Великие раскопки и великие вызовы

Леонид Левкович–Маслюк

КОМПЬЮТЕРНЫЙ ПОИСК ЗНАНИЙ СТАНОВИТСЯ ВСЕ БОЛЕЕ ЦЕННЫМ

Наука об извлечении содержания из гигантских массивов данных становится все более изощренной, а задачи, за которые берутся мастера такого поиска, — все более человечными.

Прочесыванием гор информации в поисках скрытых в них закономерностей люди занимаются уже многие века. Но только с появлением компьютеров, баз данных, локальных и глобальных сетей понятие «больших массивов» обрело нынешний смысл, а их вдумчивое сканирование, когда–то занимавшее лишь шпионов и каббалистов–мистиков, позже — социологов культуры и теоретиков медиа с их страстью к контент–анализу, превратилось в индустрию. Причем индустрию высокотехнологичную даже на фоне инфотеха. Ведь найти смысловые связи в новостной заметке, правильно ответить на элементарный вопрос — о чем она, к какому тематическому классу ее причислить, — сложнейшая, как оказалось, задача для машины. С другой стороны, даже простая для машины, но неподъемная и невыносимо тоскливая для человека задача механического сканирования текста с одновременной сортировкой имен, названий, ключевых слов часто оказывается очень и очень востребованной. А если еще и выйти за пределы мира текстов, попытаться научить компьютер понимать, о чем люди говорят (хотя бы в телефонных переговорах с туристическим бюро), что они показывают друг другу на фотографиях и видеолентах, — станет ясно, что колоссальный спрос на результаты таких исследований сталкивается с колоссальными трудностями в их реализации.

Вот где–то между этими молотом и наковальней и зародился современный датамайнинг (data mining, буквально — раскопки данных, или добыча чего–то из данных), в котором научные и индустриальные компоненты трудно разделить. В 1998 году научную зрелость этой отрасли подтвердило создание Special Interest Group (SIG), Группы особых интересов, в рамках авторитетной международной организации по компьютерным исследованиям ACM (Association for

Computing Machinery, Ассоциация по вычислительным машинам).

Что такое SIG? Вспомним о самой популярной из подобных групп — SIGGRAPH. Ежегодные мегаконференции, на которых делаются доклады, читаются лекции и демонстрируются высшие достижения компьютерной графики, анимации и сопутствующей всему этому математики, других наук и технологий, известны далеко за пределами сообщества специалистов. Другие SIG'и (сейчас их тридцать четыре, в том числе SIGART [искусственный интеллект], SIGMOD [базы данных], SIGPLAN [языки программирования], SIGSOFT [разработка ПО] и др.) не так знамениты среди широкой публики, но заслужили уважение специалистов, а проводимые ими конференции, издаваемые журналы являются индикаторами качества в своих областях.

На наши вопросы о теории и практике датамайнинга ответил **Григорий Пятецкий-Шапиро** (Gregory Piatetsky-Shapiro), основатель и председатель SIGKDD — Группы особых интересов, посвященной «открытию знаний в данных» (Knowledge Discovery in Data).

Какие новые разделы датамайнинга (ДМ) появились в последние годы? Какие из них самые перспективные для бизнеса, для исследовательской работы?

— Одно из замечательных новых полей исследований — анализ связей (link analysis). Приложения весьма обширны, от биоинформатики до выявления преступлений, от маркетинга до исследования социальных сетей. Вокруг Web 2.0 сейчас столько шума именно потому, что он очень эффективно использует веб как инструмент социальных связей, — а это придает все большую значимость анализу этих связей.

Огромный прогресс виден и в майнинге текста (большинство программных комплексов [suites] для датамайнинга теперь включают компоненты для текстмайнинга), а также в майнинге мультимедиа. И то и другое — прекрасные области для исследований.

Датамайнинг широко применяется в больших компаниях, особенно работающих в электронной коммерции. Amazon, Yahoo — примеры таких компаний (мой коллега Усама Файяд занимает должность руководителя по обработке данных [Chief Data Officer] в Yahoo, он первым в индустрии e-коммерции получил такой титул). Вот неполный список областей применения датамайнинга:

- реклама;
- биоинформатика;
- связь с клиентами (CRM);
- маркетинг;
- выявление мошенничества (fraud detection);
- e-коммерция;
- здравоохранение;
- инвестиции/ценные бумаги;
- управление производством;
- развлечения и спорт;
- телекоммуникации;
- изучение веба.

Если говорить об успехах индустрии датамайнинга, то самый яркий пример здесь — Google. Оба его сооснователя в Стэнфорде занимались исследованиями в этой области, и ранняя история самого Google связана с датамайнингом.

Рекомендации на сайте Amazon.com («покупатели, купившие/искавшие/посмотревшие X, купили также Z») привели к огромному росту продаж. Высококачественные рекомендации такого типа обеспечили успех компании Netflix, занимающейся прокатом видео.

Например, если вам понравилась знаменитая абсурдистская комедия «Монти Пайтон и священный Грааль» («Monty Python and the Holy Grail»), то вы получите от Netflix рекомендацию посмотреть «This is Spinal Tap»[1], известную пародию на документальный фильм о гастролях экстравагантной рок-группы. Netflix придает такое значение датамайнингу, что в прошлом году учредила приз в миллион долларов за улучшение алгоритма выработки рекомендаций (см. врезку).

Кто заказывает вашей фирме KDnuggets (Kdnuggets.com) датамайнинговые проекты? Насколько они масштабны (по количеству участников, ресурсам, времени выполнения)? Требуют ли разработки нового ПО специально для каждого проекта?

— Многие думают, что Kdnuggets — большая компания с веб-программистами, редакторами, менеджерами по развитию бизнеса, отделом кадров и т. п. На самом деле она состоит из одного человека — меня самого, а все ее дела я веду при помощи множества скриптов, автоматически выполняющих большинство необходимых действий.

Время от времени я получаю интересные заказы на консалтинговые проекты, которые тоже обычно выполняю самостоятельно. Главное, что требуется от кон-

1 «Пункция спинномозговой жидкости».

ЦИФРА

Агентство IDC прогнозирует, что объем цифровой информации в мире достигнет тысячи экзабайт к 2010 году, тоесь по сравнению с 2006 годом увеличится в 6 раз (1 экзабайт = 2^{60} байт, примерно миллиард гигабайт).

РАСПРЕДЕЛЕНИЕ ОБЛАСТЕЙ ПРИМЕНЕНИЯ ДАТАМАЙНИНГА СОГЛАСНО ОПРОСУ ПОСЕТИТЕЛЕЙ KDNUGGETS.COM

ИСТОКИ KDD

Как развивалась ваша карьера? Как вы заинтересовались датамайнингом?

— С детства у меня была склонность к математике, очевидно унаследованная от папы, крупного математика Ильи Пятецкого-Шапиро. Живя в Москве, я учился в известной Второй математической школе, принимал участие в математических олимпиадах — но поскольку перенял от папы лишь малую часть математического таланта, то уже в школе понял, что чистая математика не для меня. Я открыл для себя компьютеры в 1974 году, на первом курсе в Технионе, когда эмигрировал в Израиль, и сразу заинтересовался ими. Меня особенно увлекали вопросы искусственного интеллекта. Первую интересную программу я написал в 1974 году на языке АПЛ — она была предназначена для игры в «морской бой». Сыграв с ней одну партию, я безоговорочно уступил своей же программе. Желание продолжать игру исчезло — зато усилилось желание писать программы. Потом была учеба в аспирантуре в США, тоже с концентрацией на задачах искусственного интеллекта. Темой диссертации стало приложение искусственного интеллекта к работе с базами данных.

Датамайнингом я заинтересовался, работая в Лабораториях GTE (организация, подобная знаменитой Bell Labs, только поменьше) над крупными коммерческими базами данных. Оказалось, что если найти определенные правила, некоторые запросы к этим базам можно ускорить на несколько порядков. Я заинтересовался вопросом — можно ли находить такие правила автоматически, и занялся применением идей искусственного интеллекта к большим базам данных. Побывав в 1988 году на встрече (workshop) по этой теме (в рамках конференции AAAI '88), я понял, что этому мероприятию нужна более четкая фокусировка. По молодости лет я не представлял себе, каких усилий стоит организовать такую встречу, и поэтому в 1989 взялся за организацию воркшопа сам. Термин «датамайнинг» я считал недостаточно завлекательным (sexy) и вместо него предложил назвать тему «открытие знаний в базах данных» (Knowledge Discovery in Databases, KDD). Это название подчеркивало, что конечная цель работы — знания, и намекало на дух первооткрывательства, сопутствующий поиску знаний. Тогда же я начал новый проект в GTE Labs, и это был первый в мире проект по KDD.

Воркшоп прошел в 1989 году с большим успехом, и с тех пор я продолжаю работать в этой области. В 1993 году начал рассылку «Knowledge Discovery Nuggets», чтобы помочь в установлении связей между исследователями, занятыми этой проблематикой (первыми ее получили пятьдесят участников KDD-93). В 1994 году, с началом массового распространения веба, я создал один из первых сайтов по датамайнингу, из которого вырос мой нынешний сайт KDnuggets.com. Я очень рад, что вовремя сообразил, что в одиночку не потяну организацию воркшопов, и подключил к этому делу Усаму Файяда (Usama Fayyad), ставшего председателем оргкомитета KDD-94. С ним и еще несколькими коллегами мы превратили KDD в полномасштабную конференцию, а при поддержке Вон Кима (Won Kim) создали в 1998 году SIGKDD (www.kdd.org) — исследовательское общество по открытию знаний и датамайнингу. В 2007 году в Сан-Хосе (Калифорния) пройдет уже 13-я конференция KDD (www.kdd2007.com). Воркшоп KDD в 1989 году был единственным в мире, а сейчас каждый год собирается дюжина конференций и встреч по этой теме. ■

сультанта по датамайнингу, — интуиция, которая подсказывает, как найти интересные объекты в массиве данных и как при помощи существующих методов и технологий обнаруживать именно то, что принесет пользу заказчику.

К сожалению, многие успешные датамайнинговые проекты, в том числе и часть моих, связаны с деликатными вопросами бизнеса — такими, как выявление мошенничества и обмана, — и поэтому о них нельзя подробно рассказать в прессе. Однако недавно состоялся воркшоп, специально посвященный «историям успеха» технологий датамайнинга (www.dataminingcases-

ОЦЕНКА

Удачные статистические модели позволили выявить потенциальные «налоговые убежища» обеспеченных американцев объемом в сотни миллионов долларов.

tudies.com/papers.html1.html). Там были представлены статьи, против публикации которых заказчики проектов не возражали. Лучшей была признана работа Бхарата Рао (Bharat Rao) из Siemens, в которой описывалась очень интересная система. Она позволяет автоматически повысить качество лечения и ухода за пациентами кардиологических отделений благодаря тому, что извлекает важную медицинскую информацию из невнятно написанных и неточных записей в историях болезни[2].

Среди кандидатов в «Великие вызовы KDD» (см. врезку) есть задачи, близкие к тесту Тьюринга. Есть ли надежда, что техники ДМ помогут существенно продвинуться в решении такого рода классических проблем искусственного интеллекта? С другой стороны — можно ли в задачах протеомики надеяться на то, что только за счет ДМ появятся ответы на важные вопросы биологии?

— Из кандидатов в «Великие вызовы» ближе всего к Тьюринг-тесту предложение Ронена Фельдмана (Ronen Feldman) — выдвинуть в качестве вызова создание текст-майнинговых систем, которые смогут сдавать стандартные экзамены на понимание текстов, — SAT, GRE, GMAT, причем обучаться системы будут, исследуя веб.

Лично я думаю, что это вполне решаемая в течение пяти-десяти лет задача, а когда она будет решена, это полностью изменит существующую практику вступительных экзаменов.

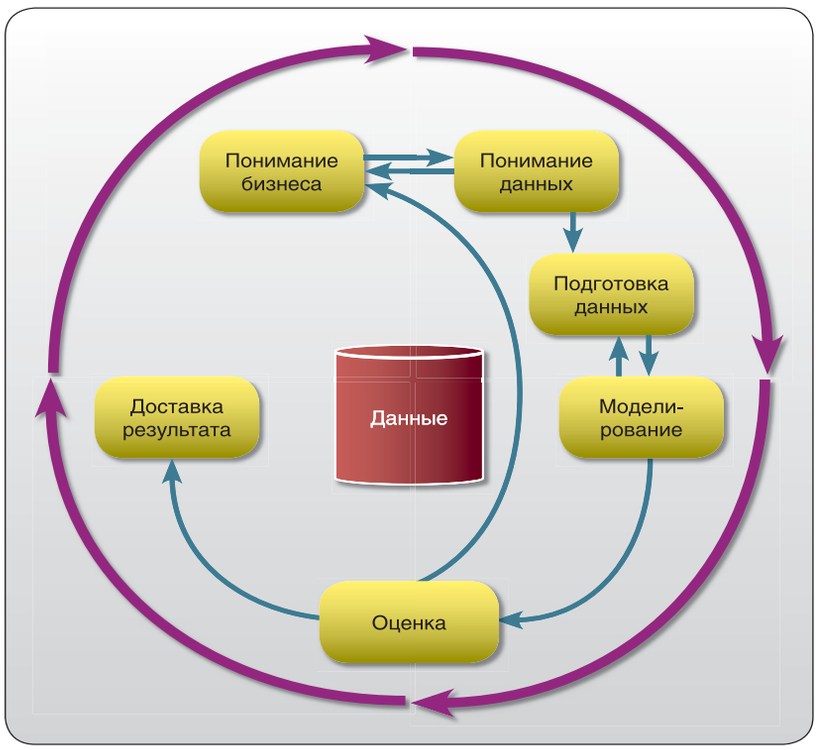
Недавно Ларри Пейдж, сооснователь Google, объявил, что Google серьезно работает над ИИ, а использование сосредоточенной там вычислительной мощности и базы знаний может серьезно ускорить движение в сторону ИИ.

2 Гм-гм. Недавно мы упоминали о том, как широко применяется распознавание речи при надиктовывании врачами историй болезни. Может быть, система Рао исправляет ошибки не только врачей, но и той системы, которая записывала их диктовку? — *Л.Л.-М.*

Для продвижения в биологии (протеомике, геномике) критически важно понимание предметной области. Однако и без инновационных алгоритмов датамайнинга прогресс там невозможен.

Как устроены системы датамайнинга? Много ли общего у этих технологий с технологиями поисковых машин типа Гугла?

— Системы датамайнинга устроены не так, как системы поиска по вебу (Google, Yahoo), поскольку датамайнинг работает обычно с цифровыми базами данных и задает другие вопросы, нежели Google. Обычно эти системы реализуют различные методы очистки и препроцессинга, а затем применяется основное ядро алгоритмов. Самые важные задачи, решаемые этими алгоритмами, — классификация, кластеризация, визуализация. Процесс датамайнинга требует множества итераций, как показано на рисунке. Важнейшая алгоритмическая часть — использование алгоритмов машинного обучения, то есть построение модели; для датамайнинговой системы это так же важно, как двигатель для спортивного автомобиля. Однако основные усилия обычно уходят на подготовку данных. Заинтересованных читателей приглашаю познакомиться с моими (свободно доступными) лекциями (www.kdnuggets.com/data_mining_ course/index.html). ■



■ ХОРОШАЯ МОДЕЛЬ ДАННЫХ ТАК ЖЕ ВАЖНА ДЛЯ ДМ, КАК ДВИГАТЕЛЬ ДЛЯ СПОРТКАРА

Кандидаты в великие

На конференции KDD–2006 (www.kdd2006.com) несколько известных исследователей в области извлечения знаний из данных предложили задачи, которые в будущем могут претендовать на роль «великих вызовов», бросаемых повседневной практикой (www.acm.org/sigs/sigkdd/explorations).

■ **ПРОВЕСТИ АННОТАЦИЮ 1000 ЧАСОВ ЦИФРОВОГО ВИДЕО В ТЕЧЕНИЕ ОДНОГО ЧАСА.** Согласно автору предложения Шабану Джерабе (Chabane Djeraba), в настоящее время это требует тысяч человеко–часов при ручной работе. Под аннотацией подразумевается краткое описание происходящего. Например, сегодня невозможно без выполненной человеком аннотации выделить в записи баскетбольного матча эпизоды атаки и обороны каждой команды. Ручная аннотация одной фотографии для Национального географического общества требует двадцать минут.

■ **ВИКИПЕДИЯ–ТЕСТ** (Lise Getoor, Лиз Гетур). По сборнику статей, созданному либо в режиме партисипативной журналистики (то есть по принципу наполнения Википедии), либо с использованием автоматических инструментов поиска линков по требуемой тематике, определить, какой из этих двух методов использовался: то есть составлен ли сборник машиной или людьми (и в каком случае качество оказалось выше)? Автор предложения указывает на связь этого вызова с другим, брошенным специалистам по сжатию информации: сжать 100 мегабайт Википедии до 18 мегабайт, не потеряв ни единого бита (за это уже назначен приз Хаттера в 50 тысяч долларов, www.hutter1.net).

■ **ОЦЕНИТЬ МИЛЛИАРД ПРОГНОЗИРУЮЩИХ МОДЕЛЕЙ** (Robert Grossman, Роберт Гроссман). В ходе многолетней практики датамайнинга было построено великое множество статистических моделей для различных типов и конкретных ансамблей данных. Во многих случаях для одних и тех же массивов данных строится несколько моделей, чтобы ухватить их характеристики разных видов. Пример: имеется информация от 833 датчиков движения транспорта в Чикаго. Задача состоит в автоматическом определении ситуаций, когда в транспортном потоке возникают аномалии, происходит что–то необычное (но не простая пробка!). Данные сегментировались по дням, часам и участкам дороги, что приводило к появлению 7x24x250 = 42000 автоматически генерируемых статистических моделей — хотелось бы значительно сократить их число! Подобная ситуация возникает и в онлайновом маркетинге (отдельная модель поведения для каждого клиента), в перспективных подходах к оценке эффективности лекарств на основе индивидуального генотипа и т. д. Так что миллиард набирается легко — вопрос в том, как радикально уменьшить это число.

■ **РАЗРАБОТКА СИСТЕМ АНАЛИЗА ТЕКСТОВ (TEXT MINING), СПОСОБНЫХ СДАТЬ ОБЫЧНЫЕ ЭКЗАМЕНЫ НА ПОНИМАНИЕ ТЕКСТА SAT, GRE, GMAT** (Ronen Feldman, Ронен Фелдман). Эту задачу с оптимизмом комментирует в своих ответах Григорий Пятецкий–Шапиро. Она покруче даже стандартного теста Тьюринга (определить, машина или человек отвечает на ваши вопросы), по поводу которого тоже было много оптимизма, в том числе и у его гениального автора. Однако не будем забывать, что этот вызов — лишь планка, которую автор предложения поднимает так высоко в надежде на достижение более приземленных практических целей: довести точность реализации реляционных запросов с нынешних 70–80% до 98–100%, причем в самой общей ситуации.

Кроме этого, был предложен еще один весьма важный вызов — функциональная аннотация белков. Однако формулировка здесь так сложна, а задач так много, что мы ограничимся лишь констатацией — это направление, датамайнинг в геномике и протеомике, тоже служит источником великих вызовов (напомним, кстати, что недавно назначен приз X PRIZE за снижение стоимости сканирования генома до 10 тысяч долларов при повышении производительности до ста геномов за десять дней). Ну а для полноты картины упомянем и конкурс, который состоится на конференции KDD–2007. Участникам предоставляется тренировочный массив данных Netflix, в котором собрано больше 100 млн. рейтингов (по пятибалльной шкале) по 18 тысячам фильмов от 480 тысяч случайно выбранных анонимных пользователей Netflix (то есть людей, бравших у Netflix DVD напрокат), с 1998 по 2005 год. Вот одна из двух задач, по которым будет проводиться состязание:

■ Дан список из 100 тысяч пар вида «номер_пользователя, номер_фильма», относящийся к 2006 году (то есть не входящий в тренировочный массив). Для каждой такой пары нужно указать вероятность, что данный пользователь хоть как–то рейтинговал данный фильм в 2006 году.

Денежные призы не предусмотрены — в отличие от основного конкурса Netflix (www.netflixprize.com). Там, чтобы заработать миллион долларов, требуется превзойти точность действующей сейчас на фирме системы рекомендаций Cinematch™ всего лишь на 10% (на исторических данных); ежегодно разыгрывается приз в скромные 50 тысяч долларов просто за самое большое уточнение прогноза. Прогноз состоит в том, чтобы угадать по предшествующим оценкам фильмов клиентами, какие из фильмов они высоко оценят в будущем. По состоянию на 14 марта 2007 года лучший результат в конкурсе Netflix уже 6,75%, то есть две трети пути к миллиону пройдено. ■