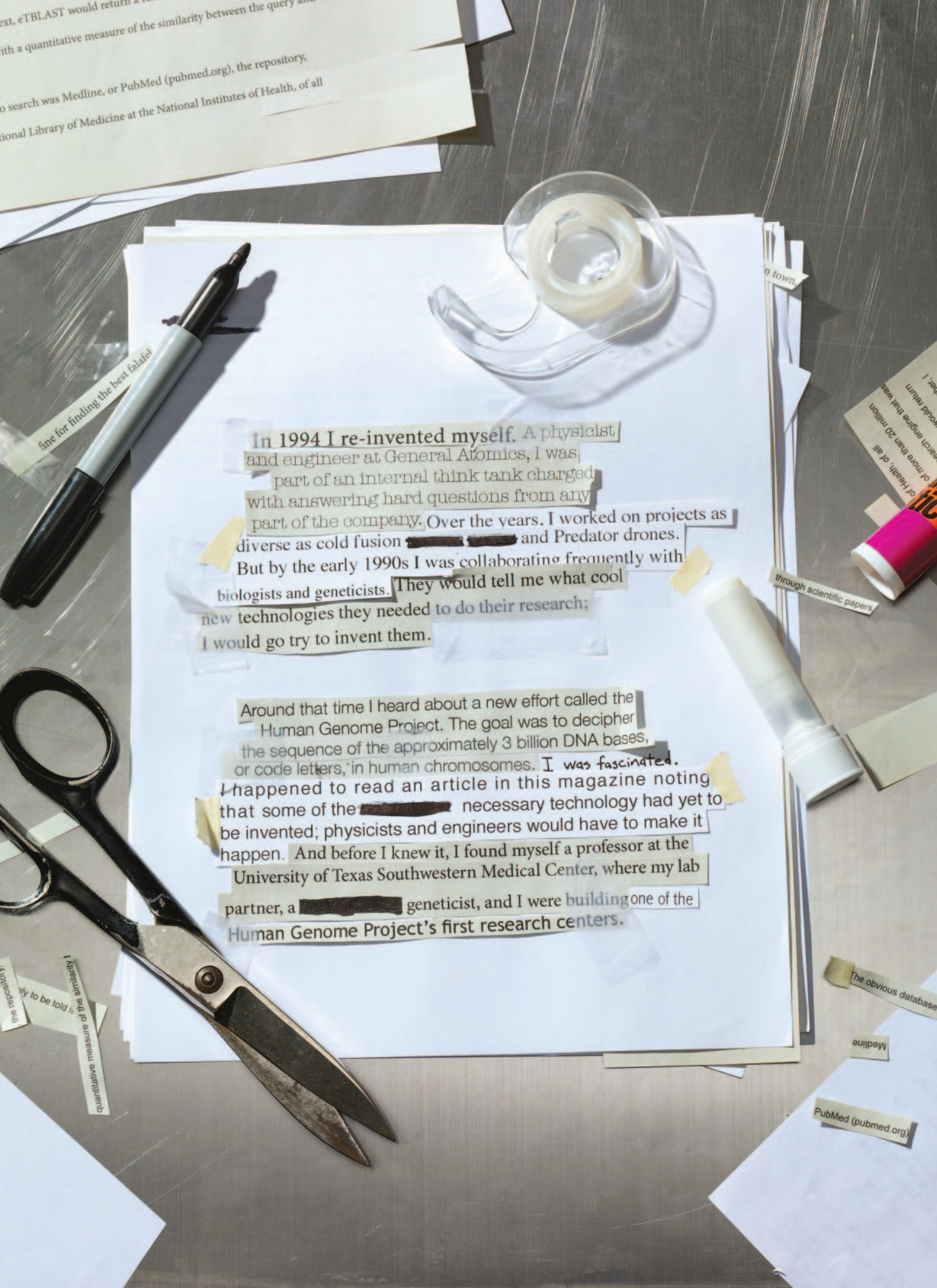
ext, eTBLAST would return a
ith a quantitative measure of the similarity between the query
o search was Medline, or PubMed (pubmed.org), the repository,
tional Library of Medicine at the National Institutes of Health, of all
In 1994 I re-invented myself. A physicist and engineer at General Atomics, I was part of an internal think tank charged with answering hard questions from any part of the company. Over the years, I worked on projects as diverse as cold fusion [redacted] and Predator drones. But by the early 1990s I was collaborating frequently with biologists and geneticists. They would tell me what cool new technologies they needed to do their research; I would go try to invent them.
Around that time I heard about a new effort called the Human Genome Project. The goal was to decipher the sequence of the approximately 3 billion DNA bases, or code letters, in human chromosomes. I was fascinated. I happened to read an article in this magazine noting that some of the [redacted] necessary technology had yet to be invented; physicists and engineers would have to make it happen. And before I knew it, I found myself a professor at the University of Texas Southwestern Medical Center, where my lab partner, a [redacted] geneticist, and I were building one of the Human Genome Project's first research centers.
fine for finding the best falafel
through scientific papers
The obvious database
Medline
PubMed (pubmed.org)
quantitative measure of the similarity

Гарольд («Скип») Гарнер

Автор этой статьи всего лишь хотел написать программу, которая помогла бы ему разобраться в медицинских терминах. Дело кончилось тем, что он обнаружил повсеместный плагиат в научных статьях и указал на потенциальные мошеннические схемы с грантами на сотни миллионов долларов

Дело об №94 ???украденных словах

С 1994 г. для меня началась новая жизнь. Работая инженером-физиком в компании *General Atomics*, я входил в группу аналитиков, которым было поручено отвечать на трудные вопросы, поступавшие изо всех подразделений компании. Годами я занимался самыми разнообразными проектами — от холодной плавки до беспилотных самолетов *Predator*, но к началу 90-х гг. стал часто сотрудничать с биологами и генетиками. Они все время говорили, что для исследований им не хватает новых передовых технологий, и я пытался эти технологии создать.

Примерно тогда же я услышал о новом проекте под названием «Геном человека». Он ставил своей целью расшифровать последовательность трех с лишним миллиардов базовых пар ДНК, или букв кода, в хромосомах человека. Я был воодушевлен. Мне попалась в журнале одна статья, где говорилось, что некоторые из необходимых технологий еще только предстоит изобрести и сделать это придется физикам и инженерам. Прежде чем я успел это осознать, я стал профессором Юго-Западного медицинского центра Техасского университета, где мы с коллегой, ученым-генетиком, начали развивать один из первых научно-исследовательских центров по расшифровке генома человека.

Там все было по-другому. Мои коллеги говорили на языке медицины, я же — на языке физики. В физике почти все описывается основными уравнениями. В медицине всеобщих уравнений не существует — там есть множество наблюдений, некое отрывочное понимание вещей и огромное количество специальных терминов. Я постоянно ходил на семинары и составлял длиннющие списки слов, которых никогда раньше не слышал, а потом часами разбирался, что они значат. Чтобы прочитать научную работу, мне приходилось держать под рукой медицинский словарь.

Удрученный своей неспособностью понять любой связный кусок текста, в подмогу самому себе я решил написать специальную компьютерную программу. Мне требовался поисковый механизм, который брал

ОБ АВТОРЕ

Гарольд («Скип») Гарнер (Harold "Skip" Garner) преподает биологию, вычислительную технику и медицину в Виргинском политехническом институте. Периодически занимается бизнесом. Соучредитель компании *HelioText*, специализирующейся на анализе текстов. Член консультативного совета *Scientific American*.

бы порцию текста и выдавал ссылки на литературу для дальнейшего чтения, рефераты и статьи, которые введут меня в курс дела по данной теме. Это оказалось нелегким делом. Поисковые системы для Всемирной паутины только начали появляться. Они успешно справлялись с поиском ресторана, где лучше всего в городе готовят фалафель, но не могли толком переварить абзац текста, содержащего несколько взаимосвязанных понятий, и выдать мне подходящие ссылки на литературу для чтения.

С помощью нескольких студентов и постдокторантов я занялся изучением анализа текстов, и вместе мы разработали программу под названием *eTBLAST* («электронный механизм поиска базового текстового локального соответствия» — *electronic Text Basic Local Alignment Search Tool*). Она была создана под влиянием программного инструмента *BLAST*, используемого для поиска в базах ДНК и первичных структур белковых молекул. Запросом для *BLAST* обычно была цепочка из 100–400 нуклеотидных оснований («букв») ДНК, в результате программа выдавала более длинные последовательности, которые включали запрашиваемый код. Запрос для *eTBLAST* — это абзац или страница текста, обычно длиной 100 или более слов. Разработать протокол поиска было труднее, чем написать программу для нахождения заданной цепочки символов, потому что поисковый механизм должен был не просто работать пословно, но еще и распознавать синонимы, сокращения и родственные понятия, выраженные другими словами, и учитывать порядок слов. В ответ на запрос в виде куска текста *eTBLAST* должна была выдавать упорядоченный список совпадений, обнаруженных в исследуемой базе данных, с указанием меры соответствия между запросом и каждым найденным фрагментом текста.

Очевидно, поиск следовало вести в базе данных *Medline* (доступной на интернет-портале *PubMed* по адресу *pubmed.org*). Это электронный архив публикаций, который поддерживает Национальная медицинская библиотека при Национальных институтах здравоохранения. Там есть информация обо всех научных исследованиях в биологии, касающихся медицины. Архив содержит заголовок и краткий обзор каждой из миллионов научных статей из тысяч журналов, публикации в которых рецензируются членами экспертного сообщества. У ресурса *Medline* был поисковый механизм по ключевым словам. Введя в строку поиска несколько слов, например «генетика рака молочной железы», получаем ворох совпадений, зачастую со ссылками на полный текст научных работ. Но, будучи новобранцем в области биомедицинских исследований, я часто даже не знал, как сформулировать поисковый запрос.

Первым версиям *eTBLAST* требовались долгие часы, чтобы прогнать по базе *Medline* абзац в несколько сотен слов для поиска совпадений. Однако программа работала. Пользуясь ею, я мог продираться сквозь заросли научных статей, абзац за абзацем разбираясь в их сути. Я мог забить в программу описание темы студенческой курсовой и быстро выйти на список литературы по данной теме. Мы с коллегами даже предлагали корпорации *Google* запустить нашу программу в коммерческий оборот, но нам сказали, что это не укладывается в бизнес-модель компании.

А дальше события приняли странный оборот. Несколько раз в заявках на тему студенческой курсовой я находил куски текста, совпадающие с текстом из других опубликованных исследований, но ссылка на них отсутствовала. Такие студенты приговаривались к нравоучительным лекциям по этике научного сообщества. Передо мной как исследователем возник вопрос, который впоследствии изменил всю мою карьеру: как много плагиата содержит специальная литература по биомедицине?

Дежавю

Когда я взялся за эту новую проблему, исследования на тему плагиата в биомедицине состояли из анонимных опросов. Согласно последнему на тот момент опросу, который мне удалось найти, 1,4% ученых сознались в незаконных заимствованиях. Но точность этой цифры зависела от честности респондентов. Программа *eTBLAST* помогла бы выяснить, не врут ли ученые.

ОСНОВНЫЕ ПОЛОЖЕНИЯ

- Углубившись в медицинскую литературу с помощью программы текстового анализа, автор обнаружил доказательства повсеместного плагиата и потенциальных финансовых лазеек. Сегодня, считает он, резкий рост количества сомнительных научных сборников облегчает публикацию плагиата.
- Анализ текста — полезный инструмент для обнаружения плагиата. Но, может быть, пришло время обсудить новую модель научно-издательской деятельности — возможно, такую, при которой ученые непрерывно редактируют общий электронный массив знаний типа «Википедии».

Поскольку у нас не было недостатка в помощниках среди студентов и имелся довольно мощный компьютер, мы случайным образом брали аннотации статей из базы *Medline* и использовали их в качестве запроса для *eTBLAST*. Компьютер сравнивал запрос с содержимым всей базы *Medline* в поисках текстовых соответствий и выдавал список совпадений. Каждое совпадение получало оценку соответствия. Сам запрос всегда находился в вершине списка — стопроцентное совпадение. Следующее успешное совпадение обычно имело ранг сходства от единиц до 30%. Однако время от времени мы обнаруживали, что второе, а иногда и третье совпадение имеет ранг, близкий к 100%. Обработав несколько тысяч запросов, мы увидели, что около 5% запросов имеют подозрительно высокую оценку совпадения. Мы перечитали эти аннотации, чтобы убедиться, что человек увидит совпадения там же, где их обнаружила программа. После этого мы начали сравнивать полные тексты статей, чьи аннотации подозрительно совпадали.

Вскоре мы стали находить примеры явного плагиата. К нашему разочарованию и даже изумлению, беспардонно заимствовались не просто отдельные фразы — целые статьи и диссертации. Разумеется, мы знали, что 1,4% ученых сознавались в плагиате. Но увидеть своими глазами, как воруются научные труды, — совсем другое дело. Работа была увлекательной, особенно для студентов. Они чувствовали себя борцами с преступностью, и в каком-то смысле так оно и было.

На следующем этапе мы расширили масштаб вычислений и анализа. Для досконального исследования нам было необходимо выполнить поиск на соответствие по каждому элементу текста разумной длины, хранящемуся в базе *Medline*, — на тот момент это составляло почти 9 млн записей, в среднем по 300 слов каждая, помноженные на те же девять миллионов операций сравнения. Работа заняла несколько месяцев и потребовала приличного объема вычислительных мощностей нашей лаборатории. Получив результаты, мы их проанализировали и разместили сведения о максимальных совпадениях в базе данных, которую мы назвали *Déjà Vu*.

Déjà Vu стала заполняться парами крайне схожих аннотаций из базы *Medline* — нашлось 80 тыс. пар, рейтинг подобия которых превышал 56%. Большинство этих пар имели высокую степень сходства по совершенно понятным причинам — например, это были публикации, дополняющие более ранние исследования, или обзоры научных конференций. Но остальные записи вызывали подозрение.

Мы отправили в журнал *Nature* статью с изложением информации о частоте незаконных заимствований и двойных публикаций (которые иногда называют автоплагиатом), с описанием содержимого базы данных *Déjà Vu* и некоторых особо ярких примеров плагиата. Редакция статью приняла, но, поскольку мы приводили некоторые аннотации в качестве примеров плагиата, редакционные юристы нашу статью зарубили. Они дали этому блестящее объяснение: определять, есть плагиат или нет, якобы уполномочены только редакторы и комитеты по соблюдению научной этики. Мы же могли только предоставить факты об объеме текстовых совпадений или сходстве между любыми двумя образцами научной литературы. В конце концов — с одобрения юристов — так мы и поступили.

Когда наш отчет был опубликован в *Nature*, случилось настоящее светопреставление. Редакторы научных сборников были удручены, ведь на них свалилась дополнительная работа. Чтобы защитить авторские права, редакторам, публиковавшим оригинальные труды, пришлось настаивать на отзыве статей, содержащих плагиат. Издатели вторичных работ, разумеется, были обескуражены. Ученые злились, потому что наши результаты с очевидностью обнажили порок в системе внутреннего рецензирования научных работ. Однако скрепя сердце все признали, что тема поднята важная и налицо серьезная проблема. Ведь ученые и практикующие врачи принимают ключевые решения на основании информации, которую они берут из литературы. А к чему могут привести решения, опирающиеся на сомнительные исследования?

В конечном итоге мы определили, что 0,1% публикаций по специальности содержат откровенный плагиат. (Мы искали только статьи, которые практически совпадают друг с другом; случаев воровства небольших фрагментов чужих статей наверняка было гораздо больше, но, поскольку наша программа вела поиск лишь по аннотациям статей, такие случаи она бы не распознала.) Около 1% работ представляли собой автоплагиат; статья одного автора могла почти дословно повторяться чуть ли не в пяти научных сборниках. Если эти цифры кажутся вам несущественными, подумайте о том, что ежегодно по биомедицинской тематике публикуется около 600 тыс. новых статей.

Вскоре мы заметили, что издательский процесс меняется. Редакторы журналов стали использовать программу *eTBLAST* для проверки предлагаемых к публикации статей на наличие плагиата. Я тоже изменился. В поисках новых путей я добавил в описание своей квалификации запись «исследователь этики науки» *(ethics researcher)*.

Моя жизнь в должности полицейского научных нравов

Это первое большое исследование плагиата было всего лишь началом. Чтобы разобраться в причинах плагиата и понять, как это влияет на науку, предстояло проделать гораздо бóльшую работу. В каких случаях воспроизведение текста представляется допустимым? Когда и почему ученые занимаются плагиатом? Может ли текстовый анализ обнаружить иные виды неэтичного поведения? Какие? Чтобы ответить на эти вопросы, мы усовершенствовали нашу программу, дополнили базы данных и взялись за новые исследования.

Дальнейшая работа привнесла неожиданные оттенки в дискуссию о плагиате. Мы обнаружили, что в некоторых случаях текстовое сходство не только приемлемо, но даже предпочтительно. Например, в методической

части научной работы, где важнейшим фактором выступает воспроизводимость результатов, типовые формулировки служат важной цели — убедить, что в исследованиях был использован стандартный протокол.

Мы обнаружили и подлинно вопиющие этические нарушения. В одном из наших исследований, опубликованном в журнале *Science*, мы взяли худшие примеры плагиата, которые удалось обнаружить, — пары статей, в которых статья *B* совпадала со статьей *A* в среднем на 86%, — и детально их проанализировали. Мы отправили по электронной почте копии этих статей с нашими примечаниями авторам и редакторам, работавшим над этими текстами, и попросили их поучаствовать в закрытом анкетировании. Мы спрашивали: знали ли они о сходстве текстов? Могут ли они это объяснить? Откликнулись 90% людей, к которым мы обратились.

Некоторые авторы рассказали о поразительных нарушениях этики. Одни признались, что копировали для себя статьи в процессе рецензирования — и написали на эти статьи плохие отзывы, чтобы они не прошли в печать. Другие сваливали вину на каких-то вымышленных студентов-медиков. Один человек сказал, что опубликовал плагиат ради шутки. У себя в стране он оказался вице-президентом национального комитета по этике. Неудивительно, что большинство ущербных статей из нашего списка впоследствии были отозваны.

И это еще не все виды нарушений этики ученого, которые мы обнаружили. В начале 2012 г. мы начали искать примеры двойных грантов, т.е. денег, полученных из разных правительственных ведомств для выполнения одной и той же работы. Мы скачали почти 860 тыс. заявок на гранты от правительства и частных организаций, включая Национальные институты здравоохранения, Национальный фонд науки, Министерство обороны, Министерство энергетики и фонд *Susan G. Komen for the Cure* (крупнейшая организация в США, занимающаяся предупреждением рака молочной железы) и ввели их в программу *eTBLAST*. Задача требовала 800 тыс. раз выполнить по 800 тыс. сравнений (это приблизительно 10^{12}), для чего была нужна производительность суперкомпьютера.

Просмотрев 1,6 тыс. наиболее похожих грантовых заявок, мы обнаружили, что 170 пар заявок содержали практически идентичные цели, задачи или гипотезы. Мы пришли к нескольким выводам: что двойные заявки на гранты подавались регулярно в течение долгого времени; что этим занимались самые престижные университеты США; что в результате сфера биомедицинских исследований в стране теряла до $200 млн в год.

Будущее научных изданий

Небольшой процент людей всегда нарушают принятые в обществе нормы, и ученые здесь не исключение. В трудные времена, когда финансирование науки сокращается, а конкуренция за научные должности обостряется, некоторые ученые ведут себя неподобающе. Бурный рост числа сомнительных, не заслуживающих доверия журналов превратил научно-издательскую деятельность в подобие ковбойской авантюры по завоеванию Дикого Запада. Найти площадку для публикации своих материалов сегодня легче, чем когда-либо, даже если эти материалы позорно скопированы.

Метод анализа текстов дает нам хороший инструмент для наведения порядка в издательской деятельности. Однако он мог бы не только вывести на чистую воду плагиаторов, но в конечном итоге принести намного больше пользы. Он мог бы способствовать появлению совершенно новых способов обмена результатами научных исследований.

Чем не соблазнительная идея — использовать модель «Википедии»: создать динамично развивающийся электронный банк информации по некоторой теме, который ученые постоянно модифицируют и улучшают. Каждая новая «публикация» представляла бы собой индивидуальный вклад в единый растущий массив знания; избыточные разделы с описаниями методик отпали бы за ненадобностью. Модель «Википедии» стала бы шагом в сторону создания централизованной базы данных научных публикаций по всем дисциплинам. Авторы и редакторы могли бы использовать метод углубленного анализа текстов для проверки оригинальности новой работы, а также для разработки надежных показателей, которые помогут оценить значение некой концепции или открытия. В идеале вместо того, чтобы измерять авторитет научной работы с помощью индекса цитирования, мы бы измеряли ее влияние на состояние научного знания в целом и даже на состояние общества.

В Технологическом университете Виргинии, куда я перебрался четыре года назад, мы изо всех сил поддерживаем *eTBLAST* в рабочем состоянии, и у этой программы все еще тысячи пользователей. Между тем мы с Ким Менье (Kim Menier), моей женой и деловым партнером, возлагаем большие надежды на компьютерный анализ текста. Мы думаем о том, как использовать описанный вид поиска текстовых совпадений (в пределах абзаца) для других целей, включая администрирование грантов, маркетинговые исследования и оценку патентной чистоты. Станет ли наш инструмент столь же успешным, как *Google*? Как знать. Но мой опыт мне подсказывает, что анализ текстов может стать подлинным откровением. Однажды с его помощью удалось доказать, что ученые так же несовершенны, как и все прочие смертные. ■

Перевод: С.В. Гогин

ДОПОЛНИТЕЛЬНЫЕ ИСТОЧНИКИ

- A Tale of Two Citations. Mounir Errami and Harold Garner in Nature, Vol. 451, pages 397–399; January 24, 2008.
- Responding to Possible Plagiarism. Tara C. Long et al. in Science, Vol. 323, pages 1293–1294; March 6, 2009.
- Systematic Characterizations of Text Similarity in Full Text Biomedical Publications. Zhaohui Sun et al. in PLOS ONE, Vol. 5, No. 9, Article No. e12704; September 15, 2010.
- Research Funding: Same Work, Twice the Money? Harold R. Garner et al. in Nature, Vol. 493, pages 599–601; January 31, 2012.