РОССИЯ ДОЛЖНА РАЗРАБОТАТЬ СОБСТВЕННЫЙ УНИВЕРСАЛЬНЫЙ ИНСТРУМЕНТ ПО ПОИСКУ НАУЧНОЙ ИНФОРМАЦИИ В ИНТЕРНЕТЕ

В.М. Московкин

БелГУ

В настоящее время в мире разработано достаточно много публичных и коммерческих поисковых инструментов, которые ведут поиск научной информации по конкретным базам данных. Среди таких публичных поисковых инструментов выделим Microsoft Academic Search (beta версия), CiteSeerX (разработан Пенсильванским университетом в 1997 г., спонсоры NSF, NASA, Microsoft), Bielefeld Academic Search Engine (BASE). Важно то, что CiteSeerX и BASE функционируют в рамках Протокола по сбору метаданных Инициативы “Открытые архивы” (OAI-PMH). Четвертый такой инструмент Scirus был недавно закрыт своим владельцем издательством Elsevier, по-видимому, из-за того, что не сумел выдержать конкуренции с остальными публичными поисковыми инструментами научной информации. Единственной компанией, которой удалось решить задачу поиска научной информации по всему интернету, является компания Google. Ее универсальный поисковый инструмент Google Scholar (запущен в ноябре 2004 г.) функционирует в рамках того же протокола OAI-PMH. Среди коммерческих поисковых инструментов научной информации пальма первенства принадлежит базам данных Web of Science и Scopus, их вместе с Google Scholar относят к международным наукометрическим базам данных. Имеется множество работ, которые, на примере различных предметных категорий и научных терминов, показывают, что Google Scholar имеет гораздо лучший их охват, по сравнению с Web of Science и Scopus. Возьмем для примера три актуальных сейчас кластера научных статей, порожденных терминами “cultural distance”, “institutional distance” и “psychic distance” и протестируем их вышеуказанными поисковыми инструментами на предмет встречаемости этих терминов в заголовках статей. Общее в этих терминах состоит в том, что расстояния между различными объектами рассчитываются с помощью евклидовой метрики. Отметим, что поисковый инструмент Microsoft Academic Search нами не использован, так как в нем отсутствует возможность вести поиск по заголовкам статей. В итоге получим следующую таблицу.

Таблица. Встречаемость избранных терминов в заголовках статей при тестировании с помощью различных поисковых инструментов. 17 июня 2015 г.

Поисковые инструменты научной информации	Cultural distance	Institutional distance	Psychic distance
Google Scholar	616	171	253
CiteSeerX	15	5	1
Bielefeld Academic Search Engine	202	50	91
Scopus	155	45	65
Web of Science	98	25	52

Как видим, для всех терминов Google Scholar дает лучшие результаты. Аналогичную ситуацию мы будем наблюдать при тестировании любых других терминов.

Помимо Google Scholar, компания Google в 2014-15 гг. разработала специализированные инструменты Google Books и Google Patents, а в 2010-2011 гг. совместно с учеными Гарвардского университета - уникальный аналитико-графический инструмент NgramViewer, о которых кратко скажем позднее.

GOOGLE SCHOLAR: ЕГО НЕДОСТАТКИ И ПЕРСПЕКТИВЫ РАЗВИТИЯ

Несмотря на уникальные его возможности, благодаря которым его относят к третьей по счету международной наукометрической базе данных после Web of Science и Scopus, он обладает рядом существенных недостатков. Например, до сих пор не разработан оператор для суммирования количества цитирований по найденному множеству публикаций. Такая задача очень часто возникает, когда мы хотим определить общее количество цитирований статей, относящихся к научной организации, автору или научному термину. Мы предлагали компании Google лет 6-7 назад решить эту задачу, наш запрос был передан в инжиниринговую группу Google Scholar, но дальнейших действий не последовало. В дальнейшем мы научились преодолевать защиту Google Scholar при массовых автоматизированных запросах и подсчетах, и для нас задача суммирования количества цитирований уже не стояла. Но если компания Google хочет позиционировать Google Scholar как наукометрический инструмент, то эта задача должна быть решена. В этой связи следует отметить, что с 2014 г. Минобрнауки РФ при мониторинге организаций, занимающихся научными исследованиями, требует определять количество публикаций, индексированных Google Scholar в текущем году, и общее количество цитирования статей, опубликованных за последние пять лет. Но, не имея

методики для подсчета публикаций и их общего цитирования с помощью Google Scholar, такую задачу решить корректно невозможно.

Далее, Google Scholar до сих пор не имеет хороших фильтров по отсеиванию публикаций с плохими метаданными и борьбе с дублированием документов. На наш взгляд, используя технологии Machine Learning и Data Mining, не предоставляет большего труда обучить Google Scholar вести поиск отдельно по журналам, входящим в базы данных Web of Science и Scopus, так как все издательства этих журналов имеют интернет-платформы или сайты. Кроме того, на «скопусовские» журналы можно заходить и через открытую платформу SCIMAGO. В идеале Google Scholar, помимо поиска академических документов по всему Интернету, должен иметь, по крайней мере, четыре опции:

1. WoS – поиск по журналам, входящих в базу данных Web of Science;
2. Scopus – поиск по журналам, входящих в базу данных Scopus;
3. DOAJ – поиск по журналам открытого доступа, зарегистрированных в Directory of Open Access Journals (DOAJ);
4. ROAR – поиск самоархивированных статей по репозитариям открытого доступа, которые размещены в Registry of Open Access Repositories (ROAR).

В последних двух случаях проблем в поиске академических документов не возникает, так как журналы и репозитарии открытого доступа вместе с Google Scholar функционируют в рамках одного Протокола по сбору метаданных инициативы «Открытые архивы» (OAI PMH).

Если все это будет реализовано, то ценность этого инструмента возрастет многократно.

ОТ GOOGLE SCHOLAR К RUSSIAN SCHOLAR

Предлагаем подумать над вышеуказанными вопросами поиска научной информации российским компьютерным компаниям, которые могли бы создать аналог Google Scholar с условным названием Russian Scholar, но с гораздо лучшими функциональными возможностями. Помимо вышесказанного, этот инструмент, естественно, может включить в свою поисковую базу информацию с платформы eLibrary с базой данных РИНЦ, а также использовать поисковые возможности Google Scholar с преодолением его системы защиты.

На основе Russian Scholar можно попытаться разработать аналог Google Ngram Viewer, который будет вести поиск по всем полнотекстовым научным документам, рассчитывать частоту встречаемости слов и словосочетаний и показывать в графическом виде их частотные тренды и осцилляции на больших временных интервалах. В отличие от Google Ngram Viewer, здесь поиск может вестись по всем форматам полнотекстовых документов, а не только по pdf-файлам книг, целых журналов и сборников, с конвертацией их в легко редактируемые форматы (Word, Excell, Text output) с помощью открытого программного обеспечения по оптическому распознаванию символов (Optical

Character Recognition, OCR).

Разрабатывать отечественный аналог Google Books смысла не имеет. Лучше будет, если крупнейшие российские библиотеки, имеющие ценные коллекции книг, особенно XV-XVIII вв., подпишут безфинансовые партнерские соглашения с компанией Google и включатся в глобальный процесс по оцифровке и индексированию всего значимого мирового книжного наследия, которое по оценкам специалистов составляет около 130 млн книг (на апрель 2013 г. количество отсканированных книг составило 30 млн). Это, конечно, не означает, что сами библиотеки не должны заниматься оцифровкой своих коллекций. Пожар в ИНИОН тому наглядный пример.

Что касается создания отечественного аналога Google Patents, то следует сказать следующее. Если когда-то удастся оцифровать весь советский и российский фонд по изобретениям, открытиям и другим объектам интеллектуальной собственности, то тогда такой аналог будет полезен.