

**Правило Парето и статистика использования
электронных журналов в университетской библиотеке**
Pareto Rule and Statistics of Digital Journals Demand in a University Library

**Правило Парето та статистика використання
електронних журналів в університетській бібліотеці**

Писляков В. В.

*Библиотеки Государственного университета — Высшей школы экономики,
Москва, Россия*

Vladimir V. Pislyakov

Library of State University — High School of Economics, Moscow, Russia

Писляков В. В.

Бібліотеки Державного університету — Вищої школи економіки, Москва, Росія

Рассмотрено эмпирическое правило Парето, описывающее отклонение от равномерности в различных процессах. Правило применено к процессу пользования электронными журналами в университетской библиотеке. Проанализированы сведения по статистике востребованности зарубежных периодических изданий из баз данных EBSCO, JSTOR и ProQuest за 2003 г., а также обобщенные данные по всем трем ресурсам. Построены графики распределения спроса, установлено, что правило Парето хорошо выполняется для процесса использования электронной периодики из онлайн-баз данных.

The presentation reviews Pareto empiric rule describing the deviation from uniformity in different processes. The rule is applied to the process of digital journals usage in a university library. Statistics of demand on foreign periodic from EBSCO, JSTOR and ProQuest is analysed for the year 2003, including combined statistics on all three resources. Charts of demand's distribution are built and it is proven that Pareto rule is well applied to the demand on the foreign periodical publications from on-line databases.

Розглянуто емпіричне правило Парето, що описує відхилення від рівномірності в різноманітних процессах. Правило застосовано до процесу користування електронними журналами в університетській бібліотеці. Проаналізовано відомості за статистикою попиту на зарубіжні періодичні видання з баз даних EBSCO, JSTOR та ProQuest за 2003 р., а також узагальнено дані за всіма трьома ресурсами. Побудовано графіки розподілення попиту, встановлено, що правило Парето добре виконується для процесу використання електронної періодики з онлайн-баз даних.

1. Формулировка правила Парето и краткий обзор литературы

В 1897 году итальянский экономист и социолог Вильфредо Парето сформулировал эмпирический принцип, согласно которому 80% материальных благ приходятся на 20% населения (соответственно, остальные 80% людей довольствуются 1/5 частью материальных богатств, имеющихся у человечества). Подобно тому как распределение Гаусса описывает целый спектр разнообразных процессов, принцип Парето применим ко многим явлениям совершенно разного порядка. Для примера можно привести следующие «отражения» данного правила:

- 20% клиентов приносят компании 80% дохода
- за 20% рабочего времени делается 80% работы
- 20% посетителей конференции «Крым-2004» обеспечивают 80% из всей явки на доклады и сообщения

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

Повторим, что данное правило эмпирическое, является попыткой приблизительно описать сложные неравномерные процессы и должно проверяться в каждом конкретном случае. Кроме того, сама пропорция может оказаться иной — не 80—20, а, например, 90—10 или 70—30. Правило Парето указывает именно на значительное отклонение от пропорции 50—50 в различных системах, а не конкретную величину отклонения.

Впервые правило в применении к библиотекам было рассмотрено в 1969 г. Р. Трюсвеллом [1], который пришел к выводу, что принцип Парето хорошо выполняется для статистики выдачи литературы: примерно 80% требований пользователей относилось к 20% библиотечного фонда.

В дальнейшем подобные исследования проводились неоднократно. Так, например, в отношении печатных периодических изданий в статье [2] авторы из Северной Каролины (США) приходят к выводу, что 80% спроса пришлось всего на 15% наименований журналов (стоимость подписки на которые при этом составила 38% от всего бюджета на периодику), причем точно такую же пропорцию 80—15 наблюдала в Канаде сотрудница библиотеки Трентского университета М. Щильяно [3]. В последнее время появились аналогичные работы по статистике использования *электронных* изданий. В масштабном исследовании деятельности консорциума OhioLINK (всего за период наблюдения было открыто более полутора миллиона электронных статей) оказалось, что спрос распределен более равномерно и 80% загруженных статей относятся к 40% журналов [4]. Сходные результаты получены в проекте РЕАК, во время которого предоставлялся доступ к 1100 электронным журналам издательства Elsevier для консорциума из 12 крупных и средних библиотек: 80% открытых файлов содержали статьи из 37% представленных журналов. 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 [5]. Наконец, в последних работах [6, 7] на основе аналогичного изучения статистического распределения использования наименований журналов и применения элементов теории нечетких множеств развивается метод выделения «ядра», «основного блока» периодики, востребованного той или иной академической организацией.

Целью настоящего доклада является проведение исследования распределения спроса на электронную периодику по журнальным наименованиям и проверка правила Парето для небольшой московской университетской библиотеки.

2. Условия исследования

Библиотека Государственного университета — Высшей школы экономики (ГУ-ВШЭ) создана в 1994 г. и является небольшой по меркам московских университетских библиотек: в настоящий момент печатный фонд составляет ок.45 тыс. наименований и ок.250 тыс. экземпляров. Три года назад началось развитие электронной подписки на базы данных зарубежной периодики, в период с апреля по ноябрь 2001 г. была осуществлена подписка на ресурсы компаний JSTOR, ProQuest и EBSCO, которые до сих пор являются «ядром» электронного фонда библиотеки и используются наиболее активно.

Необходимо сказать о «составляющих» каждого ресурса, включенного в анализ. До февраля 2003 г. ГУ-ВШЭ имел доступ лишь к одной «коллекции журналов» электронной библиотеки JSTOR — «Arts & Sciences I». В феврале подписка была пополнена коллекциями «Arts & Sciences II», «Business», «Language & Literature» и общее число журналов JSTOR, доступных в полных текстах, достигло 293-х наименований (416 с учетом смены заглавий в течение истории журнала). В течение 2003 г. также произошло расширение доступа к ресурсам компании ProQuest. В июле в подписке появились базы данных «PQ Psychology Journals» и «PQ Computing», что в сочетании с доступными прежде «ABI/Inform Global», «Academic Research Library», «PQ Education Journals» и «Social Sciences PlusText» дало ок.3000 полнотекстовых периодических изданий. Полнотекстовые ресурсы компании EBSCO, находящиеся в подписке ГУ-ВШЭ, оставались неизменными в течение периода эксперимента и включали базы данных «Academic Search Premier», «Business Source Premier», «MasterFILE Premier», «Newspaper Source», «Regional Business News» и «Health Source», в сумме ок.5800 полнотекстовых журналов.

В общей сложности в данных электронных библиотеках находятся в полнотекстовом доступе, за вычетом дубликатов, ок.7300 периодических изданий. В ГУ-ВШЭ неоднократно проводились тренинги по использованию всех перечисленных ресурсов, активность обращений к ним достаточно высока: в течение периода наблюдения (2003 г.) в совокупности было отправлено более 39000 поисковых запросов и открыто более 33000 полнотекстовых статей из EBSCO, ProQuest и JSTOR. Это позволяет проводить обобщения и обеспечивает статистическую значимость результатов исследования.

3. Методика

В двух из трех указанных ресурсов (а именно, в EBSCO и ProQuest) содержатся не только полнотекстовые журналы, но также и издания, представленные лишь аннотациями статей. Тем не менее в настоящей работе нас будут интересовать только обращения к полным текстам, т. к. именно такое использование ресурсов можно считать «полноценным», его статистика легче поддается интерпретации и более свободна от «случайных» заходов пользователей.

В настоящее время любая развитая электронная библиотека предоставляет развернутую статистику активности пользователей, оформленную в виде веб-страниц или файлов данных, готовых для импорта в Excel или СУБД. Помимо прочих статистических отчетов можно запросить распределение спроса по наименованиям использованных периодических изданий, которому и будет уделено особое внимание в настоящей работе.

Для единообразного представления результатов необходимо выбрать показатель использования, подсчитываемый статистическими модулями всех трех 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 представления статистики по использованию электронных ресурсов [8].

Журналы идентифицировались по ISSN (при наличии), издания с различными ISSN считались разными (не проводилось объединение данных при смене журналом названия и ISSN). При отсутствии ISSN идентификация проводилась по наименованию. В EBSCO все журналы без ISSN учитываются единой строкой «Unknown». Такое свойство отчетов EBSCO как преграду для полноценного анализа статистики характеризуют и коллеги из РГБ [9]. В нашем случае количество обращений к журналам без ISSN равняется 467 и составляет 4,4% от суммарной активности пользователей. Хотя данный показатель является третьим в списке наиболее спрашиваемых журналов EBSCO, небольшая доля от общего числа обращений позволяет считать, что влияние на общие выводы исследования будет минимальным. Строка «Unknown», таким образом, при обработке не учитывалась.

В качестве временного промежутка, на котором собиралась статистика обращений к базам данных, выбран весь 2003 г., с января по декабрь.

4. Результаты

В таб.1 приведены проценты журналов в каждой базе данных, из которых была открыта хотя бы одна статья. В столбце «число полнотекстовых изданий» приводятся лишь приблизительные цифры для EBSCO и ProQuest: специфика данных ресурсов, т. н. «агрегаторов», не позволяет установить с точностью до единиц число доступных журналов.

Таблица 1

	число полнотекстовых изданий	изданий, к полным текстам которых были обращения	доля востребованных изданий
EBSCO	5800	1105	19%
ProQuest	3000	1282	42%
JSTOR	416	296	71%
всего	7300	2143	29%

Как видно из таблицы, далеко не все издания из электронных библиотек используются. При этом у «агрегаторов» (EBSCO и ProQuest) меньший процент востребованных журналов. Наблюдается обратная зависимость процента востребованных изданий от суммарного их количества в ресурсе.

Ведущие пятерки для каждой электронной библиотеки и суммарное количество открытых полнотекстовых статей приведены в таб.2.

Таблица 2

журнал из EBSCO	открыто статей	журнал из ProQuest	открыто статей	журнал из JSTOR	открыто статей
Euroweek	1813	The Economist	278	American Economic Review	1376
Harvard Business Review	1045	Wall Street Journal	242	Journal of Political Economy	847
Journal of Marketing Research	344	American Sociological Review	102	Journal of Finance	698
Futures: News, Analysis & Strategies for Futures, Options & Derivatives Traders	324	Computerworld	86	Quarterly Journal of Economics	639
Journal of Management Accounting Research	297	Journal of Economic Issues	82	Review of Economic Studies	438
...
всего	10133	всего	8028	всего	13734

Далее следует преобразовать полученные данные в цифры, отражающие количество открытых статей в двух наиболее используемых журналах, в трех наиболее используемых и т. д., получая в итоге зависимость числа открытых статей от количества журналов, взятых из N верхних строк таблицы востребованности изданий. Отразим полученную зависимость на трех графиках, показанных на рис. 1.

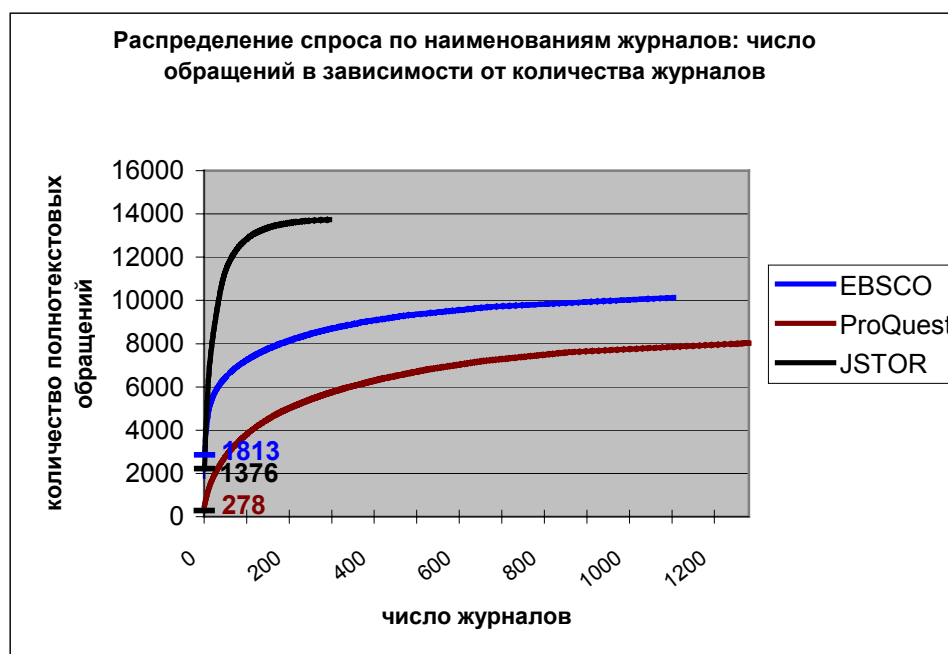


Рис. 1

Следующим шагом нормируем графики по обеим осям, выбрав 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. В результате получим графики, изображенные на рис. 2.

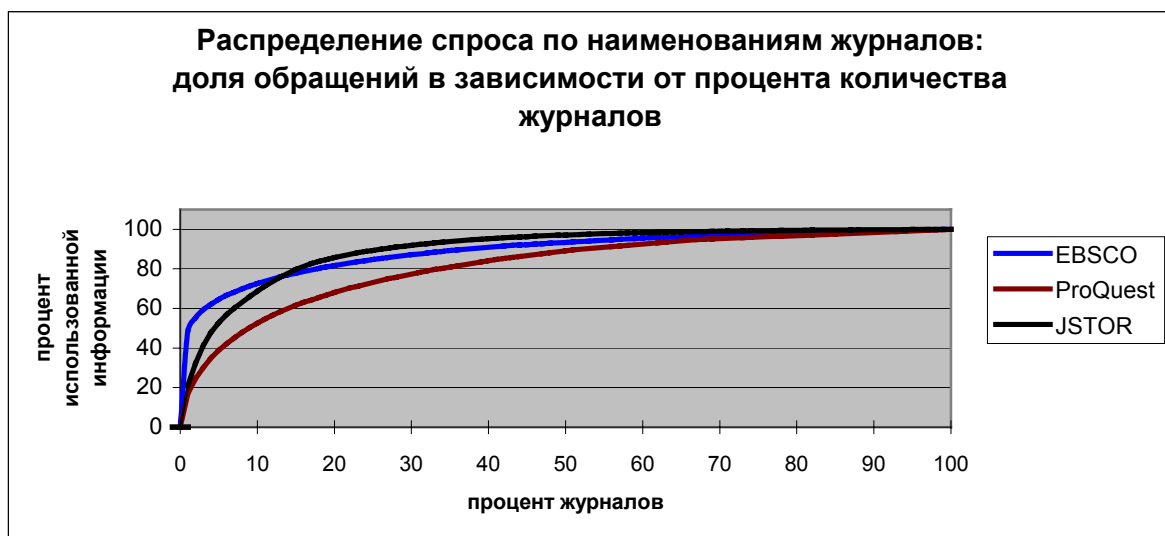


Рис. 2

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 двух ресурсов правило выполняется еще более жестко, чем в отношении 80—20. Для ProQuest, наоборот, пропорция мягкая, около 80—35 (80% востребованных статей опубликованы примерно в 35% журналов). Точные цифры приведены в таб.3.

Наконец, объединим данные по всем трем ресурсам (просуммировав показатели для тех журналов, которые представлены более чем в одной базе данных) и построим аналогичный график, рис.3.

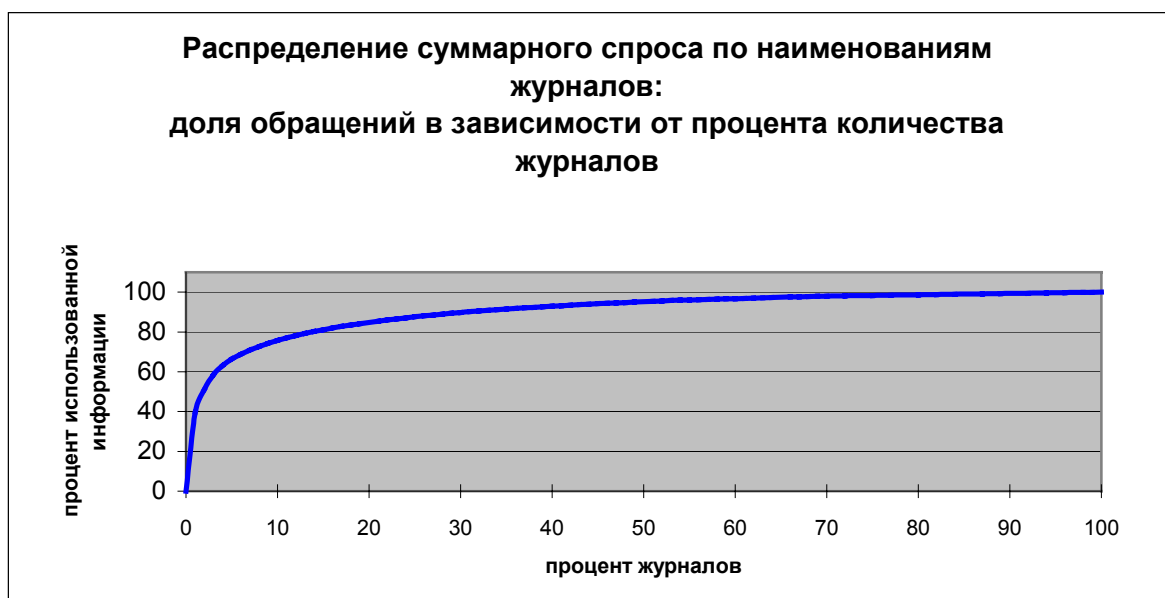


Рис. 3

Итоговые цифры сведены в таб.3. Из представленных данных можно сделать вывод, что принцип Парето не выполняется (вернее, плохо выполняется) только для одной базы данных — ProQuest. Для EBSCO и JSTOR, а также для всего информационного массива в целом пропорция Парето даже более «жесткая», чем классическое отношение 80—20. Наиболее неравновесное распределение спроса на информацию наблюдается именно для электронной подписки, рассмот-

ренной как целое: 80% открытых статей из всех трех баз данных опубликованы в 14% журналов из числа тех, которые имеются в подписке и хотя бы один раз были востребованы пользователями. Если учитывать еще и те журналы, которые не были востребованы ни разу, пропорция только еще более ужесточится.

Таблица 3

	востребовано информации	из числа журналов
EBSCO	80%	18%
ProQuest	80%	34%
JSTOR	80%	16%
все ресурсы	80%	14%

В качестве развития настоящего исследования можно предложить проведение аналогичного анализа для консорциума библиотек (например, консорциума Неикон) или для распределения статистики по пользователям: какой процент читателей университета потребляет 80% от используемых организацией полнотекстовых материалов из онлайн-баз данных. Подобные исследования могли бы прояснить, как устроен спрос на электронную периодику, насколько эффективно используются электронные журналы нашими пользователями и каких действий по повышению эффективности это требует от нас, библиотекарей.

Литература

1. *Trueswell R. L.* Some Behavioral Patterns of Library Users: The 80/20 Rule // *Wilson Library Bulletin*. Vol.43 (January 1969). P.458—461.
2. *Pongracz S., Ellern G. D., Newsome N.* Collection Development and a Long-Term Periodical Use Study: Methodology and Implications // *Serials Review*. Vol.28, Iss.1 (Spring 2002). P.38—44.
3. *Scigliano M.* Serial Use in a Small Academic Library: Determining Cost-Effectiveness // *Serials Review*. Vol.26, Iss.1 (April 2000). P.43—52.
4. *Diedrichs C. P.* E-journals: the OhioLINK experience // *Library Collections, Acquisitions, and Technical Services*. Vol.25, Iss.2 (Summer 2001). P.191—210.
5. *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.
6. *Egghe L., Rousseau R.* A Proposal to Define a Core of a Scientific Subject: A Definition Using Concentration and Fuzzy Sets // *Scientometrics*. Vol.54, No.1 (2002). P.51—62.
7. *Burrell Q. L.* Defining a Core: Theoretical Observations on the Egghe-Rousseau Proposal // *Scientometrics*. Vol.57, No.1 (2003). P.75—92.
8. Counter Code of Practice, Release 1: December 2002 // http://www.projectcounter.org/code_practice.html#start, секция 4.
9. *Афанасьева Т. Г., Давыдова Н. Р., Елисина Е. Ю., Литвинова Н. Н.* Использование сетевых ресурсов в Российской государственной библиотеке (доклад на Международной конференции «Культура: от информации к знанию». Москва, 7—9 апреля 2003 г.).