

**СОДЕРЖАНИЕ**

<b>Ван К.</b> Х-индекс: предлагаемый новый показатель для определения научного влияния ученого	3
<b>Бжезински М.</b> Степенные зависимости в распределениях ссылок: свидетельства из базы данных Scopus	8
<b>Захеди З., Костас Р., Уотерс П.</b> Насколько хорошо разработаны альтернативы? Междисциплинарный анализ наличия «альтернативных метрик» в научных публикациях ,	20

## РЕДАКЦИОННАЯ КОЛЛЕГИЯ

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

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

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

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

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

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

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

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

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

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

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

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

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

# X-индекс: предлагаемый новый показатель для определения научного влияния ученого\*

Ксяо Джун Ван

(Xiaojun WAN)

Институт вычислительной техники  
и технологии, ведущая лаборатория  
вычислительной лингвистики  
Министерства образования,  
Пекинский университет, г. Пекин, Китай

*H-индекс стал самым популярным показателем для количественного определения научного влияния ученого в различных областях науки. H-индекс определяется как наибольшее число статей с числом ссылок  $\geq h$ , и он рассматривает каждую ссылку одинаково. Однако различные ссылки, как правило, происходят из разных статей с различным влиянием и качеством, а ссылка из высоковлиятельной статьи является большим признанием целевой статьи, чем ссылка из обычной статьи. Основываясь на этом предположении, мы предлагаем новый, названный x-индексом, показатель для определения научного влияния ученого, учитывая при этом только ссылки, взятые из влиятельных статей. X-индекс определяется как наибольшее число статей с числом влиятельных ссылок  $\geq x$ , где каждая влиятельная ссылка происходит из статьи, для которой среднее число ссылок на статью ее авторов (ACNPP — Average Citation Number Per Paper)  $\geq x$ . Посредством анализа на множестве данных APS (American Physical Society — Американское физическое общество) мы обнаружили, что предлагаемый x-индекс имеет лучшую способность провести различие между обладателями наград в области физики и рядовыми учеными-физиками.*

## ВВЕДЕНИЕ

Как разумно оценить научное влияние ученого является важной и требующей внимания проблемой в областях информатики и наукометрии. Цель достижения высокого научного влияния представляется долговременной задачей ученого в его карьере, а само научное влияние ученого имеет важное значение для его карьерной деятельности — получение необходимого финансирования, продвижение по служебной лестнице и завоевание наград. Так как научные исследования или статьи, публикуемые ученым, служат его наиболее очевидными научными результатами и достижениями, то научное воздействие ученого обычно измеряется анализом влияния его опубликованных научных исследований или статей.

До сегодняшнего дня многие показатели разрабатывались для определения научного влияния различных ученых на основе их опубликованных статей в научном сообществе. Среди разнообразия показателей самым популярным является *h*-индекс [1]. *h*-индекс определяется как наибольшее число статей с числом ссылок  $\geq h$ . Он полагается на число ссылок, полученных каждой

статьей, и если ученый опубликовал больше статей с большими полученными ссылаками, тогда *h*-индекс ученого является высоким. Мы можем видеть, что *h*-индекс обращается с каждой полученной ссылкой одинаково, однако различные ссылки обычно происходят из разных цитирующих статей, некоторые из них являются более влиятельными, чем другие. Ссылка из высоковлиятельной цитирующей статьи называется влиятельной ссылкой, которая, как правило, получает большее признание цитируемой статьи, чем ссылка из обыкновенной цитирующей статьи. Поэтому мы стремимся рассматривать только влиятельные ссылки, полученные ученым, чтобы оценить его научное влияние. Мы используем среднее ACNPP (среднее число ссылок на статью) авторов для цитирующей статьи, чтобы оценить влияние статьи. На основе этого предположения мы предлагаем новый показатель, названный *x*-индексом, для определения научного влияния ученого. *x*-индекс определяется как наибольшее число статей с числом влиятельных ссылок  $\geq x$ , где каждая влиятельная ссылка происходит из цитирующей статьи, для которой среднее ACNPP ее авторов  $\geq x$ . Предлагаемый нами *x*-индекс является вариантом *h*-индекса с контролем качества ссылок, и *x*-индекс является более точным, чем *h*-индекс.

Посредством качественного анализа на множестве данных APS мы обнаружили, что предлагаемый *x*-индекс

\* Перевод Wan X. X-index: A fantastic new indicator for quantifying a scientist's scientific impact. — 2014.—  
<http://arxiv.org/ftp/arxiv/papers/1405/1405.0641.pdf>

способен лучше различить обладателей наград по физике и рядовых ученых-физиков, чем  $h$ -индекс и некоторые другие показатели.

## СВЯЗАННЫЕ/ РОДСТВЕННЫЕ РАБОТЫ

До сих пор некоторые различные показатели предлагались для оценки научного влияния ученого или исследователя на основе ссылок от статьи к статье. Самым популярным является  $h$ -индекс, определяемый как число статей с числом ссылок  $\geq h$ , который становится наиболее широко используемым показателем для характеристики научного выхода ученого [1]. Этот показатель уже применялся в практических службах, таких как Google Scholar и ISI Web of Science. В соответствии со стохастической моделью, предложенной в [2],  $h$ -индекс является приблизительно прямолинейным относительно протяженности карьеры, логарифмической скорости публикаций и логарифмической скорости ссылок. Предсказательная сила  $h$ -индекса эмпирически сравнивается с другими показателями (общим подсчетом ссылок, числом ссылок на статью и общим подсчетом статей) и демонстрирует лучший прогноз будущих научных достижений [3], чем другие показатели.

Кроме того, несколько вариантов показателей было предложено на основе  $h$ -индекса, включая  $g$ -индекс [4],  $a$ -индекс [5],  $hg$ -индекс [6],  $r$ -индекс [7],  $m$ -индекс [8],  $b_a$ -индекс [9],  $h^2$ -индекс [10],  $h'$ -индекс [11],  $h_{int}$ -индекс [12],  $ar$ -индекс [13],  $h_n$ -индекс [14],  $p$ -индекс [15],  $j$ -индекс [16] и так далее. Почти все эти показатели осуществляют некоторые модификации относительно  $h$ -индекса. Например,  $g$ -индекс придает больший вес высокоцитируемым статьям и определяется как наибольшее число  $g$  статей, которые вместе получили  $g^2$  или больше ссылок. Был проведен ряд исследований, чтобы проанализировать или сравнить некоторые из этих показателей [8,17-19]. Однако все указанные выше показатели не рассматривают качество ссылок, а ссылки, происходящие из разных статей, написанных различными учеными, обрабатываются единообразно.

## ПРЕДЛАГАЕМЫЙ НАМИ ПОКАЗАТЕЛЬ

Так как предлагаемый нами показатель является показателем в стиле Хирша, мы должны пристальнее посмотреть на  $h$ -индекс. Оригинальное определение  $h$ -индекса следующее [1]: «Ученый имеет индекс  $h$ , если  $h$  из его  $N_p$  статей имеют, по крайней мере,  $h$  ссылок каждая, а другие  $(N_p - h)$  статьи имеют  $\leq h$  ссылок каждая».

$h$ -индекс вычисляется на основе ссылок, полученных каждой опубликованной ученым статьей, и интуиция, стоящая за  $h$ -индексом, заключается в том, что если ученый имеет больше влиятельных статей, то этот ученый имеет высокий  $h$ -индекс. Это влияние статьи измеряется числом полученных ею ссылок.

Мы можем видеть, что различные ссылки, полученные статьями, обрабатываются одинаково в вычислениях  $h$ -индекса. Однако различные ссылки в статье обычно происходят из различных цитирующих статей, и эти статьи, как правило, имеют различное влияние. Например, ссылка из цитирующей статьи, написанной ведущим ученым, обычно считается большим признанием цитируемой статьи, чем ссылка из цитирующей статьи, написанной студентом младшего курса. То есть, различные ссылки в статье не имеют равного признания статьи, и разное признание может быть использовано с выгодой для определения качества научного влияния ученого.

В целях отражения различных влияний различных ссылок, полученных ученым, мы предлагаем новый показатель –  $x$ -индекс – для определения качества научного влияния ученого.

**$x$ -индекс:** Ученый имеет индекс  $x$ , если  $x$  из его  $N_p$  статей получили, по крайней мере,  $x$  влиятельных ссылок каждая, а другие  $(N_p - x)$  статьи получили  $\leq x$  влиятельных ссылок каждая, в которых влиятельная ссылка происходит из цитирующей статьи, чьи авторы имеют среднее  $ACNPP \geq x$ .

Наш предлагаемый  $x$ -индекс рассматривает только влиятельные ссылки, полученные ученым, и игнорирует все другие ссылки. Влияние ссылки измеряется научным влиянием цитирующей статьи, предложившей эту ссылку, а научное влияние статьи измеряется средним  $ACNPP$  ее авторов. Отметим, что  $ACNPP$  является простой и широко используемой метрикой для оценки научного влияния автора, и она просто вычисляется делением общего числа ссылок, полученных автором, на общее число статей этого автора.

Интуиция, стоящая за  $x$ -индексом, заключается в том, что если ученый имеет больше статей, цитируемых более влиятельными учеными, то этот ученый имеет высокий  $x$ -индекс. Как видно из этого определения,  $x$ -индекс является более точным показателем, чем  $h$ -индекс, и  $x$ -индекс ученого всегда является меньшим или равным  $h$ -индексу ученого. По сравнению с  $h$ -индексом  $x$ -индекс больше фокусируется на качестве ссылок, чем на их числе, и таким образом считается, что он имеет лучшую способность определить научное влияние ученого.

## КОЛИЧЕСТВЕННЫЙ АНАЛИЗ

В экспериментах мы используем массив данных APS, предоставленный Американским физическим обществом. Этот массив данных открыто доступен (<https://publish.aps.org/datasets>) и содержит свыше 45 тыс. статей, опубликованных с 1893 г. вплоть до 2009 г. в многообразии основных журналов по физике, включая *Physical Review Letters*, *Physical Review* и *Reviews of Modern Physics*. Он содержит основные метаданные всех статей, включая следующие области: идентификатор цифрового объекта (DOI), название, авторы, место работы авторов, журнал, том, номер, публикационная история и т.д. Этот массив также содержит все пары статей APS, цитирующих друг друга. Например, если статья А цитирует статью В, то будет запись в массиве данных, содержащая пару DOI для А и В.

Для сравнения предложенного нами  $x$ -индекса с другими показателями мы должны взять точку отсчета. Подобно авторам работы [20] мы оцениваем показатели, сравнивая их способность различать обладателей наград по физике и рядовых ученых-физиков. Обладатели наград широко признаны в физическом сообществе, и они, как правило, имеют более высокое научное влияние, чем рядовые физики. Мы ранжируем физиков в соответствии с каждым показателем, и чем больше обладателей наград ранжируется выше рядовых физиков по показателю, тем лучше показатель. Мы берем обладателей Нобелевской премии по физике и семи основных наград в области физики (медаль Больцмана, премия Дэни Хайнемана, медаль Маттеуччи, медаль им. Макса Планка, премия Сакураи, премия Вольфа и премия Энрико Ферми) после 1960 г. Эти обладатели наград, как правило, имеют более высокое научное влияние, чем рядовые физики, и ожидается, что они будут ранжированы выше, нежели рядовые физики.

В этом массиве данных имя иногда может относиться к различным авторам, а иногда различные имена могут относиться к одному и тому же автору в силу неопределенностей имен. Как говорится в работе автора [20], мы не устраняем неопределенности имен авторов, так как эта задача требует большей информации, которая не доступна в массиве APS, а применение существующих алгоритмов устранения неопределенности имен далеко от совершенства. Вместо этого мы просто сокращаем каждое полное имя автора и приводим его в каноническую форму, состоящую из фамилии и первой буквы его имени. Например, «Джон С. Матер» будет сокращено до «Дж. Матер». Таким образом, различные имена, относящиеся к одному и тому же лицу, будут обработаны правильно. Иногда различные авторы, имеющие одинаковое имя, могут быть обработаны неправильно, т.е. как одно лицо, но, к счастью, такие ситуации встречаются нечасто, особенно для обладателей наград. После такого сокращения имен всего насчитывается 183 459 авторов в массиве APS и 162 736 из них были процитированы, по крайней мере, однажды. Мы получили список из 122 обладателей Нобелевской премии по физике и 315 обладателей других основных наград по физике, а всего имеется 403 обладателя наград, так как некоторые лауреаты Нобелевской премии также завоевали другие награды по физике.

Все ученые-физики в массиве данных APS ранжированы в соответствии с каждым показателем, и если имеется связь между многочисленными физиками, мы повторно ранжируем их с помощью общих чисел ссылок на них. Сравниваем  $x$ -индекс и другие показатели в сфере информационного поиска. Мы рассматривали обладателей наград как стандартный «золотой» массив, ранжировали всех ученых-физиков согласно показателю и затем применили обычную оценочную метрику информационного поиска для оценки ранжированного списка массива «золотого» эталона. В целях оценки мы использовали следующую метрику:

**P@N**: измеряет точность верхних  $N$  ученых-физиков в ранжированном списке ( $N$  обычно равняется 10~100), что вычисляется следующим образом:

$$P @ N = \frac{\text{Число обладателей премии в верхних } N \text{ ученых-физиках}}{N}$$

**AP**: сначала определяет точность в каждой позиции, когда каждый обладатель награды найден в ранжированных результатах, а если обладатель премии не найден в массиве «золотого» стандарта, то соответствующая точность равна 0 и в итоге значения точности являются усредненными.

Помимо  $h$ -индекса мы также сравниваем предложенный нами  $x$ -индекс с ACNPP (среднее число ссылок на

статью), TCN (общее число ссылок) и TPN (общее число статей).

Сравнение результатов, основанных на различных метриках, показано в табл. 1 и 2. В табл. 1 рассматриваются только обладатели Нобелевской премии как массив «золотого» эталона, тогда как в табл. 2 и обладатели Нобелевской премии, и обладатели других наград считаются массивом «золотого» эталона.

Как видно из табл. 1 и 2, и  $x$ -индекс, и  $h$ -индекс превосходят простые показатели, включая ACNPP, TCN и TPN. Более интересно то, что ранжированные результаты, основанные на  $x$ -индексе, действуют гораздо лучше, чем результаты на основе  $h$ -индекса во всех оценочных метриках, а точность оценок даже повысилась на 50~100%. Результаты показывают, что  $x$ -индекс может использоваться, чтобы поместить большее число обладателей наград в верхнюю часть ранжированного списка. Учитывая тот факт, что обладатели наград, как правило, имеют более высокое научное влияние, чем рядовые ученые-физики,  $x$ -индекс показывает большую способность проводить различие между обладателями наград и рядовыми учеными-физиками. Следовательно,  $x$ -индекс имеет заметные преимущества перед  $h$ -индексом.

Табл. 3 приводит 20 верхних ученых-физиков, ранжированных по  $x$ -индексу и  $h$ -индексу, в массиве APS, соответственно. Оценки  $x$ -индекса и  $h$ -индекса этих физиков также представлены в таблице. В числе верхних 20 физиков, ранжированных по  $x$ -индексу, четверо являются обладателями Нобелевской премии (P. Anderson, S. Weinberg, W. Ketterle и F. Wilczek) и пятеро представляют обладателей других основных наград по физике (M. Fischer, S. Weinberg, B. Halperin, P. Zoller и F. Wilczek) и имеется всего 7 обладателей наград. Тогда как в числе верхних 20 физиков, ранжированных по  $h$ -индексу, только двое являются Нобелевскими лауреатами (P. Anderson и S. Weinberg), а четверо являются обладателями других основных наград по физике (M. Fischer, S. Weinberg, B. Halperin и P. Zoller) и имеется всего пять обладателей наград. Результаты показывают, что  $x$ -индекс может действительно больше ранжировать обладателей наград, опережая рядовых физиков.

Также мы можем наблюдать в табл. 3, что оценки  $x$ -индекса вообще более низкие, чем оценки  $h$ -индекса, это можно приписать тому факту, что  $x$ -индекс является более точным, чем  $h$ -индекс, с помощью добавления моментов относительно качества ссылок. Например,  $h$ -индекс P. Anderson 53, тогда как его  $x$ -индекс падает до 23. Несмотря на то, что S. Kim имеет самый высокий  $h$ -индекс, он выпадает из верхнего списка, ранжированного по  $x$ -индексу. В целом, предложенный нами  $x$ -индекс имеет лучшую способность показывать научное влияние ученого, чем  $h$ -индекс.

Таблица 1

### Сравнение результатов, основанных на обладателях Нобелевской премии

	$x$ -индекс	$h$ -индекс	ACNPP	TCN	TPN
<b>P@10</b>	<b>0,3</b>	0,2	0	0	0
<b>P@20</b>	<b>0,2</b>	0,1	0	0,05	0
<b>P@30</b>	<b>0,2</b>	0,0667	0	0,1	0,0333
<b>P@40</b>	<b>0,175</b>	0,05	0	0,075	0,025
<b>P@50</b>	<b>0,2</b>	0,08	0	0,06	0,02
<b>P@100</b>	<b>0,16</b>	0,06	0	0,04	0,02
<b>AP</b>	<b>0,0484</b>	0,0206	0,0046	0,0146	0,0046

Сравнение результатов, основанных на всех обладателях наград (обладатели Нобелевской премии и обладатели семи основных наград)

	<i>x</i> -индекс	<i>h</i> -индекс	ACNPP	TCN	TPN
<b>P@10</b>	<b>0,6</b>	0,3	0	0	0
<b>P@20</b>	<b>0,35</b>	0,25	0,05	0,1	0
<b>P@30</b>	<b>0,3667</b>	0,2333	0,0333	0,1333	0,0333
<b>P@40</b>	<b>0,3</b>	0,175	0,025	0,15	0,025
<b>P@50</b>	<b>0,3</b>	0,2	0,02	0,18	0,02
<b>P@100</b>	<b>0,25</b>	0,12	0,02	0,11	0,02
<b>AP</b>	<b>0,0482</b>	0,0283	0,0119	0,0255	0,0099

Таблица 3

Верхние 20 авторов в массиве данных APS

Ранжированные по <i>x</i> -индексу		Ранжированные по <i>h</i> -индексу	
имя	<i>x</i> -индекс	имя	<i>h</i> -индекс
<i>P. Anderson</i> *	23	S.Kim	53
V. Kosteletsky	23	<i>P. Anderson</i> *	53
<i>M. Fischer</i> #	22	M. Cohen	52
<i>S. Weinberg</i> *#	22	<i>M. Fischer</i> #	52
P. Lee	21	<i>S. Weinberg</i> *#	52
<i>B. Halperin</i> #	21	H. Kim	51
M. Cohen	20	C. Wang	49
J. Cirac	20	P. Lee	48
<i>P. Zoller</i> #	20	L. Pfeiffer	48
<i>W. Ketterle</i> *	20	K. West	48
D. Fischer	20	D. Kim	47
M. Kastner	20	<i>B. Halperin</i> #	47
A. Suzuki	19	R. Wagner	46
L. Pfeiffer	19	D. Brown	46
K. West	19	J. Cirac	46
J. Schrieffer	19	<i>P. Zoller</i> #	45
T. Rice	19	S. Lee	44
D. Scalapino	19	J. Wang	44
<i>F. Wilczek</i> *#	19	A. Zunger	44
X. Wen	19	S. Louie	43

\* — указывает на обладателей Нобелевской премии,

# — указывает на обладателя другой основной премии по физике

## ЗАКЛЮЧЕНИЕ И БУДУЩАЯ РАБОТА

В этой статье мы включили качество ссылки в *h*-индекс и предложили новый *x*-индекс для количественного определения научного влияния ученого. На основе количественного анализа массива APS мы обнаружили, что *x*-индекс имеет лучшую способность показывать научное влияние ученого-физика, чем *h*-индекс.

В будущем мы будем применять предложенный нами *x*-индекс для количественного определения научного влияния ученых в других дисциплинах, чтобы показать его надежность. Более того, *x*-индекс можно легко адаптировать к количественному определению научного влияния журналов и организаций.

*Благодарность.* Эта работа была поддержана Национальным фондом естественных наук Китая (№ 6110166 и № 61331011), а также Программой Китая в сфере национальных исследований и разработок в высокой технологии (Программа 863) (№ 2012AA011101).

## ЛИТЕРАТУРА

1. *Hirsch J.* An index to quantify an individual's scientific research output// Proceedings of the National Academy of Sciences. — 2005.—Vol.102, No.46. — P. 16569-16572.
2. *Burrell Q. L.* Hirsch's h-index: A stochastic model// Journal of Informetrics.— 2007.— Vol. 1, No.1.— P. 16-25.
3. *Hirsch J. E.* Does the h index have predictive power?// Proceedings of the National Academy of Sciences.— 2007.— Vol. 104, No.49.— P. 19193-19198.
4. *Egghe L.* Theory and practice of the g-index// Scientometrics.— 2006.— Vol. 69, No. 1.— P. 131-152.
5. *Jin B. H.* H-index: an evaluation indicator proposed by scientist// Science Focus.— 2006.— Vol. 1, No.1.— P. 8-9.

6. *Alonso S., Cabrerizo F. J., Herrera-Viedma E., Herrera F.* hg-index: A new index to characterize the scientific output of researchers based on the h-and g-indices// *Scientometrics*.— 2009.— Vol. 82, No.2.— P. 391-400.
7. *Jin B., Liang L., Rousseau R., Egghe L.* The R-and AR-indices: Complementing the h-index// *Chinese science bulletin*.— 2007.— Vol. 52, No.6.— P. 855-863.
8. *Bornmann L., Mutz R., Daniel H. D.* Are there better indices for evaluation purposes than the h index? A comparison of nine different variants of the h index using data from biomedicine// *Journal of the American Society for Information Science and Technology*.—2008.— Vol. 59, No.5.— P. 830-837.
9. *Van Eck N. J., Waltman L.* Generalizing the h- and g-indices// *Journal of Informetrics*.— 2008.— Vol. 2, No.4.— P. 263-271.
10. *Kosmulski M.* A new Hirsch-type index saves time and works equally well as the original h-index// *ISSI newsletter*.— 2006.—Vol. 2, No.3.— P. 4-6.
11. *Sidiropoulos A., Katsaros D., Manolopoulos Y.* Generalized Hirsch h-index for disclosing latent facts in citation networks// *Scientometrics*. — 2007.— Vol. 72, No.2.— P. 253-280.
12. *Ruane F., Tol R. S.* Rational (successive) h-indices: An application to economics in the Republic of Ireland// *Scientometrics*.— 2008.— Vol. 75, No.2.— P. 395-405.
13. *Jin B.* The AR-index: Complementing the h-index// *ISSI newsletter*. — 2007.— Vol. 3, No.1.— P. 6.
14. *Egghe L., Rousseau R.* An h-index weighted by citation impact// *Information Processing & Management*. — 2008.— Vol. 44, No. 2.— P. 770-780.
15. *Prathap G., Gupta B. M.* Ranking of Indian universities for their research output and quality using a new performance index// *Curr. Sci*.— 2009.— Vol. 97, No.6.— P. 751-752.
16. *Todeschini R.* The j-index: A new bibliometric index and multivariate comparisons between other common indices// *Scientometrics*.— 2011.— Vol. 87, No.3.— P. 621-639.
17. *Alonso S., Cabrerizo F. J., Herrera-Viedma E., Herrera F.* h-Index: A review focused in its variants, computation and standardization for different scientific fields// *Journal of Informetrics*. — 2009.— Vol. 3, No.4.— P. 273-289.
18. *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, No.3.— P. 193-203.
19. *Bar-Ilan J.* Which h-index?—A comparison of WoS, Scopus and Google Scholar// *Scientometrics*.— 2008.— Vol. 74, No. 2.— P. 257-271.
20. *Jiang X., Sun X., Zhuge H.* Graph-based algorithms for ranking researchers: Not all swans are white!// *Scientometrics*.— 2013.— Vol. 96.— P. 743-759.

# Степенные зависимости в распределениях ссылок: свидетельства из базы данных Scopus\*

**Михаил БЖЕЗИНСКИ**  
(Michał BRZEZINSKI)

Факультет экономических наук,  
Варшавский университет,  
г. Варшава, Польша

*Моделирование распределений ссылок к научным статьям является решающим моментом для понимания того, как развивается наука. Однако существует значительная эмпирическая полемика относительно того, какая статистическая модель лучше всего подходит для распределений ссылок. Данная статья касается точного эмпирического обнаружения поведения степенного закона в распределении ссылок, полученных наиболее высокоцитируемыми научными статьями. Использовался большой массив новых данных, касающихся ссылок к научным статьям, опубликованным в 1998 – 2002 гг. и взятым из Scopus. Модель степенной зависимости сравнивается с рядом альтернативных моделей, при этом используются вероятностные сравнительные признаки. Обнаружено, что гипотеза степенной зависимости отвергается для почти половины научных областей в Scopus. Для этих научных областей распределение Yule, степенной закон с экспоненциальным выключением и логарифмически нормальное распределение, кажется, лучше подходят для данных, чем чистая (строгая) модель степенной зависимости. С другой стороны, когда гипотеза степенной зависимости не отвергается, она обычно является эмпирически неразличимой от большинства альтернативных моделей. Чистая модель степенной зависимости, вероятно, представляется самой лучшей моделью только для наиболее высокоцитируемых статей в журнале «Physics and Astronomy». Кроме того, наши результаты, кажется, поддерживают теории, подразумевающие, что наиболее высокоцитируемые научные статьи следуют распределению Yule, степенной зависимости с экспоненциальным выключением или логарифмически нормальным распределением. Наши результаты также предполагают, что степенные законы в распределениях ссылок, когда они имеют место, отчитываются только за очень небольшую часть опубликованных статей (менее 1% для большинства научных областей) и что пересчетный параметр (экспонент) степенной зависимости значительно выше (примерно от 3,2 до 4,7), чем обнаружено в более старой литературе.*

## ВВЕДЕНИЕ

В наукометрии, социальной физике и других науках часто говорится, что распределения некоторых научных позиций (например, статей, ссылок), созданных некими научными источниками (например, авторами, журналами), имеют тяжелые хвосты распределе-

ния, которые могут быть смоделированы с использованием моделей степенной зависимости. Эти распределения, как потом говорится, должны соответствовать закону Лотки [1]. Примеры таких распределений включают авторскую продуктивность, встречаемость слов, ссылки, полученные статьями, узлы социальных сетей, число авторов на статью, распространение научной литературы в журналах и многое другое [2]. Действительно модели степенной зависимости широко используются во многих науках, таких как физика,

\* Перевод Brzezinski M. Power laws in citation distributions: Evidence from Scopus. – <http://arxiv.org/pdf/1402.3890.v1.pdf>

биология, науки о Земле и планетарные науки, экономика, финансы, вычислительная наука и другие [3,4]. Модели, эквивалентные закону Лотки, известны как закон Парето в экономике [5] и закон Ципфа в лингвистике [6]. Соответствующее измерение и обеспечение научных объяснений для степенных законов играет важную роль в понимании поведения многочисленных естественных и социальных феноменов.

Данная статья связана с эмпирическим определением поведения степенного закона в распределении ссылок, полученных научными статьями. Распределение ссылок в соответствии со степенным законом для высокоцитируемых статей впервые было предложено Прайсом [7], также обеспечившим механизм «кумулятивного предпочтения», который мог бы генерировать распределение степенного закона [8]. Совсем недавно растущий поток литературы по этому направлению расширился, нацеливаясь на измерение степенных зависимостей в правых хвостах распределения ссылок. В частности, Реднер [9,10] обнаружил, что правые хвосты распределения ссылок для статей, опубликованных в журнале «Physical Review» более чем за сто лет, и статей, опубликованных в 1981 г. в журналах, охваченных Scientific's Web of Science (WoS), следуют степенным законам. Последний массив данных был также смоделирован с помощью методов степенного закона Клозетом и др. [4] и Петерсоном и др. [11]. Последнее исследование также использовало данные из списка 2007 г. ныне живущих ученых-химиков с наивысшим h-индексом и «Physical Review D» за 1975-1994 г. Ван Раан [12] наблюдал, что верхняя часть распределения, примерно 18 тыс. статей, опубликованных в 1991-1998 гг. в области химии в Нидерландах, следует степенному закону в процессе распределения. Модели степенной зависимости также подходили для данных из области физики высоких энергий [13], данных для наиболее цитируемых физиков [14], данных для всех статей, опубликованных в журналах Американского физического общества за 1983-2008 гг. [15], и данных для всех статей по физике, опубликованных в 1980-1989 гг. [16].

Недавно Альбаррэн и Руис-Кастильо [17] подвергли проверке поведение степенного закона, используя большой массив данных WoS, состоящий из 3,9 млн. статей, опубликованных в 1998-2002 гг. и распределенных по 22 категориям научных областей WoS. Тот же самый массив данных также использовался для определения наличия степенных зависимостей в правом хвосте распределения ссылок, классифицированных в 219 научных подобластях WoS [18, 19]. Эти исследования предлагают самый большой существующий набор свидетельств относительно поведения степенного закона в сфере распределения ссылок. Отсюда вытекают три главных вывода. Первый – поведение степенного закона не является универсальным. Существование степенного закона не может отвергаться в данных WoS для 17 из 22 и для 140 из 219 подобластей в [17] и в [18,19], соответственно. Второй – в противоположность предшествующим исследованиям эти статьи обнаружили, что пересчетный параметр (экспонент) распространения степенного закона составляет 3,5 в большинстве случаев, тогда как более старая литература предполагает, что значение параметра находится между 2 и 3 [19]. Третий – степенные зависимости в распределении ссылок довольно-таки небольшие – в среднем они охватывают около 2 % наиболее высокоцитируемых статей в данной области науки в WoS и отвечают примерно за 13,5 % всех ссылок в данной области.

Основная цель данной статьи – использовать статистически строгий подход, чтобы ответить на эмпирический вопрос, описывает ли модель степенной зависимости наблюдаемое распределение высокоцитируемых статей наилучшим образом. Мы используем набор статистических средств для определения поведения степенного закона, предложенного Клозетом и др. [4]. Имеются два основных достоинства настоящей статьи. Во-первых, мы рассматриваем очень большой, не используемый прежде массив данных о распределении ссылок наиболее высокоцитируемых статей в ряде научных областей. Этот массив данных взят из Scopus, библиографической базы данных, созданной в 2004 г. издательством Elsevier и содержащей 2,2 млн. статей, опубликованных в 1998-2002 гг. и классифицированных в Scopus по 27 основным предметным научным областям. Большинство предыдущих исследований использовало довольно небольшие массивы данных, которые не подходили для строгого статистического определения поведения степенной зависимости. В противоположность этому наша выборка даже значительнее в отношении наиболее высокоцитируемых статей, чем сравнительно большая выборка, использованная в недавних статьях, основанных на данных WoS [17-19]. Это вытекает из того факта, что Scopus индексирует примерно на 70 % больше источников по сравнению с WoS [20, 21] и, следовательно, дает более исчерпывающий охват распределения ссылок.\*

Второе основное достоинство данной статьи состоит в том, чтобы обеспечить строгое статистическое сравнение модели степенной зависимости и ряда альтернативных моделей относительно проблемы, какое теоретическое распределение лучше подходит эмпирическим данным по ссылкам. Проблема выбора модели ранее изучалась в ряде работ в имеющейся литературе. Утверждалось, что модели, подобные растянутому экспоненциальному распределению [14], распределению Yule [8], логарифмически нормальному распределению [10, 23, 24], распределению Tsallis [25-27] или измененной степенной зависимости [15], в равной степени или даже лучше подходят для распределения ссылок, чем чистая (строгая) модель степенной зависимости. Однако предыдущие статьи либо фокусировались на одиночном альтернативном распределении, либо использовали только визуальные методы для выбора между конкурирующими моделями. Настоящая статья заполняет пробел обеспечением систематического и статистически строгого сравнения распределения со степенной зависимостью с такими альтернативными моделями, как логарифмически нормальное, экспоненциальное, растянутое экспоненциальное (Weibull) распределение, распределения Tsallis, Yule и распределения степенной зависимости с экспоненциальным выключением. Сравнение между моделями осуществлялось с использованием сравнительных признаков правдоподобия [4, 28].

\* С точки зрения измерения степенных зависимостей, наиболее важной частью распределения является правый хвост. Кажется, что база данных, используемая в данной статье, имеет лучший охват правого хвоста распределений ссылок. Наиболее высокоцитируемая статья в нашей базе данных получила 5 187 ссылок (см. табл. 2), тогда как соответствующее число для базы данных, основанной на WoS, составляет 4 461 [22]. Наша база данных далее описывается в разделе «Материалы и методы».

## МАТЕРИАЛЫ И МЕТОДЫ

### Соответствие модели степенной зависимости данным ссылок

При выборе методов для соответствия степенной зависимости распределением ссылок мы следуем Клозету и др. [4]. Эти авторы с осторожностью показывают, что в общем соответствующие методы зависят от того, являются ли данные непрерывными или дискретными. В нашем случае последнее справедливо, поскольку ссылки являются неотрицательными целыми числами. Пусть  $x$  будет числом ссылок, полученных статьей в определенной научной области. Функция плотности вероятности (probability density function – pdf) модели дискретной степенной зависимости определяется как

$$p(x) = \frac{x^{-\alpha}}{\zeta(\alpha, x_0)}, \quad (1)$$

где  $\zeta(\alpha, x_0)$  является обобщенной функцией или дзета-функцией Гурвица.  $\alpha$  является очерченным параметром распределения степенной зависимости, известным как экспонент (показатель) степенной зависимости или пересчетный параметр. Поведение степенной зависимости, как правило, обнаруживается только для значений, больших, чем некий минимум, обозначенный как  $x_0$ . В случае распределений ссылок поведение степенной зависимости было обнаружено в среднем только в 2 % верхней части всех статей, опубликованных в какой-то научной области [18, 19].

Поэтому, следовательно, должна быть вычислена граница поведения степенной зависимости,  $x_0$ , если мы хотим точно измерить, в какой части распределения ссылок применяется модель. Кроме того, нам необходим подсчет  $x_0$ , если мы желаем получить беспристрастный подсчет экспонента степенной зависимости,  $\alpha$ .

Мы вычисляем  $\alpha$ , используя подсчет максимального правдоподобия (ML - maximum likelihood). Функция логарифмического правдоподобия следующая:

$$L(\alpha) = -n \ln \zeta(\alpha, x_0) - \alpha \sum_{i=1}^n \ln x_i \quad (2)$$

Подсчет максимального правдоподобия для  $\alpha$  найден с помощью численной максимизации (2).\*

Придерживаясь Клозета и др. [4], мы используем следующую процедуру для вычисления нижней границы поведения степенной зависимости,  $x_0$ . Для каждого  $x \geq x_{min}$  мы определяем оценку максимального правдоподобия экспонента степенной зависимости,  $\alpha$ , а затем выводим хорошо известную статистику Колмогорова-Смирнова (Kolmogorov-Smirnov statistic - KS) для данных и соответствующей модели. Статистика KS определяется следующим образом:

$$KS = \max_{x \geq x_0} |S(x) - P(x; \hat{\alpha})|, \quad (3)$$

где  $S(x)$  является функцией кумулятивного распределения (cumulative distribution function – cdf) для наблюдений со значением, по крайней мере,  $x_0$ , а  $P(x, \hat{\alpha})$  является функцией распределения (cdf) для соответствующей

модели степенной зависимости для наблюдений, для которых  $x \geq x_0$ . Оценка  $\hat{x}_0$  затем выбрана в качестве значения  $x_0$ , для которого статистика KS является минимальной. Стандартные ошибки для обоих подсчитанных параметров,  $\hat{\alpha}$  и  $x_0$ , вычисляются стандартными методами с параметрической компенсацией погрешностей с 1 000 повторений опыта для увеличения объема выборки.

### СООТВЕТСТВИЕ И ПРОВЕРКА ВЫБОРА МОДЕЛИ

Следующий шаг в измерении степенных зависимостей включает проверку соответствия. Позитивный результат такой проверки (теста) позволяет сделать вывод, что модель степенной зависимости совместима с данными. И опять, следуя Клозету и др. [4], мы используем тест, основанный на полупараметрическом, с компенсацией погрешностей, подходе.\* Процедура начинается с определения соответствия модели степенной зависимости данным и подсчета статистики KS для этого соответствия,  $k$ . Затем собирается большое число массивов синтетических данных, которые следуют оригинальной модели степенной зависимости с подсчитанным  $\hat{x}_0$  выше и имеет то же самое распределение не степенной зависимости, как и оригинальный массив данных, где  $\hat{x}_0$  ниже. Тогда модель степенной зависимости подходит каждому массиву собранных данных, используя те же методы, как и для оригинального массива данных, и подсчитывается статистика KS. Часть массивов данных, для которых их собственная статистика KS больше, чем  $k$ , является значением  $p$  теста. Это представляет вероятность того, что KS статистика, подсчитанная для данных, взятых из модели степенной зависимости, подходящей для оригинальных данных, является, по меньшей мере, такой же большой, как  $k$ . Гипотеза степенной зависимости отвергается, если значение  $p$  меньше, чем некоторый выбранный порог. Следуя Клозету и др. [4], мы исключаем возможность использования модели степенной зависимости, если вычисленное значение  $p$  для этого теста меньше 0,1. В данной статье мы используем 1 тыс. собранных массивов данных.

Если тест на соответствие отвергает гипотезу степенной зависимости, мы можем сделать вывод, что степенная зависимость не обнаружена. Однако, если массив данных в достаточной степени отвечает требованиям степенной зависимости, остается вопрос, существует ли альтернативное распределение, которое в равной степени хорошо или лучше подходит для этого массива данных. Поэтому нам следует приспособить некоторые конкурирующие распределения и оценить, какое распределение дает лучшее соответствие. Для этой цели мы используем тест соотношения правдоподобия, который проверяет, действительно ли сравниваемые модели в равной степени близки к оригинальной модели по сравнению с альтернативной, той, которая ближе всего. Тест вычисляет логарифм отношения правдоподобий данных (LR – logarithm of the ratio) по двум конкури-

\* Клозет и др. [4] также обеспечивают приблизительный метод подсчета  $\alpha$  для дискретной модели степенной зависимости, предполагается, что непрерывные распределенные числа степенной зависимости округляются до ближайших целых чисел. Однако в данной статье мы используем точный подход, основанный на максимизации (2).

\* Если наши данные взяты из заданной модели, тогда мы должны использовать статистику KS в проверке соответствия, поскольку распределение статистики KS известно в таком случае. Однако, когда подлежащая рассмотрению модель не известна или когда ее параметры оценены на основе данных, являющихся нашим случаем, распределение статистики KS должно быть получено с помощью моделирования.

рующим распределениям, который негативно или позитивно зависит от того, какая модель лучше соответствует данным. Давайте конкретно рассмотрим два распределения с pdf (функция плотности вероятностей), обозначенных  $p_1(x)$  и  $p_2(x)$ . LR (логарифм отношения) определяется как:

$$LR = \sum_{i=1}^n [\ln p_1(x_i) - \ln p_2(x_i)] \quad (4)$$

Позитивное значение LR предполагает, что модель  $p_1(x)$  лучше соответствует данным. Однако условное обозначение LR может быть использовано для определения того, какая модель предпочтительна, только если LR значительно отличается от нуля. Вуонг [28] показал, что в случае не вложенных моделей нормализованное логарифмическое отношение правдоподобия (NRL – normalized log-likelihood ratio)  $NRL = n^{1/2} LR / \sigma$ , где  $\sigma$  – вычисленное стандартное отклонение LR, имеет лимитированное стандартное нормальное распределение.\* Этот результат может быть использован для вычисления значения  $p$  для теста, делающего различие между конкурирующими моделями. Если значение  $p$  маленькое (например, меньше 0,1), тогда условному обозначению LR можно, по всей видимости, доверять как показателю, какая модель предпочтительна. Однако, если значение  $p$  большое, тогда тест не способен сделать выбор между сравниваемыми распределениями.

Мы следовали Клозету и др. [4] при выборе следующих альтернативных дискретных распределений: экспоненциальное, растянутое экспоненциальное (Weibull), логарифмически нормальное, распределение Yule и степенной зависимости с экспоненциальным выключением.\*\* Большинство этих моделей рассматривалось в предыдущей литературе по моделированию распределения ссылок. В качестве другой альтернативы мы также использовали распределение Tsallis, которое тоже было предложено в качестве модели для распределения ссылок [26, 27]. Определения наших альтернативных распределений даны в табл. 1.

Мы следовали Клозету и др. [4] при выборе следующих альтернативных дискретных распределений: экспоненциальное, растянутое экспоненциальное (Weibull), логарифмически нормальное, распределение Yule и степенной зависимости с экспоненциальным выключением.\*\* Большинство этих моделей рассматривалось в предыдущей литературе по моделированию распределения ссылок. В качестве другой альтернативы мы также использовали распределение Tsallis, которое тоже было предложено в качестве модели для распределения ссылок [26, 27]. Определения наших альтернативных распределений даны в табл. 1.

Таблица 1

Определения альтернативных дискретных распределений

Название распределения	Функция плотности вероятности
Экспоненциальное	$(1 - e^{-\lambda})e^{\lambda x_0} e^{-\lambda x}$
Растянутое экспоненциальное (Weibull)	$\frac{1}{\sum_{x_0}^{\infty} (q^{x^\beta} - q^{(x+1)^\beta})} q^{(x+1)^\beta}$
Логарифмически нормальное	$\sqrt{\frac{2}{\pi\sigma^2}} \left[ \operatorname{erfc}\left(\frac{\ln x_0 - \mu}{\sqrt{2}\sigma}\right) \right]^{-1} \frac{1}{x} \exp\left[-\frac{(\ln x - \mu)^2}{2\sigma^2}\right]$
Tsallis	$\frac{1}{\sum_{x_0}^{\infty} (1 + x/\sigma)^{-\theta-1}} (1 + x/\sigma)^{-\theta-1}$
Yule	$(\alpha - 1) \frac{\Gamma(x_0 + \alpha - 1)}{\Gamma(x_0)} \frac{\Gamma(x)}{\Gamma(x + \alpha)}$
Степенная зависимость с экспоненциальным выключением	$\left(\sum_{x_0}^{\infty} x^{-\alpha} e^{-\lambda x}\right)^{-1} x^{-\alpha} e^{-\lambda x}$

*Примечание:* Распределения были нормализованы для обеспечения того, чтобы суммарная вероятность в области  $[x_0, +\infty]$  составляла 1. Дискретное логарифмически нормальное распределение аппроксимируется путем округления непрерывных логарифмически нормально распределенных действительных чисел до ближайших целых чисел. Для распределения Tsallis мы используем параметризацию, рассмотренную Шализи [29].

\* В случае вложенных моделей 2LR имеет предельное распределение хи-квадрат (chi-squared distribution).

\*\* Степенная зависимость с экспоненциальным выключением ведет себя подобно чистой модели степенной зависимости для меньших значений  $x$ ,  $x \geq x_0$ , в то время как для больших значений  $x$  она ведет себя подобно экспоненциальному распределению. Истинная модель степенной зависимости вложена в степенную зависимость с экспоненциальным выключением и по этой причине последняя всегда обеспечивает соответствие, по крайней мере, так же хорошо, как и первая.

## ДАННЫЕ

Мы используем данные ссылок из Scopus, библиографической базы данных, введенной издательством Эльзевир в 2004 г. Scopus является главным конкурентом наиболее широко используемого источника данных в литературе по моделированию распределения ссылок – Web of Science (WoS) издательства Thomas Reuters. Scopus охватывает 29 млн. записей со ссылками, идущими к 1996 г., и 21 млн. записей до 1996 г., идущих аж к 1823 г. Важное ограничение этой базы данных состоит в том, что она не включает процитированные ссылки для статей, предшествующих 1996 г. Scopus содержит 21 тыс. рецензируемых журналов от более 5 тыс. международных издателей. Она охватывает примерно на 70 % больше источников по сравнению с WoS [20], но большая часть дополнительных источников – это журналы незначительного влияния. Недавний обзор литературы показал, что достаточно обширная литература, которая сравнивает WoS и Scopus с точки зрения анализа ссылок, предлагает совокупные результаты [21]. Однако большинство исследований предполагает, что, по крайней мере, за период с 1996 г. и далее число ссылок в обеих базах данных или приблизительно одинаковое, или выше в Scopus, чем в WoS. Поэтому кажется, что Scopus составляет полезную альтернативу WoS с точки зрения моделирования распределения ссылок.

Журналы в Scopus классифицируются по четырем основным предметным областям: науки о жизни

(4,2 тыс. журналов), науки по здравоохранению (6,5 тыс. журналов), физические науки (7,1 тыс. журналов) и общественные науки, включая искусство и гуманитарные науки (7 тыс. журналов). Четыре основные предметные области далее подразделяются на 27 главных предметных областей и более 300 второстепенных предметных областей. Журналы могут классифицироваться более чем под одной предметной областью.

В данной статье анализ проводился на уровне 27 главных предметных областей науки, отраженных в Scopus.\* Из многообразия типов документов, содержащихся в Scopus, мы отобрали только статьи. С целью сравнимости с последними исследованиями, основанными на WoS [17, 18], рассматривались только статьи, опубликованные в период 1998-2002 гг.\*\* Для объяснения выбора пятилетнего интервала цитирования, общего для всех областей науки, см. [17].

Чтобы измерить поведение степенной зависимости ссылок, нам необходимы данные относительно правых хвостов распределения ссылок. С этой целью мы использовали Scopus Citation Tracker для сбора ссылок для минимума (400 000;  $x$ ) самых высокоцитируемых статей, где  $x$  – фактическое число статей, опубликованных в определенной научной области за 1998-2002 гг. Анализ проводился отдельно для каждой из 27 научных областей, классифицированных Scopus.

Дескриптивная статистика для наших массивов данных представлена в табл. 2.

Таблица 2

Дескриптивная статистика для распределения ссылок, Scopus, 1998-2002 гг., 5- летний интервал цитирования

Предметные научные области Scopus	Общее число статей	Число статей в выборке	% всех статей в выборке	Среднее число ссылок	Стандартное отклонение ссылок	Максимальное число ссылок
Сельскохозяйственные и биологические науки	372575	99804	26,8	15,17	14,36	628
Искусство и гуманитарные науки	47191	47074	99,8	1,256	3,357	91
Биохимия, генетика и молекулярная биология	636421	99819	15,7	49,09	46,29	3118
Бизнес, управление и учет	61211	61156	99,9	3,452	7,273	287
Химическое машиностроение	158673	98989	62,4	7,232	9,236	344
Химия	416660	99398	23,9	21,07	21,17	1065
Вычислительная наука	134179	99933	74,5	6,44	18,13	2737
Науки о принятии решений	27409	27393	99,9	3,467	5,496	143
Науки о Земле и планетарные науки	228197	99788	43,7	14,1	17,03	1195
Экономика, эконометрия и финансы	49645	49559	99,8	4,652	8,653	287
Энергетика	67076	66378	99,0	2,553	5,596	334
Машиностроение	439719	99765	22,7	11,77	15,83	971
Наука об окружающей среде	186898	99847	53,4	10,72	11,27	730
Иммунология и микробиология	195339	99858	51,1	22,11	25,11	926
Материаловедение	331310	99591	30,1	12,48	14,49	697
Математика	193740	99922	51,6	6,912	11,38	929

\* См. табл. 2 на предмет списка анализируемых Scopus областей науки.

\*\* Например, для статей, опубликованных в 1998 г., мы анализировали ссылки, полученные за 1998-2002 гг., а для статей, опубликованных в 2002 г., ссылки, полученные в 2002-2006 гг.

Предметные научные области Scopus	Общее число статей	Число статей в выборке	% всех статей в выборке	Среднее число ссылок	Стандартное отклонение ссылок	Максимальное число ссылок
Медицина	1191154	99823	8,4	48,55	60,14	4365
Нейронаука	445181	99886	22,4	18,97	20,39	771
Уход за больными	51283	50464	98,4	5,274	12,07	518
Фармакология, токсикология и фармацевтика	179427	99757	55,6	12,19	12,28	347
Физика и астрономия	541328	99817	18,4	24,75	31,64	3118
Психология	104449	99736	95,5	7,446	11,55	377
Общественные науки	215410	99890	46,4	6,148	8,055	519
Ветеринария	53203	53117	99,8	3,637	5,843	128
Стоматология	27470	27437	99,9	4,943	6,736	115
Профессии области здравоохранения	75491	75414	99,9	7,272	11,49	348
Многодисциплинарные области	50287	50226	99,9	30,38	76,08	5187
Все науки	6480926	2203841	34,0	14,92	27,74	5187

В некоторых случаях было менее 100 тыс. статей, опубликованных в какой-то научной области в период 1998-2002 гг., и нам надо было получить полные или почти полные распределения ссылок (см. колонки 2-4 табл. 2).\* В других случаях мы получили только часть релевантного распределения, охватывающего правый хвост и какую-то часть середины распределения. Самые маленькие порции распределений ссылок были получены для медицины (8,4 % всех статей), биохимии, генетики и молекулярной биологии (15,7 %) и физики и астрономии (18,4 %). Однако, используя данные WoS для 22 научных категорий, Альбаран и Руис-Кастильо [17] обнаружили, что степенные зависимости, как правило, отвечают только за менее, чем 2 % наиболее высокоцитируемых статей. Следовательно, кажется, что охват правых хвостов распределений ссылок в наших выборках является удовлетворительным для наших целей.

## РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

Табл. 3 представляет результаты дискретной модели степенной зависимости для наших массивов данных, состоящих из ссылок на научные статьи, опубликованные за период 1998-2002 гг. (с общим 5-летним интервалом цитирования), отдельно для каждой из 27 главных предметных научных областей, отраженных в Scopus. Последняя строка также дает результаты для всех предметных областей, объединенных под рубрикой «Все науки». Кроме подсчетов экспонента ( $\alpha$ ) и нижней границы поведения степенной зависимости ( $x_0$ ), таблица также показывает подсчитанное число и процентное отношение степенной зависимости распределенных статей, а также значение  $p$  для нашего теста относительно соответствия.

Результаты, касающиеся соответствия, предполагают, что гипотеза степенной зависимости не может отвергаться для следующих 14 научных областей Scopus: «Сельское хозяйство и биологические науки», «Биохимия, генетика и молекулярная биология», «Химическое машиностроение», «Химия», «Энергетика», «Науки об окружающей среде», «Материаловедение», «Нейронаука», «Уход за больными», «Фармакология, токсикология и фармацевтика», «Физика и астрономия», «Психоло-

\* Для всех анализируемых научных областей имелись некоторые статьи с отсутствием информации о ссылаках. Такие статьи были изъяты из наших выборок. Однако это, как правило, касалось примерно только 0,1 % наших выборок.

гия», «Профессии сферы здравоохранения» и «Многодисциплинарные области». Оставшиеся 13 научных областей Scopus, для которых модель степенной зависимости не принимается, включают гуманитарные и общественные науки («Искусство и гуманитарные науки», «Бизнес, управление и учет», «Экономика, эконометрия и финансы», «Общественные науки»), а также формальные науки («Вычислительная наука», «Науки о принятии решений», «Математика»), науки о жизни («Иммунология и микробиология», «Медицина», «Ветеринария», «Стоматология»), как и «Науки о Земле и планетарные науки» и «Машиностроение». Наилучшие соответствия для этих научных областей показаны на рисунке.

Для большинства распределений, показанных на рисунке, можно четко увидеть, что их правые хвосты распадаются быстрее, чем показывает чистая модель степенной зависимости. Это предполагает, что самые большие наблюдения для этих распределений скорее должны быть смоделированы с распределением, имеющим более слабый хвост, чем строгая модель степенной зависимости, им подобны логарифмически нормальная модель или модель степенной зависимости с экспоненциальным выключением.

Значение  $p$  для нашего теста на соответствие в случае «Все науки» составляет 0,076, что ниже нашего допустимого порога - 0,1. Однако это значение  $p$  не принимается в расчет и оно значительно выше, чем значения  $p$  для большинства из 13 научных областей Scopus, для которых мы отвергаем гипотезу степенной зависимости. По этой причине мы делаем вывод, что свидетельство не является в этом случае окончательным. Однако наш результат для графы «Все науки» находится в полном противоречии с результатом авторов работы [17], которые, используя данные WoS, обнаружили, что соответствие для определенного массива данных было очень надежным (со значением  $p = 0,85$ )\*.

\* В работе [17] гипотеза степенной зависимости считается правдоподобной для 17 из 22 научных областей WoS. Она отвергается для «Фармакологии и токсикологии», «Физики», «Сельскохозяйственных наук», «Строительства» и «Общественных наук в целом». Эти результаты не являются напрямую сравнимыми с таковыми настоящей статьи, так как Scopus и WoS используют различные системы классификации для категоризации журналов.

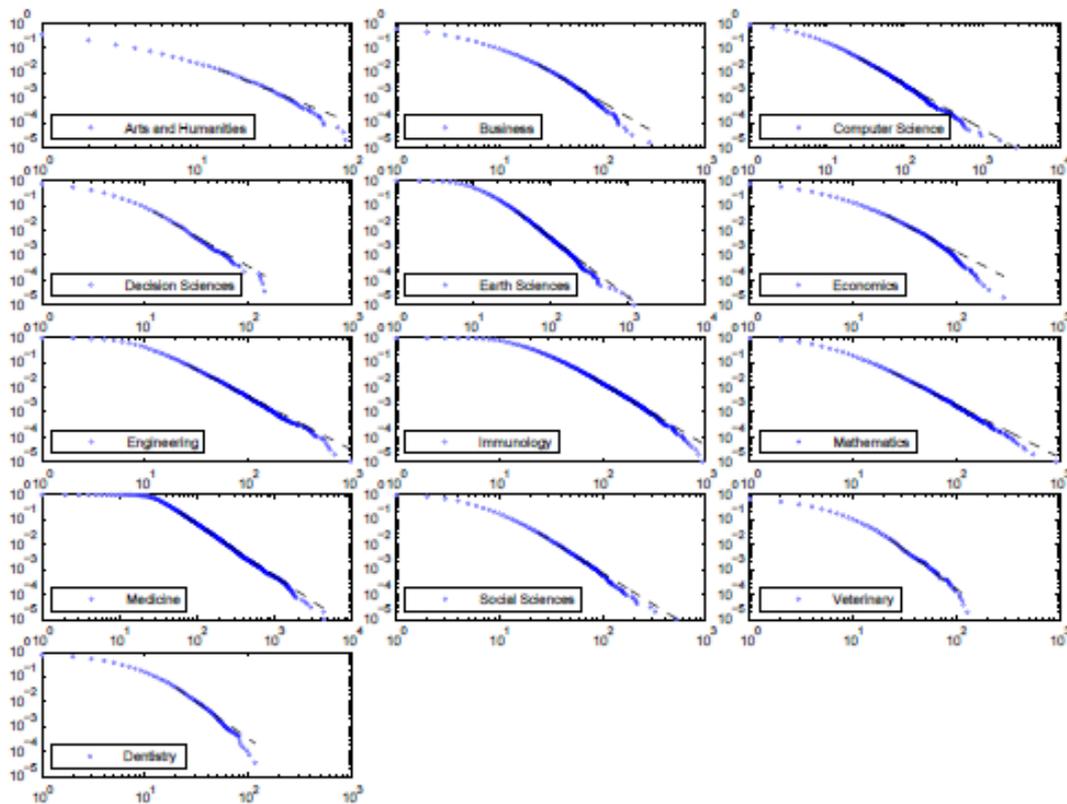


Рис. Дополнительные кумулятивные функции распределения (кружочки) и наилучшие соответствия степенной зависимости (прерывистая линия) для распределения ссылок, которые не прошли тест на соответствие, Scopus, 1998-2002 гг., 5-летний интервал цитирования.

Таблица 3

Соответствия степенной зависимости для распределений ссылок, Scopus, 1998-2002, интервал цитирования – 5 лет.

Предметные научные области Scopus	$\hat{x}_0$	$\hat{\alpha}$	Число статей, отвечающих степенной зависимости	% общего числа статей	Значение $p$
Сельскохозяйственные и биологические науки	92(15,1)	4,19 (0,25)	488	0,1	0,566
Искусство и гуманитарные науки	14 (5,4)	3,46 (0,47)	655	1,4	0,005
Биохимия, генетика и молекулярная биология	148 (28,0)	3,72 (0,13)	2813	0,4	0,175
Бизнес, управление и учет	24 (10,1)	3,4 (0,38)	1339	2,2	0,000
Химическое машиностроение	38 (6,7)	4,01 (0,19)	1418	0,9	0,099
Химия	41 (7,1)	3,4 (0,05)	8193	2,0	0,110
Вычислительная наука	26 (10,6)	2,78 (0,11)	3989	3,0	0,000
Науки о принятии решений	12 (4,0)	3,36 (0,24)	1596	5,8	0,000
Науки о Земле и планетарные науки	36 (8,9)	3,37 (0,09)	5834	2,6	0,000
Экономика, эконометрия и финансы	21 (10,2)	3,13 (0,36)	1995	4,0	0,000
Энергетика	32 (5,4)	3,91 (0,22)	356	0,5	0,825
Машиностроение	26 (9,4)	3,14 (0,09)	7986	1,8	0,000
Наука об окружающей среде	63 (10,3)	4,33 (0,22)	624	0,3	0,506
Иммунология и микробиология	78 (13,6)	3,48 (0,10)	2713	1,4	0,049
Материаловедение	43 (8,9)	3,47 (0,11)	2687	0,8	0,193

Предметные научные области Scopus	$\hat{x}_0$	$\hat{\alpha}$	Число статей, отвечающих степенной зависимости	% общего числа статей	Значение $p$
Математика	24 (4,0)	3,11 (0,06)	4152	2,1	0,012
Медицина	59 (16,3)	3,07 (0,04)	20163	1,7	0,000
Нейронаука	135 (28,4)	4,69 (0,41)	423	0,1	0,896
Уход за больными	60 (15,7)	3,68 (0,40)	439	0,9	0,256
Фармакология, токсикология и фармацевтика	56 (6,8)	4,1 (0,12)	1215	0,7	0,865
Физика и астрономия	61 (6,5)	3,35 (0,04)	5034	0,9	0,797
Психология	52 (8,8)	3,9 (0,17)	1060	1,0	0,812
Общественные науки	24 (6,4)	3,56 (0,15)	2963	1,4	0,007
Ветеринария	23 (4,0)	4,09 (0,27)	858	1,6	0,017
Стоматология	20 (2,4)	3,89 (0,18)	1012	3,7	0,011
Профессии области здравоохранения	49 (10,2)	3,85 (0,24)	942	1,2	0,352
Многодисциплинарные области	209 (40,4)	3,24 (0,14)	1147	2,8	0,100
Все науки	186(46,3)	3,45 (0,10)	6364	0,2	0,076

Примечание: Стандартные отклонения приводятся в скобках.

Подсчеты экспонента степенной зависимости для 14 научных областей Scopus, для которых степенная зависимость кажется правдоподобной гипотезой, колеблется в пределах от 3,24 до 4,59. Это хорошо согласуется с показателями авторов работы [17] и подтверждает их мнение, что истинное значение этого параметра в значительной степени выше, чем обнаруженное в более ранней литературе [9, 13, 25], которая предлагала подсчеты в пределах от 2,3 до примерно 3. Мы также подтверждаем свидетельство авторов работы [17] о том, что степенные зависимости в распределениях ссылок скорее небольшие – обычно они относятся к менее, чем 1 % общего числа статей, опубликованных в какой-либо научной области. Существуют только две области в нашем исследовании с несколько большими степенными зависимостями – «Химия» (2 %) и «Многодисциплинарные области» (2,8 %).

Сравнение между гипотезой степенной зависимости и альтернативами, использующими тест Вуонга (Vuong's test), представлено в табл. 4. Можно наблюдать, что экспоненциальная модель может быть исключена в большинстве случаев. Мы обсуждаем другие результаты в первую очередь для 13 научных областей Scopus, которые не прошли наш тест на соответствие. Для всех этих областей, за исключением «Ветеринарии», модель Yule и модель степенной зависимости с экспоненциальным выключением соответствуют данным лучше, чем чистая модель степенной зависимости в статистически значимом методе. Логарифмически нормальная модель проявляет себя лучше, чем чистая модель степенной зависимости, в 10 из обсуждаемых областей; то же самое относится к распределению Weibull в случае 5 областей. Однако эти результаты не подразумевают, что распределения, которые дают лучшее соответствие относительно данных, распределенных без учета степенной зависимости, чем строгая модель степенной зависимости, являются правдоподобными гипотезами для этих массивов данных. Необходимо и далее изучать эту проблему, используя соответствующие тесты на пригодность.

Теперь вернемся к результатам для оставшихся научных областей Scopus, которые не отвергались нашим тестом на соответствие. Гипотеза степенной зависимости, кажется, должна быть самой лучшей моделью только для «Физики и астрономии». В этом случае статистика, лежащая в основе критерия, всегда является неотрицательной, подразумевая, что модель степенной зависимости почти или даже лучше соответствует данным, чем каждая из альтернатив. Для оставшихся 13 научных областей логарифмически нормальная модель, модель Yule и модель степенной зависимости с экспоненциальным выключением всегда имеют более высокое правдоподобие, предполагая, что эти модели могут лучше соответствовать данным, нежели строгое распределение степенной зависимости. Однако только в нескольких случаях различия между моделями являются статистически важными. Для «Химии» и «Многодисциплинарных областей» модель Yule и модель степенной зависимости с экспоненциальным выключением являются предпочтительными по сравнению с чистой моделью степенной зависимости. Степенная зависимость с экспоненциальным выключением также предпочтительна в случае «Профессии в сфере здравоохранения». В других случаях значения  $p$  для теста на отношение правдоподобия больше, а это подразумевает, что нет окончательного свидетельства, которое бы позволило сделать различия между моделями распределения строгой степенной зависимости, логарифмически нормальной, Yule и степенной зависимости с экспоненциальным выключением. Сравнивая распределение степенной зависимости с распределениями Weibull и Tsallis, мы наблюдаем, что критерий статистики является позитивным в примерно половине случаев, но значения  $p$  всегда больше и ни одна модель не может быть исключена. Наши тесты относительно отношения правдоподобия поэтому предполагают, что когда степенная зависимость является правдоподобной гипотезой, соответствующей нашему тесту на пригодность, она часто неразличима от альтернативных моделей.

Тесты выбора модели для распределения сылок, Scopus, 1998-2002 гг., интервал цитирования – 5 лет

Предметные научные области, отраженные в Scopus	Значение p	Экспоненциальная		Weilbull		Логарифмически нормализованная		Tsallis		Yule		PL с включением	
		LR	p	LR	p	LR	p	LR	p	LR	p	NLR	p
Сельскохозяйственные и биологические науки	0,566	20,740	0,009	0,338	0,779	-0,096	0,782	0,054	0,890	-0,011	0,858	-0,268	0,464
Искусство и гуманитарные науки	0,005	6,287	0,457	-6,93	0,023	-6,56	0,025	-4,325	0,198	-1,38	0,000	-7,37	0,000
Биохимия, генетика и молекулярная биология	0,175	204,5	0,000	1,22	0,758	-1,12	0,473	-1,227	0,479	-0,155	0,108	-0,567	0,287
Бизнес, управление и учет	0,000	34,390	0,034	-9,60	0,013	-9,24	0,013	-7,279	0,065	-1,39	0,000	-9,98	0,000
Химическое машиностроение	0,099	69,480	0,001	-0,021	0,994	-0,972	0,480	0,025	0,990	-0,358	0,187	-0,78	0,211
Химия	0,110	736,0	0,000	7,48	0,262	-2,67	0,204	1,290	0,687	-0,999	0,060	-3,31	0,010
Вычислительная наука	0,000	736,0	7,48	-7,05	0,248	-8,80	0,035	-6,719	0,132	-2,00	0,000	-5,23	0,001
Науки о принятии решений	0,000	609,4	0,000	-6,71	0,046	-6,81	0,048	-0,275	0,956	-2,66	0,000	-5,91	0,001
Науки о Земле и планетарные науки	0,000	459,7	0,000	-4,69	0,451	-7,52	0,045	-4,928	0,264	-1,95	0,000	-5,69	0,001
Экономика, эконометрия и финансы	0,000	45,080	0,021	-21,6	0,000	-20,4	0,000	-17,027	0,002	-2,68	0,000	-22,9	0,000
Энергетика	0,825	20,630	0,065	0,357	0,789	-0,072	0,838	0,347	0,690	-0,023	0,884	-0,119	0,625
Машиностроение	0,000	825,5	0,000	-	-	-7,98	0,032	-0,763	0,877	-2,71	0,000	-7,52	0,000
Науки об окружающей среде	0,506	26,730	0,104	0,003	0,999	-0,422	0,685	-0,333	0,793	-0,114	0,334	-0,18	0,547
Иммунология и микробиология	0,049	170,3	0,000	-1,85	0,539	-2,48	0,176	-1,111	0,496	-0,268	0,076	-3,98	0,005
Материаловедение	0,193	233,4	0,000	2,02	0,610	-1,02	0,460	-0,034	0,987	-0,412	0,178	-0,850	0,192
Математика	0,012	414,8	0,000	-1,54	0,784	-4,97	0,083	-0,264	0,943	-1,56	0,007	-5,19	0,001
Медицина	0,000	2740,0	0,000	-	-	-7,78	0,043	-4,566	0,309	-2,03	0,000	-5,62	0,001
Нейронаука	0,896	11,920	0,072	-0,018	0,987	-0,178	0,726	-0,066	0,888	-0,020	0,637	-0,285	0,451
Уход за больными	0,256	21,520	0,012	-0,284	0,803	-0,372	0,580	-0,048	0,936	-0,045	0,565	-0,733	0,226
Фармакология, токсикология и фармацевтика	0,865	47,520	0,000	-0,361	0,844	-0,747	0,449	-0,002	0,999	-0,148	0,337	-1,24	0,115
Физика и астрономия	0,797	706,2	0,000	19,5	0,006	0,048	0,646	0,954	0,495	0,091	0,771	0,000	1,000
Психология	0,812	53,220	0,000	0,186	0,920	-0,460	0,562	0,129	0,904	-0,112	0,475	-0,791	0,208
Общественные науки	0,007	173,3	0,000	-3,56	0,366	-4,27	0,114	0,0744	0,983	-1,43	0,007	-4,21	0,004

Предметные научные области, отраженные в Scopus	Значение p	Экспоненциальная		Weibull		Логарифмически нормализованная		Tsallis		Yule		PL с включением	
		LR	p	LR	p	LR	p	LR	p	LR	p	NLR	p
Ветеринария	0,017	38,090	0,000	0,841	0,598	-0,183	0,677	1,953	0,330	-0,047	0,874	-0,542	0,298
Стоматология	0,011	11,830	0,200	-6,60	0,025	-6,26	0,028	-3,714	0,257	-1,28	0,000	-7,14	0,000
Профессии области здравоохранения	0,352	38,620	0,001	-0,944	0,599	-1,10	0,352	-0,395	0,760	-0,192	0,189	-1,63	0,071
Многодисциплинарные области	0,100	98,560	0,001	-1,37	0,595	-1,67	0,339	-1,497	0,377	-0,067	0,069	-1,44	0,090
Все науки	0,076	672,3	0,000	18,30	0,009	-0,125	0,797	-0,007	0,922	-0,054	0,625	-0,240	0,488

*Примечание:* Вторая колонка дает значение  $p$  для гипотезы, что данные следуют модели степенной зависимости. “ - ” означает, что максимальная правдоподобная статистика, используемая в качестве оценки, не приближается к пределу. Позитивные значения логарифмически правдоподобного соотношения LR (log-likelihood ratio) или нормализованного логарифмически правдоподобного соотношения NLR (normalized log-likelihood ratio) указывают, что модель степенной зависимости предпочтительнее альтернативы.

В общем наши результаты показывают, что свидетельство в пользу поведения степенной зависимости правых хвостов распределений ссылок является довольно слабым. Для приблизительно половины изучаемых научных областей Scopus гипотеза степенной зависимости отвергается. Другие распределения, особенно распределение Yule, распределение степенной зависимости с экспоненциальным выключением и логарифмически нормальное распределение, вероятно, должны соответствовать данным из этих научных областей лучше, чем чистая модель степенной зависимости. С другой стороны, когда гипотеза степенной зависимости не отвергается, она, как правило, эмпирически не различима от всех альтернатив, за исключением экспоненциального распределения. Чистая модель степенной зависимости, кажется, является преобладающей над всеми альтернативными моделями только для наиболее высокоцитируемых статей в «Физике и астрономии». Наши результаты предполагают: теории, подразумевающие, что наиболее высокоцитируемые научные статьи, следующие распределению Yule, степенной зависимости с экспоненциальным выключением или логарифмически нормальному распределению, могут иметь несколько большую поддержку в данных, чем теории, предсказывающие поведение строгой степенной зависимости.

## ЗАКЛЮЧЕНИЕ

Мы использовали большой массив новых данных по ссылкам к научным статьям, опубликованным в период 1998-2002 гг., из Scopus, чтобы эмпирически проверить поведение степенной зависимости правых хвостов распределений ссылок. Нами было обнаружено, что гипотеза степенной зависимости отвергается для примерно половины научных областей Scopus. Для оставшихся научных областей распределение степенной зависимости является вероятной моделью, но различия между степенной зависимостью и альтернативными моделями, как правило, статистически незначительны. Статья также подтверждает недавно полученные результаты авторов работы [17], что степенные зависимости в распределениях ссылок, когда они являются правдоподобными, отвечают только за небольшую часть опубликованных статей (менее 1 % для большинства научных областей) и что экспонент степенной зависимости в значительной мере выше, чем обнаружено в более старой литературе.

*Благодарность.* Мне бы хотелось с большой благодарностью признать использование программного обеспечения Matlab and R, сопровождающего статьи Клозета и др. [4] и Шализи [29]. За любые оставшиеся погрешности ответственность несу я как автор данной статьи.

## ЛИТЕРАТУРА

1. Lotka A. The frequency distribution of scientific productivity // Journal of Washington Academy Sciences. – 1926. - Vol. 16 – P. 317–323.
2. Egghe L. Power laws in the information production process: Lotkaian informetrics. - Oxford: Elsevier, 2005.
3. Newman M.E. Power laws, Pareto distributions and Zipf's law// Contemporary Physics. – 2005. – Vol. 46. – P. 323–351.
4. Clauset A., Shalizi C.R. Power-law distributions in empirical data// SIAM review. – 2009. – Vol. 51. – P. 661–703.
5. Gabaix X. Power laws in economics and finance// Annual Review of Economics. – 2009. -Vol. 1. – P. 255–294.
6. Baayen R.H. Word frequency distributions. - Dordrecht: Kluwer, 2001.
7. de Solla Price D. Networks of scientific papers// Science. – 1965. – Vol. 149. – P. 510–515.
8. de Solla Price D. A general theory of bibliometric and other cumulative advantage processes// Journal of the American Society for Information Science. – 1976. – Vol. 27. – P. 292–306.
9. Redner S. How popular is your paper? An empirical study of the citation distribution// The European Physical Journal B. – 1998. – Vol. 4. – P. 131–134.
10. Redner S. Citation statistics from 110 years of Physical Review// Physics Today. – 2005. - Vol. 58. – P. 49–54.
11. Peterson G.J., Presse' S., Dill R.A. Nonuniversal power law scaling in the probability distribution of scientific citations// Proceedings of the National Academy of Sciences. – 2010. – Vol. 107. – P. 16023–16027.
12. Van Raan A.F. Statistical properties of bibliometric indicators: Research group indicator distributions and correlations// Journal of the American Society for Information Science and Technology. – 2006. – Vol. 57. – P. 408–430.
13. Lehmann S., Lautrup D., Jackson A. Citation networks in high energy physics// Physical Review E. – 2003. – Vol. 68. – P. 026113.
14. Laberrere J., Sornette D. Stretched exponential distributions in nature and economy: Fat tails with characteristic scales// The European Physical Journal B. – 1998. – Vol. 2. – P. 525–539.
15. Eom Y.H., Fortunato S. Characterizing and modeling citation dynamics// PloS One. -2011. - Vol. 6: e24926.
16. Golosovsky V., Solomon S. Runaway events dominate the heavy tail of citation distributions //The European Physical Journal. Special Topics. – 2012. – Vol. 205. – P. 303–311.
17. Albarrán P., Ruiz-Castillo J. References made and citations received by scientific articles// Journal of the American Society for Information Science and Technology. - 2011. – Vol. 62. – P. 40–49.
18. Albarrán P., Crespo J.A., Otruño I., Ruiz-Castillo J. The skewness of science in 219 sub-fields and a number of aggregates// Scientometrics. - 2011. – Vol. 88. – P. 385–397.
19. Albarrán P., Crespo J.A., Otruño I., Ruiz-Castillo J. The skewness of science in 219 sub-fields and a number of aggregates.//Working paper 11-09, Universidad Carlos III, 2011.
20. López-Illescas C., de Moya-Anegón F., Moed H.F. Coverage and citation impact of oncological journals in the Web of Science and Scopus// Journal of Informetrics. – 2008. – Vol. 2. – P. 304–316.
21. Aghaei Chadegani A., Salebi H., Md Yunus M., Farhadi H., Foolladi M., et al. A comparison between two main academic literature collections: Web of Science and Scopus databases// Asian Social Science. – 2013. – Vol. 9. – P. 18–26.

22. *Li Y., Ruiz-Castillo J.* The impact of extreme observations in citation distributions// Technical report, Universidad Carlos III, Departamento de Economía, 2013.
23. *Stringer M.J., Sales-Pardo M., Amaral L.A.N.* Effectiveness of journal ranking schemes as a tool for locating information// PLoS One 3. – 2008: e1683.
24. *Radicchi F., Fortunato S., Castellano C.* Universality of citation distributions: Toward an objective measure of scientific impact//Proceedings of the National Academy of Science. – 2008. – Vol. 105. – P.17268-17272.
25. *Tsallis C., de Albuquerque M.P.* Are citations of scientific papers a case of nonextensivity?//The European Physical Journal B. – 2000. – Vol.13. – P.777-780.
26. *Anastasiadis A.D., de Albuquerque M.P., Mussi D.B.* Tsallis q-exponential describes the distribution of scientific citations - a new characterization of the impact//Scientometrics. -2010. – Vol. 83. – P. 205-218.
27. *Wallace M.L., Larivière V., Gingras Y.* Modeling a century of citation distributions// Journal of Informetrics.-2009. – Vol. 3. – P. 296-303.
28. *Vuong Q.H.* Likelihood ratio tests for model selection and non-nested hypotheses//Econometrica. – 1989. – Vol.57. – P.307-333.
29. *Shalizi C.R.* Maximum likelihood estimation for q-exponential (Tsallis) distributions. – Technical report, arXiv preprint math/0701854, 2007.

# Насколько хорошо разработаны альтметрии? Междисциплинарный анализ наличия «альтернативных метрик» в научных публикациях\*

**Зохране ЗАХЕДИ**  
(Zohreh ZAHEDI)

**Родриго КОСТАС**  
(Rodrigo COSTAS)

**Пол УОТЕРС**  
(Paul WOUTERS)

Центр научных и технологических исследований, Лейденский университет, г. Лейден, Нидерланды

*Проводится анализ наличия и возможностей альтметрии для изучения библиометрии и эффективности. С использованием средства на основе сети Imapact Story собраны метрики для 20 тыс. произвольных публикаций, взятых из Web of Science. Изучались как наличие и классификация альтметрии в массиве публикаций по областям, типам документов и годам изданий, так и степень, с которой альтметрия соотносится с показателями ссылок. Основным результатом исследования является то, что источником альтметрии, предоставляющим большую часть метрик, служит программа Mendeley с метриками по читателям для 62,6% всех исследуемых публикаций, тогда как другие источники представляют только маргинальную информацию. Что касается ссылок, то умеренная корреляция Спирмана ( $r=0,49$ ) была найдена между подсчетами читателей по Mendeley и показателями ссылок. Обсуждаются другие возможности и ограничения этих показателей и намечаются направления дальнейших исследований.*

## ВВЕДЕНИЕ

Метрики на основе ссылок и рецензирование имеют давнюю традицию и широко применяются в оценке исследований. Анализ ссылок является популярным и полезным подходом измерения в контексте научной политики и управления исследованиями. Ссылки обычно рассматриваются в качестве уполномоченного «научного влияния» [1]. Однако ссылки не свободны от ограничений [2,3], они только измеряют ограниченный аспект качества (то есть влияния на научную публикацию других авторов) [4,5], их действительное значение широко обсуждается [6], и они также вызывают технические и концептуальные ограничения [7,8]. С другой стороны, рецензирование или оценка коллегами также является важным инструмен-

том и часто рассматривается как золотой стандарт в оценке качества исследования [1, 9-12], но он также имеет свои собственные ограничения и противоречия [13,14]. Более того, и ссылки, и рецензирование обычно считаются частичными показателями «научного влияния» [4], и ни одна метрика также не может в достаточной степени раскрыть полное влияние исследования [15]. С учетом этих ограничений сочетание рецензирования с «многометричным» подходом предлагается как необходимое для оценки исследования [16] наряду с идеей «информированного рецензирования», предложенной авторами [17].

Однако недостатки этих более традиционных подходов в оценке исследования привели к предложениям новых метрик, которые могли обеспечивать «новые, более широкие и быстрые» измерения влияния, нацеленные на дополнение традиционных метрик для ссылок [18]. Это предложение использовать и применять так называемые «альтернативные показатели» в оценке научного влияния столкнулось с научными спорами, и эти новые метрики, как ожидается, должны не только преодолеть некоторые ограничения предыдущих подходов, но и обеспечить новое представление в оценке исследования [19-21].

\* Перевод Zahedi Z., Costas R., Wouters P. How well developed are altmetrics? A cross-disciplinary analysis of the presence “alternative metrics” in scientific publications// Proceedings of the 14th International Society of Scientometrics and Informetrics Conference, Vienna, Austria (pp. 876-884). — 2014.— <http://arxiv.org/ftp/arxiv/papers/1404/1404.1301.pdf>

Эти альтернативные метрики относятся к более «нестандартным» измерениям оценки исследования [22], включая такие метрики, как использование анализа данных (загрузка и просмотр подсчетов) [23-27], анализы сетевых ссылок и взаимосвязей [28-33] или анализ социальных сетей [34]. Важность сети как богатого источника для измерения влияния научных публикаций и ее возможности охватить недостатки существующих метрик в оценке исследования также признаются предыдущими исследованиями. Например, научное свидетельство использования найденных в сети публикаций рассматривается как дополнение к метрикам ссылок, а также как предсказатель вероятных ссылок [35] и релевантности для областей с меньшим числом ссылок [36]. В этом смысле более традиционные метрики на основе ссылок, хотя широко используемые и применяемые в оценке исследования, не способны измерить электронное влияние научной литературы (например, через Фейсбук, Твиттер, системы управления библиографической информацией, блоги или вики), а также им не хватает способности измерить влияние научных результатов, отличных от журнальных статей или трудов конференций, игнорируя другие результаты, такие как базы данных, программное обеспечение, слайды, персональные блоги и т. д. Таким образом, ученые, публикующиеся онлайн и в форматах, отличающихся от журнальных статей, в реальности не извлекают пользы от метрик данных на основе ссылок.

Рост этих новых метрик формировался с предположением о так называемой «альтметрии» или метриках социальных медиа, введенных в 2010 г. Примом и его коллегами [37] в качестве альтернативного способа более широкого измерения научных влияний в социальной сети через различные средства [18, 37]. Точнее, альтметрия охватывает упоминания научных результатов в соцмедиа, новостных медиа и средствах управления ссылками. Это развитие понятия альтметрии сопровождалось ростом разнообразия средств, нацеленных на отслеживание в «реальном времени»\* влияния научных результатов путем изучения долей, лайков, комментариев, обзоров, обсуждений, книжных закладок, сохранений, твитов и упоминаний научных публикаций и источников в соцмедиа [38]. Среди этих средств находятся F1000 (<http://f1000.com>), PLOS Article-Level-Metrics (ALM) (<http://article-level-metrics.plos.org/>), Altmetric.com ([www.altmetric.com/](http://www.altmetric.com/)), Plum Analytics ([www.plumanalytics.com/](http://www.plumanalytics.com/)), Impact Story\*\* ([www.impactstory.org/](http://www.impactstory.org/)), CiteULike ([www.citeulike.org/](http://www.citeulike.org/)) и Mendeley ([www.mendeley.com/](http://www.mendeley.com/)). Эти средства на основе сети покрывают и отслеживают широкий диапазон результатов ученого путем сбора данных по альтметрии в большом разнообразии источников. В следующем разделе кратко рассматриваются предыдущие исследования по альтметрии, которые воспользовались этими средствами.

\* Будучи непосредственно доступными, сравниваются со ссылками, которым требуется время на обработку.

\*\* Ранее известное как общее влияние (Total Impact); мы используем аббревиатуру IS в этом исследовании по отношению к Impact Story. Обзор средств отслеживания научного влияния см. в [38].

## ОПИСАНИЕ

Изучение альтметрии находится на ранней стадии, но некоторая работа уже проделана. Особенности средств альтметрии в общем [39] и их надежности как источников влияния оценки изучались в ряде работ. Например, авторы [40] анализировали сильные, слабые стороны и полезность двух средств управления ссылками для оценки исследования. Их выводы показывают, что по сравнению с CiteULike средство Mendeley кажется должно быть более многообещающим для оценки будущих исследований. Авторы [38] сравнили особенности 16 средств на основе сети и изучили их возможности измерения влияния для целей оценки реальных исследований. Они пришли к выводу, что хотя эти новые средства являются обнадеживающими для научной оценки благодаря их нынешним ограничениям и пределам, они кажутся более полезными для самоанализа, чем для систематического измерения влияния на различных уровнях агрегации.

Авторы [27] исследовали реакции ученых на недавно представленные препринты в архиве arXiv.org, показывая, что соцмедиа могут быть важным фактором в определении научного влияния статьи. Анализ средств управления социальными ссылками, сравниваемый со ссылками, широко изучался в этой области, в частности сравнения ссылок и подсчетов читателей в Mendeley, в большинстве случаев показывающей умеренную и значительную корреляцию между двумя метриками [18,40,41-45]. Также сообщались сведения о слабых корреляциях между тегами пользователей и книжными закладками (как показателями журнального использования, восприятия) и ссылками, наблюдаемыми для физических журналов [46]. В случае F1000 было обнаружено, что и подсчеты пользователей в Mendeley, и факторы статей F 1000 в статьях по геномике и генетике соотносятся со ссылками, и они связаны с импакт факторами журналов [47].

Некоторые другие исследования основаны на предположении, может ли альтметрия использоваться как предсказывающий инструмент ссылок. Например, в случае F1000 найдено, что рекомендации имеют относительно более низкую предсказательную силу при указании высокой цитируемости по сравнению с оценками журнальных ссылок [48]. Также сделано предположение, что на уровне статьи твиты могут предсказать высокоцитируемые статьи в течение первых 3 дней публикации [49], хотя эти результаты подверглись критике со стороны автора [50], и больше исследований должно изучаться с этой точки зрения. Кроме того, большинство статей, получивших ссылки в блогах ближе ко времени их публикации, являются более высокоцитируемыми, чем статьи без таких ссылок в блогах [51].

Выше упомянутые предыдущие исследования использовали альтметрию как источник новых данных и изучали связь между альтметрией и влиянием ссылок. Большинство этих исследований основано на журналах, таких как *Nature & Science* [40]; *JASIST* [42], *Information System Journal* [43], статьях, опубликованных сообществом по библиометрии и наукометрии [45,52], *PLoS* и других медицинских и биомедицинских журналах в PubMed [18,44,45].

Однако по нашим сведениям сегодня предпринимается мало усилий, чтобы изучить наличие альтметрии в различных научных областях, а также для срав-

нительно протяженных временных периодов. Таким образом, это исследование является одним из первых в анализе относительно большой выборки публикаций, принадлежащих к различным областям, типам документов и годам издания. Данная статья опирается на исследования авторов [38].

Наша главная цель в этой статье заключается в том, чтобы представить исследовательский анализ данных по альтметрии, полученных с помощью Impact Story (IS), фокусируясь на связи альтметрии со ссылками публикаций из различных областей науки, социальных и гуманитарных наук. Для этого изучается степень, в которой статьи имеют альтметрию, полученную через различные источники данных, найденных с помощью IS и связи между альтметрией и ссылками на эти статьи. При изучении этих проблем выдвинуты два следующих научных вопроса:

Каковы наличие и распределения альтметрии в IS по типам документов, предметным областям и годам издания относительно исследуемой выборки?

Имеется ли какая-либо связь между альтметрией, найденной в IS, и показателями ссылок в исследуемой выборке? Другими словами, в какой степени альтметрия IS соотносится с показателями ссылок?

## МЕТОДОЛОГИЯ ИССЛЕДОВАНИЯ

В этом исследовании мы концентрируемся на Impact Story (IS). Несмотря на свою по-прежнему раннюю стадию («Бета версия»), IS сегодня является одним из наиболее популярных средств на основе сети с некоторыми возможностями обеспечения целей научной оценки [38]. IS группирует «данные влияния из многих источников и отображает их в одном отчете, делая его быстрым и легким для просмотра влияния широкого диапазона научного результата» (<http://impactstory.org/faq>). Он принимает в качестве входных данных различные типы идентификаторов публикаций (например, DOI, идентификаторы URL в PubMed и т.д.). Они обрабатываются различными внешними службами, чтобы собрать метрики, связанные с определенным «артефактом» (например, публикацией). Итоговый отчет на основе сети создается IS, который показывает влияние этих «артефактов» в соответствии с разнообразием метрик, таких как число читателей, книжных закладок, твитов, упоминаний, долей, мнений, загрузок, персональных блогов и ссылок в Mendeley, CiteULike, Twitter, Wikipedia, Figshare, Dryad, Scienceseeker, PubMed и Scopus\*.

Для этого исследования мы составили произвольную выборку 20 тыс. публикаций с идентификаторами DOI (опубликованных в 2005 – 2011 гг.) из всех дисциплин, охватываемых Web of Science (WoS). Публикации были произвольно выбраны с использованием команды «New ID ()» SQL [53, с. 193]. Массив данных по альтметрии обрабатывался в течение последней недели апреля 2013 г. Данные по альтметрии были получены автоматически через Impact Story REST API\*\*, затем загружались отве-

ты, обеспечивающие поисковые запросы с использованием DOI. Используя этот интерфейс программного приложения, мы смогли загрузить данные по альтметрии быстрее (один запрос в 18 секунд) по сравнению со сбором данных вручную, проведенным нами в предыдущем исследовании\*. Файлы загружались посредством поискового запроса в API отдельно в формате JAVA Script Object Notations (JSON) на основе индивидуального DOI и анализировались с помощью использования библиотеки JAVA внутри программного обеспечения SAS\*\*. Наконец, данные были преобразованы в формат Comma Separated value (CSV) и сопоставлены по DOI с внутренней версией CWTS из Web of Science на предмет возможности добавления к ним других библиометрических данных. Итоговый список публикаций дал 19 772 DOI (из 20 тыс.) после сопоставления\*\*\*. Основываясь на этой таблице, мы изучили распределение альтметрии по предметным областям, типам документов и годам издания (см. табл. 5,2 и 7, соответственно). Подсчитывались показатели ссылок, и итоговые файлы были импортированы в IBM SPSS Statistics 21 для дальнейшего статистического анализа.

Для проверки надежности массива нашей выборки мы сравнили распределение публикаций по основным областям науки в нашей выборке с тем, что есть в базе данных Web of Science (рис. 1) в тот же период и только теми публикациями, у которых имеется DOI. Как можно видеть, это распределение публикаций нашей выборки в основном похоже на распределение публикаций в целой базе данных WOS, поэтому мы можем считать, что наша выборка представляет многодисциплинарность этой базы данных.

---

са, использующего GET (DOIs), и получения требуемого ответа из Impact Story.

\*В предыдущем исследовании сбор данных производился вручную непосредственно через сетевой интерфейс IS. Вручную IS позволяла собрать альтметрию для 100 DOI за один поиск и максимум 2000 DOI за поиск в течение дня с целью избежать засорения пределов его API, подробнее см. в [61].

\*\* Дополнительная функция из этой «рутинной процедуры», которая является средой разработки java, добавлена к среде SAS (системы статистического анализа) для грамматического разбора и чтения формата JSON и возврата данных как объекта.

\*\*\* Из IS один DOI был пропущен. Мы также обнаружили, что 301 DOI был неправилен в WoS (включая экстра характеристики, делающие их несопоставимыми, поэтому они были исключены из анализа). Также 61 оригинальный DOI из WoS указывал на 134 различные публикации WOS (т.е. дублируя DOI). Это значит, что 74 публикации были дубликатами. Учитывая тот факт, что не было систематического способа определения того, какая была правильной (т.е. действительно получившей альтметрию), мы включили их все в анализ с одинаковыми оценками альтметрии, приведшими к:  $20000 - 1 - 301 + 74 = 19772$  итоговым публикациям. Вообще этот процесс показал, что только 1,8% изначально произвольно отобранных DOI имел те же проблемы, таким образом демонстрируя, что DOI является подходящим идентификатором публикаций, хотя и несвободным от ограничений (т.е. ошибок в записях данных DOI, технических ошибок при распознавании DOI в API, а также существования множественных идентификаторов публикаций в источниках данных, приводящих к некоторым ошибкам в полном массиве альтметрии для этих публикаций).

\* Весь список источников см. на сайте <http://impactstory.org/faq>

\*\* A REpresentational State Tranfer (REST) (ful) API (интерфейс прикладных базовых систем) применялся для создания запро-

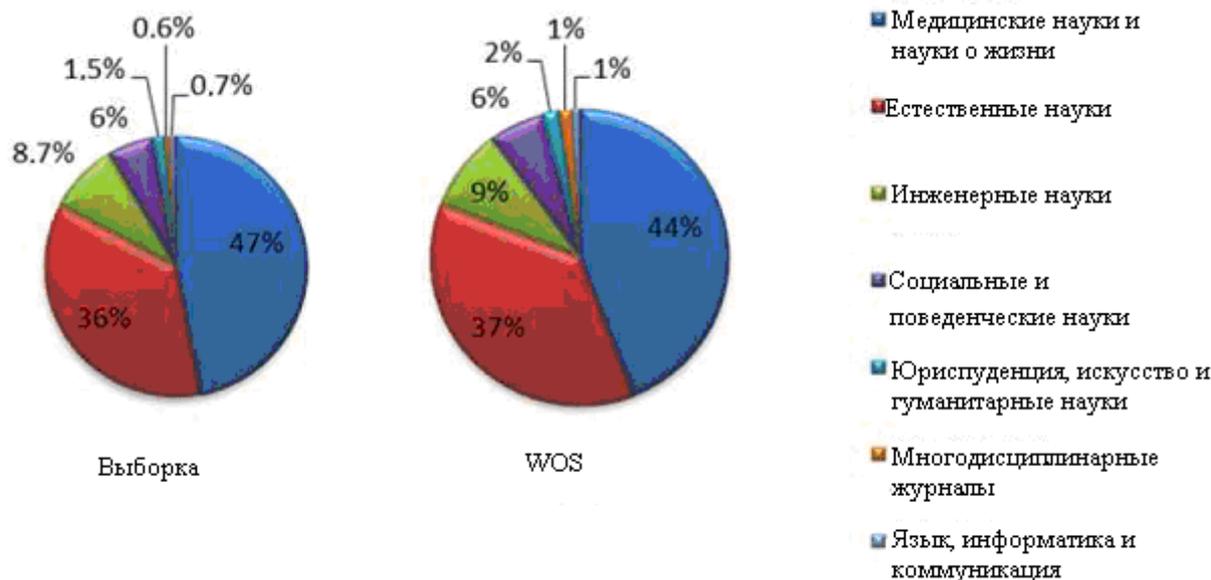


Рис. 1. Распределение публикаций по основным областям науки: выборка в сравнении с целой базой данных

## РЕЗУЛЬТАТЫ И ОСНОВНЫЕ ВЫВОДЫ

На первом месте мы представляем результат нашего исследовательского анализа наличия IS альтметрии в 19 772 публикациях WOS, изданных между 2005-2011 гг. Затем мы изучаем степень, в которой статьи представлены в источниках данных вообще, а также по типам документов, предметным областям и годам издания. В конце, сравниваем связи (корреляцию) между IS альтметрией и ссылками на эти статьи.

### НАЛИЧИЕ IS АЛЬТМЕТРИИ ПО ИСТОЧНИКАМ ДАННЫХ

В нашей выборке наличие IS альтметрии по публикациям различается в каждом источнике данных. Из 19 722 публикаций 12 380 (62,6%) статей имеют, по крайней мере, одного читателя\* в Mendeley, 324 (1,6%) статьи имеют, по меньшей мере, одно сообщение в Твиттере, 289 (1,4%) статей имеют хотя бы одно упоминание в Википедии, 72 (0,3%) статьи имеют, по крайней мере, одну книжную закладку в Delicious и 7413 (37,4%) статей имеют как минимум одну ссылку в PubMed. Только одна статья в выборке имеет метрики из PLoS ALM\*\*. Опираясь на эту предварительную проверку, мы решили исключить некоторые метрики из нашего исследования: показатели PLoS Alm из-за их низкой частоты, так как они доступны только для журналов PLoS и поэтому их наличие в нашей выборке незначительно, и ссылки на основе PubMed, поскольку они ограничены только науками сферы здравоохранения и относятся к ссылкам, которые мы подсчитаем непосредственно на

основе Web of Science. Мы также решили суммировать метрики, исходящие из Твиттера («Topsy tweets» и «Topsy influential tweets»), учитывая их относительно низкую частоту. В результате в этом исследовании были проанализированы данные из Mendeley, Wikipedia, Twitter и Delicious.

Табл. 1 показывает число и доли статей с наличием и без IS альтметрии, распределенных по проценту статей с метриками (исключая метрики PLOS ALM и PubMed). На основе табл. 1 наше главное наблюдение состоит в том, что для этой выборки основным источником для альтметрии является Mendeley, с метриками по читателям для 62,6% всех исследуемых публикаций. Но для других источников данных (Twitter, Wikipedia и Delicious) наличие метрик по публикациям – очень низкое, с более чем 98% статей без метрик. Таким образом, ясно, что их возможное использование для оценки влияния научных публикаций является все еще органичным, особенно при рассмотрении многогодичного и многодисциплинарного массива данных, такого как исследуемый здесь.

Таблица 1

#### Наличие IS альтметрии из источников данных

Источник данных	Статьи с метриками	%	Статьи без метрик	%
Mendeley	12380	62,6	7392	37,3
Twitter	324	1,6	19448	98,3
Wikipedia	289	1,4	19483	98,6
Delicious	72	0,3	19700	99,7

\* Это значит, что публикации без каких-либо метрик были исключены из данного анализа.

\*\* Это была единственная статья PLoS, охваченная нашей выборкой.

## НАЛИЧИЕ IS АЛЬТМЕТРИИ ПО ТИПАМ ДОКУМЕНТОВ

Что касается типов документов, то из 19772 публикаций в выборке имеется 16740 (84,7%) статей, 944 (4,7%) обзорных статей, 487 (2,4%) писем и 1601 (8%) нецитируемых\* документов. Табл. 2 показывает охват выбранных публикаций относительно типов документов по каждому источнику данных. В соответствии с табл. 2, 81,1% (766) обзорных статей, 66,3% (11094) статей, 25,1% писем и 24,9% (398) нецитируемых документов в выборке сохранены (читаются) в Mendeley. В Твиттере 3,4% (32) обзорных статей, 1,9 % (30) нецитируемых документов, 1,5% (255) статей и 1,4% (7) писем имеют твиты. Относительно Википедии 4,6% (43) обзорных статей, 1,4% (230) статей и менее 1% других типов документов (писем и нецитируемых документов) упоминаются, по крайней мере, один раз. Следовательно, Mendeley имеет самый большой охват всех источников данных в этой выборке, (81,1% обзорных статей и 66,3% статей в выборке охвачены им).

Мы также изучили общие числа читателей Mendeley, твитов, упоминаний и книжных закладок для каждого типа документов, охваченных в выборке (т.е. не только число публикаций с метриками, но и частоту этих метрик). Табл. 3 показывает результат общей суммы и среднего числа оценок альтметрии на тип документа, представленных различными источниками данных. На основе табл. 3 и рис. 2 статьи, как правило, имеют самые высокие значения чисел читателей, твитов и книжных закладок (более 77,5% всех оценок альтметрии принадлежат статьям), за которыми следуют\* обзорные статьи, нецитируемые документы и письма (менее 18% оценок альтметрии принадлежат другим типам) во всех источниках данных. Но учитывая среднее число метрик на публикации\*\*, можно заметить, что Mendeley собирает большую часть метрик на все типы документов, чем все другие источники данных. Также в Mendeley обзорные статьи привлекали внимание большинства читателей на публикации (в среднем имеется ~ 14 читателей на обзорную статью), чем во всех других источниках данных.

Таблица 2

Охват публикаций с различными типами документов по разным источникам данных

Тип док.	публ.		Mendeley		Twitter		Wikipedia		Delicious	
Статья	16740	84,7%	11094	66,3%	255	1,5%	230	1,4%	56	0,3%
Обзор	944	4,7%	766	81,1%	32	3,4%	43	4,6%	7	0,7%
Письмо	487	2,4%	122	25,1%	7	1,4%	4	0,8%	3	0,6%
Нецитируемый	1601	8,0%	398	24,9%	30	1,9%	12	0,7%	6	0,4%
<b>Всего</b>	<b>19772</b>	<b>100</b>	<b>12380</b>	<b>62,6%</b>	<b>324</b>	<b>1,6%</b>	<b>289</b>	<b>1,4%</b>	<b>72</b>	<b>0,3%</b>

Таблица 3

Распределение IS альтметрии по типам документов в различных источниках данных

Тип документа	Число публикаций	Mendeley			Twitter			Wikipedia			Delicious		
		Число читателей	%	Среднее число	Число сообщений	%	Среднее число	Число упоминаний	%	Среднее число	Число книжных закладок	%	Среднее число
Статья	16740	82553	83,3	4,9	3020	94,5	0,18	292	77,5	0,02	213	87,3	0,01
Обзор	944	12730	12,9	13,4	78	2,4	0,08	68	18,0	0,07	7	2,9	0,01
Нецитируемый	487	3301	3,3	2,0	76	2,4	0,05	13	3,4	0,01	14	5,7	0,01
Письмо	1601	466	0,5	0,9	21	0,7	0,04	4	1,1	0,01	10	4,1	0,02
<b>Всего</b>	<b>19772</b>	<b>99050</b>	<b>100</b>	<b>5,0</b>	<b>3195</b>	<b>100</b>	<b>0,16</b>	<b>377</b>	<b>100</b>	<b>0,02</b>	<b>244</b>	<b>100</b>	<b>0,01</b>

\* Нецитируемый тип документа соответствует всем типам документов в WOS, отличным от статьи, письма или обзора (например, рецензии на книги, материалы от редакции и т.д.).

\* В Delicious статьи, нецитируемые документы, письма и обзорные статьи имеют самое высокое число метрик в определенном порядке.

\*\* Среднее число метрик на публикации подсчитывалось делением общих чисел метрик из каждого источника данных на общее число публикаций в выборке. Например, в Mendeley среднее число читателей на публикацию равняется  $99050/19772 \approx 5$ .

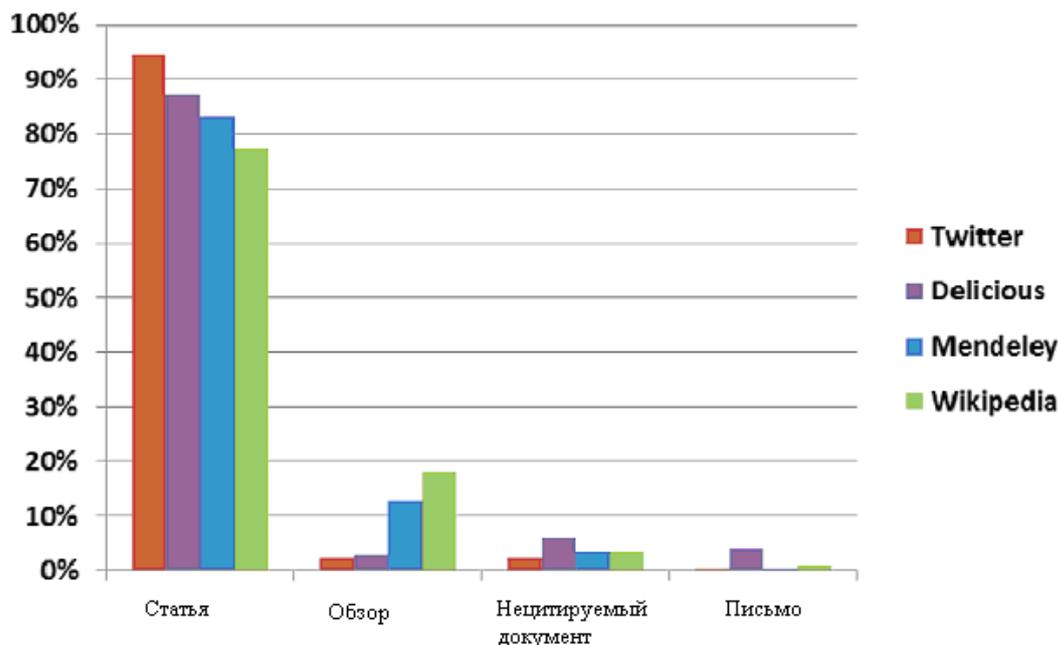


Рис. 2. Распределение IS альтметрии по типам документов

### НАЛИЧИЕ IS АЛЬТМЕТРИИ ПО ПРЕДМЕТНЫМ ОБЛАСТЯМ КЛАССИФИКАЦИИ NOWT

Для этого анализа мы использовали (высокую) классификацию NOWT (Netherlands Observatory of Science and Technology), имеющую 7 основных дисциплин, разработанную CWTS\*. Табл. 4 показывает процентное соотношение публикаций, имеющих по крайней мере одну метрику (т.е. статьи хотя бы с одним читателем в Mendeley, одной книжной закладкой в Delicious, одним сообщением в Твиттере или одним упоминанием в Википедии) по этим основным областям.\*\* Согласно этим результатам, междисциплинарные публикации ранжированы как самые высокие во всех источниках данных. Основным источником для данных по альтметрии в нашей выборке является Mendeley с самой высокой долей для междисциплинарных областей, которые включают такие журналы, как *Nature*, *Science* или *PNAS*. 80% публикаций в этой области, 73% публикаций из медицинских наук и наук о жизни\*\*\* и 68% публикаций из социальных и поведенческих наук имеют, по крайней мере, одного читателя Mendeley. Среди других источников данных междисциплинарные публикации также ранжированы как самые высо-

кие, но с более низким наличием публикаций с метриками. Относительно верхних трех областей с самым высоким процентным соотношением альтметрии Википедия имеет схожую модель с Mendeley: 7% публикаций из междисциплинарной области, 2% публикаций из медицинских наук и наук о жизни и 2% публикаций из социальных и поведенческих наук имеют по крайней мере одно упоминание в Википедии. Что касается Твиттера, то 7% публикаций из междисциплинарной области, 3% публикаций из социальных и поведенческих наук и 2% публикаций из медицинских наук и наук о жизни являются тремя верхними областями, которые имеют, по меньшей мере, одно сообщение в Твиттере. В Delicious только 1% публикаций из междисциплинарной области, языка, информатики и коммуникации и социальных и поведенческих наук имеют хотя бы одну книжную закладку, тогда как другие области — меньше 1% альтметрии.

К тому же были подсчитаны общие оценки читателей в Mendeley, сообщений в Твиттере, упоминаний и книжных закладок по каждой дисциплине в выборке. Рис. 2 показывает, что распределения IS альтметрии по различным предметным областям не являются равномерными. Медицинские науки и науки о жизни, а также естественные науки получили самую большую долю альтметрии во всех источниках данных. Вообще во всех источниках данных более 30% альтметрии собрано за счет публикаций из области Медицинских наук и наук о жизни и более 23% альтметрии являются публикациями из областей естественных наук. Что касается других областей, то каждая область получала менее 10% всей альтметрии. При сравнении различных источников данных в условиях соотношения альтметрии во всех областях возникают различные модели: области медицинских наук и наук о жизни пропорционально привлекли наибольшее внимание в Википедии, затем следуют Mendeley, Твиттер и Delicious, хотя в случае естественных наук, Delicious, Твиттер, Mendeley и Википедия соответственно получили наибольшее внимание в порядке перечисления; кроме того, что касается Mendeley, то как социальные и поведенческие науки, так и инженерные науки пропорционально получили самое большое внимание, чем все другие области.

\* В предыдущем исследовании мы использовали (среднюю) классификацию NOWT с 14 предметными областями. Подробнее см. сайт: [http://nowt.merit.unu.edu/docs/NOWT-WTI\\_2010.pdf](http://nowt.merit.unu.edu/docs/NOWT-WTI_2010.pdf)

\*\* Здесь публикации могут принадлежать ко многим предметным категориям.

\*\*\* Согласно глобальному исследовательскому отчету, проведенному Mendeley (<http://www.mendeley.com/global-research-report/#.UjwfTsangk>), охват Mendeley в различных областях является следующим: самый высокий охват имеют публикации из области биологии и медицины (31%), за ними следуют физические науки и математика (16%), инженерия и материаловедение (13%), вычислительная техника и информатика (10%), психология, лингвистика, и образование (10%), деловое администрирование, экономика и исследования операций (8%), юриспруденция и другие социальные науки (7%) и философия, искусство и литература, а также другие гуманитарные области (5%).

Охват публикаций различными предметными областями NOWT в различных источниках данных

Высокие предметные категории NOWT	Общее число публикаций	%	Mendeley	%	Википедия	%	Твиттер	%	Delicious	%
Многодисциплинарные журналы	216	47	172	80	15	7	16	7	3	1
Медицинские науки и науки о жизни	15637	36	11353	73	284	2	301	2	67	0,4
Социальные и поведенческие науки	1878	6	1268	68	32	2	58	3	11	1
Естественные науки	11935	8,7	6554	55	103	1	123	1	34	0,3
Инженерные науки	2885	0,6	1558	54	7	1	9	0,3	2	0,1
Язык. Информатика и коммуникация	241	0,7	123	51	2	1	1	0,4	3	1
Юриспруденция. Искусство и гуманитарные науки	488	1,5	190	39	8	2	7	1	0	0
		100								

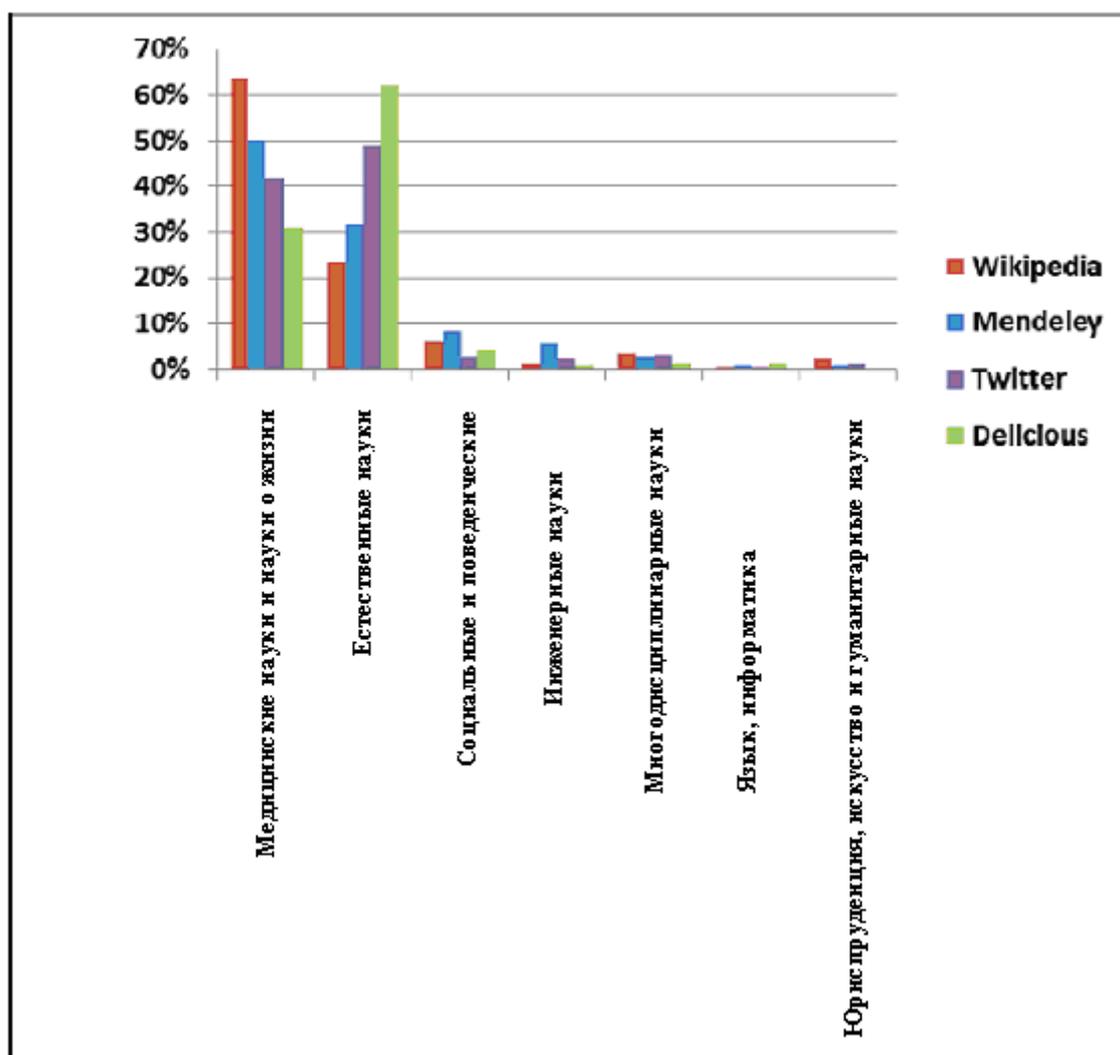


Рис. 3. Распределение IS альтметрии по предметным областям NOWT

Распределение IS альтметрии на предметные области NOWT в различных источниках данных

Предметная категория NOWT	Число читателей в Mendeley	%	Упоминания в Википедии	%	Книжные закладки в Delicious	%	Сообщения в Твиттере	%
Медицинские науки и науки о жизни	86347	50	371	64	92	31	1958	42
Естественные науки	54481	32	136	23	186	62	2317	49
Социальные и поведенческие науки	14102	8	35	6	35	6	112	2
Инженерные науки	9800	6	7	1	2	1	100	2
Многодисциплинарные журналы	4521	3	20	3	3	1	144	3
Язык. Информатика и коммуникация	1492	0,9	2	0,3	4	1	1	0
Юриспруденция, искусство и гуманитарные науки	1297	0,8	13	2	0	0	72	2
		100		100		100		100

#### СРАВНЕНИЕ ЧИСЛА ССЫЛОК НА СТАТЬИ (CPP – CITATIONS PER PAPERS) И ЧИСЛА ЧИТАТЕЛЕЙ НА СТАТЬИ (RPP – READERSHIPS PER PAPERS) ПО ОБЛАСТЯМ

Хотя измерение влияния научных публикаций в социальных медиа является очень важным, до сих пор неясно для каких целей научные публикации упоминаются в соц-медиа и средствах управления ссылками, таких как Mendeley, в управлении социальными книжными закладками – Delicious, в Википедии и Твиттере – различными пользователями/учеными; в частности неясно, можно ли рассматривать эти упоминания как меры любого типа «влияния» публикаций. В случае Mendeley предполагается, что публикации сохраняются в библиотеках пользователей для безотлагательного или более позднего прочтения и, вероятно, также для дальнейшего цитирования.

В любом случае важно знать, сколько альтметрии по сравнению со ссылками получила каждая публикация и каковы модели в различных предметных областях. Учтя, что не все научные публикации одинаково охватываются базами данных по цитированию, а также существование дисциплинарных различий в рамках ссылок, которые значительно варьируются между областями, интересно изучить сравнительное соотношение альтметрии и ссылок на публикации, чтобы увидеть, какие области могут выиграть от наличия большей плотности оценок альтметрии (т.е. оценок альтметрии на статью), чем от плотности ссылок. Так как Твиттер, Википедия и Delicious показали в общем очень низкое наличие альтметрии на статью, мы фокусируемся здесь только на Mendeley. И среднее число читателей Mendeley на статьи (RPP) и ссылки на статьи (CPP) в WOS по различным предметным областям NOWT были подсчитаны и проанализированы (рис. 4). Для подсчета ссылок (исключая самоцитирования) мы использовали переменный интервал ссылок от года публикации до 2012 г. Также переменный «читательский интервал» был рассмотрен для Mendeley, подсчитывались читатели от года публикации статьи до последней недели апреля 2013 г. В этот анализ мы также включили публикации без каких-либо альтметрий (ссылок или читателей в Mendeley). Результат (рис. 4, отражающий RPP) показы-

вает, что вообще многодисциплинарные журналы имеют самые большие значения как RPP, так и CPP, а юриспруденция, искусство и гуманитарные науки имеют самые низкие значения. Что касается таких областей, как многодисциплинарные журналы, медицинские науки и науки о жизни, естественные и инженерные науки, то значение CPP выше, чем RPP, хотя для таких областей, как социальные и поведенческие науки, язык, информатика и коммуникация, а также юриспруденция, искусство и гуманитарные науки RPP превышает CPP. Последнее является интересным результатом, который может предполагать релевантность Mendeley для исследования публикаций в социальных и гуманитарных науках, часто плохо представленных ссылками [59]. В целях дальнейшей проверки различий между RPP и CPP мы расширили тот же тип анализа для всех 248 отдельных предметных категорий WOS и получили, что 167 из 248 предметных категорий WOS имеют значения CPP выше, чем значения RPP. Большинство областей с высокими значениями CPP по сравнению с RPP принадлежат естественным наукам (145), 18 – социальным наукам и 4 – искусству и гуманитарным наукам. С другой стороны, 72 области представлены более высокими значениями RPP, чем CPP (среди них 31 происходит из социальных наук, 27 из естественных наук и 13 из искусства и гуманитарных наук)\*. Следовательно, мы можем прийти к выводу, что ссылки являются более доминирующими, чем подсчет читателей, в частности в областях (которые также являются областями с самым большим охватом в базах данных по цитированию), тогда как, с другой стороны, много подобластей из социальных наук, искусства и гуманитарных наук получили пропорционально больше читателей на статью, чем ссылок на статью. Это можно рассматривать как возможность для областей с более низким охватом в базах данных по цитированию (таких, как WOS) в получении выгоды от Mendeley с точки зрения наличия большего влияния читателей, чем влияния ссылок, хотя это нуждается в дальнейшем изучении.

\* Для 9 областей (8 областей из сферы искусства и гуманитарных наук и 1 область из сферы естественных наук) значения CPP и RPP были точно такими же.

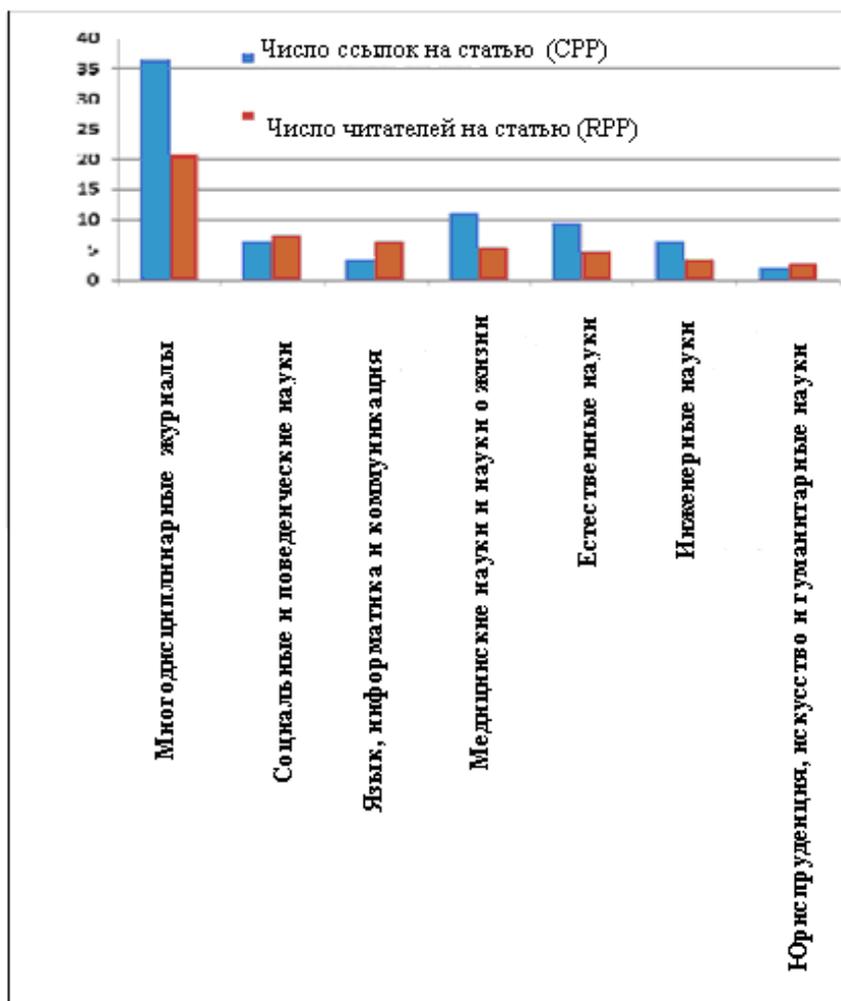


Рис. 4. Сравнение CPP и RPP в Mendeley по предметным областям

## АНАЛИЗ ТЕНДЕНЦИЙ IS АЛЬТМЕТРИИ ПО ГОДАМ ИЗДАНИЯ

Табл. 6 показывает анализ тенденций числа и доли публикаций в выборке по источникам альтметрии. Относительно годов издания доля публикаций колеблется от 10% в 2005 г. до 18% в 2011г. Охват различных источников также показан в этой таблице. В нашей выборке Mendeley имеет пик в своей доле публикаций с некоторыми читателями в 2009 г. (66%) и минимум в 2011 г. (57%), хотя общее число публикаций с некоторыми читателями Mendeley растет на протяжении всего периода, за исключением 2011 г., когда имеется небольшое снижение по сравнению с 2010 г. Твиттер достигает пика в 2011 г. (4%) и минимума в ранних годах (почти 1% в период 2005-2009 гг.). Упоминания в Википедии составляют 2% от всех публикаций, изданных в 2005 – 2008 гг., и 1% от всех публикаций, напечатанных в 2010 – 2011 гг. Что касается Delicious, то пик приходится на 2007 г. и 2011 г., а минимум на 2005 г., также публикации 2008, 2009 и 2010 гг. имеют одинаковое наличие в Delicious. Вообще кажется, что Твиттер и Delicious лучше стремятся охватить недавние публикации, чем старые, хотя, как правило, значения очень низкие.

Наличие общих оценок альтметрии (т.е. не только публикаций с альтметрией, но и их общего подсчета) также было подсчитано с целью узнать ее направлен-

ность со временем. В соответствии с табл. 7, она (альтметрия) сильно отличается в различных источниках данных. Например, что касается Википедии и Mendeley, то публикации 2006 и 2009 гг. собрали большую часть упоминаний (20%) и читателей (17%), соответственно. В случае Mendeley и Википедии мы отметили спад в количестве альтметрии в последние два года.

И в Delicious, и в Твиттере публикации 2008 г. получили самую большую долю альтметрии. В случае Delicious 50% книжных закладок и в случае Твиттера 34% сообщений (твитов) относятся к публикациям, изданным в 2008 г. Сравнение значений альтметрии в каждом году по различным источникам данных показывает, что в этой выборке самые старые и самые последние публикации в Твиттере имеют большую часть альтметрии (твитов) (26% твитов должны быть публикациями 2005\* г. и 2011 г., соответственно), а также последние публикации (2009-2010 гг.) (рис. 5) имеют большую часть альтметрии (читателей) в Mendeley.

\* В 2005 г. две наиболее сообщаемые в Твиттере статьи происходят из области физики, они получили больше половины всех сообщений в Твиттере в этом году (472 твита), таким образом показывая сильно искаженное распределение.

Охват публикаций с различными годами издания по разным источникам данных

Год издания	Публикации	%	Mendeley	%	Википедия	%	Delicious	%	Твиттер	%
2005	2006	10	1263	63	39	2	3	0,1	17	1
2006	2405	12	1491	62	58	2	4	0,2	6	0,2
2007	2682	14	1702	63	41	2	13	0,5	16	1
2008	2858	14	1799	63	46	2	11	0,4	34	1
2009	3039	15	2001	66	43	1	12	0,4	31	1
2010	3228	16	2099	65	37	1	13	0,4	62	2
2011	3548	18	2020	57	25	1	16	0,5	158	4

Таблица 7

Распределение IS альтметрии по году издания

Год издания	Публикации	%	Mendeley	%	Википедия	%	Delicious	%	Твиттер	%
2005	2006	10	10814	11	48	13	51	21	835	26
2006	2405	12	12658	13	77	20	4	2	20	1
2007	2682	14	13739	14	58	15	14	6	102	3
2008	2858	14	14299	14	67	18	122	50	1072	34
2009	3039	15	16922	17	50	13	21	9	145	5
2010	3228	16	16305	16	43	11	14	6	198	6
2011	3548	18	14239	14	34	9	18	7	823	26
		100		100		100		100		100

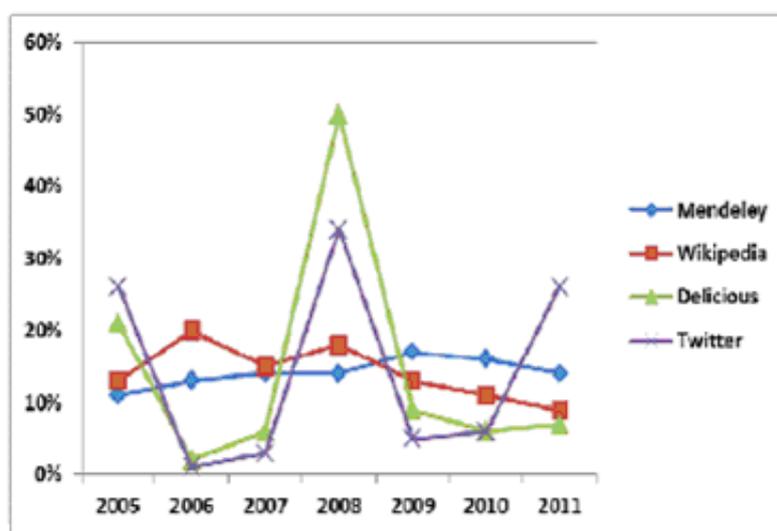


Рис. 5. Распределение IS альтметрии по годам издания

## ОТНОШЕНИЯ МЕЖДУ IS АЛЬТМЕТРИЕЙ И ПОКАЗАТЕЛЯМИ ССЫЛОК

В этом разделе мы более подробно изучаем отношения между IS альтметрией и показателями ссылок. Следуя стандартному подсчету показателей в CWTS (ср. [60]), мы вычислили для всех публикаций следующие показатели ссылок: *оценка ссылки* (CS), т.е. число ссылок на публикации; *нормализованная оценка ссылки* (NCS), т.е. число ссылок на публикации с нормализацией для различий в областях и годах издания; *оценка журнальной ссылки* (JCS), т.е. среднее число ссылок, полученное всеми публикациями в этом издании журнала; *нормализованная журнальная оценка* (NJS), т.е. среднее число ссылок, полученных всеми публикациями в этом журнале, нормализованное различиями областей и годом издания. Для

подсчета показателей влияния, как объяснено выше, мы использовали переменный интервал ссылок (т.е. ссылки до 2012 г.), за исключением самоцитирований. Результат этого факторного анализа, корреляционного анализа и влияния публикаций с наличием альтметрии или без нее будет представлен в следующих разделах.

## ФАКТОРНЫЙ АНАЛИЗ IS АЛЬТМЕТРИИ И БИБЛИОМЕТРИЧЕСКИХ ПОКАЗАТЕЛЕЙ

Исследуемый факторный анализ был проведен с использованием версии 21 SPSS, чтобы больше узнать о соответствующей структуре, связях между переменными и измерениями переменных (табл. 8). Анализ главных компонент (PCA – Principal Component Analysis) об-

наружил наличие 2 основных компонент или измерений с собственными значениями (eigenvalues), превышающими 1, которые объясняют 58% всех расхождений. В первом измерении преобладают библиометрические показатели. Читатели Mendeley и упоминания в Википедии также включены в это измерение, хотя подсчеты читателей Mendeley имеют самую высокую загрузку в этом измерении двух показателей. Второе измерение больше относится к метрикам соцмедиа, показывая, что Твиттер и Delicious сильно коррелируются. Эти результаты предполагают, что переменные в каждой группе могут представлять одинаковые понятия.

Таблица 8

**Факторный анализ переменных**

Матрица А повернутых компонент		
	Компонент	
	1	2
<b>CS</b>	<b>0,837</b>	0,005
<b>NCS</b>	<b>0,752</b>	0,009
<b>JS</b>	<b>0,745</b>	-0,11
<b>NJS</b>	<b>0,720</b>	-0,15
<b>Mendeley</b>	<b>0,680</b>	0,008
<b>Википедия</b>	<b>0,297</b>	0,009
<b>Delicious</b>	0,003	<b>0,954</b>
<b>Твиттер</b>	0,004	<b>0,954</b>
Метод извлечения: Метод главных компонент		
Метод преобразования: Варимакс с нормализацией Кайзера		
Показаны загрузки, больше, чем 0,1		
Объяснено 58% всех расхождений		
а. Преобразование сводилось в 4 взаимодействия		

**КОРРЕЛЯЦИИ МЕЖДУ IS АЛЬТМЕТРИЕЙ И БИБЛИОМЕТРИЧЕСКИМИ ПОКАЗАТЕЛЯМИ**

Чтобы преодолеть техническое ограничение SPSS при вычислении корреляции Спирмана для крупных массивов данных\*, во-первых, осуществились ранжирования переменных, вычисленных с использованием процедуры Data>rank cases, а затем – корреляция Пирсона, выполненная на ранжированных переменных; этот метод обеспечивает корреляцию Спирмана относительно первоначальных переменных. Табл. 9 показывает результат корреляционного анализа среди различных источников данных по альтметрии и оценок ссылок и оценок журнальных ссылок и их 95% интервалы доверия (подсчитанные с использованием технологии Bootstrapping, примененной в SPSS). В соответствии с этой таблицей, показатели ссылок больше коррелируются между собой, чем с альтметрией. Вообще, прямые показатели ссылок (т.е. CS и NCS) соотносятся лучше между собой, чем с показателями журнального влияния (JS и NJS), хотя корреляции между двумя группами явно высоки. Mendeley соотносится с Википедией ( $r=0,08$ ), а Твиттер соотносится с Delicious ( $r=0,12$ ), что отвечает результатам факторного анализа, но корреляции значений являются очень низкими. По сравнению с показателями ссылок Mendeley имеет самую высокую корреляционную оценку со ссылками (средняя корреляция  $r=0,49$ ) среди всех источников альтметрии. Другие источники альтметрии показывают очень слабую или незначительную корреляцию с показателями ссылок.

Таблица 9

**Корреляционный анализ ранжированных значений переменных**

	<b>NCS</b>	<b>JS</b>	<b>NJS</b>	<b>Mendeley</b>	<b>Википедия</b>	<b>Delicious</b>	<b>Твиттер</b>
<b>CS</b>	0,886 (0,882-0,89)	0,762 (0,756-0,769)	0,557 (0,547-0,567)	0,497 (0,485-0,508)	0,94 (0,08-0,108)	0,011 (-0,005-0,027)	0,025 (0,01-0,039)
<b>NCS</b>		0,528 (0,516-0,538)	0,6 (0,59-0,609)	0,467 (0,455-0,478)	0,074 (0,059-0,087)	0,019 (0,002-0,035)	0,054 (0,037-0,068)
<b>JS</b>			0,711 (0,702-0,718)	0,44 (0,428-0,452)	0,09 (0,075-0,105)	-0,003 (-0,018-0,012)	-0,003 (-0,018-0,011)
<b>NJS</b>				0,427 (0,415-0,439)	0,058 (0,044-0,072)	0,012 (-0,005-0,028)	0,39 (0,023-0,053)
<b>Mendeley</b>					0,083 (0,067-0,099)	0,031 (0,015-0,047)	0,07 (0,055-0,084)
<b>Википедия</b>						0,021 (-0,001-0,049)	0,056 (0,025-0,087)
<b>Delicious</b>							0,125 (0,073-0,185)

\* Подсчет анализа корреляции Спирмана в SPSS для крупных массивов данных дает эту ошибку: «Слишком много случаев для доступного хранения», для преодоления этого ограничения мы следовали процессу, упомянутому в тексте. Подробнее см на сайте: <http://www-01.ibm.com/support/docview.wss?uid=swg2146714>

## ВЛИЯНИЕ ПУБЛИКАЦИЙ С АЛЬТМЕТРИЕЙ ИЛИ БЕЗ НЕЕ

В этом разделе мы изучаем различия во влиянии между публикациями с альтметрией и без нее. Основная идея заключается в том, чтобы увидеть, стремятся ли публикации с альтметрией иметь больше влияния ссылок, чем публикации без нее. Табл. 10 представляет библиометрические показатели и их 95% интервалы доверия (подсчитанные с использованием технологии Bootstrapping, примененной в SPSS). Например, в соответствии со средними значениями можно наблюдать, что публикации с метриками имеют вообще более высокие оценки ссылок по сравнению с публикациями без метрик во всех источниках данных (хотя в ряде случаев интервалы доверия показывают некоторое перекрытие, таким образом требование более вы-

сокого влияния для этих случаев является менее сильным и, вероятно, более влиятельным за счет резко выделяющихся значений).

Фокусируясь на числе читателей Mendeley на публикацию и рассматривая их влияние как измеренное NCS и NJS, мы можем видеть, как публикации стремятся увеличиться во влиянии ссылок по мере роста числа читателей (рис. 6). Этот эффект является достаточно сильным, особенно для среднего числа ссылок на публикацию, но это менее важно для показателя NJS. Такой же результат обнаружен авторами [48] для связи между рекомендациями из F1000, ссылками и журнальным влиянием. В своем исследовании они обнаружили, что в среднем публикации с большим числом рекомендаций также имеют большее влияние ссылок и журнала.

Таблица 10

Сравнение NCS и NJS публикаций с альтметрией и без нее

		С метриками				Без метрик			
		CS	JS	NCS	NJS	CS	JS	NCS	NJS
<b>Mendeley</b>	<b>N</b>	12380	12380	12380	12380	7392	7392	7392	7392
	<b>Медиана</b>	5	6,53	0,72	1,02	1	1,76	0,10	0,53
Интервал доверия	<b>Самый низкий</b>	4	6,4	0,69	1,01	0,5	1,67	0,08	0,51
	<b>Самый высокий</b>	5	6,69	0,74	1,04	1	1,89	0,12	0,55
<b>Википедия</b>	<b>N</b>	289	289	289	289	19483	19483	19483	19483
	<b>Медиана</b>	12	13,87	1,18	1,18	2	4,43	0,47	0,86
Интервал доверия	<b>Самый низкий</b>	9	11,91	0,97	1,07	2	4,32	0,46	0,85
	<b>Самый высокий</b>	14	15,2	1,35	1,31	3	4,57	0,49	0,87
<b>Твиттер</b>	<b>N</b>	324	324	324	324	19448	19448	19448	19448
	<b>Медиана</b>	4	3,6	1	1,1	3	4,53	0,47	0,86
Интервал доверия	<b>Самый низкий</b>	3	3,1	0,86	0,97	2	4,39	0,46	0,85
	<b>Самый высокий</b>	5	4,74	1,29	1,27	3	4,62	0,49	0,87
<b>Delicious</b>	<b>N</b>	72	72	72	72	19700	19700	19700	19700
	<b>Медиана</b>	3	3,99	0,89	1,07	3	4,52	0,48	0,86
Интервал доверия	<b>Самый низкий</b>	2	2,34	0,52	0,76	2	4,38	0,46	0,85
	<b>Самый высокий</b>	6	5,55	1,57	1,33	3	4,62	0,49	0,87

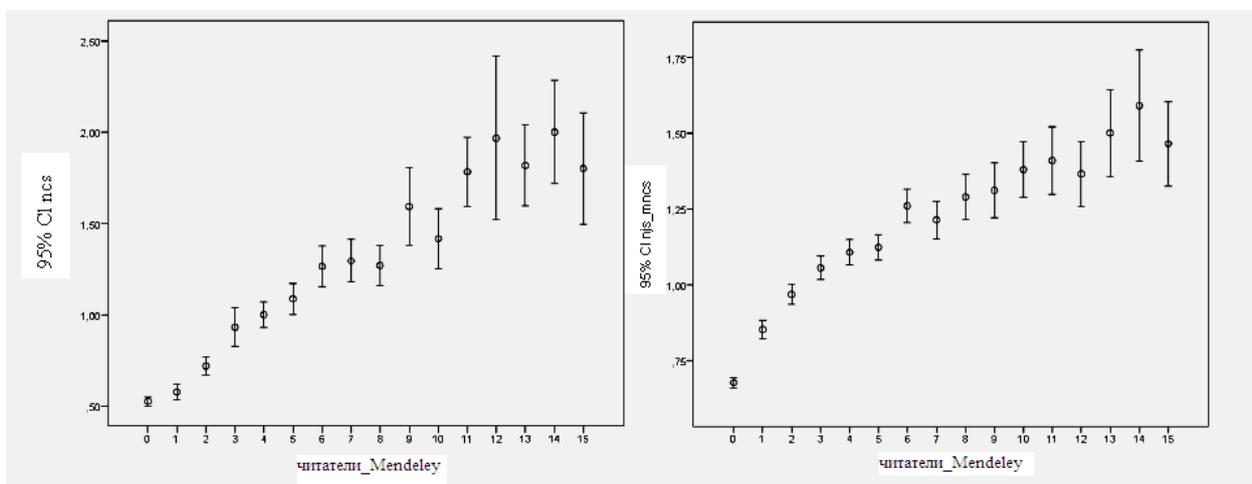


Рис. 6. Связь между числом читателей Mendeley и влиянием ссылок и журналов

## ОБСУЖДЕНИЕ И ВЫВОДЫ

В этой статье мы использовали Impact Story (IS)\* для сбора альтметрии в массиве произвольно отобранных публикаций. IS является интересным открытым источником по сбору альтметрии, однако, мы также видим некоторые важные ограничения\*\* в особенности относительно скорости и возможности сбора данных и их форматирования. Мы выявляем различные результаты, сравнивая наши текущие результаты с представленными в предыдущем исследовании [61], в большей степени благодаря отличающейся методологии сбора данных (вручную в противовес автоматической) и сбору данных в различных временных периодах, как произошло с нашими двумя исследованиями, где в первом исследовании Mendeley был представлен только в примерно 37% публикаций\*\*\*, а теперь – в более, чем 60%\*\*\*\*. Эта ситуация также указывает на необходимость в прозрачности средств относительно того, как собираются данные и каковы их ограничения. Это означает, что важным естественным дальнейшим шагом должна быть истинная оценка надежности данных, найденных через различные источники данных по альтметрии (как было сделано для Google Scholar - ср. [54]). Это признание качества, надежности и прочности средств альтметрии является необходимым, чтобы иметь возможность применять альтметрию для целей серьезной научной оценки. Для полного внедрения этих средств в регулярные процессы научной оценки они должны отвечать обязательным требованиям по качеству данных, прозрачности и показателю надежности и общей значимости, как отмечали авторы [38] в своем исследовании средств альтметрии. Более того, результаты этого исследования основаны на охватываемых WOS публикациях, значит, важно помнить об ограничениях этой базы данных, учитывая охват ею некоторых областей, язык и форматы публикаций [55-58].

В итоге, учитывая исследовательский характер и тот факт, что в основном были получены одни и те же результаты в двух массивах данных, можно предположить, что наши результаты являются устойчивыми и надежными для наших целей. Вообще наше исследование показывает, что Mendeley является главным и более полезным источником для данных по альтметрии. Mendeley имеет самый большой охват и долю альтметрии по сравнению с Твиттером, Википедией и Delicious для изученных публикаций. Из 19772 публикаций все 12380 случаев (62,6%) имели по крайней мере одного читателя в Mendeley. Предыдущие исследования также показали, что Mendeley является наиболее исчерпывающим источником данных по альтметрии [18,52] в основном для публикаций из области библиотековедения и информа-

тики: 97,2% – охват для статей JASIST, опубликованных в 2001–2011 гг. [42]; 82% – охват для статей, опубликованных учеными в *Scientometrics* [52], и 82 % литературы по библиометрии [45]; что касается многодисциплинарных журналов, таких как *Nature* и *Science* (93% и 94% статей, опубликованных этими журналами в 2007 г.) [40], и более 80% публикаций PLoS ONE [18], то они также охвачены Mendeley. Относительно типов документов обзоры и статьи были пропорционально наиболее читаемыми, совместно используемыми, понравившимися или отмеченными книжными закладками форматами, если сравнить их с нецитируемыми документами и письмами по всем источникам данных. Многодисциплинарные области (т.е. область, где включены такие журналы, как *Nature*, *Science* или *PNAS*) являются наиболее представленными в источниках данных по альтметрии, но в рамках распределения альтметрии по различным областям более 30% альтметрии собраны публикациями из медицинских наук и наук о жизни и более 23% альтметрии относятся к публикациям из областей естественных наук. Сравнение доли и распределения IS альтметрии в различных областях среди разных источников данных показывает и разнообразные модели, в частности в Mendeley как социальные и поведенческие, так и инженерные науки соответственно получили самое пристальное внимание по сравнению со всеми другими областями. Рассматривая сылки и число читателей на публикацию, многодисциплинарные журналы имеют самую высокую, а юриспруденция, искусство и гуманитарные науки – самую низкую плотность как сылок, так и числа читателей на публикацию. Однако, по нашим наблюдениям, имеется более высокая плотность читателей на статью, чем сылок на статьи в ряде областей социальных и гуманитарных наук. Данные этого наблюдения предполагают, что подсчеты читателей в Mendeley могут иметь некоторое дополнительное значение в поддержке оценки и анализе этих областей, которые традиционно хуже представлены показателями сылок (ср. [59]). Другим объяснением для этих областей с более низкой долей читателей, чем сылок, может быть тот факт, что Mendeley является относительно новым средством и до сих пор еще не широко используемым и адаптированным среди ученых из всех дисциплин. Кроме того, расхождения в поведении сылок и читателей и практики среди областей можно также объяснить этими различиями. В любом случае этот аспект нуждается в дальнейшем анализе.

Наш анализ тенденций показывает, что особенно публикации с читателями в Mendeley со временем растут, хотя имеется небольшой спад в числе читателей и доле публикаций с читателями Mendeley в последние два года. Наиболее правдоподобным объяснением этого является то, что аккумуляция читателей занимает некоторое время. По нашим сведениям не имеется информации об «истории читателей» публикаций (помимо того факта, что число читателей может существенно снижаться по мере удаления или изменения своих библиотек пользователями) и поэтому у нас нет результатов по скорости чтения. Это значит, что мы не знаем, когда статья определенного года достигает своего пика у читателей. Весьма вероятно, что сбор читателей для публикаций также занимает некоторое время, и это является причиной того, почему для самых недавних публикаций число читателей накапливается медленнее по сравнению с более старыми публикациями, имеющими больше времени для аккумуляции читателей. Дальней-

\*Impact Story находился в первоначальной стадии разработки (т.е. в «Бета версии») на момент развития этого исследования.

\*\* Современные ограничения сегодняшнего дня см. на сайте IS: [http://impactstory.org/faq#toc\\_3\\_11](http://impactstory.org/faq#toc_3_11)

\*\*\* Временной интервал между первым и вторым сборами данных составлял 6 месяцев, и данные собирались вручную в противоположность второму исследованию, проведенному автоматически с использованием требований RESTAPI.

\*\*\*\* Причиной этих различий могут быть изменения/улучшения при идентификации публикаций в Mendeley (например, путем слияния версии той же самой статьи, определения больших DOI, увеличения числа пользователей в Mendeley и т.д.).

шее исследование должно также концентрироваться на раскрытии этого аспекта.

Использование корреляции Спирмана для соотнесения читателей Mendeley с показателями влияния ссылок показало умеренную корреляцию ( $r=0,49$ ) между двумя переменными, также найденными в предыдущих исследованиях [18,42]. Это демонстрирует, что чтение и цитирование являются родственными, хотя все еще различными видами деятельности, которые стоит изучать. Согласно результату сравнения влияния публикаций с альтметрией и без нее с их оценками ссылок, можно также прийти к выводу, что в целом публикации с большей альтметрией также стремятся иметь больше прямых ссылок и быть опубликованными в журналах более высокого влияния. Этот вопрос о возможной предсказуемости ссылок с помощью оценок альтметрии будет изучаться в последующем исследовании.

Наконец, хотя ссылки и альтметрия (в частности читатели Mendeley) проявляют умеренное позитивное отношение, все еще не ясно, каково качество данных по альтметрии и какой тип распределения влияния они могут представлять. Так как альтметрия пока еще находится в развитии, на данный момент у нас до сих пор нет четкого определения возможных значений оценок альтметрии. Другими словами, основной вопрос относительно того, что означает альтметрия, все еще не решен. С этой точки зрения также необходимо знать мотивации перед использованием этих источников данных, например в случае Mendeley: что это отражает, когда элемент сохранен/добавлен несколькими пользователями в свои библиотеки? Также, что означает то, что элемент упоминается в Википедии, CiteULike, Твиттере и любой другой платформе социальных медиа? Касается ли это одного и того же или различного измерения по сравнению со ссылкой? В том же ряду помимо изучения, в какой степени различные публикации представлены в Mendeley и других средствах соцмедиа и каковы их отношения с влиянием ссылок, нам необходимо изучить, для каких целей и почему эти платформы как раз используются различными учеными. Более того, исследование качества и надежности данных по альтметрии, найденных различными провайдерами альтметрии, все еще необходимо, прежде чем любая интерпретация и вероятные реальные использования для этих данных будут разработаны. Эта информация в сочетании с оценкой надежности и общей значимости данных и средств по альтметрии прольет больше света на значения альтметрии и поможет раскрыть скрытые величины альтметрии в последующих исследованиях.

*Благодарность.* Это исследование является расширенной версией нашего исследования, описанного в статье, представленной на 14-й конференции Международного общества по наукометрии и инфоформетрии, состоявшейся 15 – 19 июля 2013 г., в Вене, Австрия. Выражаем благодарность команде Impact Story за их поддержку в работе с Impact Story API. Данная работа частично выполнена при поддержке проекта EU FP7 ACUMEN (Соглашение по гранту: 266632). Авторы благодарят Эрика ван Вийка из CWTS за его огромную помощь в управлении данными по альтметрии. Авторы также признательны Людо Валтману из CWTS за полезные предложения и рекомендации анонимным рецензентам журнала.

## ЛИТЕРАТУРА

1. Moed H.F. Citation analysis in research evaluation.— Berlin/ Heidelberg/New York: Springer, 2005.

2. MacRoberts M. H., MacRoberts B. R. Problems of citation analysis: A critical review// Journal of the American Society for Information Science.— 1989.— Vol. 40.— P. 342–349.

3. Nicolaisen J. Citation Analysis// Annual Review of Information Science and Technology. — 2007.— Vol. 41.— P. 609–641.

4. Martin B. R., Irvine J. Assessing basic research: Some partial indicators of scientific progress in radio astronomy// Research Policy. — 1983. — Vol. 12. — P. 61–90.

5. Bornmann L., Leydesdorff L. The validation of (advanced) bibliometric indicators through peer assessments: A comparative study using data from InCites and F1000// Journal of Informetrics.— 2013.— Vol. 7, No.2.— P. 286–291.

6. Wouters P. The Citation Culture, PhD thesis. — University of Amsterdam, 1999.

7. Seglen P.O. Why the impact factor of journals should not be used for evaluating research// British Medical Journal. — 1997. — P. 314–497.

8. Bordons M., Fernandez M.T., Gomez I. Advantages and limitations in the use of impact factor measures for the assessment of research performance// Scientometrics. — 2002. — Vol. 53. — P. 195–206.

9. Thevall M. Weak benchmarking indicators for formative and semi-evaluative assessment of research// Research Evaluation. — 2004.—Vol. 13, No.1.— P. 63–68.

10. Butler L., McAllister I. Evaluating university research performance using metrics// European Political Science. — 2011. — Vol. 10, No.1. — P. 44–58.

11. Taylor J. The assessment of research quality in UK. Universities: Peer review or metrics?// British Journal of Management.— 2011.— Vol. 22, No.2.— P. 202–217.

12. Hicks D., Melkers J. “Bibliometrics as a Tool for Research Evaluation”// Handbook on the Theory and Practice of Program Evaluation/ Ed. Al Link & Nick Vornatas. — Edward Elgar, 2012. — [http://works.bepress.com/diana\\_hicks/31](http://works.bepress.com/diana_hicks/31)

13. Moed H. F. The future of research evaluation rests with an intelligent combination of advanced metrics and transparent peer review// Science and Public Policy. — 2007.— Vol. 34, No.8.— P. 575 – 583.

14. Benos D. J., Bashari E., Chaves J. M., et al. The ups and downs of peer review// Advances in Physiology Education. — 2007. — Vol. 31. — P.145–152.

15. Bollen J, Van de Sompel H., Hagberg A., Chute R. A Principal Component Analysis of 39 Scientific Impact Measures// PLoS ONE. —2009. — Vol. 4, No. 6 : e6022. DOI:10.1371/journal.pone.0006022.

16. Rousseau R., Ye F. A multi-metric approach for research evaluation// Chinese Science Bulletin. —2012. — P. 10– 12.

17. Nederhof A. J., Van Raan A. F. J. Peer review and bibliometric indicators of scientific performance: A comparison of cum laude doctorates with ordinary doctorates in physics// Scientometrics. — 1987. — Vol. 11, No.5-6. — P. 333–350.

18. Priem J., Pivovar H., Hemminger B. H. Altmetrics in the wild: Using social media to explore scholarly impact. — 2012. — ArXiv: 1203.4745v1.

19. Priem J., Hemminger B.H. Scientometrics 2.0: Toward new metrics of scholarly impact on the social Web// First Monday. — 2010. — No. 15. — <http://firstmonday.org/>

htbin/cgiwrap/bin/ojs/index.php/fm/article/view/2874/2570

20. *Galligan F., Dyas-Correia S.* Altmetrics: Rethinking the way we measure// *Serials Review*. — 2013. — Vol. 39, No.1. — P. 56–61.

21. *Bornmann L.* Is there currently a scientific revolution in Scientometrics?// *Journal of the American Society for Information Science & Technology*.— 2013.— <http://www.lutz-bornmann.de/icons/impactrevolution.pdf>

22. *Torres-Salinas D., Cabezas-Clavijo A., Jimenez-Contreras E.* Altmetrics: New indicators for scientific communication in web 2.0// *Comunicar*.—2013. — <http://arxiv.org/ftp/arxiv/papers/1306/1306.6595.pdf>

23. *Blecic D.* Measurements of journal use: An analysis of the correlations between three methods// *Bulletin of the Medical Library Association*.—1999. — Vol. 87.— P. 20–25.

24. *Duy J., Vaughan L.* Can electronic journal usage data replace citation data as a measure of journal use? An empirical examination// *The Journal of Academic Librarianship*. — 2006. — Vol. 32, No.5.— P. 512–517.

25. *Rowlands I., Nicholas D.* The missing link: Journal usage metrics// *Aslib Proceedings*. — 2007. — Vol. 59, No.3. — P. 222–228.

26. *Bollen J., Van de Sompel H., Rodriguez M. A.* Towards usage-based impact metrics// *Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL)*, New York, USA. — 2008.— <http://arxiv.org/pdf/0804.3791v1.pdf>.

27. *Shuai X., Pepe A., Bollen J.* How the scientific community reacts to newly submitted preprints: Article downloads, Twitter mentions, and citations. 2012. — <http://arxiv.org/abs/1202.2461v1>.

28. *Smith A. G.* A tale of two web spaces; comparing sites using web impact factors// *Journal of Documentation*. — 1999. — Vol. 55, No.5. — P. 577–592.

29. *Thehwall M.* Extracting macroscopic information from web links// *Journal of American Society for Information Science and Technology*. — 2001. — Vol. 52, No.13. — P. 1157–1168.

30. *Vaughan L., Shaw D.* Bibliographic and web citations: What is the difference?// *Journal of the American Society for Information Science and Technology*. — 2003. — Vol. 54, No.14. — P. 1313–1322.

31. *Thehwall M.* Bibliometrics to webometrics//*Journal of Information Science*. — 2008. — Vol. 34, No.4. — P. 605–621.

32. *Thehwall M.* A history of webometrics// *Bulletin of the American Society for Information Science and Technology*. — 2012. — Vol. 38, No.6. — P. 18–23.

33. *Thehwall M.* Journal impact evaluation: A webometric perspective// *Scientometrics*. — 2012. — Vol. 92, No.2. — P. 429–441.

34. *Haustein S.* Multidimensional journal evaluation// *Proceedings of the 11th International Conference on Science and Technology Indicators* (pp. 120–122), Leiden, the Netherlands. — 2010.

35. *Brody T., Harnad S., Carr L.* Earlier web usage statistics as predictors of later citation impact// *Journal of the American Society for Information Science*. — 2006. — Vol. 57. — P. 1060–1072.

36. *Armbruster C.* Access, usage and citation metrics: What function for digital libraries and repositories in re-

search evaluation? — 2007. — [http://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=1088453](http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1088453)

37. *Priem J., Taraborelli D., Groth P., Neylon C.* Altmetrics: A manifesto. — 2010. — <http://altmetrics.org/manifesto/>

38. *Wouters P., Costas R.* Users, narcissism and control: Tracking the impact of scholarly publications in the 21st century. — Utrecht: SURF foundation, 2012. — <http://www.surffoundation.nl/nl/publicaties/Documents/Users%20narcissism%20and%20control.pdf>

39. *Zhang Y.* Comparison of select reference management tools// *Medical reference services quarterly*. — 2012. — Vol. 31, No.1. — P. 45–60.

40. *Li X., Thehwall M., Giustini D.* Validating online reference managers for scholarly impact measurement// *Scientometrics*. — 2012. — Vol. 91, No.2. — P. 461–471.

41. *Henning V.* The top 10 journal articles published in 2009 by readership on Mendeley. — 2010. — <http://www.mendeley.com/blog/academic-features/the-top-10-journal-articles-published-in-2009-byreadership-on-mendeley/>.

42. *Bar-Ilan J.* JASIST@ Mendeley// *ACM Web Science Conference Workshop on Altmetrics*, Evanston, IL, 21 June 2012. —2012. — <http://altmetrics.org/altmetrics12/bar-ilan/>

43. *Schlögl C., Gorraiz J., Gumpenberger C., Jack K., Kraker P.* Download vs. citation vs. readership data: The case of an information systems journals/ *J. Gorraiz, E. Schiebel, C. Gumpenberger, M. Hörlesberger, & H. Moed* (eds.)// *Proceedings of the 14th International Society of Scientometrics and Informetrics Conference*, Vienna, Austria (pp. 626–634). — Wien: Facultas Verlags und Buchhandels AG, 2013.

44. *Thehwall M., Haustein S., Larivière V., Sugimoto C.R.* Do altmetrics work? Twitter and ten other social web services// *PLoS ONE*. — 2013. — Vol.8, No.5: e64841.

45. *Haustein S., Peters I., Bar-Ilan J., Priem J., Shema H., Terliesner J.* Coverage and adoption of altmetrics sources in the bibliometric community// *14th International Society of Scientometrics and Informetrics Conference* (pp. 1–12). *Digital Libraries*. — 2013. — <http://arxiv.org/abs/1304.7300>

46. *Haustein S., Siebenlist T.* Applying social bookmarking data to evaluate journal usage// *Journal of Informetrics*. — 2011. — Vol. 5. — P. 446–457.

47. *Li X., Thehwall M.* F1000, Mendeley and traditional bibliometric indicators// *Proceedings of the 17th International Conference on Science and Technology Indicators*. Montréal, Canada (pp. 451–551). — 2012.

48. *Waltman L., Costas R.* F1000 recommendations as a potential new data source for research evaluation: A comparison with citations// *Journal of the Association for Information Science and Technology*. — 2013. — DOI: 10.1002/asi.23040

49. *Eysenbach G.* Can tweets predict citations? Metrics of social impact based on twitter and correlation with traditional metrics of scientific impact// *Journal of Medical Internet Research*. — 2011. — Vol.13, No.4: e123.

50. *Davis P. M.* Tweets, and our obsession with alt metrics// *The Scholarly Kitchen*. — 2012. — <http://scholarlykitchen.sspnet.org/2012/01/04/tweets-and-our-obsession-with-alt-metrics/>

51. *Shema H., Bar-Ilan J., Thehwall M.* Do blog citations correlate with a higher number of future citations? Research blogs as a potential source for alternative metrics/ *J.*

Gorraiz, E. Schiebel, C. Gumpenberger, M. Hörlesberger, H. Moed (eds.)// Proceedings of the 14th International Society of Scientometrics and Informetrics Conference, Vienna, Austria (pp. 604-611). — Wien: Facultas Verlags und Buchhandels AG, 2013.

52. Bar-Ilan J., Haustein S., Peters I., Priem J., Sbema H., Terliesner J. Beyond citations: Scholars' visibility on the social Web// Proceedings of the 17th International Conference on Science and Technology Indicators, Montreal, Quebec. —2012. — <http://arxiv.org/abs/1205.5611/>

53. Forta B. Sams teach yourself SQL in 10 minutes. — USA: Sams Publishing, 2008.

54. López-Cózar E.D., Robinson-García Nicolas, Torres Salinas D. Manipulating Google Scholar citations and Google Scholar metrics: Simple, easy and tempting. — 2012. — <http://arxiv.org/abs/1212.0638>

55. Moed H. F. New developments in the use of citation analysis in research evaluation// Archivum immunologiae et therapiae experimentalis. — 2009. — Vol. 57, No.1. — P. 13–18.

56. Van Raan A. F. J., Van Leeuwen T. N., Visser M. S. Severe language effect in university rankings: Particularly Germany and France are wronged in citation-based rankings// Scientometrics. — 2011. — Vol. 88, No.2. — P. 495–498.

57. Archambault É., Larivière V. (2006). The limits of bibliometrics for the analysis of the social sciences and humanities literature, International Social Science Council: World social sciences report 2010: Knowledge divides. — Paris: UNESCO, 251–254.

58. Torres-Salinas D., Robinson-García N., Campanario J. M., López-Cózar, E. D. Coverage, field specialisation and the impact of scientific publishers indexed in the Book Citation Index// Online Information Review. — 2013. — Vol. 38, No.1. — P. 24–42.

59. Nederhof A.J. Bibliometric monitoring of research performance in the social sciences and the humanities: A review// Scientometrics. — 2006. — Vol. 66, No. 1. — P. 81–100.

60. Waltman L., Van Eck N.J., Van Leeuwen T.N., Visser M.S., Van Raan A.F.J. Towards a new crown indicator: Some theoretical considerations// Journal of Informetrics. — 2011. — Vol. 5, No. 1.— P. 37–47.

61. Zabedi Z., Costas, R., Wouters P. How well developed are Altmetrics? Cross disciplinary analysis of the presence of 'alternative metrics' in scientific publications (RIP)//J. Gorraiz, E. Schiebel, C. Gumpenberger, M. Hörlesberger, & H. Moed (eds.) //Proceedings of the 14th International Society of Scientometrics and Informetrics Conference, Vienna, Austria (pp. 876-884).— Wien: Facultas Verlags und Buchhandels AG, 2013.

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

Ответственный за выпуск *Л. В. Кобзева*

Компьютерная верстка *М. А. Филимонова*

ИД № 04689 от 28.04.2001 г.

Подписано в печать 11.09.2014 г.

Бумага «Хегох». Формат 60x841/8. Гарн. литер. Печать цифровая

Усл. печ. л 4,50 Уч.-изд. л. 5,15 Тираж 36 экз.

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

Тел. (499) 155-44-95