На правах рукописи

Писляков Владимир Владимирович

# ИНФОРМЕТРИЧЕСКОЕ МОДЕЛИРОВАНИЕ ПРОЦЕССА ОБРАЩЕНИЯ К ЭЛЕКТРОННЫМ ИНФОРМАЦИОННЫМ РЕСУРСАМ

Специальность 05.13.18 – Математическое моделирование, вычислительные методы и комплексы программ

Автореферат
диссертации на соискание ученой степени
кандидата физико-математических наук

Казань – 2008

**Работа выполнена в НИИ математики и механики им. Н. Г. Чеботарева Казанского государственного университета**

**Научный руководитель:** доктор физ.-мат. наук, профессор,
заслуженный деятель науки РТ
Елизаров Александр Михайлович

**Официальные оппоненты:** доктор физ.-мат. наук, профессор
Сотников Александр Николаевич

доктор технических наук, профессор
Захаров Вячеслав Михайлович

**Ведущая организация:** Всероссийский институт научной и технической информации Российской академии наук (ВИНИТИ РАН), г. Москва

Защита состоится «31» октября 2008 года в 14-00 на заседании Диссертационного совета Д 212.079.01 в Казанском государственном техническом университете им. А. Н. Туполева по адресу: 420111, Казань, ул. К. Маркса, д. 10

С диссертацией можно ознакомиться в научной библиотеке Казанского государственного технического университета им. А. Н. Туполева

Автореферат разослан «29» сентября 2008 г.

Ученый секретарь диссертационного совета
доктор физ.-мат. наук, профессор П. Г. Данилаев

## ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

**Актуальность работы.** Статистический анализ научной и образовательной деятельности получает всё большее распространение как в мировой, так и в отечественной практике[1]. 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*.

Хорошо известным 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.

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. Поэтому актуальной проблемой информетрического моделирования, проводимого на конкретном информационном процессе, является определение, насколько применим к рассматриваемому процессу, насколько корректно и точно выполняется тот или иной информетрический закон, каковы особенности его действия 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. Настоящее исследование связано с проблемой применимости информетрических моделей к процессу использования электронной информации.

---

[1] Main Science and Technology Indicators. – Vol. 2008/1. – OECD, 2008. – 105 p.; *National Science Board.* Science and Engineering Indicators – 2008. – Vol. 1–2. – Arlington, VA: National Science Foundation, 2008; *Pislyakov V.* Assessing the Relative Standing of Russian Science through a Set of Citation and Publication Indicators / V. Pislyakov, L. Gokhberg // Excellence and Emergence. Book of Abstracts. 10th International Conference on Science and Technology Indicators. – Vienna: ARC, 2008. – P. 400–403; Индикаторы науки: 2007. Статистический сборник. – М.: ГУ-ВШЭ, 2007. – 341 с.; Индикаторы образования: 2007. Статистический сборник. – М.: ГУ-ВШЭ, 2007. – 174 с.

Таким образом, актуальность работы заключается, прежде всего, в том, что она соединяет информетрические и библиометрические методы, набирающие силу и авторитет в исследованиях науки и образования, и электронные информационные ресурсы, завоевывающие в наше время всё большую популярность и получающие всё более широкое распространение. Тем самым 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.

**Целью** диссертационного 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 (ГУ-ВШЭ);
- получение в процессе приближения моделей к эксперименту основных их коэффициентов;
- анализ применимости информетрических законов для моделирования спроса на электронные документы и исследование с точки зрения информетрии особенностей, характерных для такого рода спроса;
- 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 ГУ-ВШЭ. Сбор и анализ этой статистики производился на протяжении всего 2004 г. Таким образом, имеется «живой» экспериментальный набор данных информетрического процесса, обладающий достаточным объемом для репрезентативности результатов информетрического моделирования (всего зафиксировано 41959 обращений к статьям из 2590 различных изданий, размещенных в базах данных).

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

Среди авторов, на чьи методологические разработки опирается настоящее диссертационное исследование, необходимо упомянуть С. Д. Хайтуна, А. И. Яблонского, B. C. Brookes, Q. L. Burrell, L. Egghe, E. Garfield, S. Naranan, R. Rousseau, D. J. Urquhart, B. C. Vickery.

**Положения, выносимые на защиту:**

- методология применения основных информетрических законов (Брэдфорда, Парето, Леймкулера, Лотки, Ципфа) для математического моделирования процесса обращения читателей к электронным изданиям;
- 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.

**Научная новизна** настоящего исследования определяется несколькими аспектами. 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, в которых проводилось бы информетрическое моделирование на таком солидном материале. Наконец, использование 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 в России крайне редко, либо вообще не рассматривались.

**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 ценность** диссертационного исследования заключается в возможности применения его методики и основных выводов 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.

**Наиболее существенные научные и практические результаты, полученные лично соискателем:**

- собрана годичная статистика использования электронных ресурсов в Государственном университете – Высшей школе экономики;
- уточнена методика Л. Эгге определения оптимальных параметров информетрической модели Брэдфорда;
- предложено два метода оценки ядра литературы по кривой Брэдфорда;
- получены оптимальные коэффициенты моделей Брэдфорда, Парето, Леймкулера, Лотки, Ципфа в применении к экспериментальным данным о процессе обращения к электронным онлайновым источникам;
- предложен и обоснован метод корректировки данных и трактовки понятия числа источников с заданной продуктивностью для информетрической модели Лотки;
- предложен и обоснован метод сглаживания графика модели Ципфа и корректного отображения ранга источников информетрического процесса с малой продуктивностью;
- проведено сравнение характеристик кривых Леймкулера и индексов Джини, полученных численным интегрированием и приближением аналитической моделью соответственно;
- двумя различными методами выделено компактное информационное ядро читательского спроса, определяющее основные направления информационной поддержки научной и образовательной деятельности вуза.

**Апробация работы.** Результаты диссертации по мере их получения докладывались и обсуждались на семинаре Отделения математического моделиро-

вания 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., Государственный университет – Высшая школа экономики, руководители проф. Ф. Т. Алескеров и проф. В. В. Подиновский), на международных конференциях «SCIENCE ONLINE: электронные информационные ресурсы для науки и образования» (2003, 2004, 2005 и 2007 гг.) и «Крым: Библиотеки и информационные ресурсы в современном мире науки, культуры, образования и бизнеса» (2004 и 2005 гг.).

Результаты проведенного исследования были использованы при формировании концепции комплектования электронными ресурсами библиотеки ГУ-ВШЭ. В 2007 г. за успехи в трудовой деятельности автор диссертации был отмечен Благодарностью министра экономического развития и торговли.

**Публикации.** Результаты исследования опубликованы в 5 работах: в четырех статьях в журналах из Перечня, рекомендованного ВАК для публикации результатов диссертационных работ, одна из статей перепечатана в виде главы в коллективной монографии. Кроме того, результаты опубликованы в ряде сборников трудов конференций. Список публикаций приведен в конце автореферата.

**Содержание, структура и объем работы.** Диссертационная работа состоит из введения, четырех глав, заключения, списка литературы и трех приложений; содержит 7 таблиц и 20 рисунков. Общий объем диссертации 155 страниц. Библиографический список состоит из 134 наименований работ отечественных и зарубежных авторов.

Автор выражает признательность своему научному руководителю, доктору физ.-мат. наук профессору А. М. Елизарову, благодаря сотрудничеству и плодотворным дискуссиям с которым данный труд смог увидеть свет, а также директору библиотеки ГУ-ВШЭ Н. Ю. Максимовой, осуществлявшей неизменную поддержку усилий автора всё время написания диссертации.

## СОДЕРЖАНИЕ РАБОТЫ

Во **введении** производится постановка проблемы, определяется понятие информетрического моделирования и излагается краткая история вопроса. Подчеркивается основная проблема математического моделирования социальных процессов — вопрос о переносимости моделей с процессов одного рода на другой. 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 связано с изменением масштаба исследования (пример: переход от изучения рассеяния по журналам статей, написанных исследовательской лабораторией, к рассеянию публикаций целой страны), сменой дисциплинарной области, в которой проис-

ходит 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).

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[2]). 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. Данное разделение источник — продукт является фундаментальным и неотъемлемым от понятия об информационном процессе производства. 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.

Далее во введении отмечается, что все законы информетрии, с использованием которых будет проводиться информетрическое моделирование, исходно были открыты на информационных процессах иной природы, чем те, применимость к которым исследуется в диссертационной работе. В диссертации моделируется 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. Сказанное определяет один из главных фокусов диссертационного исследования — вопрос о том, насколько корректно можно совершить перенос информетрического моделирования от «исходных» информационных процессов, давших рождение изучаемым моделям, к процессу обращения к документам, информационным ресурсам.

Кроме того, особое внимание уделяется тому месту, которое занимает процесс использования информации вообще и электронной информации в частности. Этот процесс является вторичным информационным процессом: прежде чем информацию использовать, ее необходимо создать. Это позволяет исследователям говорить о моделировании спроса на информационные источники как об особом виде информетрии: «линейной трехмерной информетрии» (linear

---

[2] *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; *Egghe L.* Introduction to Informetrics: Quantitative Methods in Library, Documentation and Information Science / L. Egghe, R. Rousseau. – Amsterdam e. a.: Elsevier Science, 1990. – P. 292, 313.

three-dimensional informetrics[3]). «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 опубликованные статьи; финальными продуктами — цитирования, полученные данными статьями. В настоящем исследовании в качестве источников фигурируют онлайновые журналы, в качестве финальных продуктов — загрузки читателями полнотекстовых статей из этих журналов.

**Первая глава «Электронные издания и процесс их использования»** содержит обзор методов, применяемых в работе, и характеристик объектов настоящего исследования.

В параграфе 1.1 изложены классификация, структура и особенности использования различных хранилищ электронных научных документов. При этом особое внимание уделено онлайновым базам данных — источникам, размещенным во всемирной сети Интернет.

Параграф 1.2 посвящен определению процесса обращения к электронным изданиям как объекта информетрического моделирования. Рассмотрены различные типы информационного процесса производства и определено место среди них объекта настоящего исследования.

В параграфе 1.3 изложены особенности инструментария, при помощи которого в работе собраны экспериментальные статистические данные, использованные далее в процессе информетрического моделирования. Эта статистика получена из отчетов систем анализа обращений пользователей к онлайновым информационным ресурсам. Важным свойством собранных таким образом данных являются их точность и, одновременно, низкая ресурсоемкость получения. Это позволяет работать с точными данными, собранными за значительный период времени (в настоящей работе — за год).

Параграф 1.4 подробно описывает базы данных, включенные в настоящее исследование. Это научные электронные ресурсы EBSCO, JSTOR, ProQuest и ScienceDirect. В общей сложности в них имелось на момент исследования около 13400 полнотекстовых периодических изданий, при этом за время сбора данных 2590 из них были хотя бы один раз востребованы пользователями (т. е. из них была открыта хотя бы одна публикация). Со ссылкой на аргументацию К. Л. Беррела[4] вводится важный методологический принцип, которому далее следует вся работа: в рассмотрение включаются только те журналы, к которым

---

[3] *Egghe L.* Power Laws in the Information Production Process: Lotkaian Informetrics. – Amsterdam e. a.: Elsevier, 2005. – P. 161–163.

[4] *Burrell Q. L.* The 80/20 Rule: Library Lore or Statistical Law? // Journal of Documentation. – 1985. – Vol. 41, No. 1. – P. 24–39; *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.

были обращения, остальные выносятся за рамки данного исследования и не учитываются ни в каких выкладках.

Во **второй главе «Моделирование процесса обращения к электронным информационным источникам при помощи информетрического закона Брэдфорда»** рассмотрена модель Брэдфорда в двух ее формулировках — вербальной и графической.

Параграф 2.1 содержит теоретические основы информетрического моделирования при помощи закона Брэдфорда. «Вербальный» вариант закона Брэдфорда формулируется так: выделим множество журналов, в которых появляются статьи, посвященные некоторой выбранной научной тематике, и упорядочим их в порядке убывания количества этих статей; тогда это упорядоченное множество журналов можно разбить на три зоны так, чтобы в каждой было одинаковое количество статей по заданной теме. При этом, как утверждает закон Брэдфорда, числа журналов в соответствующих зонах будут относиться друг к другу как $1:q:q^2$, где $q$ — некоторое число большее единицы (коэффициент Брэдфорда). Закон может быть органичным образом распространен на случай $n$ зон (их называют зонами Брэдфорда).

На практике закон Брэдфорда, как и любая информетрическая модель, редко выполняется в совершенстве. Поэтому можно выбирать различные значения для размера первой зоны Брэдфорда $S_1$, числа зон $n$ и коэффициента Брэдфорда $q$ и получать более или менее хорошие разбиения на зоны Брэдфорда — в большей или меньшей степени удовлетворяющие закону. Отсюда возникает задача определения таких параметров модели Брэдфорда, чтобы она максимально хорошо приближала исходные экспериментальные данные. При этом необходимо, чтобы искомые параметры могли быть найдены из «макроскопических» характеристик набора данных, а именно из общего числа журналов S, общего числа статей I, числа источников с минимальной продуктивностью $s_1$ и числа статей, произведенных журналом с максимальной продуктивностью $i_{max}$.

Рассматривается два решения данной задачи. Метод А. И. Яблонского[5] дает систему решений

$$q = \frac{S}{s_1}; \quad S_1 = s_1(q-1)/i_{max} + q; \quad n = I/(S - s_1)$$

(при этом делаются предположения о подчинении распределения статей закону Ципфа – Мандельброта и о попадании в последнюю зону Брэдфорда только журналов с одной статьей).

Метод Л. Эгге[6] исходит только из выполнения закона Лотки, полагает свободу в выборе числа зон $n$ и приводит к следующей оценке параметров модели Брэдфорда:

---

[5] *Яблонский А. И.* Модели и методы исследования науки. – М.: Эдиториал УРСС, 2001. – С. 349–354.

$$q = \left(e^{\gamma} s_1\right)^{1/n}; \quad S_1 = S\frac{q-1}{q^n - 1}, \tag{1}$$

где γ — постоянная Эйлера – Маскерони.

Л. Эгге подробно описывает процесс построения зон Брэдфорда, однако рекомендует при определении размера первой зоны $S_1$ из (1) округлять ее размер строго в меньшую сторону. Автором настоящего диссертационного исследования приводятся теоретические обоснования предпочтительности округления в сторону ближайшего целого, что позднее демонстрируется на практике.

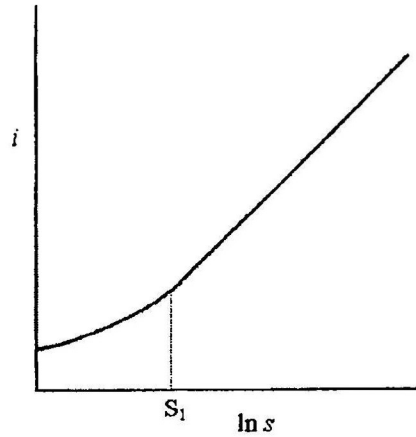


Рис. 1. Библиограф: график зависимости
кумулятивного числа статей $i$ в $s$ первых журналах от ln $s$

Графическая формулировка закона Брэдфорда говорит о том, что построенная в полулогарифмических координатах кривая зависимости кумулятивного числа статей от кумулятивного числа журналов (журналы упорядочены по убыванию продуктивности) будет иметь вид, показанный на рис. 1. Такую кривую называют библиографом.

Библиограф будет вогнутым до некоторого значения $s = S_1$, после которого перейдет в прямую. При этом, согласно С. Брэдфорду, именно $S_1$ соответствует границе первой зоны Брэдфорда. В более поздних исследованиях было отмечено, что нередко на конце библиограф отклоняется вниз от прямой, образуя т. н. Groos Droop[7].

---

[6] *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.

[7] *Groos O. V.* Bradford's law and Keenan-Atherton data // American Documentation. – 1967. – Vol. 18, No. 1. – P. 46

Далее со ссылками на соответствующую литературу показано, что закон Брэдфорда выполняется в различных дисциплинарных областях и на информационных процессах разнообразной природы: на рассеянии статей по журналам, авторам и странам публикации; на распределении сделанных цитирований по журналам и полученных цитирований по авторам и по публикациям. Особое внимание уделено исследованиям применимости закона Брэдфорда к процессу использования информации, к чтению документов. Приводится более десяти различных работ на данную тему, показывающих применимость модели Брэдфорда в различных ситуациях — при выдаче статей по межбиблиотечному обмену, при обработке информационных запросов специалистами справочной службы, при чтении реферативных изданий, при книговыдаче и в работе службы доставки документов.

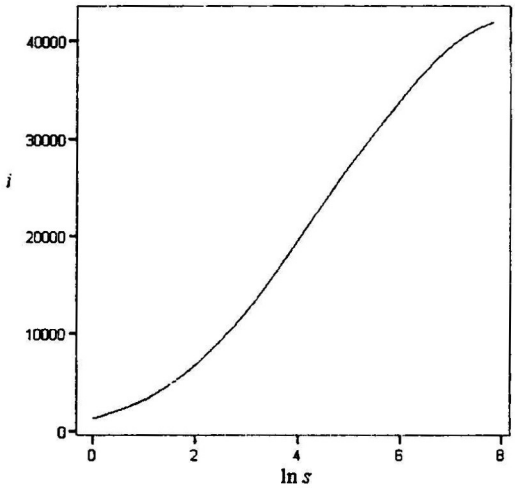


Рис. 2. Библиограф, построенный по массиву экспериментальных данных

В параграфе 2.2 изложенные теоретические основания закона Брэдфорда применены к информетрическому моделированию на основе экспериментальных данных о статистике обращений читателей ГУ-ВШЭ к онлайн-периодике. Сначала строится библиограф и проверяется графическая формулировка закона. Выясняется, что библиограф имеет классический вид, с ярко выраженным Groos Droop (рис. 2). Таким образом, можно констатировать, что графический вариант закона Брэдфорда выполняется.

Установлено, что вербальная формулировка в ее простейшем варианте разбиения на три зоны не выполняется: отношения чисел журналов в последовательных зонах Брэдфорда меняется более чем в три раза (таблица 1).

Далее исследуется более развитый метод деления множества журналов на $n$ зон Брэдфорда, предложенный Л. Эгге и проанализированный в параграфе 2.1. Тестируется различное число $n$, от 3 до 12 (последнее является вычисленным ограничением сверху числа зон $n$ в методе Л. Эгге для данного случая). При этом используется описанная выше корректировка метода Эгге в части округления величины первой зоны. Выясняется, что на всех зонах сразу вербальная формулировка выполняется плохо: при фиксированном $q$ максимальное различие между числами статей в разных зонах колеблется от 1,9 до 4,7 раз. Однако в средних зонах имеются участки, на которых число статей в зонах почти не меняется. Так, при разбиении на 12 зон (таблица 2) три зоны VI–VIII содержат практически одинаковое число статей: отклонение составляет менее 7%.

Таблица 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)

| 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 |

Сделан вывод о том, что закон Брэдфорда выполняется на тех зонах, где библиограф имеет практически прямолинейную форму. Ввиду того что характерной особенностью библиографа, построенного по исследуемым экспериментальным данным, является сильный Groos Droop в области высоких рангов (в правой части библиографа), прямолинейный участок не является ярко выраженным, что объясняет плохое выполнение вербальной формулировки закона Брэдфорда.

Согласно литературе[8], сильный Groos Droop может являться следствием политематичности исследуемой области или недостаточно полного охвата всех журналов заданной тематики при построении библиографа. В нашем случае могут иметь место оба эффекта, если их перенести на процесс использования литературы: исследуемое множество онлайновых источников не исчерпывает всех информационных потребностей читателей вуза (отсюда «неполнота» учтенной библиографии), а включенные в анализ онлайновые библиотеки существенно политематичны, равно как и интересы читателей ГУ-ВШЭ. Именно сочетанием

[8] *Egghe L.* Reflections on a deflection: A note on different causes of the Groos droop / L. Egghe, R. Rousseau // Scientometrics. – 1988. – Vol. 14, No. 5–6. – P. 493–511; *Brookes B. C.* Bradford's law and the bibliography of science // Nature. – 1969. – Vol. 224. – P. 953–956.

этих причин может объясняться особенно сильная деформация библиографа в области малоспрашиваемых изданий. Отсюда же, по предлагаемой интерпретации, следует недостаточно удовлетворительное выполнение вербальной формулировки закона Брэдфорда.

В параграфе 2.3 информетрическая модель Брэдфорда применена к выделению «ядра» периодических изданий — наиболее важных и спрашиваемых журналов, представляющих ключевой интерес для читательской аудитории вуза. Согласно Брэдфорду и его последователям, ядром считается либо первая зона в вербальной формулировке модели, либо точка перехода библиографа в прямую в ее графическом варианте. Так как в данном исследовании установлено, что лучше выполняется графическая формулировка, то именно по ней определяется ядро читательского спроса.

Таблица 2. Разбиение на зоны Брэдфорда скорректированным методом Эгге, $n$=12

| зона Брэдфорда | число журн-в | $q$ | число статей | зона Брэдфорда | число журн-в | $q$ | число статей |
|---|---|---|---|---|---|---|---|
| I | 1 | – | 1306 | VII | 49 | 1,88 | 5055 |
| II | 2 | 2,00 | 2105 | VIII | 93 | 1,90 | 4723 |
| III | 4 | 2,00 | 3058 | IX | 176 | 1,89 | 4292 |
| IV | 7 | 1,75 | 3580 | X | 339 | 1,93 | 3874 |
| V | 13 | 1,86 | 4210 | XI | 646 | 1,91 | 2951 |
| VI | 26 | 2,00 | 5075 | XII | 1234 | 1,91 | 1730 |

В отсутствие общепринятого способа определения точки перехода библиографа от вогнутой части к прямой, автором настоящего исследования предложены два подхода к решению данной задачи. Один использует приближение кривой полиномом третьей степени и поиск точки перегиба, второй — обращение к разбиению на зоны Брэдфорда методом Эгге и определение нижней границы тех зон, где удовлетворительно выполнятся вербальная формулировка закона Брэдфорда (следовательно, участка, на котором библиограф имеет квазипрямолинейный вид). Первый метод является оценкой размера ядра сверху и дает как результат 77 журналов, второй, по всей видимости, более точен и выделяет 28 ядерных изданий. Наконец, найденное ядро анализируется по различным базам данных — это позволяет установить роль каждого ресурса в ядре электронной коллекции.

**Глава 3 «Моделирование процесса обращения к электронным информационным источникам при помощи информетрического принципа Парето»** рассматривает закон (принцип, правило) Парето и, в его контексте, понятия о кривой Леймкулера – Лоренца и индексе Джини.

В параграфе 3.1 сформулирован эмпирический принцип Парето 80/20, который в общем виде звучит как «20% причин отвечают за 80% следствий». Проведен обзор библиографии по закону Парето в применении к процессу использования информации. Отмечено, что из всех информетрических закономерностей, охваченных настоящей диссертационной работой, принцип Парето — единственный закон, о приложении которого к использованию *электронных* документов автору известно из существующих научных публикаций.

Параграф 3.2 посвящен определению понятий кривой Лоренца (Леймкулера – Лоренца), которая отражает зависимость кумулятивной доли статей, приходящихся на кумулятивную долю журналов, и одной из мер концентрации распределений, индекса Джини $G$. Последний в явном виде связан с площадью под кривой Лоренца, изменяется от 0 до 1 и указывает на максимальную неравномерность распределения при $G=1$ и абсолютно равномерное распределение при $G=0$. Получена формула для оценки $G$ численными методами, при помощи приближения кривой Лоренца трапециями и вычисления площади под ней.

Кроме того, в этом параграфе изложен оригинальный метод определения ядра журналов (для распределения статей по журналам) при помощи индекса Джини, предложенный современными исследователями[9]. В нем последовательно обнуляются числа статей для $N-s$ наименее продуктивных изданий и ненулевые значения остаются лишь для $s$ наиболее продуктивных журналов. При этом индекс Джини возрастает с уменьшением $s$ (т. е. при уменьшении количества журналов с ненулевым числом относящихся к ним статей). Для определения границы ядра предложено использовать величину

$$m_s = \frac{G_s - G}{G_1 - G},$$

названную авторами упомянутого метода «коэффициентом принадлежности к ядру», которая убывает с ростом $s$, изменяясь от 1 до 0. Авторы рассматриваемого метода предлагают выбирать некоторый уровень (в 90 или 95 процентов) и считать ядром те журналы, для которых $m_s$ превышает соответственно 0,9 или 0,95 (в результате получается «90%-ядро» и «95%-ядро»). Сильная сторона такого подхода заключается в том, что он учитывает реальную степень отличия журналов, входящих в ядро, от остальных изданий, а также их роль в неравномерности всего распределения.

---

[9] *Egghe L.* A proposal to define a core of a scientific subject: A definition using concentration and fuzzy sets / L. Egghe, R. Rousseau // Scientometrics. – 2002. – Vol. 54, No. 1. – P. 51–62; *Burrell Q. L.* Defining a core: Theoretical observations on the Egghe-Rousseau proposal // Scientometrics. – 2003. – Vol. 57, No. 1. – P. 75–92.

В параграфе 3.3 правило Парето проверено на экспериментальных данных о спросе на онлайновую периодику в ГУ-ВШЭ. Сначала кривые Леймкулера – Лоренца построены для четырех электронных ресурсов по отдельности. Оказалось, что правило Парето выполняется с разной степенью успешности: от 80/18 (80% обращений приходится на 18% наименований периодики) для JSTOR до 80/31 для ProQuest (таблица 3). Затем построена кривая для всех четырех ресурсов, рассмотренных как единый информационный массив. Выяснилось, что в этом случае принцип Парето имеет даже более сильную форму: 80% обращений пользователей относится к 14,7% журналов, или 83% — к 17% (пропорция в классическом виде: 83/17).

Таблица 3. Выполнение правила Парето и величина индекса Джини для четырех ресурсов и суммарной электронной подписки

| Ресурс | пропорция Парето | 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** |

Индексы Джини для четырех отдельных ресурсов и общего массива найдены численными методами, изложенными в параграфе 3.2. Самая большая концентрация наблюдается для всех ресурсов как целого ($G$=0,80 — «сводная» кривая Леймкулера покрывает площадь 0,9, т. е. 90% от возможного), самая слабая — для базы данных ProQuest ($G$=0,65). Кроме того, выяснено, что имеет место прямая зависимость: чем больше индекс Джини, тем «строже» исполняется закон Парето. Это не является необходимым условием и означает, что в данном случае чем раньше кривая Леймкулера для ресурса (или их общей совокупности) пересечет ординату в 80%, тем большая площадь под ней находится на всем ее протяжении.

В параграфе 3.4 метод определения ядра журналов при помощи индекса Джини, изложенный в параграфе 3.2, использован для рассмотрения усеченной выборки с обнуленными значениями для ряда малопродуктивных изданий. Как и раньше, в данном методе публикация статей в журналах заменяется на их открытие пользователями — в этом состоит методика адаптации данного метода от процесса производства к процессу обращения к информации. Для суммарной подписки выяснено, что 90%-ядро состоит из 89 журналов (3,4% от общего числа), 95%-ядро — из 37 (1,4%). Можно констатировать, что по порядку полученной величины оба способа — по Брэдфорду и по Джини – Парето — дают одинаковый размер ядра. Получение схожих результатов концептуально разными 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.

**Глава 4 «Моделирование процесса обращения к электронным информационным источникам при помощи других информетрических законов»** охватывает еще три информетрических закона, при помощи которых осуществляется моделирование процессов обращения к электронным онлайновым ресурсам.

Параграф 4.1 посвящен применению модели, базирующейся на законе Леймкулера, который задает вид одноименной кривой выражением

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

где $x$ — доля наиболее продуктивных журналов, $y$ — доля статей, содержащихся в этих журналах, а $\beta > 0$ — эмпирический коэффициент, который необходимо найти при информетрическом моделировании.

Подбор коэффициента $\beta$ проводится при помощи нелинейной регрессии и метода наименьших квадратов. Получено очень хорошее приближение экспериментальных данных регрессионной кривой (коэффициент детерминации равен 0,937), что доказывает применимость модели Леймкулера. При этом $\beta$=5850.

Для дополнительной проверки модели, с ее помощью в аналитическом виде найдена площадь под кривой Леймкулера и оценен индекс Джини, который получился равным 0,77. Это вновь хорошо согласуется со значением $G$=0,80, полученным ранее в параграфе 3.3 численными методами.

В параграфе 4.2 рассмотрено применение информетрической модели Лотки к процессу обращения к электронным документам. Закон Лотки заключается в том, что если берется некоторое множество ученых и изучается, сколько статей написал каждый из них, то число $s_i$ ученых, написавших ровно $i$ публикаций, будет обратно пропорционально некоторой степени $i$:

$$s_i = Ai^{-\alpha}.$$

Показатель степени $\alpha$ при этом обычно близок к двум.

Методика переноса данного закона на ситуацию исследования спроса на онлайн-периодику находится в русле общего подхода настоящей диссертационной работы: заменим число авторов числом журналов, а число написанных авторами публикаций — числом открытых из данных журналов статей.

Для определения коэффициента распределения Лотки $\alpha$ построена зависимость числа журналов $s_i$, из которых открыто ровно $i$ статей, от $i$ (рис. 3). График показан в двойных логарифмических координатах, что позволяет по наклону регрессионной прямой определить $\alpha$. Здесь, однако, имеется серьезная проблема, которая вообще свойственна закону Лотки: если $i$ велико, то лишь для редких $i$ $s_i \neq 0$. Более того, журналов, из которых открыто фиксированное, при-

чем большое, число статей, всегда будет немного, 1 – 2, и поэтому на графике соответствующие им точки вытягиваются по ординатам $\ln 1 = 0$ и $\ln 2$. При этом таких точек много, более 100, в связи с чем они оказывают существенное влияние на построенную регрессионную прямую. В итоге регрессионное приближение получается не очень удачным (R-квадрат равен 0,74), а $\alpha \approx 0{,}96$, что крайне мало для закона Лотки.

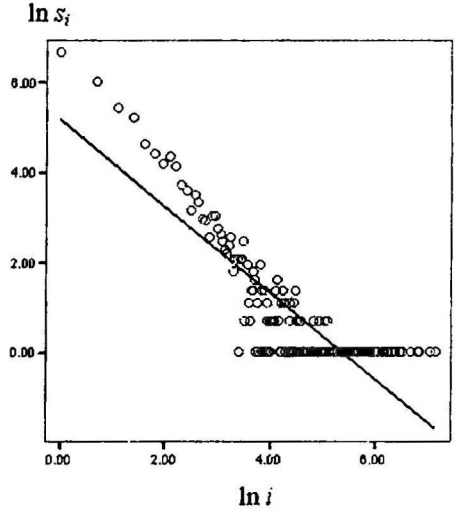


Рис. 3. Приближение экспериментальных данных законом Лотки

Автором диссертации предложено два различных способа коррекции данных, чтобы уйти от упомянутых проблем модели. В первом подобно тому, как поступал С. Наранан[10], из набора данных убираются точки такой продуктивности, что источников, ей соответствующих, только 1 или 2. Тогда остается 51 точка из 155, регрессионная кривая дает $\alpha \approx 1{,}37$, а коэффициент детерминации резко вырастает до 0,96. Второй способ состоит в том, что дискретные значения для журналов с большим числом открытых статей заменяются на аналог непрерывного распределения с соответствующей плотностью. Если имеются $s_j \neq 0$ и $s_i \neq 0$, $i > j$, причем $\forall k$, $j < k < i$, выполняется $s_k = 0$ (т. е. источников с продуктивностью больше $j$ и меньше $i$ нет), то на графике точки с координатами $(i; s_i)$ переносятся в точки $\left(\frac{i+j}{2} + 0{,}5;\ \frac{s_i}{i-j}\right)$.

[10] *Naranan S.* Power Law Relations in Science Bibliography – A Self-consistent Interpretation // Journal of Documentation. – 1971. – Vol. 27, No. 2. – P. 83–97.

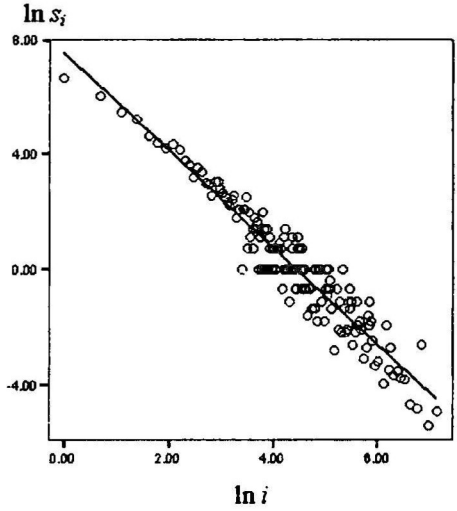


Рис. 4. Приближение скорректированных экспериментальных данных законом Лотки

Этот метод, в отличие от предыдущего, учитывает, например, что в области средней продуктивности источники с некоторым заданным $i$ встречаются чаще, чем в зоне максимальной продуктивности (меньше значений $i$ с $s_i = 0$), — эффект, полностью игнорируемый при простом исключении ряда источников с высокой продуктивностью. Для нового скорректированного распределения регрессия показана на рис. 4, а оценка $\alpha \approx 1{,}69$ уже близка к классической (R-квадрат равен 0,92). Можно говорить, что после необходимой корректировки данных в области точек с высокой продуктивностью модель Лотки хорошо описывает исследуемый нами процесс обращения к электронным информационным ресурсам.

Наконец, в параграфе 4.3 к процессу использования электронной информации применена информетрическая модель Ципфа.

Открытый в лингвистике закон Ципфа утверждает, что частота встречаемости слова $f$ обратно пропорциональна некоторой степени ранга $r$ этого слова (места в рейтинге слов, упорядоченных по частоте встречаемости), причем показатель степени близок к единице. Вновь вместо частоты встречаемости слова рассмотрим число открытых статей из журнала, а ранг слова заменим рангом журнала — номером в списке журналов, выстроенных в порядке убывания числа открытых из них статей.

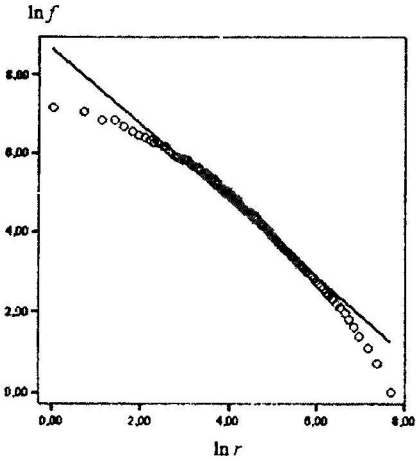


Рис. 5. Приближение скорректированных экспериментальных данных законом Ципфа

В двойных логарифмических координатах построена зависимость $f(r)$, для данного случая — зависимость числа открытых статей от ранга журнала. В результате получен показатель степени 1,35. Однако, в отличие от закона Лотки, который с трудом применим без корректировки данных к журналам с большой продуктивностью, закон Ципфа имеет недостатки в области источников с малой продуктивностью: число журналов с одной или двумя открытыми статьями велико и все они, по логике, должны иметь один ранг. Для отражения этого факта проведем логичную замену: источники, имеющие одинаковую продуктивность, отразим одной точкой с соответствующей продуктивностью и рангом, равным среднему рангу всех точек. Построим новый график $f(r)$ в двойных логарифмических координатах (рис. 5).

Наклон регрессионной прямой дал показатель степени в зависимости $f(r)$ 0,97, т. е. $f$ обратно пропорционально $r^{0,97}$. Это практически точное совпадение с классическим видом закона Ципфа. При этом R-квадрат регрессии равен 0,96. Таким образом, для наших экспериментальных данных информетрическая модель Ципфа применима в ее классическом варианте.

**Заключение** диссертационной работы подводит итоги информетрического моделирования процесса обращения к электронным информационным ресурсам при помощи различных законов, содержит выводы о пригодности большинства из них к такого рода исследованиям и указывает перспективы дальнейшего применения информетрии к моделированию процессов использования научной информации.

## СПИСОК ОПУБЛИКОВАННЫХ РАБОТ

***Глава в монографии:***

1. **Писляков В. В.** Использование онлайновых ресурсов и управление электронной подпиской в библиотеке ГУ-ВШЭ // Университетские инновации: опыт Высшей школы экономики / Я. И. Кузьминов, В. В. Радаев, А. А. Яковлев и др.; под ред. Я. И. Кузьминова. – М.: ГУ-ВШЭ, 2006. – С. 160–178.

***Статьи в журналах:***

2. **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.

3. **Писляков В. В.** Спрос на электронные журналы в университетской библиотеке: работает ли правило Парето? // Научно-техническая информация. Сер. 1. – 2005. – № 12. – С. 27–32.

4. **Зибарева И. В.** Библиометрический анализ журнала «Успехи химии» / И. В. Зибарева, В. В. Писляков, Т. Н. Теплова, О. М. Нефёдов // Вестник Российской академии наук. – 2008. – Т. 78, № 6. – С. 490–499.

5. **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.

***Материалы конференций:***

6. **Писляков В. В.** Правило Парето и статистика использования электронных журналов в университетской библиотеке // Библиотеки и информационные ресурсы в современном мире науки, культуры, образования и бизнеса: Труды Одиннадцатой международной конференции «Крым-2004». – http://www.gpntb.ru/win/inter-events/crimea2004/disk/doc/276.pdf.

7. **Писляков В. В.** Системы сбора и анализа статистики использования электронных ресурсов: сравнительный обзор // Библиотеки и информационные ресурсы в современном мире науки, культуры, образования и бизнеса: Труды Двенадцатой международной конференции «Крым-2005». – http://www.gpntb.ru/win/inter-events/crimea2005/disk/51.pdf.


Отпечатано с готового оригинал-макета
в типографии Издательства
Казанского государственного университета
Тираж 100 экз. Заказ 65/9

420008, ул. Профессора Нужина, 1/37
тел.: 231-53-59, 292-65-60

10-