статей приходится на 14,7% наименований — пропорция получается более сильная, чем для каждого ресурса в отдельности.

Для того чтобы получить «наглядное» отношение Парето, сумма членов которого будет равна 100, построим отрезок с координатами (0; 100), (100; 0) и найдем точку его пересечения с графиком на рис. 3.3.3. Визуаль-

но (а также из точных численных данных) следует, что в данном случае имеет место соотношение «83/17»: 83% открытых статей находятся в 17% использованных хотя бы один раз журналов.

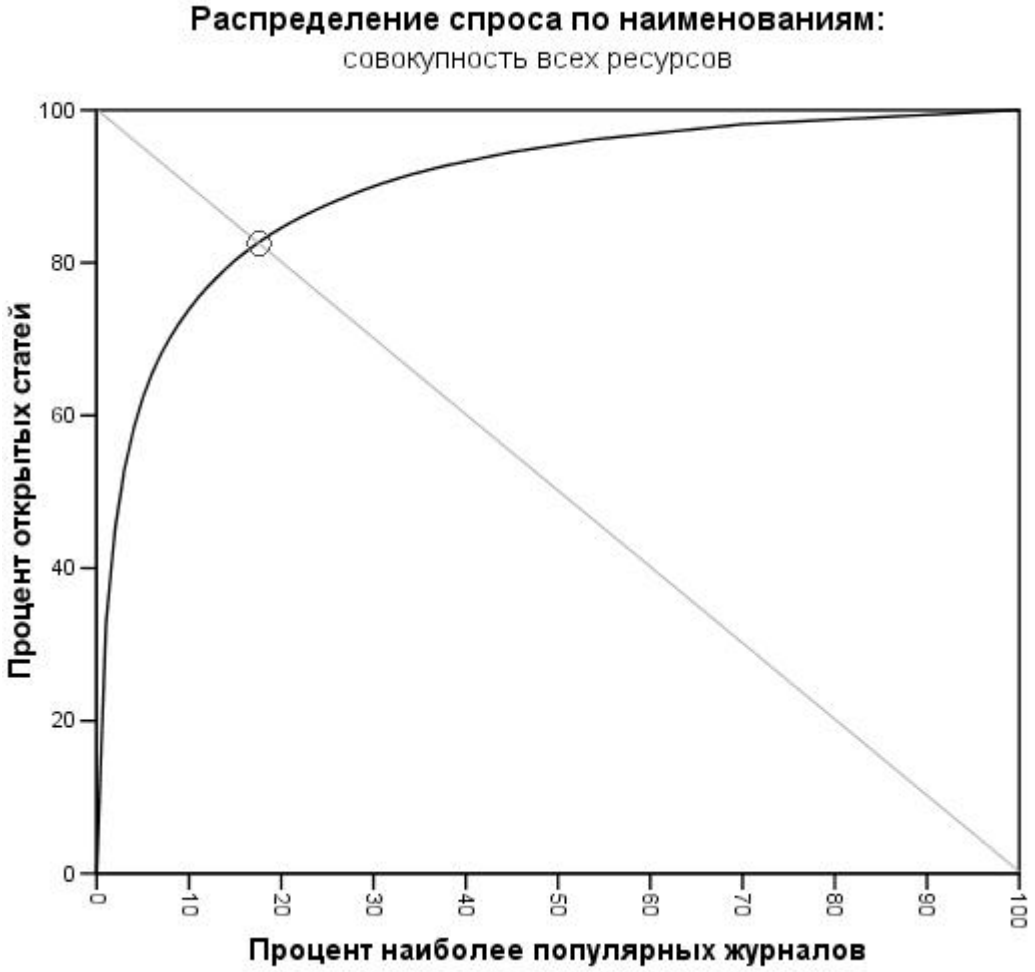

Рис. 3.3.3. Зависимость кумулятивного процента открытых статей от процента наиболее востребованных журналов (суммарно по всем четырем ресурсам).

Итоговые показатели сведены в таб. 3.3.1. Из представленных данных можно сделать вывод, что принцип Парето выполняется не самым лучшим образом для баз данных ProQuest и EBSCO. В случае ScienceDirect пропорция уже ближе к классическому отношению «80/20», а для JSTOR она даже более «сильная». Если же рассмотреть всю электронную подписку как еди-

ный информационный массив, то пропорция Парето демонстрирует здесь наибольшую неравновесность и концентрацию спроса на более популярных журналах: 80% открытых статей из всех четырех баз данных опубликованы в 14,7% журналов из числа тех, которые имеются в подписке и хотя бы один раз были востребованы пользователями. Обратим внимание, что если бы мы учитывали и те журналы, которые не были востребованы ни разу, пропорция в каждом случае еще более ужесточилась бы.

	востребовано информации	из числа изданий	индекс Джини
EBSCO	80%	28%	0,695
ProQuest	80%	31%	0,649
JSTOR	80%	18%	0,780
ScienceDirect	80%	25%	0,709
все ресурсы	**80%**	**14,7%**	**0,804**

Таб. 3.3.1. Выполнение правила Парето и величина индекса Джини для четырех ресурсов и суммарной электронной подписки

Хотя это не является необходимым условием, в нашем случае наблюдается прямая зависимость: чем больше индекс Джини, тем «строже» исполняется закон Парето. Это означает, что чем раньше кривая Леймкулера для ресурса (или их общей совокупности) пересечет ординату в 80%, тем большая площадь под ней находится на всем ее протяжении. Для полной совокупности всех ресурсов индекс Джини равен 0,80. Это означает, что «сводная» кривая Леймкулера покрывает площадь 0,9, т. е. 90% всего квадрата. Это очень серьезная степень концентрации спроса.

Стоит подробнее остановиться на обнаруженном эффекте: объединение всех ресурсов дает бóльшую концентрацию, чем каждый ресурс в отдельности. В теории этот факт не является необходимым, и возможны как варианты, соответствующие такой тенденции, так и демонстрирующие об-

ратную картину. Так, в уже упоминавшейся работе [124] по графикам Fig. 3 можно определить, что для четырех рассматриваемых там ресурсов пропорции Парето составляют около 80/25, 80/35, 80/35 и 80/45. Для общего же массива электронных изданий, как указывают авторы, выполняется пропорция 80/34,45. Таким образом, усиления концентрации при объединении ресурсов не происходит. Заметим, впрочем, что исследование [124] отличается от нашего, во-первых, тем, что рассматриваемые электронные базы данных являются платформами четырех *издательств*, следовательно, присутствие одних и тех же изданий в двух различных ресурсах практически исключено. Кроме того, речь идет об использовании информационных продуктов целым консорциумом из 11 библиотек.

Если вспомнить работу [72], последняя деталь может показаться немаловажной: именно вся совокупность подписчиков базы данных Emerald Management Xtra дает такое использование ресурса, что принцип Парето существенным образом не выполняется: 80 процентов загруженных статей пришлись на 47,4% наименований журналов, представленных в продукте. Также и в статье [54] сообщается о том, что агрегация использований отдельных институтов в общую статистику по консорциуму NERL делает использование наименований журналов более «гладким». Если три различные группы университетов, входящих в консорциум, — медицинские, «крупные» и «иные» — давали пропорции соответственно 90/20, 80/20 и 83/20, то по консорциуму как целому получался результат 75/20 (в каждом случае пропорция бралась от всех 206 журналов, входящих в пакет доступа, а не от использованных каждой группой изданий). Можно осторожно предположить, что соединение статистики использования различных подписчиков ресурса чаще «размывает» концентрацию. В ГУ-ВШЭ мы наблюдаем обратный случай: рассмотрение всех ресурсов сразу делает выпуклее правило Парето. Но это именно различные и разнотипные ресурсы,

а не различные пользователи или группы пользователей. Здесь, вполне возможно, и заключается влияющая на результат разница, хотя теоретические исследования и построение информетрических моделей в данном направлении, конечно, должны быть продолжены для получения более обоснованных выводов.

Итак, мы проверили правило Парето в применении к процессам обращения к электронным информационным ресурсам. Было установлено, что принцип Парето выполняется удовлетворительно, а в ряде случаев даже сильнее, чем классическое соотношение 80/20 (80% информации почерпнуто из 20% источников). При этом наиболее «контрастная» пропорция наблюдается при соединении всех ресурсов, рассматриваемых как единый информационный массив: в этом случае отношение может быть записано в виде 83/17 (83% загруженных статей находятся в 17% журналов, востребованных хотя бы единожды), в отличие от классического отношения 80/20.

Какие отсюда следуют практические выводы? Например, значит ли это, что комплектаторам целесообразно отказаться от значительного массива выписываемых электронных изданий, не вошедших в верхние 20%, — тех, чья востребованность, как видно, незначительна по сравнению с остальными? Является ли уровень таких изданий однозначно низким? Вполне очевидно, что в случае университетской библиотеки это не так: в число 20% лидеров, как правило, попадают издания, отвечающие запросам более широкой аудитории, в то время как узкоспециализированные журналы могут иметь небольшую общую популярность, однако представлять ключевой интерес для преподавателя или сотрудника, работающего в соответствующей отрасли.

Кроме того, следует иметь в виду, что при подписке на базы данных обычно невозможно выбирать журнал за журналом: договор чаще всего заключается на доступ к целой «коллекции» изданий, сформированной са-

мим вендором. Это неизбежно приводит к приобретению «информационного шума», пользующегося пониженным спросом.

Также стоит напомнить о принципиальной «устойчивости», свойственной большинству информетрических законов, о которой говорит А. Букстейн [26; 27]. В данном случае это означает, что если из совокупной статистики (2590 изданий) убрать все журналы, кроме первых 380, отвечающих за 80% спроса, то в оставшемся множестве вновь будет наблюдаться значительная концентрация спроса, хотя и в меньшей степени: 80% от спроса на все эти издания (т. е. 64% от суммарного) придется на 141 журнал, что составляет 37% от выделенных 380 изданий. Отметим, однако, что по соображениям здравого смысла эта концентрация не может сохраняться при многократном повторении аналогичного процесса «усечения» числа учитываемых изданий: едва ли, если выделить 5 наиболее спрашиваемых журналов (возьмем предельный случай), на первый из них будет приходиться около 80% спроса, а остальные четыре будут делить оставшиеся 20%. В нашем случае, как можно видеть из таблицы Приложения 1, на ведущий журнал приходится лишь 25% использования от всей лидирующей пятерки. Аналогичный результат следует, например, из исследования [115] статистики востребованности различных журналов в службе доставки документов Subito за 2003 г. В нем берутся 500 наиболее востребованных журналов (из суммарного числа 750000, доступных в Subito), на которые приходится 22,1% спроса. При этом кривая Леймкулера для этих журналов оказывается почти прямой линией (график Fig. 1 в статье), т. е. концентрация минимальна, а 80% из выделенных 22,1% запросов приходится примерно на 70% из 500 ведущих журналов.

Тем не менее, в ситуации ограниченного бюджета библиотеки, когда так или иначе приходится выбирать между различными ресурсами и, вероятно, отказываться от каких-то из них, анализ востребованности баз дан-

ных в свете правила Парето может служить ценным ориентиром при принятии решений о составе подписки и стать одним из инструментов, используемых при формировании политики комплектования фонда библиотеки онлайновыми информационными источниками.

В качестве развития исследований настоящей главы можно предложить проверить информетрический принцип Парето в ином аспекте — на распределении числа обращений к онлайновым ресурсам не по наименованиям журналов, а по отдельным пользователям. «Пользователями» могут быть в таком случае как индивидуальные читатели, так и коллективные подписчики, например, библиотеки в случае консорциума библиотек. Можно поставить вопросы: какой процент организаций-участников консорциума отвечает за 80% от суммарного числа обращений к электронным ресурсам со стороны всего консорциума? Какая доля читателей университета формирует 80% от общего спроса на материалы из онлайновых баз данных? Это тема отдельного исследования, однако можно предположить, что правило Парето найдет свое место и при ответе на данные вопросы. Так или иначе, концентрация спроса в таком информационном процессе неизбежна, что подтверждают некоторые более ранние опыты. Например, в [52] сообщается, что 10 наиболее активных клиентов службы фотокопирования в корпоративной библиотеке биотехнологической компании (2% от общего числа идентифицированных пользователей) обеспечили 43% общего спроса на копирование статей. Полученные результаты интересно было бы сверить с исследованием принципа Парето для распределения спроса консорциумов по *наименованиям* источников. Как отмечалось, существуют исследования ([72; 124]), показывающие неприменимость или плохую применимость принципа в таких случаях.

3.4. Ядро изданий при моделировании процесса обращения к электронным информационным источникам с применением принципа Парето и индекса Джини

Теперь на имеющемся у нас экспериментальном массиве определим ядро источников (наиболее спрашиваемые издания) методом Л. Эгге и Р. Руссо [67], описанным выше. Воспользуемся формулой (3.2.5), чтобы определить «коэффициент принадлежности к ядру» m_s для востребованных изданий.

Константа G в (3.2.5) равна 0,804 — см. таб. 3.3.1. Константа G_1 определяется величиной индекса Джини при условии, что всё распределение сконцентрировано на одном, самом используемом журнале (т. е. в таком случае было бы 1306 обращений к журналу American Economic Review — и 0 к остальным 2589 источникам). Очевидно, что подсчет индекса Джини для такого распределения даст величину, очень близкую к единице, в нашем случае $G_1 = 0,9996$.

Теперь, зная константы G и G_1 и вычислив индекс Джини «усеченного» распределения G_s для каждого источника с порядковым номером *s*, по формуле (3.2.5) построим зависимость коэффициента принадлежности к ядру m_s от номера источника *s* (рис. 3.4.1).

Прежде всего заметим, что при построении графика была допущена некоторая вольность: т. к. многие журналы с большими номерами *s* имеют одну и ту же продуктивность (778 изданий с одной прочтенной из них статьей, 416 — с двумя и т. д.), данная функция, вообще говоря, должна иметь ступенчатый вид: коэффициент принадлежности к ядру не может быть разным для источников с одинаковой продуктивностью. Мы позволили себе сгладить данную кривую, т. к. нас будет интересовать собственно

ядро, которое находится в области высокой продуктивности, где журналы с одинаковым числом открытых из них статей встречаются редко.

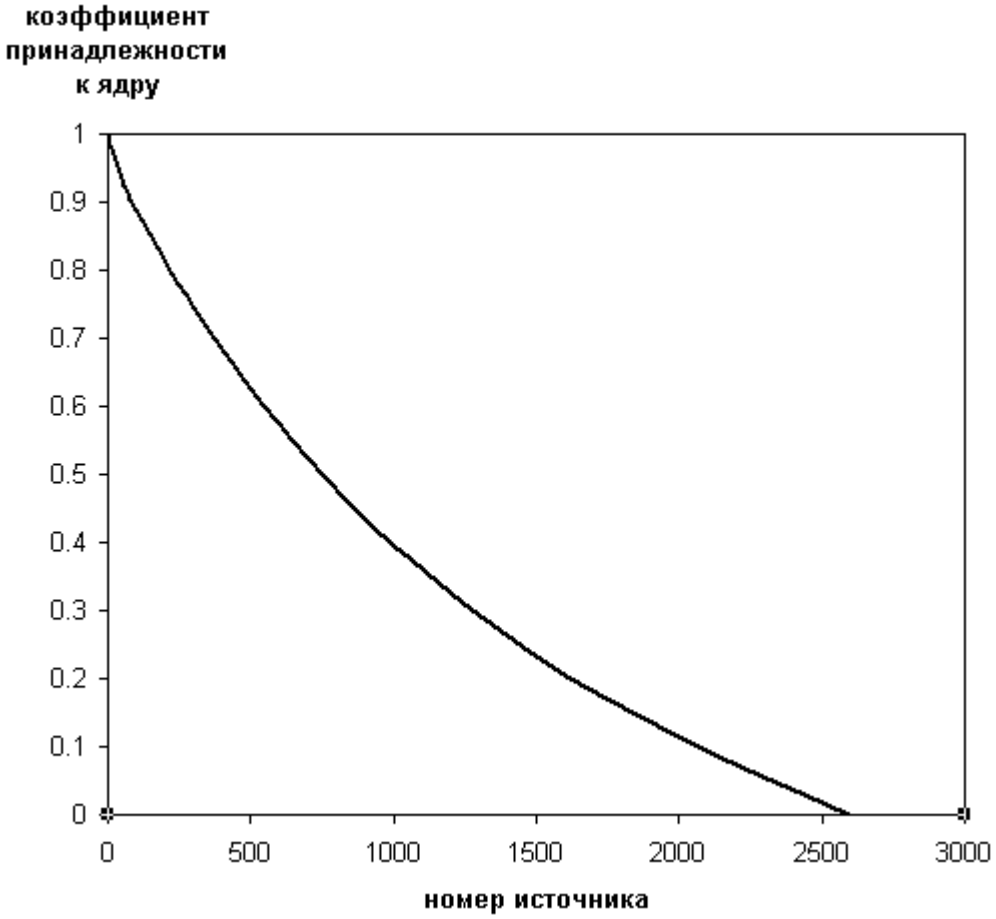

Рис. 3.4.1. Коэффициент принадлежности к ядру (см. [67]) для журналов, по результатам обращения к ним читателей ГУ-ВШЭ

Далее, график несколько вогнут. Напомним (см. параграф 3.2), что метод Эгге-Руссо по выделению «ядра» журналов на основе индексов Джини состоит в выборе некоторого порогового значения $m^{\min}$, по которому отсекается ядро: все источники с $m_s \geq m^{\min}$ входят в ядро, все остальные — нет. Если бы зависимость $m_s(s)$ имела вид прямой, то метод не отличался бы от простого выбора критерия, какой процент наиболее используемых журна-

лов считать ядром: выбор $m^{\min} = 0,90$ был бы эквивалентен отбору в ядро 10% наиболее спрашиваемых источников, выбор $m^{\min} = 0,95$ — отбору 5% самых читаемых изданий и т. д. Значит — ввиду того что график рис. 3.4.1 имеет вогнутость — в нашем случае метод Эгге-Руссо нетождествен простому «арифметическому» способу выделения «ядра» и усовершенствует его.

Авторы [67] рекомендуют брать в качестве критерия для попадания в ядро изданий $m^{\min} = 0,90$, рассматривая также вариант $m^{\min} = 0,95$ (получающиеся при этом выборки называют 90%-ядром и 95%-ядром соответственно). При помощи точных данных о значениях функции $m_s(s)$ определим 90%-ядро для наших экспериментальных данных.

На четырех журналах с порядковыми номерами с 86 по 89, в каждом из которых открыто по 89 статей, функция $m_s(s)$ переходит отметку 0,9: $m_s(86) = 0,901$; $m_s(89) = 0,898$ (эта разница не видна на рис. 3.4.1 из-за сглаженности кривой: на самом деле величина m_s для источников с одной продуктивностью должна быть одинаковой). Следовательно, границей 90%-ядра можно считать 89-й журнал. Аналогичным образом определяется, что 95%-ядро состоит из 37 журналов ($m_s(37) = 0,951$; $m_s(38) = 0,9496$).

Итак, 90%-ядро состоит из 89 наиболее спрашиваемых журналов, или 3,4% от их общего числа, а 95%-ядро — из 37, или 1,4%. Мы видим, что критерий на коэффициент принадлежности к ядру отсекает большее число журналов, чем простой арифметический выбор 10% или 5% наиболее читаемых изданий.

Сравнив эти результаты с полученными в параграфе 2.3.2, где ядро читательского спроса определялось при помощи информетрического закона Брэдфорда, можно констатировать, что по порядку полученной величины оба способа — по Брэдфорду и по Джини-Парето — дают одинаковый

размер ядра. Разные вариации метода с использованием закона Брэдфорда давали ядро в 77 или 28 журналов. Разные критерии на коэффициент принадлежности к ядру в настоящей главе давали 89 или 37 изданий. Примерное совпадение двух различных методов тем интереснее, что в параграфе 2.3.2 используется «точный» подход, с определением границы квазипрямолинейного участка экспериментального библиографа, в то время как используемый в [67] и в настоящем разделе метод не свободен от произвола: границу в 0,90 или 0,95 для коэффициента принадлежности к ядру авторы [67] установили «из опытных соображений», путем обобщения поведения кривой $m_s(s)$ для различных наборов информетрических данных. Получение схожих результатов концептуально различными методами информетрического моделирования позволяет с оптимизмом смотреть на вопрос выделения ядра наиболее спрашиваемых изданий для процесса использования электронных ресурсов, указывая на то, что наличие такого ядра подразумевается самим процессом, а не зависит критическим образом от информетрической модели.

Еще одно интересное наблюдение можно сделать, вернувшись к определенному выше «90%-ядру» (наиболее адекватному и рекомендуемому, по мнению авторов методики). Оно получилось состоящим из 89 изданий, каждое из которых содержит не менее 89 статей. Здесь это, разумеется, только случайность, однако такой результат удивительным образом совпадает с подходом, который в другом разделе информетрии — анализе цитирований — связан с понятием индекса Хирша, или *h*-индекса [86]. Последний определяется, например, для ученого как такое максимальное число *h*, что у ученого имеется *h* статей, которые цитировались не менее *h* раз. Индекс Хирша введен только в 2005 г., сейчас получает всё более широкое распространение и подвергается всестороннему изучению (например, [68; 28; 53]). Если пару «источник — продукт» сменить со «статья —

цитирование» на «статья — прочтение», то подход Хирша становится применим к процессу использования литературы. При этом *h* источников (журналов), имеющих не менее *h* открытых статей, становятся особенно важными: именно они учитываются при подсчете индекса Хирша (условно говоря, остальные журналы могли не читать вообще, *h*-индекс от этого бы не изменился), составляя «*h*-core», *h*-ядро. Возможно, это еще одна мера оценки ядра читательского спроса. Исследования в данном направлении нам пока неизвестны (в работе [68], тем не менее, показано, что индекс Хирша существует и единственен для любого обобщенного «информационного процесса производства»), и приложение *h*-индекса к статистике чтения, безусловно, заслуживает более внимательного рассмотрения. Здесь также хотелось бы отметить, что в нашем случае *h*-ядро набора экспериментальных данных в точности совпало бы с 90%-ядром, определенным с помощью понятий о кривых Леймкулера, принципе Парето и индексе Джини.

4. Моделирование процесса обращения к электронным информационным источникам при помощи других информетрических законов

4.1. Закон Леймкулера

Закон Леймкулера тесно связан с законом Брэдфорда, он был выведен Ф. Леймкулером [94] из «вербальной формулировки» последнего и фактически задает форму «кривой Брэдфорда» («библиографа»), но не в полулогарифмических, а в обычных линейных координатах. Кроме того, закон также тесно связан с принципом Парето и кривой Лоренца.

Построим библиограф в линейных координатах и нормируем его, ограничив единицей размеры по вертикали и горизонтали. Для этого по оси абсцисс отложим долю источников (журналов), упорядоченных от наиболее продуктивных к наименее продуктивным, а по оси ординат — соответствующую кумулятивную долю продуктов (статей), приходящуюся на эти журналы. Так как речь идет о долях, то весь график уложится в квадрат, ограниченный осями координат и прямыми $x = 1$ и $y = 1$. Начинаться он будет в точке (0; 0), а заканчиваться — в точке (1; 1) (см. на рис. 4.1.1 график, изображенный тонкой линией). На самом деле в силу своего определения полученная кривая (которую называют кривой Леймкулера) абсолютно аналогична по смыслу кривой Лоренца и является ее зеркальным отражением относительно точки (0,5; 0,5). Различие состоит лишь в способе сортировки источников (или представителей населения Лоренцем): от наиболее продуктивных к наименее, как в случае кривой Леймкулера, или от наименее — к наиболее продуктивным, как в случае кривой Лоренца. Первый более употребим в информетрии, второй — в экономике, а полу-

ченную кривую нередко называют обобщенно кривой Леймкулера-Лоренца.

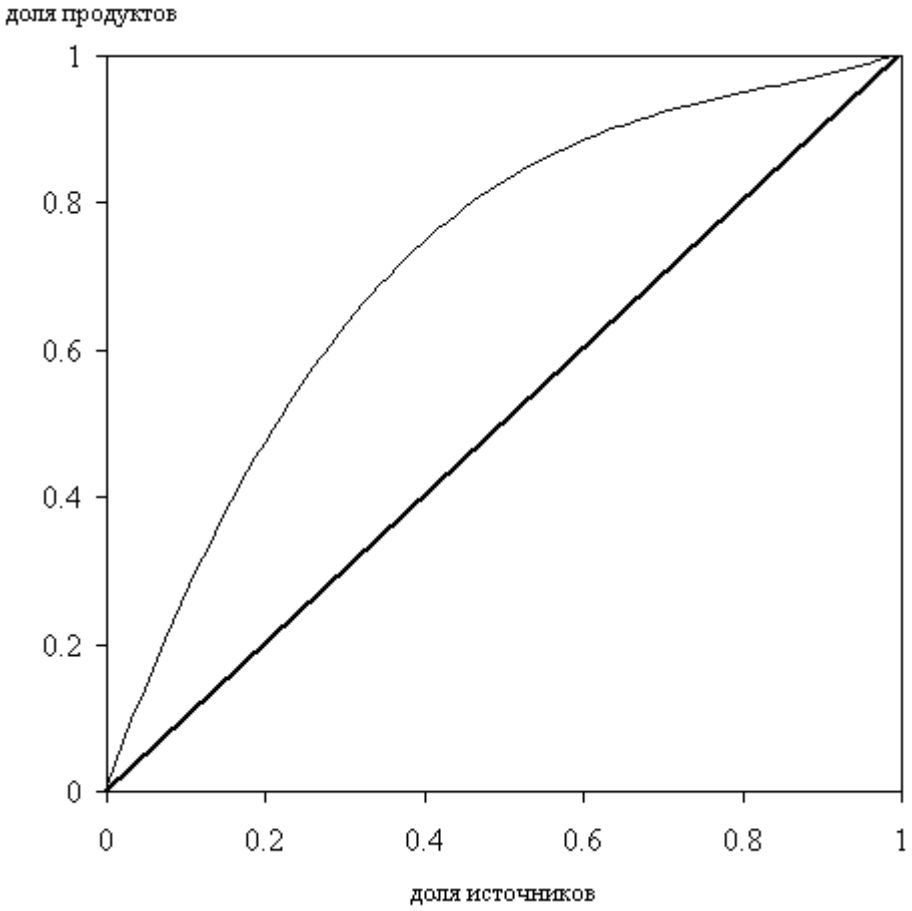

Рис. 4.1.1. Кривая Леймкулера

Закон (формула) Леймкулера определяет в аналитической форме вид кривой Леймкулера для различных информетрических процессов (в первую очередь, как обычно, для рассеяния статей по журналам). Закон гласит, что кривую Леймкулера можно приблизить формулой

$$y = \frac{\ln(1+\beta x)}{\ln(1+\beta)}, \qquad (4.2.1)$$

где x — доля наиболее продуктивных источников, y — доля продуктов, содержащихся в этих источниках, а $\beta > 0$ — эмпирический коэффициент, который необходимо найти при информетрическом моделировании. Как вид-

но из формулы (4.2.1), когда x пробегает значения от 0 до 1, y также монотонно возрастает от 0 до 1. Так как вторая производная

$$y'' = -\frac{\beta^2}{(1+\beta x)^2} \Big/ \ln(1+\beta) < 0,$$

то функция на всем отрезке выпуклая, следовательно, ее график лежит выше прямой $y = x$ («прямой абсолютно равномерного распределения»), как и должна вести себя кривая Леймкулера.

Отметим, что, исходя из (4.2.1), в абсолютных (не нормированных) линейных координатах библиограф также должен подчиняться логарифмическому закону

$$i = \mathrm{I}\frac{\ln(1+\beta' s)}{\ln(1+\beta' \mathrm{S})}, \qquad (4.2.2)$$

где s — число наиболее продуктивных источников, i — число содержащихся в них продуктов, S, в соответствии с обозначениями, принятыми в параграфе 2.1, — общее число источников, I — общее число продуктов, $\beta' = \beta/\mathrm{S}$ — эмпирический коэффициент.

Применим закон Леймкулера к информетрическому моделированию использования электронных информационных ресурсов в ГУ-ВШЭ. Для этого построим кривую Леймкулера для экспериментальных данных обращения к ресурсам и при помощи нелинейной регрессии подберем параметр β в формуле (4.2.1) так, чтобы аппроксимирующая кривая лучше всего совпадала с экспериментальной (как обычно, используется метод наименьших квадратов). Результат моделирования показан на рис. 4.1.2. Начальное значение коэффициента β выбрано равным двум, а ход итераций нелинейной регрессии суммирован в Приложении 3.

Мы видим, что приближение формулой Леймкулера получилось достаточно хорошим, что подтверждает и коэффициент детерминации (R

квадрат), который равен 0,937. Коэффициент β в данной модели получился равным 5850. При этом 95%-доверительный интервал составляют значения от 5430 до 6270.

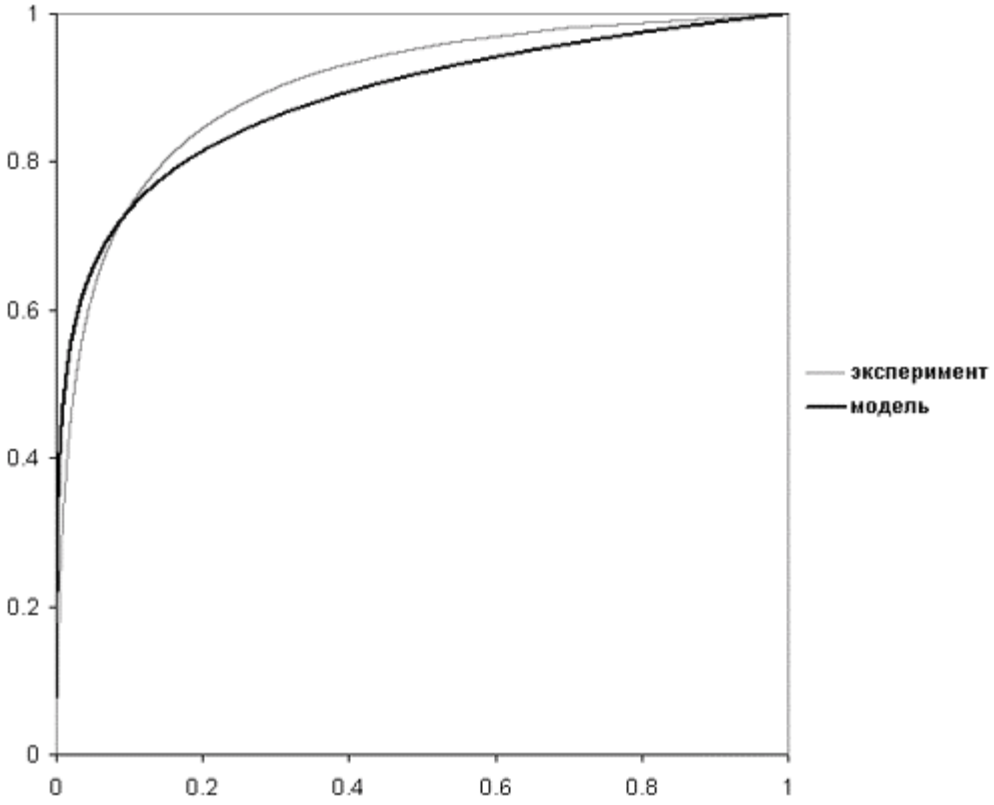

Рис. 4.1.2. Приближение экспериментальной кривой формулой Леймкулера

Так как кривая Леймкулера — это отраженная кривая Лоренца, то с ее помощью также можно рассчитать индекс Джини *G*. В этом случае формула (3.2.1) перейдет в следующую:

$$G = 2L - 1, \qquad (4.2.3)$$

где *L* — площадь под кривой Леймкулера. Вместо экспериментальной кривой (с ее помощью мы получили бы тот же результат, что и с кривой Лоренца) возьмем график, аппроксимированный формулой Леймкулера, и аналитически вычислим коэффициент Джини при заданном значении β. Использовав (4.2.1), получим:

$$L = \int_0^1 y(x)\mathrm{d}x = 2\int_0^1 \frac{\ln(1+\beta x)}{\ln(1+\beta)}\mathrm{d}x = \\ = \left(\frac{(x+1/\beta)\ln(1+\beta x) - x}{\ln(1+\beta)} \right)\Bigg|_0^1 = 1 + \frac{1}{\beta} - \frac{1}{\ln(1+\beta)} \tag{4.2.4}$$

При $\beta \to 0$ площадь L под аппроксимирующей кривой стремится к 0,5 ввиду того, что (раскладываем логарифм по Тейлору в окрестности единицы)

$$\frac{1}{\beta} - \frac{1}{\ln(1+\beta)} \approx \frac{1}{\beta} - \frac{1}{\beta - \beta^2/2} = -\frac{1}{2-\beta} \to -0{,}5\,.$$

Сама кривая при этом, очевидно, стремится к прямой $y = x$ (к равномерному распределению). При $\beta \to \infty$ площадь L стремится к единице, а кривая Леймкулера принимает форму уголка, образованного отрезками (0; 0) – (0; 1) и (0; 1) – (1; 1). Это наиболее неравномерное распределение (когда один источник содержит все продукты).

Подставим в (4.2.4) полученное при моделировании значение $\beta = 5850$ и получим, что площадь под аппроксимирующей кривой равна $L = 0{,}885$. Индекс Джини, согласно (4.2.3), будет равен $G = 0{,}770$. Это хорошо соотносится с полученным ранее точным (определенным численно в параграфе 3.3) значением G для экспериментальной кривой: $G = 0{,}804$. Полученное отличие, составляющее менее 5%, еще раз подтверждает хорошую применимость закона Леймкулера к информетрическому моделированию обращения к электронным источникам. Заметим также, что, несмотря на довольно широкий 95%-доверительный интервал у β ($\beta = 5850 \pm 420$), зависимость G от β при таких больших значениях коэффициента невелика ($G'_\beta(5850) \approx 2{,}2 \cdot 10^{-6}$), и если вычислить G при крайних значениях β из до-

верительного интервала, то получим $G = 0,770 \pm 0,002$, т. е. практически стабильный результат.

4.2. Закон Лотки

Закон Лотки был открыт А. Дж. Лоткой хронологически раньше закона Брэдфорда, в 1926 г. В классической работе [97] использовались два набора данных: публикации по химии, включенные в реферативный журнал «Chemical Abstracts» за десятилетие 1907 – 1916 гг. (только авторы, чьи фамилии начинались на A и B), и физические статьи, охваченные «Auerbach's Geschichtstafeln der Physik».

Закон Лотки заключается в том, что если взять множество ученых и посмотреть, сколько статей написал каждый из них, то число s_i ученых, написавших ровно *i* статей, будет обратно пропорционально квадрату *i*:

$$s_i = \frac{A}{i^2}. \qquad (4.2.5)$$

Например, если в некоем рассматриваемом множестве авторов научных трудов будет найдено 100 ученых, написавших только одну статью, то ученых, написавших две статьи, будет 25, три статьи — 11, четыре — 6 и т. д. Наконец, будет только один автор, который опубликовал 10 научных работ.

Для иллюстрации закона Лотки чаще всего строят график с двумя логарифмическими осями, где по оси абсцисс отложена продуктивность источников, а по оси ординат — число источников с такой продуктивностью. При идеальном выполнении закона график будет представлять собой прямую с тангенсом угла наклона, равным –2.

Так как из (4.2.5) очевидно, что $A = s_1$, то это равенство часто пишут в виде

$$s_i = s_1 / i^2 \,. \qquad (4.2.6)$$

В общем случае показатель степени в (4.2.5) может быть отличен от двойки (даже сам А. Лотка получил значения 1,89 и 2,02 для двух наборов своих данных), в этом случае обобщенный вид закона Лотки записывают так:

$$s_i = \frac{A}{i^{\alpha}} \,. \qquad (4.2.7)$$

Показатель степени α в законе Лотки — важная характеристика распределения «источники — продукты». В обзоре различных информетрических наборов данных С. Наранана [104] минимальное зафиксированное значение этого коэффициента $\alpha = 1,4$ (для распределения числа цитирований журнала American Journal of Human Genetics по цитирующим журналам, на материале [76]), максимальное — $\alpha = 2,8$ (для распределения статей по числу полученных за один год цитирований, на материале [112]). В аналогичном и более масштабном обзоре С. Д. Хайтуна [16, с. 252–299] найдено еще большее значение $\alpha = 3,9$ для распределения наиболее цитируемых ученых по числу ссылок на них (данные из [74]; здесь мы ограничиваемся только теми информетрическими наборами, представленными С. Д. Хайтуном, где можно чётко выделить пары «источники — продукты»).

Обратим внимание на то, что при строгом выполнении закона Лотки (с показателем степени 2) можно получить оценку снизу для доли источников с продуктивностью, равной единице (т. е. доли авторов, написавших только одну статью, или, в нашем случае, доли журналов, из которых была прочитана только одна публикация). Суммарное число источников S равно

$$S = \sum_1^{i_{\max}} s_i = \sum_1^{i_{\max}} \frac{s_1}{i^2} = s_1 \sum_1^{i_{\max}} \frac{1}{i^2} \,. \qquad (4.2.8)$$

Так как $\sum_{1}^{i_{\max}} \frac{1}{i^2} < \sum_{1}^{\infty} \frac{1}{i^2}$, а последняя бесконечная сумма, в свою очередь, равна $\frac{\pi^2}{6}$, то мы получим оценку для s_1:

$$s_1 > \frac{6}{\pi^2} S$$

или для доли источников с минимальной продуктивностью

$$\frac{s_1}{S} > \frac{6}{\pi^2} \approx 0,61. \qquad (4.2.9)$$

Можно получить аналогичное выражение для ожидаемой доли источников с минимальной продуктивностью в случае обобщенного закона Лотки в виде

$$S = \sum_{1}^{i_{\max}} s_i = \sum_{1}^{i_{\max}} \frac{s_1}{i^\alpha} = s_1 \sum_{1}^{i_{\max}} \frac{1}{i^\alpha},$$

то есть

$$\frac{s_1}{S} = \frac{1}{\sum_{1}^{i_{\max}} \frac{1}{i^\alpha}} > \frac{1}{\sum_{1}^{\infty} \frac{1}{i^\alpha}}.$$

Так как $\sum_{1}^{\infty} \frac{1}{i^\alpha} = \varsigma(\alpha)$, где ς — дзета-функция Римана, то при достаточно больших $i_{\max}$ получаем оценку для доли наименее продуктивных источников:

$$\frac{s_1}{S} \approx \frac{1}{\varsigma(\alpha)}. \qquad (4.2.10)$$

Закон Лотки неоднократно применялся к совершенно различным информетрическим объектам, см., например, [16]. Библиографию можно найти как в указанной работе С. Д. Хайтуна, так и в публикации [128]. Отметим, что о важности закона Лотки среди других информетрических законов

свидетельствует появление нового термина «лоткаианская информетрия» [63].

Одной из первых работ, применивших закон Лотки к выдаче периодики, является статья С. Наранана [104]. На материале полученных ранее в [79] данных по месячной статистике обращений к журналам в медицинской библиотеке Allen Memorial, С. Наранан строит в двойных логарифмических координатах стандартный график продуктивность/число источников с такой продуктивностью (в данном случае: число выдач/число журналов с таким количеством выдач), получает хорошую линейную зависимость и подтверждает выполнение закона Лотки для спроса на периодику.

Но не всегда исследования спроса на литературу идеально подтверждали закон Лотки. В работе [36] проводится анализ семилетней статистики книговыдач в Хиллмановской библиотеке Питтсбургского университета и строится график в полулогарифмических координатах, показывающий, что это распределение скорее удовлетворяет логнормальному закону (т. е. логарифм плотности вероятности распределен по Гауссу), чем закону Лотки.

Применим закон Лотки к статистике обращений к электронным информационным ресурсам в ГУ-ВШЭ. Для этого построим график, на котором по оси абсцисс отложим продуктивность источников, т. е. число открытых статей из журналов, а по ординате — количество источников с соответствующей продуктивностью, т. е. число журналов с ровно таким числом открытых из них статей. Ровно одна статья была открыта из 778 журналов, ровно две статьи — из 416 и т. д. (см. Приложение 1). При этом для исследования степенной зависимости сделаем обе оси логарифмическими. Результат вместе с регрессионной прямой показан на рис. 4.2.1.

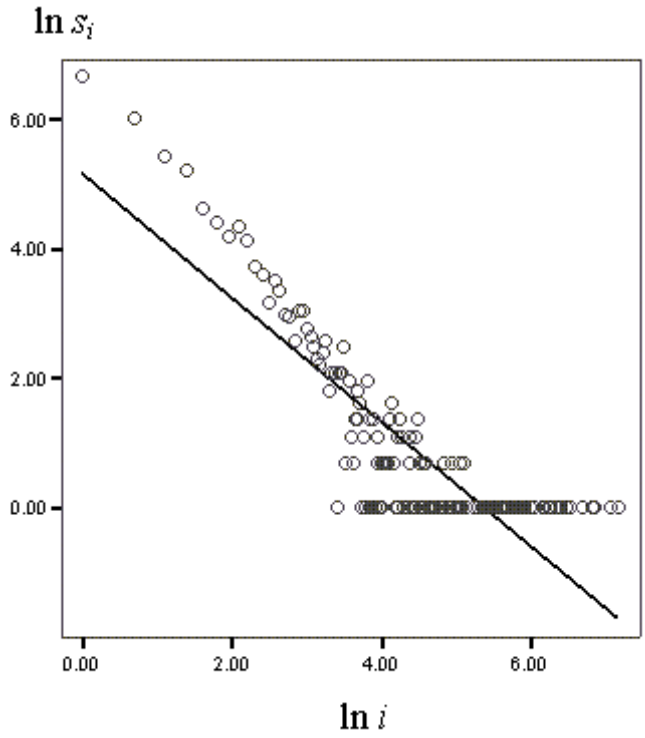

Рис. 4.2.1. Приближение экспериментальных данных законом Лотки (оценка: $\alpha \approx 0,96$)

Мы видим, что оценка при помощи линейной регрессии дает значение показателя степени в (4.2.7) приближенно равное 0,96. Это крайне малое значение, не наблюдаемое в большинстве исследований, и можно было бы говорить о плохой применимости закона Лотки в нашем случае. Кроме того, коэффициент детерминации R-квадрат невелик, 0,74, что говорит о слабом приближении зависимости регрессионной прямой.

Однако заметим, что серьезные помехи в аппроксимацию вносят точки с большой продуктивностью. Журналов, из которых открыто фиксированное, причем большое, число статей, всегда будет немного, $1-2$, и поэтому на графике эти точки вытянулись по ординатам $\ln 1 = 0$ и $\ln 2$. Причем этих точек много, более 100, в связи с чем они оказывают существенное влияние на построенную регрессионную прямую.

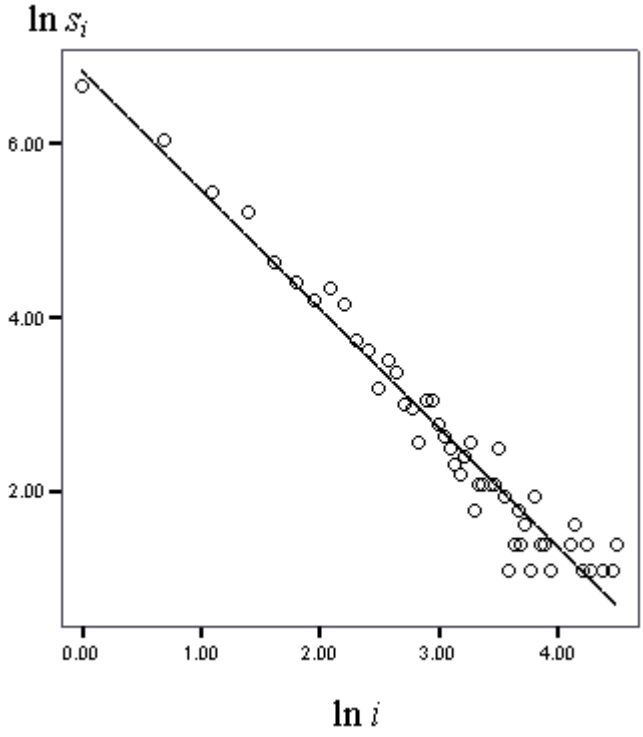

Рис. 4.2.2. Приближение экспериментальных данных законом Лотки после отсеивания источников с высокой продуктивностью (оценка: $\alpha \approx 1{,}37$)

То, что при информетрическом моделировании с использованием закона Лотки негативную роль играют источники с высокой продуктивностью, плохо ложась на модель, отмечается во многих исследованиях. Для того чтобы избежать такого влияния, можно поступить подобно С. Наранану [104], при аппроксимации эмпирических данных из [79] отбросившему несколько точек с максимальной продуктивностью. Уберем из нашего набора данных точки, с такой продуктивностью, что источников, ей соответствующих, только 1 или 2. Тогда останется 51 точка из 155, результат показан на рис. 4.2.2.

В новом варианте аппроксимации прежде всего отметим высокий коэффициент детерминации: R-квадрат равен 0,96, что говорит об очень хорошем приближении данных регрессионной прямой. Визуально сравнив рис. 4.2.1 и рис. 4.2.2, обратим внимание на более «равномерный» учет то-

чек всех продуктивностей в новой регрессии (на рис. 4.2.1, до отсечения точек с минимальными ординатами, регрессию «забивали» именно эти точки).

Также и значение коэффициента α из (4.2.7) получается более привычным для информетрии, 1,37. Хотя наклон регрессионной прямой все равно мал для выполнения «классического» закона Лотки, но он уже близок к значению, обнаруженному в других исследованиях спроса на библиотечные материалы (С. Наранан получает в [104] $\alpha = 1{,}5$). Модель Лотки можно считать вполне приемлемой для наших экспериментальных данных.

Дальнейшее усечение данных, например отсев точек с $s_i = 3$ и $s_i = 4$, уже не приведет к существенному изменению результатов: показатель α изменится менее чем на один процент.

Однако примененное простое исключение из рассмотрения ряда источников с высокой продуктивностью все-таки содержит в себе изъян. Мы никак не учитываем, что, например, в области средней продуктивности источники с некоторым заданным *i* встречаются чаще, чем в зоне максимальной продуктивности (меньше значений *i* таких, что $s_i = 0$). Попробуем учесть и этот фактор, для чего проведем корректировку данных.

Пусть у нас есть $s_i \neq 0$ и $s_j \neq 0$, $i > j$, причем $\forall k$, $j < k < i$, выполняется $s_k = 0$ (т. е. источников с продуктивностью больше *j* и меньше *i* нет). Перенесем на графике точки с координатами $(i; s_i)$ в точки $\left(\frac{i+j}{2} + 0{,}5;\ \frac{s_i}{i-j} \right)$. Теперь:

— останутся на своих местах точки, соответствующие «плотному» фрагменту графика (область низкой продуктивности, где для каждого значения *i* имеется источник с такой продуктивностью; имен-

но для выполнения данного условия в абсциссу добавлен коэффициент 0,5);

— точки, соответствующие источникам, у которых нет «соседей слева» по продуктивности (т. е. нет источников с продуктивностью на единицу меньше) переместятся по абсциссе примерно на половину интервала до следующей продуктивности, которой соответствует хотя бы один источник;

— ординаты этих точек станут тем меньше, чем больше отсутствует «соседей слева по продуктивности».

Фактически, мы «размазываем» число журналов, соответствующих продуктивности *i*, по всему интервалу влево до следующего значения продуктивности, которой соответствует хотя бы один источник. Это проиллюстрировано на рис. 4.2.3. Здесь имеется ровно один источник с продуктивностью 15 и отсутствуют источники с продуктивностями от 11 до 14. При этом ордината источника с продуктивностью 15 изменяется с 1 на 0,2 (значение 1 «делится» между продуктивностью 15 и четырьмя пустыми продуктивностями), а абсцисса сдвигается, чтобы поместить точку приблизительно в середину пустого интервала.

Например, в случае наших данных имеется один источник с продуктивностью 1306, следующая непустая продуктивность — 1166. До предлагаемой корректировки мы имеем нулевое число источников в интервале продуктивностей [1167; 1305], затем один источник в точке с продуктивностью 1306. Корректировка заменит эту ситуацию на $1/140 = 0,0071$ источник с продуктивностью 1236,5. Напомним, что какие-либо перемещения касаются только источников с высокой продуктивностью: в нашем случае первая точка, которая совершит перемещение, соответствует продуктивности 58, т. к. отсутствует источник с продуктивностью 57. Все точки, соот-

ветствующие продуктивностям источников с 1 до 56, остаются на своем прежнем месте.

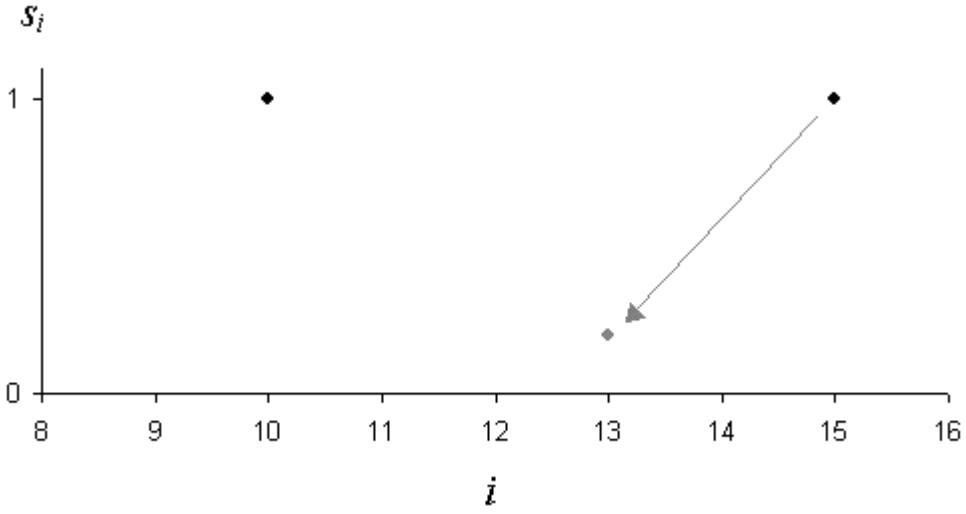

Рис. 4.2.3. Схема преобразования данных для оптимизированного приближения эксперимента законом Лотки

Полученный набор скорректированных экспериментальных данных вновь изобразим в двойных логарифмических координатах и приблизим регрессионной прямой (рис. 4.2.4). Прежде всего мы видим, что облако в зоне высокой продуктивности ведет себя в целом точно так же, как и в области низкой. Это очень важный результат: нашей корректировкой удалось «продлить» облако в область отрицательных ординат (мы ввели «дробную продуктивность», логарифм которой меньше нуля) таким образом, что это оказалось логичным продолжением исходных (не затронутых корректировкой) экспериментальных данных. Это является подтверждением обоснованности используемого преобразования.

Интересен и результат аппроксимации. Угловой коэффициент регрессионной прямой соответствует коэффициенту α в (4.2.7), равному 1,69. Это уже близко к классическому показателю $\alpha = 2$. Корректирующей подготовкой данных нам удалось продолжить облако точек ниже оси абсцисс,

избавиться от «рядов», в которые выстраивались точки с малым s_i на рис. 4.2.2 и особенно на рис. 4.2.1 (тем самым они «приподнимали» регрессионную кривую), но при этом оставить эти точки на графике, использовав их значения для аппроксимации. Можно уверенно говорить, что после необходимой корректировки данных в области точек с высокой продуктивностью модель Лотки хорошо описывает исследуемый нами процесс обращения к электронным информационным ресурсам. Коэффициент детерминации (R-квадрат) равен 0,92, что также свидетельствует о хорошей аппроксимации (это немногим менее полученного ранее значения 0,96, притом что мы имеем в 3 раза больше точек, чем на рис. 4.2.2).

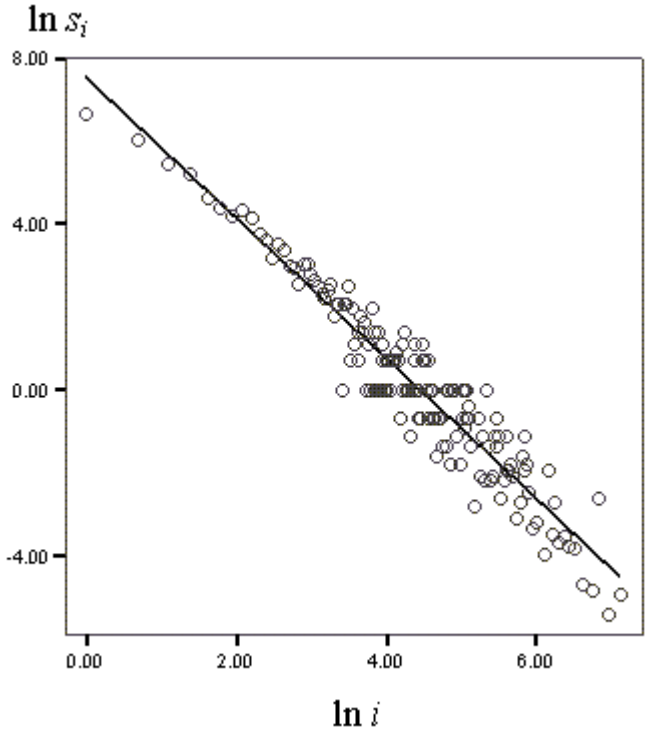

Рис. 4.2.4. Приближение экспериментальных данных законом Лотки после корректировки показателей источников с высокой продуктивностью (оценка: $\alpha \approx 1,69$)

С теоретической точки зрения проведенное нами преобразование данных является приближенной заменой дискретного распределения s_i непре-

рывной функцией плотности вероятности появления источника с заданной продуктивностью. Сопоставление двух форм закона Лотки проводит Л. Эгге [63, p. 378–386; 64], приходя к выводу, что показатель степени в дискретном распределении (4.2.7) можно корректно оценить из непрерывной формы закона (при небольших α).

Заметим, что корректировка данных проводилась нами также иными способами: без смещения +0,5 по абсциссе; с размещением строго в середине между двумя непустыми продуктивностями значения, равного среднему арифметическому их ординат, отнесенного к расстоянию между ними, т. е. $(i;\ s_i) \to \left(\frac{i+j}{2};\ \frac{s_i + s_j}{2(i-j)} \right)$. Показатель степени α при этом оставался практически неизменным.

Наконец, обратимся к оценке доли числа источников с минимальной продуктивностью, основанной на формулах (4.2.9) и (4.2.10). В нашем случае $s_1 = 778$ и $S = 2590$, т. е. $\frac{s_1}{S} = 0{,}30$, что более чем в 2 раза меньше, чем ограничение для «классического» закона Лотки (4.2.9). Это неудивительно, т. к. показатель степени α у нас, как мы видели, отличен от 2. Если применить обобщенное выражение (4.2.10), подставив в него найденное регрессионным приближением после проведенной корректировки экспериментальных данных $\alpha = 1{,}69$, то получим оценку $\frac{s_1}{S} > 0{,}48$, что также велико относительно эксперимента.

Используем (4.2.10) в обратную сторону, для оценки показателя распределения α. В нашем случае $\frac{s_1}{S} = 0{,}30$, чему соответствует $\alpha = 1{,}34$. Этот результат находится в почти точном согласии с регрессионной оценкой α, полученной путем учета всех экспериментальных точек, за исклю-

чением продуктивностей, которым соответствуют только 1 или 2 источника (рис. 4.2.2). Там было получено $\alpha = 1{,}37$.

Последний вывод, безусловно, имеет ценность, однако его не следует считать свидетельством ущербности предложенного метода корректировки данных перед аппроксимацией моделью Лотки (рис. 4.2.4). Число источников с минимальной продуктивностью s_1 в эмпирических данных может флуктуировать. Это следует учитывать и не полагать, что такие оценки обязаны выполняться со всей строгостью. Также не стоит их считать основным «проверочным условием» выполнения закона Лотки и правильности найденного показателя степени α.

Можно сказать, что метод простого усечения (рис. 4.2.2) хорошо решает частную задачу — согласование модели с числом наименее продуктивных источников. В то же время более сложный подход, продемонстрированный на рис. 4.2.4, учитывает все экспериментальные данные и дает более содержательную оценку α, характеризующую экспериментальный информетрический массив в его целостности.

4.3. Закон Ципфа

Закон Ципфа, в отличие от других информетрических законов, которых касается настоящая диссертационная работа, сначала был открыт в лингвистике. Первенство открытия не так просто установить, чаще всего в контексте истории закона Ципфа цитируют работы [70; 51; 134]. В дальнейшем он применялся к разнообразным распределениям, в том числе не имеющим отношения к информетрии (например, к распределению производства цемента по союзным республикам СССР или бакалейных фирм Великобритании по числу служащих [16, p. 283, 285]).

Суть лингвистического закона Ципфа можно сформулировать так: если в некотором корпусе текстов выстроить слова в порядке частоты их встречаемости, приписав каждому слову «ранг» (порядковый номер в этом упорядоченном от наиболее используемых к наименее используемым словам списке), то произведение ранга слова на частоту его употребления приближенно будет константой.

Если обозначить ранг r, а количество употреблений слова с рангом r — f_r, то закон Ципфа гласит, что

$$r \times f_r = const . \qquad (4.3.1)$$

Так же как и в случае закона Лотки (см. параграф 4.2), имеется «классический» вариант закона Ципфа вида (4.3.1), но есть также и обобщенный его вариант, когда степенная зависимость f_r от r не является простой гиперболой, а выражается формулой

$$f_r = \frac{const}{r^{\delta}}, \qquad (4.3.2)$$

где δ — некоторое число, как правило близкое к единице (мы вводим обозначение показателя степени δ, что несколько необычно для литературы по закону Ципфа, для того чтобы растождествить его с показателем степени α из закона Лотки вида (4.2.7)). Еще более общая форма известна под именем закона Ципфа-Мандельброта:

$$f_r = \frac{const}{\left(r + a\right)^{\delta}}, \qquad (4.3.3)$$

где δ опять близко к единице.

Применим закон Ципфа к моделированию процесса обращения к электронным источникам при помощи следующей замены пары «источники — продукты». Вместо ранга слова будет фигурировать порядковый номер журнала в списке востребованных изданий, выстроенных в порядке убыва-

ния числа обращений к ним. Вместо частоты употребления слов — количество открытых полнотекстовых статей.

Построим в двойных логарифмических координатах зависимость $f(r)$, т. е. в нашем случае — зависимость числа открытых статей от ранга журнала. Результат показан на рис. 4.3.1. Тангенс наклона регрессионной кривой равен $-1{,}35$, таким образом в (4.3.2) коэффициент δ для нашего случая равен 1,35. Коэффициент детерминации R-квадрат равен 0,96 (очень хорошее приближение).

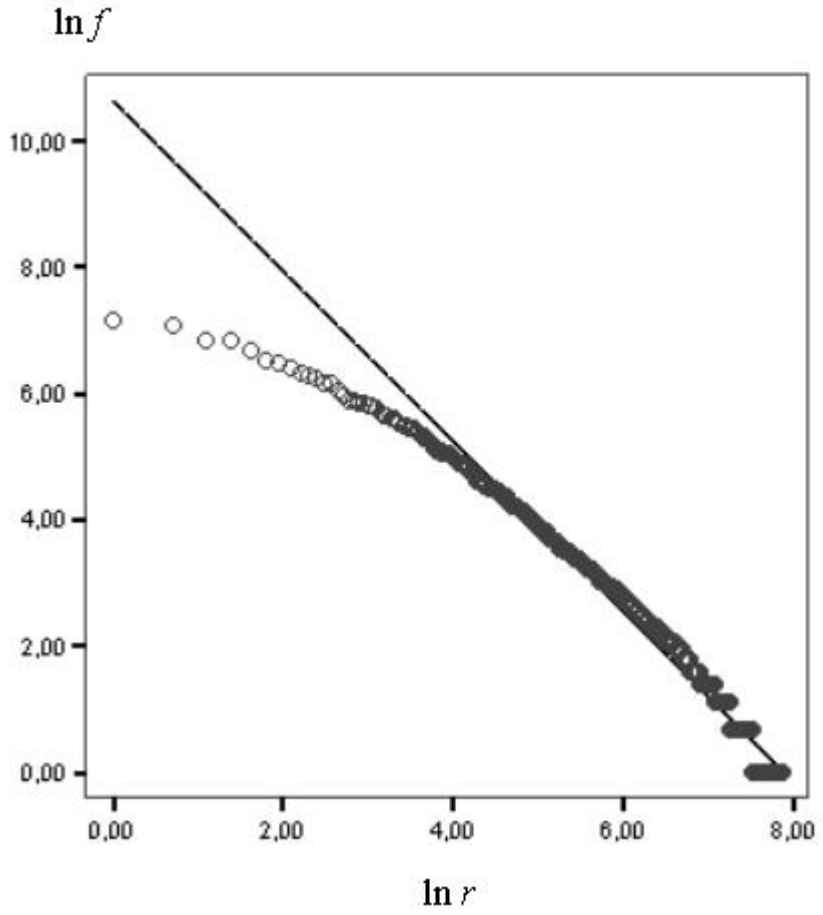

Рис. 4.3.1. Приближение экспериментальных данных законом Ципфа (оценка: $\delta \approx 1{,}35$)

Однако обратим внимание на краевой эффект на рис. 4.3.1: в области высоких рангов точки начинают выстраиваться в линии, параллельные оси абсцисс. Действительно, число журналов с одной или двумя открытыми

статьями велико и они имеют одинаковую ординату. Более того, если строго подходить к вопросу, все эти точки должны иметь одинаковый ранг: они абсолютно равнозначны и внутри каждого такого множества порядок сортировки источников с одинаковой продуктивностью абсолютно произвольный.

Здесь мы столкнулись со стандартной проблемой аппроксимации графика, отражающего информетрическое моделирование с применением закона Ципфа. В отличие от закона Лотки, который с трудом применим без корректировки данных о журналах с большой продуктивностью, закон Ципфа имеет недостатки в области источников с малой продуктивностью (область высоких рангов). График имеет ступенчатый вид, причем внутри каждой ступеньки сортировка журналов произвольная. Помимо сложностей теоретического обоснования такой аппроксимации, недостатком является то, что на регрессию решающее влияние будут оказывать как раз точки из области высоких рангов (в силу большого своего количества). Это видно и из рис. 4.3.1: регрессионная прямая проходит близко к точкам, соответствующим источникам с высоким рангом и сильно удалена от наиболее продуктивных источников. Получается, что первая десятка журналов, пользующихся максимальным спросом, практически не влияет на аппроксимацию (и, соответственно, на получение степенного коэффициента распределения Ципфа).

Здесь, как и в случае с законом Лотки, необходима корректировка данных (как было сказано выше, не в области много-, а в области малопродуктивных источников). Проведем следующую замену: источники, имеющие одинаковую продуктивность, отразим одной точкой с соответствующей продуктивностью и рангом, равным среднему рангу всех этих источников. Иными словами, пусть имеются ранги r_1 и r_2, $r_2 > r_1$, при этом $\forall r$, $r_1 \le r \le r_2$, выполняется $f(r) = f(r_1) = f(r_2)$, но $f(r_1 - 1) > f(r_1)$ и

$f(r_2+1) < f(r_2)$ (т. е. отрезок $[r_1;\ r_2]$ соответствует одной «ступеньке» графика, подобного рис. 4.3.1). Тогда заменим все точки $(r;\ f(r))$, $r_1 \le r \le r_2$, одной единственной точкой $\left(\frac{r_1+r_2}{2};\ f(r_1)\right)$.

Так, в нашем случае имеется, например, 778 журналов, из которых была открыта всего одна статья, эти источники имеют ранги с 1813 по 2590. Заменим их (множество точек с ординатой равной $\ln 1 = 0$ на рис. 4.3.1) одной точкой с координатами $(2201{,}5;\ 1)$ (в обычных координатах, в логарифмических добавится натуральный логарифм). Проведя аналогичные замены для всех источников с одинаковой продуктивностью, построим новый график *f*(*r*) в двойных логарифмических координатах (рис. 4.3.2).

Прежде всего мы видим, что новая регрессионная прямая одинаково учитывает точки как высоких, так и низких рангов. Так как новый график имеет выпуклый вид, то аппроксимирующая прямая, проходя ближе всего к точкам среднего ранга, дает примерно средний наклон касательных к графику (как видно, в области низких рангов касательная более пологая, в области высоких — более крутая): «усредняющая» роль регрессии выполняется лучше, чем на исходном рис. 4.3.1. Коэффициент детерминации (R-квадрат) при этом не изменился, он вновь равен 0,96: очень хорошее приближение регрессией.

Наклон регрессионной прямой на этот раз меньше (что ожидаемо: возросла роль низких рангов, где касательная к графику имеет меньший наклон). Угловой коэффициент прямой равен $-0{,}97$, следовательно степенной показатель δ в (4.3.2) для построенной модели равен 0,97, и это означает, что $f(r)$ обратно пропорциональна $r^{0{,}97}$. Это практически точное

совпадение с классическим видом закона Ципфа (в формулировке которого этот показатель равен единице).

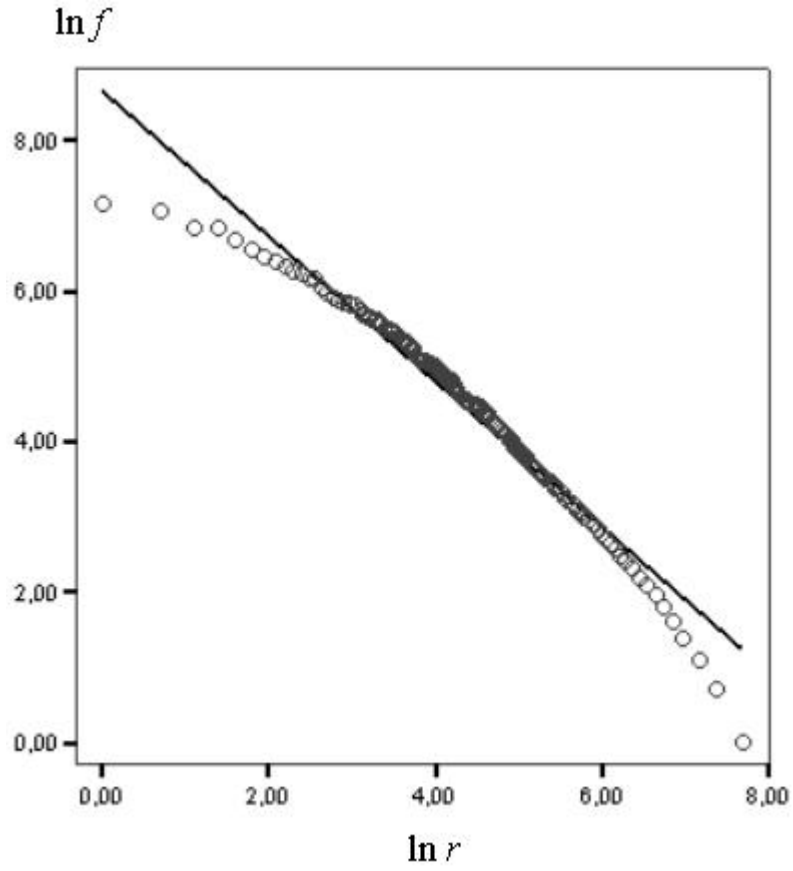

Рис. 4.3.2. Приближение экспериментальных данных законом Ципфа после корректировки ранга источников с низкой продуктивностью (оценка: $\delta \approx 0{,}97$**)**

Можно говорить, что для наших экспериментальных данных информетрическая модель Ципфа применима в ее классическом варианте. Особенно хорошо закон выполняется, как можно видеть из графика рис. 4.3.2, в середине распределения, в области источников со средними рангами.

Заключение

В настоящей диссертационной работе рассмотрены и исследованы на практике основные информетрические законы — Брэдфорда, Парето, Леймкулера, Лотки, Ципфа. Они применяются для математического моделирования процесса обращения к электронным информационным ресурсам — экспериментального массива, полученного из статистики использования онлайновых баз данных в Государственном университете – Высшей школе экономики.

В связи с тем, что исследуемые информетрические законы исходно были открыты для информационных процессов иного рода, их применение к использованию информации требовало выработки внятных методологических принципов. Эти принципы были взяты из концепции «информационного процесса производства»: производилась замена пар «источники — продукты» на те источники и продукты, которые фигурируют в изучаемом нами процессе. Например, в случае закона Брэдфорда пара «журналы — статьи, опубликованные в журналах» заменялась на «(электронные) журналы — статьи, открытые из журналов». Аналогичные переходы осуществлены для других законов. Также изменились и величины, которые законы связывают между собой: например, в случае закона Лотки вместо переменных «число статей — число авторов, написавших ровно столько статей» исследовалась зависимость «число открытых статей — число журналов, из которых открыто ровно столько статей». Это составило методику применения информетрических законов к моделированию использования информации.

Одним из ключевых направлений исследования было изучение применимости информетрических моделей к процессу обращения к электронным документам. Получены следующие результаты:

— Модель Брэдфорда в ее вербальной формулировке применима недостаточно удовлетворительно: только при разбиении на большое число зон Брэдфорда в середине распределения имеется участок, где закон выполняется хорошо.

— Закон Брэдфорда в графической форме выполняется хорошо: построенная кривая Брэдфорда («библиограф») имеет предсказанный законом вид. Прямолинейный участок, однако, не столь ярко выражен, отсюда, вероятно, плохое выполнение вербальной формы закона.

— Принцип Парето 80/20 достаточно хорошо выполняется для каждого из изучаемых электронных ресурсов в отдельности. При объединении ресурсов в единый информационный массив выполнение закона становится еще более очевидным и принцип 80/20 переходит в более «концентрированный» вид 83/17.

— Модель Леймкулера хорошо приближает экспериментальные данные, построенная на их основе кривая Леймкулера близка к предсказанной, при использовании оптимальных коэффициентов модели коэффициент детерминации достигает 0,94 (очень хорошее приближение).

— Модель Лотки также хорошо выполняется для исследуемого процесса, однако требует специальной корректировки данных, без которой хорошо известная проблема аппроксимации данных в области источников с высокой продуктивностью препятствует применимости закона (степенной коэффициент Лотки оказывается очень малым, 0,96). Предложенная в настоящей работе корректировка

данных позволяет, не отбрасывая при аппроксимации ни одного показателя высокопродуктивных источников, получить осмысленный массив, готовый к аппроксимации, при этом коэффициент Лотки становится равным 1,69, что свидетельствует об удовлетворительной применимости закона.

— Закон Ципфа также выполняется, для данных в исходной форме получается приемлемый коэффициент распределения 1,34. Однако здесь также предложено видоизменение данных, увеличивающее их осмысленность и корректность приближения регрессионной прямой. В результате оказалось, что закон Ципфа для исследуемого процесса выполняется очень хорошо, причем в своей классической форме: степенной коэффициент очень близок к единице (0,97).

Итак, резюмируя полученные результаты по всем моделям, можно сказать, что в целом информетрические законы хорошо применимы для процесса обращения к электронной информации.

Еще одной центральной темой, рассмотренной в настоящей диссертационной работе, был вопрос об определении «ядра спроса», т. е. наиболее важных информационных источников, играющих ключевую роль в информационном обеспечении научного и образовательного процессов в вузе. Выделить такое ядро удалось при помощи двух различных информетрических моделей — с использованием закона Брэдфорда и индекса Джини. Для обеих моделей определялись более «жесткий» (компактный) и «мягкий» варианты ядра. Важным полученным результатом стало совпадение полученных размеров ядра по порядку величины вне зависимости от используемой модели: в случае «мягкого» варианта это 77 (Брэдфорд) и 89 (Джини) журналов, в случае более «жесткого» — 28 и 37 изданий. Для самого компактного варианта ядра исследована роль каждого электронного ресурса в его формировании и установлено, что имеется важное разделе-

ние: некоторые ресурсы содержат большое число ядерных журналов, но при этом они «неэксклюзивны» в предоставлении доступа к этим изданиям, другие вносят мало журналов в ядро, но эти журналы имеются только в данном конкретном ресурсе и более нигде.

Фактически, все информетрические законы говорят о большой концентрации, наблюдаемой в «информационных процессах производства». И раз эти законы применимы к использованию электронных ресурсов, значит и здесь наблюдается высокая концентрация спроса. Однако рекомендации, которые следуют из полученных в настоящем исследовании результатов, должны быть «позитивными», а не «негативными»: следует наращивать представленность в фонде и объем доступа к ядерным журналам и изданиям, пользующимся высоким спросом, а не, например, отказываться от закупки всех журналов, не вошедших в ядро или в 20% спроса.

Именно в таком ключе результаты настоящего диссертационного исследования применяются в концепции формирования электронной подписки ГУ-ВШЭ. Например, было замечено, что три журнала Американской экономической ассоциации, вошедшие в самое компактное ядро спроса, представлены в электронном виде в ГУ-ВШЭ с двухгодичным «эмбарго» (искусственной задержкой между выходом печатной версии и предоставлением электронной копии, установленной издательством для баз данных посредников). Эта ситуация была исправлена дополнительной подпиской на эти журналы без эмбарго — напрямую у издательства.

По результатам настоящей работы можно обрисовать дальнейшие направления исследований. Прежде всего, требуется провести аналогичный анализ на базе статистики использования онлайн-ресурсов в других вузах и организациях иного типа (исследовательских институтах, публичных библиотеках и т. д.), чтобы набрать значительный массив информетрических исследований и данных о применимости информетрического моделирова-

ния к процессу обращения к электронным информационным ресурсам. Особенно интересно сравнить данные по отдельным организациям с результатами, полученными на уровне крупных консорциумов, объединяющих большое число институциональных подписчиков. Далее, к перспективным направлениям можно отнести математический и эмпирический анализ двух раздельных ступеней процесса использования электронной информации: ее производства и последующего обращения к ней. В настоящей работе эти стадии не разделялись, исследовался процесс в его целостности, однако такое более детальное изучение может пролить свет на законы концентрации и рассеяния спроса на электронные источники и внести вклад в дальнейшее развитие онлайн-информетрии.

Список литературы

1. *Горькова В. И.* Информетрия (количественные методы в научно-технической информации) // Итоги науки и техники. Сер. Информатика. Т. 10. М.: ВИНИТИ, 1988. 328 с.
2. *Гохберг Л. М.* Статистика науки. М.: ТЕИС, 2003. 478 с.
3. *Зибарева И. В., Писляков В. В., Теплова Т. Н., Нефёдов О. М.* Библиометрический анализ журнала «Успехи химии» // Вестник Российской академии наук. 2008. Т. 78, № 6. С. 490–499.
4. Индикаторы инновационной деятельности: 2007. Статистический сборник. М.: ГУ-ВШЭ, 2007. 398 с.
5. Индикаторы науки: 2008. Статистический сборник / Н. В. Городникова, С. Ю. Гостева, В. В. Писляков и др. Под ред. Л. М. Гохберга, А. Л. Кевеша, Я. И. Кузьминова и др. М.: ГУ-ВШЭ, 2008. 336 с.
6. Индикаторы образования: 2007. Статистический сборник. М.: ГУ-ВШЭ, 2007. 174 с.
7. *Писляков В. В.* Анализ контента ведущих электронных ресурсов актуальной зарубежной периодики: Препринт WP2/2002/02. М.: ГУ ВШЭ, 2002. 32 с.
8. *Писляков В. В.* Спрос на электронные журналы в университетской библиотеке: работает ли правило Парето? // Научно-техническая информация. Сер. 1. 2005. № 12. С. 27–32.
9. *Писляков В. В.* Использование онлайновых ресурсов и управление электронной подпиской в библиотеке ГУ-ВШЭ // Университетское управление: практика и анализ. 2006. № 4 (44). С. 47–56.
10. *Писляков В. В.* Использование онлайновых ресурсов и управление электронной подпиской в библиотеке ГУ-ВШЭ // Университетские ин-

новации: опыт Высшей школы экономики / Под ред. Я. И. Кузьминова. М.: ГУ-ВШЭ, 2006. С. 160–178.

11. *Писляков В. В.* Зачем создавать национальные индексы цитирования? // Научные и технические библиотеки. 2007. № 2. С. 65–71.
12. *Писляков В. В.* Методы оценки научного знания по показателям цитирования // Социологический журнал. 2007. № 1. С. 128–140.
13. *Писляков В. В.* Моделирование процесса обращения к электронным информационным источникам на основе информетрического закона Брэдфорда // Ученые записки Казанского государственного университета. Серия Физико-математические науки. 2007. Т. 149, кн. 2. С. 116–127.
14. *Писляков В. В.* Российский индекс научного цитирования: pro et contra // НТИ-2007. Материалы конференции. М.: ВИНИТИ, 2007. С. 243–244.
15. *Писляков В. В., Любушко Е. Э.* Анализ научно-информационной деятельности (чтение, публикации, цитирование) ученых института катализа им. Г. К. Борескова СО РАН // Катализ в промышленности. 2007. № 3. С.55–63.
16. *Хайтун С. Д.* Наукометрия: состояние и перспективы. М.: Наука, 1983. 344 с.
17. *Яблонский А. И.* Модели и методы исследования науки. М.: Эдиториал УРСС, 2001. 400 с.
18. *Bador P., Ben Romdhane M., Lafouge T.* European Pharmaceutical Journals: Relationship between Demand and Indexation — The Example of the Main French Document Supplier // Canadian Journal of Information and Library Science. 2002/03. Vol. 27, No. 2. P. 33–55.
19. *Bagust A.* A Circulation Model for Busy Public Libraries // Journal of Documentation. 1983. Vol. 39, No. 1. P. 24–37.

20. *Bates M. J.* Speculations on browsing, directed searching, and linking in relation to the Bradford distribution // Emerging Frameworks and Methods: Proceedings of the 4th International Conference on Conceptions of Library and Information Science (CoLIS4), Seattle, WA, USA, July 21–25, 2002 / Ed. by *H. Bruce e. a.* Greenwood Village, Colo.: Libraries Unlimited, 2002. P. 137–149.
21. *Bernal J. D.* Preliminary analysis of pilot questionnaire on the use of scientific literature // The Royal Society Scientific Information Conference, 21 June – 2 July 1948: Report and Papers Submitted. London: Royal Society, 1948. P. 589–637.
22. *Blecic D. D.* Monograph use at an academic health sciences library: The first three years of shelf life // Bulletin of the Medical Library Association. 2000. Vol. 88, No. 2. P. 145–151.
23. *Boig F. S., Loftman K. A.* Domestic and Foreign Periodicals in the Field of Petroleum Chemistry: A Statistical Analysis // Oil and Gas Journal. 1949 (21 April). Vol. 47. P. 199–208.
24. *Bollen J., Van de Sompel H., Smith J. A., Luce R.* Toward alternative metrics of journal impact: A comparison of download and citation data // Information Processing and Management. 2005. Vol. 41, Iss. 6. P. 1419–1440.
25. *Bonitz M., Schmidt P.* Transition from the macrolevel to the microlevel of information at rank distribution investigations of the report literature of an international information system // Scientometrics. 1982. Vol. 4, No. 4. P. 283–295.
26. *Bookstein A.* Informetric distributions, part II: Resilience to ambiguity // Journal of the American Society for Information Science. 1990. Vol. 41, Iss. 5. P. 376–386.

27. *Bookstein A.* Informetric distributions. III. Ambiguity and randomness // Journal of the American Society for Information Science. 1997. Vol. 48, Iss. 1. P. 2–10.
28. *Bornmann L., Daniel H.-D.* What do we know about the h index? // Journal of the American Society for Information Science and Technology. 2007. Vol. 58, Iss. 9. P. 1381–1385.
29. *Bradford S. C.* Sources of information on specific subjects // Engineering. 1934. Vol. 137. P. 85–86.
30. *Bradford S. C.* Documentation. London: Crosby Lockwood & Son, 1948.
31. *Britten W. A.* A use statistic for collection management: The 80/20 rule revisited // Library Acquisitions: Practice & Theory. 1990. Vol. 14, Iss. 2. P. 183–189.
32. *Brody T., Harnad S., Carr L.* Earlier web usage statistics as predictors of later citation impact // Journal of the American Society for Information Science and Technology. 2006. Vol. 57, Iss. 8. P. 1060–1072.
33. *Brookes B. C.* The derivation and application of the Bradford-Zipf distribution // Journal of Documentation. 1968. Vol. 24, No. 4. P. 247–265.
34. *Brookes B. C.* Bradford's law and the bibliography of science // Nature. 1969. Vol. 224. P. 953–956.
35. *Brookes B. C.* The complete Bradford-Zipf 'bibliograph' // Journal of Documentation. 1969. Vol. 25, No. 1. P. 58–60.
36. *Bulick S., Sabor W. N., Flynn R. R.* Circulation and in-house use of books // Use of Library Materials: the University of Pittsburgh study. N. Y.: M. Dekker, 1979. P. 9–55.
37. *Bulick S.* Book Use as a Bradford-Zipf Phenomenon // College & Research Libraries. 1978. Vol. 39. P. 215–219.
38. *Burrell Q. L.* A Simple Stochastic Model for Library Loans // Journal of Documentation. 1980. Vol. 36. P. 115–132.

39. *Burrell Q. L.* Alternative Models for Library Circulation Data // Journal of Documentation. 1982. Vol. 38, No. 1. P. 1–13.
40. *Burrell Q. L.* The 80/20 Rule: Library Lore or Statistical Law? // Journal of Documentation. 1985. Vol. 41, No. 1. P. 24–39.
41. *Burrell Q. L.* The Gini index and the Leimkuhler curve for bibliometric processes // Information Processing and Management. 1992. Vol. 28, No. 1. P. 19–33.
42. *Burrell Q. L.* Defining a core: Theoretical observations on the Egghe-Rousseau proposal // Scientometrics. 2003. Vol. 57, No. 1. P. 75–92.
43. *Burrell Q. L.* «Type/token-taken» informetrics: Some comments and further examples // Journal of the American Society for Information Science and Technology. 2003. Vol. 54, Iss. 13. P. 1260–1263.
44. *Burrell Q. L., Cane V. R.* The Analysis of Library Data // Journal of the Royal Statistical Society, Series A. 1982. Vol. 145. P. 439–463.
45. *Campbell M.* A Survey of the Use of Science Periodicals in Wolverhampton Polytechnic Library // Research in Librarianship. 1974. Vol. 5. P. 39–71.
46. *Chen C.-C.* The Use Patterns of Physics Journals in a Large Academic Research Library // Journal of the American Society for Information Science. 1972. Vol. 23, Iss. 4. P. 254–270.
47. *Chung Y.* Bradford distribution and core authors in classification systems literature // Scientometrics. 1994. Vol. 29, No. 2. P. 253–269.
48. *Cline G. S.* Application of Bradford's Law to Citation Data // College & Research Libraries. 1981. Vol. 42, No. 1. P. 53–61.
49. *Cole P. F.* The Analysis of Reference Question Records as a Guide to the Information Requirements of Scientists // Journal of Documentation. 1958. Vol. 14, No. 4. P. 197–207.
50. *Cole P. F.* A New Look at Reference Scattering // Journal of Documentation. 1962. Vol. 18, No. 2. P. 58–64.

51. *Condon E.* U. Statistics of vocabulary // Science. 1928. Vol. 67, No. 1733. P. 300.
52. *Cooper M. D., McGregor G. F.* Using article photocopy data in bibliographic models for journal collection management // Library Quarterly. 1994. Vol. 64, No. 4. P. 386–413.
53. *Costas R., Bordons M.* The h-index: Advantages, limitations and its relation with other bibliometric indicators at the micro level // Journal of Informetrics. 2007. Vol. 1, Iss. 3. P. 193–203.
54. *Davis P. M.* Patterns in Electronic Journal Usage: Challenging the Composition of Geographic Consortia // College and Research Libraries. 2002. Vol. 63, No. 6. P. 484–497.
55. *Diedrichs C. P.* E-journals: the OhioLINK experience // Library Collections, Acquisitions, and Technical Services. Vol. 25, Iss. 2 (Summer 2001). P. 191–210.
56. *Donohue J. S.* Understanding scientific literatures: a bibliometric approach. London, 1973.
57. *Drott M. C., Griffith B. C.* An Empirical Examination of Bradford's Law and the Scattering of Scientific Literature // Journal of the American Society for Information Science. 1978. Vol. 29, Iss. 5. P. 238–246.
58. *Egghe L.* The Dual of Bradford's Law // Journal of the American Society for Information Science. 1986. Vol. 37, Iss. 4. P. 246–255.
59. *Egghe L.* The duality of informetric systems with applications to the empirical laws // Journal of Information Science. 1990. Vol. 16, No. 1. P. 17–27.
60. *Egghe L.* Applications of the Theory of Bradford's Law to the Calculation of Leimkuhler's Law and to the Completion of Bibliographies // Journal of the American Society for Information Science. 1990. Vol. 41, Iss. 7. P. 469–492.

61. *Egghe L.* Type/Token-taken informetrics // Journal of the American Society for Information Science and Technology. 2003. Vol. 54, Iss. 7. P. 603–610.
62. *Egghe L.* Positive reinforcement and 3-dimensional informetrics // Scientometrics. 2004. Vol. 60, No. 3. P. 497–509.
63. *Egghe L.* Power Laws in the Information Production Process: Lotkaian Informetrics. Amsterdam e. a.: Elsevier, 2005. 428 p.
64. *Egghe L.* Relations Between the Continuous and the Discrete Lotka Power Function Bibliographies // Journal of the American Society for Information Science. 2005. Vol. 56, Iss. 7. P. 664–668.
65. *Egghe L., Rousseau R.* Reflections on a deflection: A note on different causes of the Groos droop // Scientometrics. 1988. Vol. 14, No. 5–6. P. 493–511.
66. *Egghe L., Rousseau R.* Introduction to Informetrics: Quantitative Methods in Library, Documentation and Information Science. Amsterdam e. a.: Elsevier Science Publishers, 1990.
67. *Egghe L., Rousseau R.* A proposal to define a core of a scientific subject: A definition using concentration and fuzzy sets // Scientometrics. 2002. Vol. 54, No. 1. P. 51–62.
68. *Egghe L., Rousseau R.* An informetric model for the Hirsch-index // Scientometrics. 2006. Vol. 69, No. 1. P. 121–129.
69. *Eldredge J. D.* The vital few meet the trivial many: Unexpected use patterns in a monographs collection // Bulletin of the Medical Library Association. 1998. Vol. 86, No. 4. P. 496–503.
70. *Estoup J. B.* Gammes Sténographiques. Paris: Institut Sténographique, 1916.
71. *Eto H., Candelaria P. M.* Applicability of the Bradford Distribution to International Science and Technology Indicators // Scientometrics. 1987. Vol. 11, No. 1–2. P. 27–42.

72. *Evans P., Peters J.* Analysis of the dispersal of use for journals in Emerald Management Xtra (EMX) // Interlending and Document Supply. 2005. Vol. 33, No. 3. P. 155–157.
73. *Fleming T. P., Kilgour F. G.* Moderately and Heavily Used Biomedical Journals // Bulletin of the Medical Library Association. 1964. Vol. 52, No. 1. P. 234–241.
74. *Garfield E.* The 250 Most-Cited Primary Authors, 1961–1975. Part I. How the Names Were Selected // Current Comments. 1977. No. 49. P. 5–15.
75. *Garfield E.* Bradford's law and related statistical patterns // Current Contents. 1980. No. 19. P. 5–12.
76. *Garfield E., Sher I. H.* New Factors in the Evaluation of Scientific Literature Through Citation Indexing // American Documentation. 1963. Vol. 14, No. 3. P. 195–201.
77. *Gini C.* Variabilità e Mutabilità: Contributo allo studio delle distribuzioni e delle relazioni statistiche. Bologna: Cuppini, 1912. 156 p.
78. *Goffman W., Warren K. S.* Dispersion of Papers among Journals based on a Mathematical Analysis of Two Diverse Medical Literatures // Nature. 1969. Vol. 221. P. 1205–1207.
79. *Goffman W., Warren K. S.* Bradford's Law and Library Acquisitions // Nature. 1970. Vol. 226. P. 922–923.
80. *Goodman D.* The Criteria for Open Access // Serials Review. 2004. Vol. 30, No. 4. P. 258–270.
81. *Gorman M.* A core collection of graphic novels: Here are 30 titles every library serving young people should own // School Library Journal. 2002. Vol. 48, Iss. 8. P. 44.
82. *Groos O. V.* Bradford's law and Keenan-Atherton data // American Documentation. 1967. Vol. 18, No. 1. P. 46.

83. *Hardesty L.* Use of library materials at a small liberal arts college // Library Research. 1981. Vol. 3. P. 261–282.
84. *Herdan G.* Type-Token Mathematics: A Textbook of Mathematical Linguistics. Hague: Mouton, 1960. 448 p.
85. *Hindle A., Worthington D.* Simple Stochastic Models for Library Loans // Journal of Documentation. 1980. Vol. 36, Iss. 3. P. 209–213.
86. *Hirsch J. E.* An index to quantify an individual's scientific research output // PNAS. 2005. Vol. 102, No. 46. P. 16569–16572.
87. *Hoffert B.* Read any good poetry lately? Fifty core titles form U. S. poets plus three Nobelists who have influenced U. S. poetry // Library Journal. 1997. Vol. 122, Iss. 6. P. 94–95.
88. *Hooper B.* Core Collection: China Yesterday and Tomorrow // Booklist. 2006. Vol. 102, Iss. 16. P. 39.
89. Journal Report 1: Number of Successful Full-Text Article Requests by Month and Journal // The COUNTER Code of Practice. Journals and Databases: Release 2. P. 9–11. [Электрон. ресурс]. Режим доступа: http://www.projectcounter.org/r2/COUNTER_COP_Release_2.pdf.
90. *Juran J. M.* The Non-Pareto Principle; Mea Culpa // Quality Progress. 1975. Vol. 8. P. 8.
91. *Kendall M. G.* The Bibliography of Operational Research // OR. 1960. Vol. 11, No. 1/2. P. 31–36.
92. *Kingman J. F. C. e. a.* Discussion of the Paper by Mr Burrell and Professor Cane // Journal of the Royal Statistical Society, Series A. 1982. Vol. 145. P. 463–471.
93. *Lawani S. M.* Bradford's Law and the Literature of Agriculture // International Library Review. 1973. Vol. 5, Iss. 3. P. 341–350.
94. *Leimkuhler F. F.* The Bradford Distribution // Journal of Documentation. 1967. Vol. 23. P. 197–207.

95. *Lockett M. W.* The Bradford distribution. A review of the literature, 1934–1987 // Library and Information Science Research. 1989. Vol. 11, Iss. 2. P. 21–36.
96. *Lorenz M. O.* Methods of measuring the concentration of wealth // Journal of the American Statistical Association. 1905. Vol. 9, No. 70. P. 209–219.
97. *Lotka A. J.* The frequency distribution of scientific productivity // Journal of the Washington Academy of Sciences. 1926. Vol. 16. P. 317–323.
98. *Lynch C. A.* Institutional Repositories: Essential Infrastructure for Scholarship in the Digital Age // Portal: Libraries and the Academy. 2003. Vol. 3. No. 2. P. 327–336.
99. Main Science and Technology Indicators. Vol. 2008/1. OECD, 2008. 105 p.
100. *Mankin C. J., Bastille J. D.* An analysis of the differences between density-of-use ranking and raw-use ranking of library journal use // Journal of the American Society for Information Science. 1981. Vol. 32, Iss. 3. P. 224–228.
101. *Meadows A. J.* The citation characteristics of astronomical research literature // Journal of Documentation. 1967. Vol. 23, No. 1. P. 28–33 (перепеч. в: Journal of Documentation. 2004. Vol. 60, No. 6. P. 597–600).
102. *Moed H. F.* Statistical relationships between downloads and citations at the level of individual documents within a single journal // Journal of the American Society for Information Science and Technology. 2005. Vol. 56, Iss. 10. P. 1088–1097.
103. *Morton D. J.* Analysis of Interlibrary Requests by Hospital Libraries for Photocopied Journal Articles // Bulletin of the Medical Library Association. 1977. Vol. 65, No. 4. P. 425–432.
104. *Naranan S.* Power Law Relations in Science Bibliography – A Self-consistent Interpretation // Journal of Documentation. 1971. Vol. 27, No. 2. P. 83–97.

105. *National Science Board.* Science and Engineering Indicators 2008. Vol. 1–2. Arlington, VA: National Science Foundation, 2008.

106. *Nicolaisen J., Hjørland B.* Practical potentials of Bradford's law: A critical examination of the received view // Journal of Documentation. 2007. Vol. 63, No. 3. P. 359–377.

107. *Nisonger Th. E.* Journals in the Core Collections: Definition, Identification, and Applications // Serials Librarian. 2007. Vol. 51, No. 3–4. P. 51–73.

108. *Pareto V.* Cours d'Économie Politique. Vol. 2. Lausanne: 1 Université de Lausanne, 1897.

109. *Pislyakov V.* Comparing two «thermometers»: Impact factors of 20 leading economic journals according to Journal Citation Reports and Scopus // Scientometrcis. 2009 (accepted for publication).

110. *Pislyakov V., Gokhberg L.* Assessing the relative standing of Russian science through a set of citation and publication indicators // Excellence and Emergence. Book of Abstracts. 10th International Conference on Science and Technology Indicators. Vienna: ARC, 2008. P. 400–403.

111. *Praunlich P., Kroll M.* Bradford's Distribution: A New Formulation // Journal of the American Society for Information Science. 1978. Vol. 29, Iss. 2. P. 51–55.

112. *Price D. J. de S.* Networks of Scientific Papers // Science. 1965. Vol. 149, No. 3683. P. 510–515.

113. *Qiu L., Tague J.* Complete or incomplete data sets. The Groos Droop investigated // Scientometrics. 1990. Vol. 19, No 3–4. P. 223–237.

114. *Salaün J.-M., Lafouge T., Boukacem C.* Demand for scientific articles and citations: An Example from the Institut de l'information scientifique et technique (France) // Scientometrics. 2000. Vol. 47, No. 3. P. 561–588.

115. *Schloegl C., Gorraiz J.* Document delivery as a source for bibliometric analyses: the case of Subito // Journal of Information Science. 2006. Vol. 32, Iss. 3. P. 223–237.

116. *Scigliano M.* Serial Use in a Small Academic Library: Determining Cost-Effectiveness // 2000. Serials Review. Vol. 26, No. 1. P. 43–52.

117. *Seaman D.* Core Collection: Climate Change // Booklist. 2006. Vol. 103. Iss. 7. P. 19.

118. *Sennyey P., Ellern G. D., Newsome N.* Collection Development and a Long-Term Periodical Use Study: Methodology and Implications // Serials Review. 2002. Vol. 28, No. 1. P. 38–44.

119. *Sharma H. P.* Download plus citation counts — a useful indicator to measure research impact // Current Science. Vol. 92, No. 7. P. 873.

120. *Stevens S. R.* Mapping the literature of cytotechnology // Bulletin of the Medical Library Association. 2000. Vol. 88 (2). P. 172–177.

121. *Tonta Y., Al U.* Scatter and obsolescence of journals cited in theses and dissertations of librarianship // Library & Information Science Research. 2006. Vol. 28, Iss. 2. P. 281–296.

122. *Tonta Y., Ünal Y.* Scatter of journals and literature obsolescence reflected in document delivery requests // Journal of the American Society for Information Science and Technology. 2005. Vol. 56, Iss. 1. P. 84–94.

123. *Trueswell R. L.* Some Behavorial Patterns of Library Users: The 80/20 Rule // Wilson Library Bulletin. 1969. Vol. 43. P. 458–461.

124. *Urbano C. e. a.* The use of consortially purchased electronic journals by the CBUC (2000–2003) // D-Lib Magazine. 2004. Vol. 10, No. 6.

125. *Urquhart D. J.* Use of Scientific Periodicals // Proceedings of the International Conference on Scientific Information. Washington, D. C. November 16–21, 1958. Washington: National Academy of Sciences — National Research Council, 1959. Vol. 1. P. 287–300.

126. *Urquhart D. J., Bunn R. M.* A National Loan Policy for Scientific Serials // Journal of Documentation. 1959. Vol. 15, No. 1. P. 21–37.

127. *Vickery B. C.* Bradford's law of scattering // Journal of Documentation. 1948. Vol. 4, No. 3. P. 198–203.

128. *Vlachý J.* Frequency distributions of scientific performance. A bibliography of Lotka's law and related phenomena // Scientometrics. 1978. Vol. 1, No. 1. P. 107–130.

129. *Weislogel J.* Elsevier Science Digital Libraries Symposium II: a conference report // Library Collections, Acquisitions, and Technical Services. Vol. 23, Iss. 4 (Winter 1999). P. 459–467.

130. *Wender R.* Hospital Journal Title Usage Study // Special Libraries. 1975. Vol. 66, No. 11. P. 532–537.

131. *Wood D. N., Bower C. A.* Survey of Medical Literature Borrowed from the National Lending Library for Science and Technology // Bulletin of the Medical Library Association. Vol. 57, No. 1. P. 47–63.

132. *Wood D. N., Bower C. A.* The use of social science periodical literature // Journal of Documentation. 1969. Vol. 25. P. 108–122.

133. *Yablonsky A. I.* On fundamental regularities of the distribution of scientific productivity // Scientometrics. 1980. Vol. 2, No. 1. P. 3–34.

134. *Zipf G. K.* Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology. Cambridge, Mass.: Addison-Wesley, 1949. 573 p.

Приложения

1. Таблица значений для исследования закона Брэдфорда

А: журналов	Б: статей в журнале	В: всего журналов	Г: всего статей	Д: ln(В)
1	1306	1	1306	0
1	1166	2	2472	0,693147
1	939	3	3411	1,098612
1	925	4	4336	1,386294
1	798	5	5134	1,609438
1	691	6	5825	1,791759
1	644	7	6469	1,94591
1	601	8	7070	2,079442
1	567	9	7637	2,197225
1	527	10	8164	2,302585
1	512	11	8676	2,397895
1	480	12	9156	2,484907
1	473	13	9629	2,564949
1	420	14	10049	2,639057
1	395	15	10444	2,70805
1	366	16	10810	2,772589
1	354	17	11164	2,833213
1	348	18	11512	2,890372
1	345	19	11857	2,944439
1	338	20	12195	2,995732
1	333	21	12528	3,044522
1	318	22	12846	3,091042
1	296	23	13142	3,135494
1	288	24	13430	3,178054

1	282	25	13712	3,218876
1	275	26	13987	3,258097
1	272	27	14259	3,295837
1	265	28	14524	3,332205
1	256	29	14780	3,367296
1	242	30	15022	3,401197
1	239	31	15261	3,433987
1	235	32	15496	3,465736
1	233	33	15729	3,496508
1	230	34	15959	3,526361
1	222	35	16181	3,555348
1	213	36	16394	3,583519
1	209	37	16603	3,610918
1	208	38	16811	3,637586
1	199	39	17010	3,663562
1	196	40	17206	3,688879
1	188	41	17394	3,713572
1	186	42	17580	3,73767
1	169	43	17749	3,7612
2	165	45	18079	3,806662
1	162	46	18241	3,828641
1	160	47	18401	3,850148
1	159	48	18560	3,871201
1	158	49	18718	3,89182
1	156	50	18874	3,912023
2	155	52	19184	3,951244
1	150	53	19334	3,970292
1	149	54	19483	3,988984
1	147	55	19630	4,007333
2	141	57	19912	4,043051
1	135	58	20047	4,060443
1	134	59	20181	4,077537
1	133	60	20314	4,094345

1	132	61	20446	4,110874
1	126	62	20572	4,127134
2	125	64	20822	4,158883
1	123	65	20945	4,174387
1	122	66	21067	4,189655
1	118	67	21185	4,204693
1	114	68	21299	4,219508
1	112	69	21411	4,234107
1	110	70	21521	4,248495
1	108	71	21629	4,26268
1	103	72	21732	4,276666
1	101	73	21833	4,290459
1	100	74	21933	4,304065
2	98	76	22129	4,330733
2	96	78	22321	4,356709
1	95	79	22416	4,369448
1	93	80	22509	4,382027
1	92	81	22601	4,394449
2	91	83	22783	4,418841
2	90	85	22963	4,442651
4	89	89	23319	4,488636
3	87	92	23580	4,521789
1	86	93	23666	4,532599
1	84	94	23750	4,543295
1	83	95	23833	4,553877
1	81	96	23914	4,564348
2	80	98	24074	4,584967
3	79	101	24311	4,615121
1	78	102	24389	4,624973
1	77	103	24466	4,634729
1	74	104	24540	4,644391
1	73	105	24613	4,65396
3	72	108	24829	4,682131

4	69	112	25105	4,718499
1	68	113	25173	4,727388
3	67	116	25374	4,75359
1	66	117	25440	4,762174
2	64	119	25568	4,779123
5	63	124	25883	4,820282
4	61	128	26127	4,85203
2	59	130	26245	4,867534
2	58	132	26361	4,882802
2	56	134	26473	4,89784
1	55	135	26528	4,905275
2	54	137	26636	4,919981
1	53	138	26689	4,927254
2	52	140	26793	4,941642
3	51	143	26946	4,962845
1	50	144	26996	4,969813
4	49	148	27192	4,997212
1	48	149	27240	5,003946
4	47	153	27428	5,030438
1	46	154	27474	5,036953
7	45	161	27789	5,081404
1	44	162	27833	5,087596
3	43	165	27962	5,105945
1	42	166	28004	5,111988
5	41	171	28209	5,141664
4	40	175	28369	5,164786
6	39	181	28603	5,198497
4	38	185	28755	5,220356
2	37	187	28829	5,231109
3	36	190	28937	5,247024
7	35	197	29182	5,283204
2	34	199	29250	5,293305
12	33	211	29646	5,351858

8	32	219	29902	5,389072
8	31	227	30150	5,42495
1	30	228	30180	5,429346
8	29	236	30412	5,463832
8	28	244	30636	5,497168
6	27	250	30798	5,521461
13	26	263	31136	5,572154
11	25	274	31411	5,613128
9	24	283	31627	5,645447
10	23	293	31857	5,680173
12	22	305	32121	5,720312
14	21	319	32415	5,765191
16	20	335	32735	5,814131
21	19	356	33134	5,874931
21	18	377	33512	5,932245
13	17	390	33733	5,966147
19	16	409	34037	6,013715
20	15	429	34337	6,061457
29	14	458	34743	6,126869
33	13	491	35172	6,196444
24	12	515	35460	6,244167
37	11	552	35867	6,313548
42	10	594	36287	6,386879
63	9	657	36854	6,487684
76	8	733	37462	6,597146
67	7	800	37931	6,684612
81	6	881	38417	6,781058
102	5	983	38927	6,890609
183	4	1166	39659	7,061334
230	3	1396	40349	7,241366
416	2	1812	41181	7,502186
778	1	2590	41959	7,859413

Таблица представляет собой свод статистики использования различных электронных документов в библиотеке ГУ-ВШЭ в 2004 г. Все наименования источников (журналов) упорядочены по убыванию числа открытых из них полнотекстовых статей. Значения столбцов:

А: количество источников с определенным (одинаковым) числом открытых из них статей;

Б: количество статей, открытых в соответствующем числе источников из столбца А;

В: скользящая сумма количества источников из столбца А;

Г: скользящая сумма общего количества статей (т.е. сумма произведений значений из А и Б);

Д: натуральный логарифм кумулятивного числа источников (столбца В).

2. Ядро читательского спроса на периодику, определенное при помощи информетрической модели Брэдфорда

название журнала	**открыто статей из журнала**	**читался в базах данных**
American Economic Review	1306	EBSCO+JSTOR+PQ
American Journal of Sociology	1166	EBSCO+JSTOR+PQ
Journal of Finance	939	EBSCO+JSTOR
Euroweek	925	EBSCO+PQ
Journal of Political Economy	798	EBSCO+JSTOR+PQ
Economic Journal	691	EBSCO+JSTOR
Harvard Business Review	644	EBSCO
Journal of Health Economics	601	SD
American Sociological Review	567	JSTOR+PQ

Quarterly Journal of Economics	527	EBSCO+JSTOR
Journal of Public Economics	512	SD
Econometrica	480	JSTOR+PQ
Economist	473	EBSCO+PQ
Journal of Economic Literature	420	EBSCO+JSTOR+PQ
Journal of Economic Perspectives	395	EBSCO+JSTOR+PQ
Journal of Marketing	366	EBSCO+PQ
Journal of Business	354	EBSCO+JSTOR+PQ
Journal of Financial Economics	348	SD
Management Accounting Research	345	SD
Sociology of Education	338	JSTOR+PQ
Strategic Management Journal	333	JSTOR+PQ
Public Opinion Quarterly	318	EBSCO+JSTOR+PQ
Management Science	296	EBSCO+JSTOR+PQ
Academy of Management Review	288	EBSCO+JSTOR+PQ
European Economic Review	282	SD
Review of Economics & Statistics	275	EBSCO+JSTOR
Journal of the American Statistical Association	272	EBSCO+JSTOR+PQ
Annual Review of Sociology	265	EBSCO+JSTOR+PQ

В таблице выделены журналы, образующие «ядро читательского спроса», определенной в параграфе 2.3.2 методом информетрического моделирования при помощи закона Брэдфорда, с использованием комбинации графического и вербального его вариантов. Всего в ядре 28 журналов.

В третьем столбце добавлена информация о тех онлайновых базах данных, из которых читались статьи соответствующего журнала: EBSCO, JSTOR, ProQuest (PQ), ScienceDirect (SD) или их сочетание.

3. Итерации нелинейной регрессии при приближении экспериментальных данных законом Леймкулера

номер итерации	**остаточная сумма квадратов**	**β**
1.0	338.430	2.000
1.1	151.406	13.531
2.0	151.406	13.531
2.1	85.077	36.594
3.0	85.077	36.594
3.1	62.089	59.656
4.0	62.089	59.656
4.1	41.870	105.781
5.0	41.870	105.781
5.1	32.205	151.906
6.0	32.205	151.906
6.1	22.478	244.156
7.0	22.478	244.156
7.1	14.374	428.655
8.0	14.374	428.655
8.1	10.725	613.154
9.0	10.725	613.154
9.1	7.295	982.153
10.0	7.295	982.153
10.1	4.764	1720.151
11.0	4.764	1720.151
11.1	3.374	3196.147
12.0	3.374	3196.147
12.1	3.020	5073.433
13.0	3.020	5073.433
13.1	3.002	5808.484

14.0	3.002	5808.484
14.1	3.002	5851.369
15.0	3.002	5851.369
15.1	3.002	5850.216
16.0	3.002	5850.216
16.1	3.002	5850.247

Отражен процесс подбора программой SPSS 14.0 коэффициента β формулы (4.2.1) при приближении моделью Леймкулера экспериментальных данных, см. параграф 4.1. Начальное значение коэффициента равно двум, найденное в регрессионном приближении — 5850.