ОРГАНИЗАЦИЯ ИНФОРМАЦИОННОЙ РАБОТЫ

УДК 004.891:378.4(04/07)

В. М. Московкин

Возможности использования поисковой машины Google Scholar для оценки публикационной активности университетов

Изучены возможности использования поисковой машины Google Scholar для оценки публикационной активности университетов и рассмотрена процедура такой оценки с помощью запросов на англоязычные названия университетов. Построены публикационные структуры за 2008 г. для десяти избранных ведущих университетов мира, включая МГУ им. М. В. Ломоносова. Сделано сравнение публикационной активности рассматриваемых университетов в 2007 г. на основе баз данных по цитируемости Института научной информации США (Web of Knowledge) и поисковой машины Google Scholar (GS-publications).

Ключевые слова: *поисковая машина Google Scholar, университетская публикационная активность, открытый доступ, индексы цитируемости, публикационные структуры*

В настоящее время в зарубежной научной литературе, посвященной наукометрическим методам исследования, возник кластер публикаций, связанный с использованием поисковой машины Google Scholar при проведении таких исследований. Отмечается, что до последнего времени единственным всеохватывающим источником данных по цитированию были базы данных Института научной информации США (ISI Citation Indexes) [1].

Несмотря на критическое отношение к этим базам данных, по причине некоторого их несовершенства и отсутствия других, они уже давно широко используются за рубежом в научном менеджменте.

Относительно недавно возникли две альтернативы этим базам данных — коммерческая поисковая система Scopus, развитая крупнейшим издательством научной периодики Elsevier и свободно доступная поисковая машина Google Scholar [1].

В работах [2, 3] показано, что Google Scholar покрывает гораздо большее количество научных документов, по сравнению с базами данных Института научной информации США, и тем самым вносит огромный вклад в движение по открытому доступу к результатам научных исследований.

В работе [4] отмечается, что поисковая машина Google Scholar обеспечивает новый метод обнаружения потенциально релевантных статей по данной проблематике за счет идентификации статей, которые были процитированы в других работах. Поэтому важной особенностью этой поисковой машины является то, что исследователи могут использовать ее для отслеживания взаимных связей между авторами, цитирующими статьи на подобную тему, а также для определения частоты, с которой другие авторы цитируют конкретную статью (за счет опции "cited by"). Здесь же сделано заключение, что поисковая машина Google Scholar обеспечивает свободную альтернативу и дополнения к другим индексам цитирования.

Базы данных Института научной информации США индексируют около одной трети от общего количества рецензируемых научных журналов, которых сейчас насчитывается около 25 тысяч. При этом Google Scholar и Google Books индексируют гораздо больше научных документов, но пока не в состоянии достичь полного их покрытия, так как только 15% текущего ежегодного научного выхода представлено публикациями открытого доступа (OA — publications) [5].

Наш обзор наукометрических исследований показал отсутствие работ, которые бы изучали публикационную структуру и вебометрические оценки университетского научного выхода с помощью поисковой машины Google Scholar.

При изучении такого выхода мы обратили внимание на то, что его нельзя качественно получить с помощью измерения откликов на URL-адреса сайтов университетов. Для постсоветских университетов часто появляется большое количество нерелевантных откликов в виде различной административной информации (решения ученого совета, ректората и др.). Для западных университетов возникает ситуация, когда, например, при приблизительно одинаковой публикационной активности ученых американских (Гарвардский и Чикагский университеты) и британских (Кембриджский

и Оксфордский университеты) университетов первые имели на порядок больше откликов на запросы из URL — адресов[1] хотя запросы на названия этих университетов дали преимущества британским университетам[2].

Причина этого состоит, на наш взгляд, в качестве организации информации на сайте. Так, множество откликов на сайт Гарвардского университета связано с наличием научного интернет-магазина (sciencemag.org), а при запросе сайта Чикагского университета (site: uchicago.edu) первая тысяча откликов, разрешенная для показа поисковой машиной Google Scholar, идет на статьи превосходно составленной коллекции журналов университета ("Chicago Journals"), размещенной на платформе uchicago.press.

Очевидно, что на сайте университетов представлены далеко не все публикации их ученых, а для постсоветских университетов вообще отсутствует практика размещения опубликованных научных статей на их сайтах.

В связи с вышеизложенным, мы решили испытывать с помощью поисковой машины Google Scholar не URL-адреса сайтов университетов, как это делается испанской кибермеетрической лабораторией при расчете вебометрического рейтинга университетов мира (www.webometrics.info), а их общепринятые англоязычные названия. Эксперименты с ведущими университетами мира показали хорошую релевантность такого поиска. В первую очередь Google Scholar находит статьи, размещенные на онлайновых платформах крупнейших издательств таких, как Elsevier, Springer, Blackweel, Wiley и др., т. е. "конвертируемые" статьи, входящие в базы данных Института научной информации США. Кроме того, эта поисковая машина хорошо находит статьи из онлайновых журналов и университетских репозитариев открытого доступа.

Отметим также, что Google Scholar в результате своего поиска включает дополнительно небольшой процент научных монографий, предоставляемых Google Books.

Нам также удалось показать, что релевантность расширенного поиска с точной фразой возрастает в направлении: отсутствие ограничений на области наук и интервалы времени→задание областей наук→задание одновременно областей наук и временных интервалов поиска.

Помимо общего количества статей в данной области знаний (7 областей), полученных на запрос англоязычного названия конкретного университета, Google Scholar дает значения общего числа ссылок на каждую статью с возможностью просмотра названий научных работ, цитирующих данную статью (с помощью опции "by cited"). Наши контакты с командой Google Scholar показали, что пока отсутствует процедура, которая позволяла бы суммировать цитирования по всей совокупности найденных статей, но команда Google Scholar с интересом восприняла идею разработки такой процедуры. При ее реализации возникает возможность рассчитывать полноценный вебометрический рейтинг научно-публикационной активности университетов мира. При расчете такого рейтинга возникает проблема идентификации всех общепринятых названий университетов. Например, для университетов франкоговорящих провинций Канады необходимо использовать общепринятые франко- и англоязычные названия, для европейских университетов неанглоговорящих стран, помимо англоязычных названий, необходимо использовать все их основные иностранноязычные названия в соответствии с принятыми в этих странах языками. Для постсоветских стран следует учитывать перманентный процесс переименования университетов. При работе с поисковой машиной Google Scholar мы отметили флуктуацию откликов на запросы названий университетов, что связано с возможным временным отсутствием доступа, исключением дублирующих или нерелевантных откликов и др. Поэтому при расчете итогового кумулятивного вебометрического показателя целесообразно, на наш взгляд, использовать сглаживающие процедуры (рассчитывать усредненный временной тренд).

Мы полагаем, что со временем, по мере активизации создания университетских репозитариев открытого доступа, будет возрастать вероятность дублирования откликов на запросы URL-адресов сайтов университетов, так как в таких репозитариях будут размещаться (самоархивироваться) ранее опубликованные статьи (в основном, в виде авторских PDF-файлов).

Трудно заранее сказать, насколько эффективно поисковая машина Google Scholar будет справляться с нарастающим масштабом дублирования статей.

В качестве экспериментов с поисковой машиной Google Scholar нами выбраны девять зарубежных университетов, которые занимали наиболее высокие позиции по количеству опубликованных статей (входящих в базы данных SCI и SSCI Института научной информации США) в китайском и тайванском рейтингах университетов мира 2008 г. Для сравнения выбран ведущий постсоветский университет — МГУ им. М. В. Ломоносова. Публикационные структуры 2008 г. для этих университетов в количественном и процентном соотношении, а также укрупненная публикационная структура, показаны в табл. 1 и 2. В шапках этих таблиц приведены основные названия университетов, по которым осуществлялся расширенный поиск с точной фразой. Обращенные названия этих университетов (например, University of Chicago—Chicago University) также учитывались в запросах Google Scholar за исключением университета Джона Гопкинса, Калифорнийского, Токийского и Московского университетов. Наибольшие доли откликов для обращенных названий наблюдались для университетов из Чикаго, Кембриджа и Оксфорда. Отклики на обращенное название Токийского университета часто приводили к другим университетам (Tokyo University of Agriculture — Technology, Science), и поэтому не учитывались в суммарных оценках. Данные табл. 2 рассчитаны нами на основе процентного распределения данных табл. 1. Например, для Гарвардского университета доля публикаций в области наук о жизни составила: 9,7+5,7=15,4%. Из табл. 2 следует, что преобладают научные школы социально-экономического и гуманитарного

[1]site: harvard.edu дает 1 310 000 документов, site: uchicago.edu — 60 400, site: ox.ac.uk — 8090, site: cam.ac.uk — 9330, измерения проводились нами в начале января 2009 г.

[2]расширенный поиск с точной фразой (advance search, with exact phrase)

Таблица 1

Публикационная структура для избранных крупнейших университетов мира за 2008 г., полученная с помощью поисковой машины Google Scholar 22.01.2009 г.

Области наук	Stanford University	Harvard University	Columbia University	University of California-Berkeley	Johns Hopkins University
1. Biology, Life Sciences, and Environmental Science	2346/8,8	3102/9,7	2977/12,2	1670/14,2	2240/10,1
2. Business, Administration, Finance, and Economics	2686/10,1	5802/18,1	2210/9,0	1090/9,3	1470/6,7
3. Chemistry and Materials Science	1118/4,2	822/2,5	935/3,8	980/8,3	584/2,6
4. Engineering, Computer Science, and Mathematics	4455/16,7	2161/6,7	1680/6,9	2370/20,1	2200/9,9
5. Medicine, Pharmacology, and Veterinary Science	5236/19,7	1832/5,7	4391/18,0	698/5,9	5180/23,5
6. Physics, Astronomy, and Planetary Science	2112/7,9	1591/5,0	1523/6,2	1950/16,6	1630/7,4
7. Social Sciences, Arts, and Humanities	8682/32,6	16813/52,3	10729/43,9	3010/25,6	8780/39,8
Всего	26635/100	32123/100	24445/100	11768/100	22084/100

Области наук	Chicago University	Cambridge University	Oxford University	University of Tokyo	Moscow State University
1. Biology, Life Sciences, and Environmental Science	4453/8,6	17920/13,1	15100/12,3	2460/19,3	560/14,8
2. Business, Administration, Finance, and Economics	6850/13,3	11759/8,6	10520/8,6	376/2,9	57/1,5
3. Chemistry and Materials Science	736/1,4	7340/5,3	8167/6,7	1720/13,5	957/25,2
4. Engineering, Computer Science, and Mathematics	1914/3,7	20230/14,7	9703/7,9	1910/15,0	583/15,4
5. Medicine, Pharmacology, and Veterinary Science	4870/9,4	9670/7,1	19970/16,3	2320/18,2	71/1,9
6. Physics, Astronomy, and Planetary Science	4017/7,8	19190/14,0	7290/6,0	3260/25,6	1380/36,4
7. Social Sciences, Arts, and Humanities	28830/55,8	50980/37,2	51570/42,2	703/5,5	180/4,8
Всего	51670/100	137089/100	122320/100	12749/100	3788/100

Примечание: в числителе количество публикаций, в знаменателе — %

направлений в Гарвардском и Чикагском университетах. В университетах Калифорнии, Токио и Москвы наблюдается обратная картина. Научные школы в области наук о жизни наиболее весомо представлены в университетах Токио и Джона Гопкинса. Постсоветская публикационная структура, представленная научным выходом МГУ им. М. В. Ломоносова, характерна явным преобладанием "конвертируемых" естественнонаучных и технических публикаций, а, следовательно, и научных школ, естественнонаучного и технического направления (за исключением медико-биологических научных школ).

Таблица 2

Укрупненная публикационная структура для избранных крупнейших университетов мира за 2008 г., полученная с помощью поисковой машины Google Scholar 22.01.2009 г. (%)

Укрупненные области наук	Stanford University	Harvard University	Columbia University	University of California-Berkeley	Johns Hopkins University
Естественные и технические науки, кроме наук о жизни	28,8	14,2	16,9	45,0	19,9
Науки о жизни	28,5	15,4	30,2	20,1	33,6
Социально-экономические и гуманитарные науки Искусство	42,7	70,4	52,9	34,9	46,5
Всего	100	100	100	100	100

Укрупненные области наук	Chicago University	Cambridge University	Oxford University	University of Tokyo	Moscow State University
Естественные и технические науки, кроме наук о жизни	12,9	34,0	20,6	54,1	77,0
Науки о жизни	18,0	20,2	28,6	37,5	16,7
Социально-экономические и гуманитарные науки Искусство	69,1	45,8	50,8	8,4	6,3
Всего	100	100	100	100	100

Таблица 3

Публикационная активность крупнейших избранных университетов мира, полученная на основе данных тайванского рейтинга университетов мира и поисковой машины Google Scholar, 2007 г.

Университеты	Количество статей			GS-publications/ Thomson-Reuter
	Thomson-Reuter		GS-publications	
	%	абс. значение		
Harvard University	100	11 221	46 768	4,2
University of Tokyo	62,51	7 014	15 495	2,2
Johns Hopkins University	52,38	5 878	27 124	4,6
University of California - Berkeley	47,67	5 349	14 571	2,7
Stanford University	47,87	5 372	35 320	6,6
Columbia University	43,10	4 836	32 990	6,8
Oxford University	39,60	4 444	142 344	32,0
Cambridge University	39,35	4 416	173 060	39,2
Chicago University	34,78	3 903	73 016	18,7
Moscow State University	28,05	3 148	5 021	1,6

Сопоставим теперь публикационную активность рассматриваемых университетов, полученную на основе баз данных по цитируемости Института научной информации США (ISI) и поисковой машины Google Scholar (GS). С этой целью мы обратились к тайванскому рейтингу университетов мира (Ranking of Scienific Papers for World Universities). В этом рейтинге имеется показатель Current Articles, который представляет собой годовое количество публикаций, полученных на основе баз данных SCI и SSCI (Thomson — Reuter). В рейтинге университетов мира за 2008 г. этот пока-

затель рассчитан на 2007 г. Его максимальное значение, взятое за 100%, имело место для Гарвардского университета и равнялось 11 221 статей.[3] На основе максимальной величины этого показателя нами пересчитаны его абсолютные значения для всех остальных университетов (табл. 3). За этот же год рассчитаны и количества научных статей, полученные описанным ранее способом с помощью поисковой машины Google Scholar (GS-publications в табл. 3). Дополнительно, в табл. 3 рассчитано превышение вебометрического показателя университетской публикационной активности над традиционным ее показателем. Как видим, это отношение варьирует достаточно сильно. В то же время логично предположить, что отношение общего количества публикаций к "конвертируемым" публикациям (Thomson–Reuter) для различных университетов является приблизительно постоянной величиной, т. е. между этими показателями должна быть достаточно хорошая линейная корреляция. Отсутствие такой корреляции между показателями Thomson–Reuter и GS-publications говорит только о плохом Web-представлении публикаций для университетов, у которых отношение GS-publications/Thomson–Reuter является заниженным.

Для более корректных расчетов в показатель Thomson–Reuter необходимо включать статьи из базы данных A&HCI, так как поисковая машина Google Scholar охватывает такие статьи.

Следует отметить, что в шанхайском рейтинге университетов мира показателю Current articles (тайванский рейтинг) полностью соответствует показатель PUB, но он прямо не может использоваться для пересчета абсолютных значений университетских публикаций, входящих в базы данных SCI и SSCI, так как для социально-экономических статей использовался повышающий коэффициент 2.

Таким образом, нами показана возможность количественной оценки публикационной активности университетов с помощью поисковой машины Google Scholar, подтверждены результаты зарубежных исследований по более широкому охвату научных публикаций этой поисковой машиной по сравнению с базами данных Института научной информации США и построены публикационные структуры для десяти избранных ведущих университетов мира. Дальнейшее развитие данного подхода должно идти по пути вычленения книжных публикаций (метка "Book") и ссылок (метка "Citation") в откликах поисковой машины Google Scholar, несмотря на небольшой процент этих откликов. Но эта работа, вместе с подсчетом общего количества ссылок по всем найденным академическим документам (опция "by sited") может быть проделана только в сотрудничестве с командой Google Scholar.

СПИСОК ЛИТЕРАТУРЫ

1. Judit Bar-Ilan. Which h-index? — A Comparison of WoS, Scopus and Google Scholar // Scientometrics.— 2008.— Vol. 74, № 2.— P. 257–271.

2. Kayvan Kousha, Mike Thelwall. Google Scholar citations and Google Web/URL citations: A multi-discipline esploratory analysis // Journal of the American Sodiety for Information Science and Technology.— 2007.— Vol. 58, № 7.— P. 1055–1065.

3. Kayvan Kousha, Mike Thelwall. Sources of Google Scholar citation outside the Science Citation Index: A comparison between four science disciplines // Scientometrics.— 2008.— Vol. 74, № 2.— P. 273–294.

4. Alireza Noruzi. Google Scholar: The New Generation of Citation Indexes // Libri.— 2005.— Vol. 55, № 4.— P. 170–180.

5. Tim Brody, Les Carr, Yves Gingras, Chawki Hajjem, Stevan Harnad, Alma Swan. Incentivizing the Open Access Research Web Publication — Archiving, Data-Archiving and Scientometrics // CTWatch Quarterly.— 2007 (August).

Материал поступил в редакцию 20.03.09.

[3] Абсолютное значение показателя "Current articles" было любезно предоставлено нам Ru-rong Hsiao (Chief of Performance Evaluation Section HEEACT, Совет по оцениванию и аккредитации вузов Тайваня).