КОНТУРЫ

В слове - сила

Формализованный анализ книжных текстов приводит к неожиданным открытиям

В декабре 2004 года компания Google подписала соглашение с пятью крупнейшими библиотеками мира (Нью-Йоркской публичной библиотекой, библиотеками университетов Гарварда, Стэнфорда, Мичигана и Оксфорда) по оцифровыванию их книжных коллекций (Google Books Library Project). В настоящее время количество таких библиотек превысило 40. За семилетний период было оцифровано более 5 млн книг (15 млн томов). Основная масса оцифрованных книг издана с 1520 по 2008 год. Отсюда, основываясь на законе об авторском праве США, можно получить интервал, в котором находятся книги в "статусе общественного достояния": 1520-1923 годы. До 1923 года в мире было опубликовано около 6 млн книг (18% книг из World Cat). Эти данные приведены в статье Эдгара Джонса (Edgar Jones) "Google Books as a General Research Collection", опубликованной в журнале "Library Resources & Technical Services" за 2010 год. Компания Google, в первую очередь, нацелена на оцифровывание этих старых и редких книг, на которые истекли авторские права.

Известный научный обозреватель Brian Hayes из "American Scientist" в своей статье "Bit Lit" (май-июнь 2011 года) отмечает, что книги могут служить не только для чтения. Так, слова можно считать, сортировать, сравнивать, классифицировать, искать закономерности в их распределении и т.д. Все эти способы формализованного анализа позволяют извлекать новое знание из текста. К этой идее пришли ученые Гарвардского университета совместно с аналитиками компании Google. Чтобы распознать слова, они использовали OCR-процесс (Optical Character Recognition - оптическое узнавание букв), а учитывая ограничения на авторские права, из всего текста "нарезали" одно-, двух-... пятисловные словосочетания. В итоге из всего массива книг было получено около 500 млрд словосочетаний, которые названы "n-grams". Для каждого года и каждого "n-gram" определены перечень и страницы книг, в которых это словосочетание обнаружено, а также частота его встречаемости. Редкие словосочетания, которые встречаются менее 40 раз в обнаруженных книгах, отсеиваются.

Данная методология с ее приложениями в культурологии и лингвистике была опубликована большим коллективом авторов во главе с Yean-Baptist Michel и Erez Lieberman Aiden из Гарварда в январе 2011 года в журнале "Science". Развитый в этой статье, на базе Google Books, аналитико-поисковый инструмент "Ngram Viewer" был запущен в августе 2010 года. Он может использоваться в изучении текстуальных изменений в структуре языка и оценке культурологических трендов.

Авторы этой работы позиционировали ее как новую область знаний, назвав "культуромикой" по аналогии с "геномикой". Точно так же, как крупномасштабные коллекции ДНК-последовательностей приводят к новым биологическим структурам, так и масштабные лингвистические данные могут помочь в анализе человеческой культуры и ее трендов.

С целью показать возможности этого инструмента для отечественных исследователей мы провели серию экспериментов. Посмотрим вначале, как ведут себя тренды и колебания встречаемости в английском языке географо-планетарных терминов "cosmography", "cosmology", "cosmogony", astrology", "geography", "astronomy", "topography", "geology", "cartography". Записывая через запятую эти термины в поисковой строке Ngram Viewer, мы сразу же получаем временные графики нормализованных частот (f). Такая частота для произвольного словосочетания, состоящего не более чем из пяти слов, в заданный момент времени определяется процентным отношением его встречаемости к встречаемости всех словосочетаний данной размерности. Этот показатель вводится с целью нейтрализации "лингвистической" инфляции, под которой понимается постоянный рост печатной продукции.

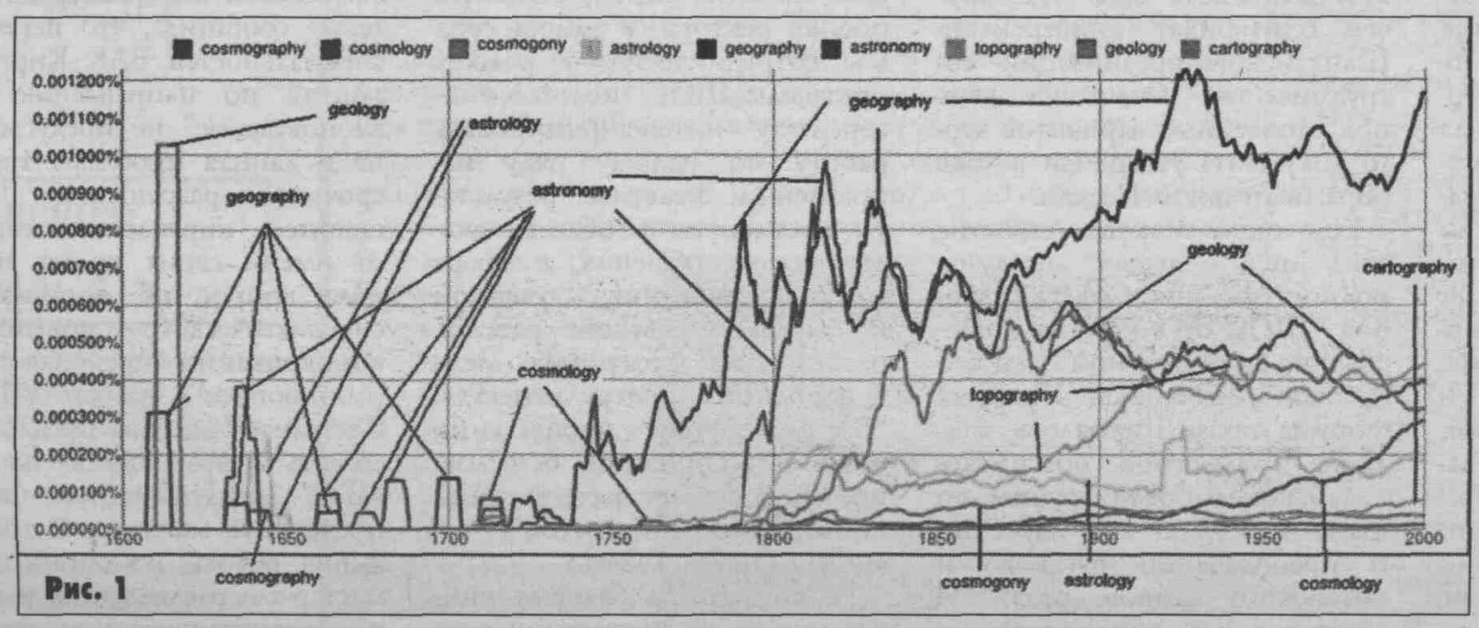

Рис. 1

В нашем случае в интервале 1600-2008 годов мы получим 10 графиков для вышеуказанных терминов (рис. 1). С XVII до середины XVIII века наблюдаются случайные всплески нормализованной частоты их встречаемости, что обусловлено нерепрезентативностью выборки книг в этом промежутке времени.

Появление непрерывных кривых нормализованных частот встречаемости со второй половины XVIII века говорит о начале формирования репрезентативной выборки книг. Разработчики инструмента Ngram Viewer считают, что такая выборка охватывает полностью XIX-XX века.

Для понимания вышеуказанных тенденций необходимо знать, на каком фоне происходило их формирование. До выхода обобщающего труда Бернхарда Варения "География генеральная" (1650 год), открывшего новую эпоху в развитии географии, фундаментальное знание в вышеуказанных областях знания опиралось на древнегреческие и древнеримские труды Птолемея, Страбона, Эратосфена и Аристотеля, дошедшие до европейцев благодаря византийским и арабским переводам. Византийские, арабские, а позднее и западноевропейские ученые значительно продвинули это знание в прикладном плане для целей картографии и навигации. Стимулом для развития космографии и географии в Средние века служило развитие мореплавания, вылившееся позднее в эпоху Великих географических открытий.

Труд Б.Варения - первая попытка определить предмет и содержание географии, основываясь на новых данных о Земле в эпоху Великих географических открытий. Он был основным учебным пособием по географии и смежным дисциплинам более 100 лет. После Варения огромный вклад в методологию и теорию географии внес Иммануил Кант.

В конце XVIII века география стала отделяться от космографии, астрономии и физики и в свете эволюционных воззрений в естествознании стала утрачивать свою значимость. Она вступила в полосу кризисов, которая ознаменовалась распадом географии на множество отраслевых наук. В этот период во всем естествознании главенствовали небесная механика Исаака Ньютона и детерминизм Пьера-Симона Лапласа. В это же время Михаил Ломоносов, намного опередив развитие западноевропейской географической мысли, предложил исторический принцип единства и связи различных компонентов природы.

Следующая за Варением и Кантом эпоха связана с немецкой географической школой, а именно, с Александром Гумбольдтом и его пятитомным сочинением "Космос" (1845-1862) и Карлом Риттером и его 19-томным трудом "Землеведение в отношении к природе и к истории человека, или Всеобщая сравнительная география" (1817-1859).

В конце XIX века большого расцвета достигла французская географическая школа благодаря трудам Элизе Реклю. Отметим, что, если в XV-XVII веках нужды мореплавания и открытия новых земель были стимулом для развития географии и других смежных наук, то в XVIII-XIX веках их место занимают потребности капиталистического способа производства.

Не останавливаясь на истории географической мысли XX века, из всего вышеизложенного можно заключить, что расцвет современной географической науки продолжается более 3,5 столетий, начиная от Б.Варения до настоящего времени. Несмотря на кризис в теоретической географии в период дифференциации географического знания, который наблюдался с конца XVIII до середины XIX века, в целом этот период находился на восходящей волне развития географии в связи с мощным развитием капитализма.

На фоне развития западноевропейской географической мысли, отраженного в книгопечатной продукции, и были сформированы временные ряды нормализованных частот встречаемости названий геолого-географических и планетарных наук (рис. 1), среди которых доминирует география, потом следуют астрономия и геология.

Интересны синхронные колебания их частот на протяжении большей части XIX века. Возможно, это связано с вышеуказанной полосой кризисов в период дифференциации знаний. Поставив задачу определить, какие литературные источники дают максимальный вклад в частотные пики этих колебаний, мы приблизимся к пониманию их происхождения. Но сразу отмечу, что эти источники не прямо связаны с трудами всех вышеперечисленных ученых, в основном это различные компиляции, справочники, энциклопедии, каталоги, библиографические списки и др.

Формальный анализ построенных графиков показывает, что термин "geography" имеет непрерывную кривую с возрастающим трендом начиная с 1700 года с пиком в 1795 году (f=0,0008%), с 1802 по 1850 год наблюдаются четыре пика с периодом в 16 лет, с 1870 года - возрастающий тренд до 1930 года (за этот 60-летний период нормализованная частота возросла в 2 раза - с 0,0006 до 0,0012). Далее видим еще две волны с максимумами в 1970 и 2000 годах.

Нормализованная частота для термина "astronomy" с 1810 по 1870 год близко повторяла колебания такой частоты для термина "geography", а после 1870 года она пошла с ниспадающим трендом, что наблюдалось до 1920 года.

Для термина "cosmology" мы видим медленный рост нормализованной частоты начиная с начала XIX века. В конце XVIII - начале XIX века наблюдался всплеск интереса к астрологии, аналогичное происходило и в 1920-е годы. Возрастающий тренд нормализованной частоты для термина "topography" наблюдался с 1880 по 1913 год, после чего она колебалась в окрестности f=0,0004%.

В качестве второго эксперимента мы выбрали имена наиболее значимых первооткрывателей и протестировали их с помощью инструмента Ngram Viewer на русском языке (рис. 2). Анализ показал, что полученная временная частотная динамика публикаций в XIX веке сформирована изданиями географического, исторического, археологического и других обществ, Академии наук, энциклопедиями и периодическими изданиями. Например, 13 мая 1896 года Географическое общество широко отмечало открытие морского пути в Индию, и это обусловило первый всплеск интереса к персоне Васко да Гама. Второй максимальный всплеск интереса к нему, а также ко всем остальным первооткрывателям пришелся на послевоенные годы. Последняя волна повышенного интереса связана с 500-летием открытия Индии. До революции в царской России пропагандировались исключительно иностранные первооткрыватели, что видно на примере Афанасия Никитина, популярность которого становится доминирующей лишь с конца 1930-х годов.

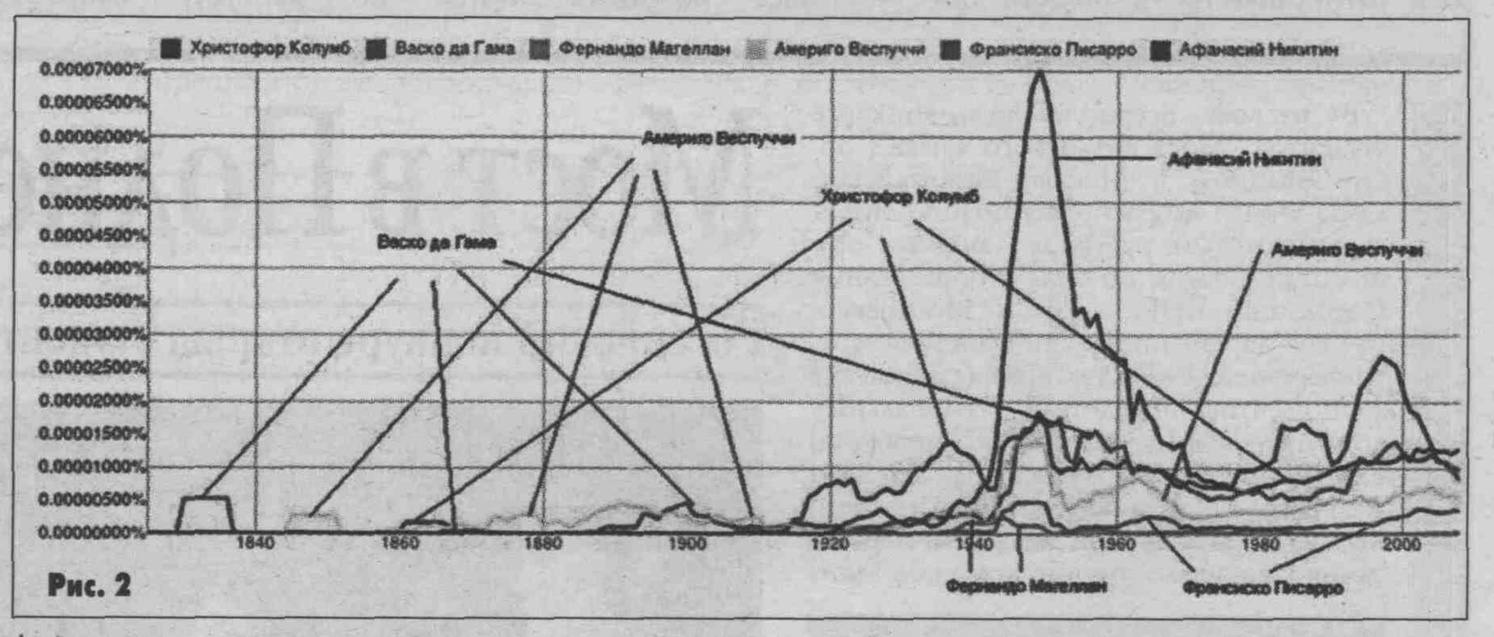

Рис. 2

Таким образом, мы видим, какие уникальные возможности открываются перед учеными в исследовании различных научных трендов на достаточно длинных промежутках времени. Эти возможности будут только улучшаться в связи с амбициозной целью компании Google по оцифровыванию всех значимых книг.

Владимир МОСКОВКИН,
доктор географических наук
Харьков - Белгород