УДК [004.774:004.891]:001 8

В.М. Московкин

Построение кластеров результатов исследований с помощью специализированных инструментов Google

С помощью возможностей поисковых машин Google Scholar, Google Books и Google Patents построены процедуры для количественной идентификации кластеров результатов исследований (статей, книг и патентов), порождаемых произвольными научными терминами. Эти процедуры апробированы на терминах «gene transfer» и «gene synthesis». Построены временные ряды публикационной и патентной активности для этих терминов и введен коэффициент патентоёмкости результатов исследований.

Ключевые слова: *Google Scholar, Google Books, Google Patents, кластер результатов исследований, gene transfer, gene synthesis, коэффициент патентоёмкости результатов исследований*

Под кластером результатов исследований, порожденным произвольным научным термином, мы будем понимать совокупность научных статей, книг и патентов, в которых встречается данный термин. Такие кластеры будем понимать, как в широком, так и в узком смысле. В первом случае предполагается, что данный термин имеется в текстах научных публикаций и патентов, а во втором – в их заголовках. Для построения таких кластеров могут применяться специализированные инструменты Google – Google Scholar, Google Books и Google Patents.

Google Scholar, запущенный в ноябре 2004 г., уже широко используется для наукометрического анализа, включая анализ откликов на запросы научных терминов [1-6]. Возможности продвинутого наукометрического анализа терминов с помощью Google Books (запущен в декабре 2004 г.) открылись недавно после запуска в августе 2010 г. аналитико-поискового инструмента Ngram Viewer [7, 8], хотя отсутствуют работы по использованию стандартных возможностей Google Books в таком анализе. Google Patents, запущенный в декабре 2006 г., практически не применяется в патентометрическом анализе, за исключением работы [9], в которой такой анализ проводился для ведущих университетов и транснациональных компаний.

Из изложенного следует, что кластер результатов исследований (research output cluster) разбивается на три отдельных кластера: кластер научных статей (research paper cluster), кластер научных книг (research book cluster) и кластер патентов (patents cluster). При этом под кластером первичных результатов исследований (primary research output cluster) будем понимать совокупность кластеров научных статей и патентов.

Далее опишем методику построения (или идентификации) кластеров результатов исследований на примере терминов *gene transfer* и *gene synthesis*.

КОЛИЧЕСТВЕННАЯ ИДЕНТИФИКАЦИЯ КЛАСТЕРОВ НАУЧНЫХ СТАТЕЙ И КНИГ

Тестирование произвольных терминов с помощью Google Scholar необходимо проводить в расширенном поиске с точной фразой, при этом для уменьшения информационного шума целесообразно поставить отметки перед всеми предметными категориями. Последняя операция автоматически исключит добавление патентов в результаты поиска. Для построения широкого кластера научных публикаций необходимо проводить поиск по всей статье (anywhere in the article), а для построения узкого кластера – по названиям статей (in the title of the article). В обоих случаях в результаты поиска можно включить цитируемые статьи (include citation) или, по крайней мере, абстракты (at least summaries). В категорию «citation» попадают статьи из списков литературы (references) ранее оцифрованных и индексированных статей. Проделав такие эксперименты с нашими двумя терминами, получим табл. 1.

Из табл. 1 видно, что в количественном отношении кластер научных статей, порожденный термином *gene transfer*, на несколько порядков превышает кластер научных статей, порожденный термином *gene synthesis*. Если для широкого кластера научных публикаций такое превышение составило 95-98 раз, то для узкого кластера научных публикаций – 45-50 раз.

Google Scholar позволяет идентифицировать наиболее цитируемые и наиболее ранние индексируемые статьи для произвольных терминов. Такая идентификация для изучаемых нами терминов приведена в табл.2. Из нее видим, что, в среднем, наиболее цитируемые статьи по тематике генетического трансфера цитировались на порядок больше, чем аналогичные статьи по тематике генетического синтеза. Анализ наиболее ранних цитируемых статей в обоих кластерах публикаций показывает, что они зародились практически одновременно в начале сороковых годов XX в., но наиболее релевантные статьи (в названиях которых

Таблица 2

Наиболее цитируемые и наиболее ранние индексируемые статьи для терминов *gene transfer* и *gene synthesis*, 28.12.2011 г.

Термин	Наиболее цитируемые статьи		Наиболее ранние индексируемые статьи	
	Поиск по всей статье	Поиск по названиям статей	Поиск по всей статье	Поиск по названиям статей
Gene transfer	1.O. Boussif, F. Lezoualch. A versatile... cited by 3361 2. J.A.Wolfs... Direct gene transfer cited by 3078 3. A.D.Miller... Improved retroviral... cited by 1859	1.O. Boussif, F. Lezoualch. A versatile... cited by 3361 2. J.A.Wolfs.. Direct gene transfer cited by 3078 3. A.D.Miller... Improved retroviral... cited by 1859	1. T.H.Goodspeed... Amphidiploidy...//The Botanical Review..., 1942 cited by 24 2. S.Spiedelman... Maintenance 1945 cited by 43 3. J.T.Patterson...A new... 1946 cited by 61	1. P.D.Skaar, Alan Garen. The orientation and extent..., 1956 cited by 31 2. R.Moav. Inheritance...1958 cited by 8 3. R.Weinbery. Gene transfer.-1960 cited by 7
Gene synthesis	1. M.H.Caruthers. Gene synthesis machines //Science..., 1985 cited by 560 2. .Prodromou...Recursive PCR..., 1992 cited by 245 3 J.Tian...Accurate multiplex...2004 cited by 239	1. M.H.Caruthers. Gene synthesis machines //Science..., 1985 cited by 560 2. .Prodromou...Recursive PCR..., 1992 cited by 245 3.J.Tian...Accurate multiplex...2004 cited by 239	1. A.H.K.Petric. Protein Synthesis in Plants// Biological Rev...1943 cited by 15 2.H.J.Muller...Pelgrim Trust Lecture: The Gene... 1947 cited by 89 3.K.C.Stern Nucleoproteins and gene structure...1947 cited by 16	1.T.Kasai. Regulation of gene specific RNA synthesis in bacteriophage T. 4 //J. of Molecular Biology. -1969 cited by 51 2.J.R.Murphy...Synthesis... 1974 cited by 58 3.K.L.Agarwal. The synthesis... 1975 cited by 3

встречаются рассматриваемые термины) из первого кластера появились в интернете на 10-15 лет раньше, чем во втором кластере.

Чтобы оценить качество кластера результатов исследований, помимо размера этого кластера (количество публикаций), следует ввести суммарный показатель цитируемости публикаций этого кластера. Например, для кластера научных статей можно ввести вектор (N, C), где N – количество статей кластера, C – суммарная цитируемость этих статей.

Таблица 1

Тестирование терминов *gene transfer* и *gene synthesis* с помощью Google Scholar, 28.12.2011 г.

Термин	Количество откликов			
	Поиск по всей статье		Поиск по названиям статей	
	включая цитирование	по крайней мере, абстракты	включая цитирование	по крайней мере, абстракты
gene transfer	539 000	502 000	16 100	12 100
gene synthesis	5 520	5 280	325	269

Таблица 3

Тестирование терминов *gene transfer* и *gene synthesis* с помощью Google Books, 28.12.2011 г.

Термин	Количество откликов	
	Поиск с точной фразой по всей книге	Поиск с точной фразой по названиям книг
gene transfer	580	162
gene synthesis	493	21

Так как в Google Scholar отсутствует опция для расчета суммарного цитирования, то его можно оценить вручную, например, при подсчете цитирования первых 100 откликов. Проделав это для наших терминов для условия последнего столбца табл. 1, мы получили на уровень 09.02.2012 г. следующие два вектора: (12 300, 67632); (272, 4866), где первый соответствует термину gene transfer, а второй - gene synthesis. Отношение количества статей в этих векторах равняется 12 300/272 ≈ 50, а отношение ссылок на них – 67 632/4 866 ≈ 14, т.е. можно сделать вывод, что интенсивность цитирования

статей во втором кластере шло выше, хотя первый кластер является более мощным.

Google Books позволяет вести расширенный поиск с точной фразой. В отличие от Google Scholar, в нем нет опций «cited by», «citation» и предметных категорий.

Помимо книг, Google Books индексирует целые тома журналов и сборников (Magazine). Имеется возможность вести поиск только по книгам, включая поиск с точной фразой по всей книге и аналогичный поиск по названиям книг. В последнем случае тестируемый термин необходимо брать в кавычки. Проделав такое тестирование с нашими терминами, мы построили табл. 3, сопоставимую с табл. 1. В случае поиска по названиям статей и книг из табл. 1 и табл. 3 следует, что количество статей для термина *gene transfer* превышает количество книг на два порядка, а для термина *gene synthesis* такое превышение было на один порядок.

КОЛИЧЕСТВЕННАЯ ИДЕНТИФИКАЦИЯ КЛАСТЕРОВ ПАТЕНТОВ

Тестирование произвольных терминов с помощью Google Patents можно проводить двумя способами:

- расширенный поиск с точной фразой заявок на патенты (Applications) и выданных патентов (Issued patents);
- расширенный поиск заявок на патенты и выданных патентов в строке «названия патентов» (Title).

В первом случае поиск осуществляется по всему описанию патента, во втором – по его названию. Чтобы поиск по названию патента происходил с точной фразой, необходимо термин брать в кавычки. Проделав такие эксперименты с нашими двумя терминами, получим табл. 4. Важно отметить, что Google Patents, в отличие от Google Scholar, дает некоторое количество релевантных откликов, отличных от их общего числа. Поэтому в табл. 4 мы видим заниженные показатели, причем здесь количество заявок на патенты для термина *gene transfer* в названии документа меньше количества выданных патентов, а должно быть наоборот. Наши эксперименты с различными названиями университетов и транснациональных компаний [9] показали, что точность откликов возрастает по мере дробления временных интервалов. Отметим, что в табл. 4 эксперименты проводились без ограничения на временные интервалы.

Таблица 4

Тестирование терминов *gene transfer* и *gene synthesis* с помощью Google Patents, 28.12.2011 г.

Термин	Количество откликов, найденных при расширенном поиске			
	с точной фразой		в названиях патентов	
	Заявки на патенты	Выданные патенты	Заявки на патенты	Выданные патенты
Gene transfer	480	411	123	134
Gene synthesis	418	402	9	4

Google Patents так же, как и два других инструмента Google, позволяет строить временные ряды заявок на патенты и выданных патентов для произвольных терминов.

Такие временные ряды, построенные на ежегодной основе для изучаемых терминов, приведены в табл. 5. Здесь индикатор «заявки на патенты» («applications») рассчитывается по индикатору «дата подачи заявки» («filling date»). Расширенный поиск с точной фразой производился с использованием строки «названия патента» (Title).

Как видим, общее количество заявок на патенты для термина *gene transfer* (271) в 2,2 раза превышает количество заявок на патенты в табл. 4 (123), для выданных патентов это превышение было меньше (в 1,1 раза).

Таблица 5

Динамика откликов (патентов) для терминов *gene transfer* и *gene synthesis*, 28.12.2011 г.

Год	Gene transfer		Gene synthesis	
	Заявки на патенты	Выданные патенты	Заявки на патенты	Выданные патенты
1980	1	0	0	0
1981	0	0	0	0
1982	1	0	1	0
1983	1	1	0	0
1984	1	0	0	0
1985	0	1	0	0
1986	0	0	1	1
1987	2	1	2	0
1988	1	0	1	0
1989	0	1	1	1
1990	0	0	1	0
1991	1	1	0	1
1992	1	1	0	2
1993	1	1	0	0
1994	16	2	0	0
1995	16	0	1	0
1996	11	2	1	0
1997	19	7	1	0
1998	15	14	0	1
1999	19	21	0	1
2000	14	11	1	1
2001	27	15	0	0
2002	31	11	1	0
2003	22	15	3	0
2004	19	3	1	0
2005	15	9	2	1
2006	8	12	3	1
2007	9	3	1	1
2008	9	10	3	3
2009	6	2	2	1
2010	5	5	0	3
Всего	271	149	27	18

Резкие скачки в росте заявок на патенты для первого термина наблюдались в 1994 и 2001 гг., а в росте выданных патентов – в 1997 и 1998 гг.

СРАВНИТЕЛЬНЫЙ АНАЛИЗ ДИНАМИКИ НАУЧНЫХ СТАТЕЙ, КНИГ И ПАТЕНТОВ

Google Scholar, Google Books и Google Patents позволяют строить временные ряды научных публикаций для произвольных терминов. Такие временные ряды, построенные на пятилетних интервалах для изучаемых нами терминов, приведены в табл. 6. При подсчете количества научных статей использовалась опция "at least summaries".

Из табл. 6 видим, что максимальный прирост количества научных статей на смежных пятилетних интервалах времени для термина *gene transfer* происходил в 1970-е гг. и первую половину 1980 гг. (в 3-4 раза). В дальнейшем этот прирост уменьшился до 1,2 раза (отношение количества откликов за 2001-2005 гг. к их количеству за предыдущие 5 лет) с последующим спадом. Максимальное количество откликов достигло 3760 в период с 2001 по 2005 гг. Для термина *gene synthesis* наблюдалась одна полная волна с максимумом количества откликов в интервале 1986-1990 гг. и наметилась вторая волна с максимумом после 2005 г.

Для книг максимум откликов наблюдался в интервале 1996-2000 гг. (см. табл. 6).

В отличие от Google Scholar, Google Books дает большие расхождения при запросах с ограничениями времени и без ограничений. Для термина «gene transfer» такое расхождение было в 4 раза (669/162=4,1) (см. табл. 3, 6).

Для кластера первичных результатов исследований введем коэффициент патентоёмкости исследований как отношение количества выданных патентов к количеству научных статей. Рассчитаем такие коэффициенты для терминов *gene transfer* и *gene synthesis* для пятилетних временных интервалов на основе данных табл. 6 (табл. 7).

Таблица 7

Коэффициенты патентоёмкости для кластера первичных результатов исследований, в названиях которых имеются термины *gene transfer* и *gene synthesis*, 28.12.2011 г.

Интервал времени	Gene transfer	Gene synthesis
1981-1985	0,009	0,0
1986-1990	0,004	0,038
1991-1995	0,004	0,077
1996-2000	0,018	0,188
2001-2005	0,014	0,039
2006-2010	0,012	0,101
1981-2010	0,013	0,074

Как видим (см. табл. 7), значения коэффициента патентоёмкости для первичных результатов исследований по тематике генетического синтеза приблизительно на порядок превышают аналогичные значения коэффициентов для исследований по генетическому (генному) трансферу. Это связано с тем, что публикационная активность по тематике генетического трансфера на один-два порядка превышает такую активность для генетического (генного) синтеза.

В заключение отметим, что контент-анализ метаданных в откликах Google Scholar позволяет изучать частоты встречаемостей названий организаций, где выполнены исследования, и журналов, где они опубликованы, для произвольных промежутков времени.

Таблица 6

Динамика научных статей, книг и патентов, в заголовках которых имеются термины *gene transfer* и *gene synthesis*, 28.12.2011 г.

Интервал времени	Gene transfer				Gene synthesis			
	Научные статьи	Книги	Заявки на патенты	Выданные патенты	Научные статьи	Книги	Заявки на патенты	Выданные патенты
1956-1960	3	1	0	0	0	0	0	0
1961-1965	9	3	0	0	0	0	0	0
1966-1970	4	1	0	0	1	0	0	0
1971-1975	17	2	0	0	3	1	0	0
1976-1980	66	10	1	0	4	0	0	0
1981-1985	213	27	3	2	13	1	1	0
1986-1990	497	71	3	2	53	2	6	2
1991-1995	1160	117	35	5	39	3	1	3
1996-2000	3050	172	78	55	16	1	3	3
2001-2005	3760	148	114	53	26	3	7	1
2006-2010	2670	117	37	32	89	11	9	9
Всего	**11499**	**669**	**271**	**149**	**244**	**22**	**27**	**18**

СПИСОК ЛИТЕРАТУРЫ

1. Norouzi A. Google Scholar: The New Generation of Citation Indexes // Libri. - 2005. - Vol. 55. - P. 170-180.
2. Robinson M.L., Wusteman J. Putting Google Scholar to the test: a preliminary study // Program: electrónic library and information systems . - 2007. - Vol.41, № 1. - P. 71-80.
3. Aalst J. Using Google Scholar to estimate the impact of journal articles in education // Educational Researcher.- 2010 . - Vol. 39, № 5. - P. 387-400.
4. Mastrangelo G. et al. Literature search on risk factors for sarcoma: Pub. Med and Google Scholar may be complementary sources // BMC Research Notes. – 2010. - № 3. – P. 131. [Электрон. ресурс]. – URL: http://www.biomedcentral.com/1756-0500/3/131.
5. Walters W.H.. Comparative recall and precision of simple and expert searches in Google Scholar and eight other databases // portal: Libraries and the Academy.- 2011. -Vol. 11, № 4. - P. 971-1006.
6. Московкин В.М. Имитационная экспертная система выбора университетов для обучения // НТИ. Сер.2. – 2009. - № 10. – С. 19 – 21; Moskovkin V.M. Simulation expert system for making students' college decisions // Automatic Documentation and Mathematical Linguistics. – 2009. – Vol. 43, № 5. – P. 292-295. [Электрон. ресурс]. – URL: http://www.springerlink.com/content/14788m08u7q7745x/fulltext.pdf.
7. Michel J.-B.et al. Quantitative analysis of culture using millions of digitized books // Science.- 2011. - Vol. 331, № 1. - P. 176 – 182.
8. Hayes B. Bit Lit // American Scienfist.- 2011.- Vol. 99, N 3. - P.190-194.
9. Moskovkin V.M. Open access to scientific knowledge and feudalism knowledge: Is there a connection? // Webology. – 2011. – Vol. 8, № 1. [Электрон. ресурс]. – URL: http://www.webology org/2011/v8n1/a83.html.

Материал поступил в редакцию 29 03 12.

Сведения об авторе

МОСКОВКИН Владимир Михайлович – доктор географических наук, профессор кафедры мировой экономики Белгородского государственного университета, профессор кафедры экологии и неоэкологии Харьковского национального университета имени В.Н. Каразина

E-mail: Moskovkin@bsu.edu.ru