ПРЕДИСЛОВИЕ

Написать предисловие к этому изданию для меня исключительно приятно. Прошло уже довольно много лет, как Евгений Соломатин познакомил нас с платформой PolyAnalyst, созданной и постоянно совершенствуемой талантливым коллективом российской компании «Мегапьютер». А затем была принята и Программа «Приоритет-2030», что позволило открыть в Президентской академии лабораторию интеллектуальной аналитики. Ее сотрудники активно используют PolyAnalyst; более того, они были первыми читателями и доброжелательными критиками данного учебного пособия.

Вместе с командой компании «Мегапьютер» и ее вдохновляющим лидером Сергеем Ананяном нами были сделаны замечательные проекты. Они касались решения наиболее острых задач в государственных проектах развития, реализации практических отраслевых решений и сервисов, запуска образовательных программ.

Анализ данных, их обогащение, интерпретация результатов, применение визуальных онтологий, поддержка 16 иностранных языков, построение интерактивных дашбордов, встроенная поддержка наиболее распространенных методов машинного обучения, проектирование аналитических решений без знания языка программирования, язык XPDL для описания и поиска паттернов в текстах — лишь часть задач, доступных для решения с применением платформы PolyAnalyst.

Поздравляю авторов платформы с выходом этого издания! Конечно, оно не заменит классический «F1» продукта, однако даст возможность любому человеку (даже не имеющему специальных знаний) приступить к работе с системой и довольно быстро получить впечатляющие результаты по анализу данных.

Мне также доставляет удовольствие выразить благодарность Томскому государственному университету и Университетскому консорциуму исследователей больших данных в лице Вячеслава Гойко. «Мегапьютер» является индустриальным партнером Консорциума, а Слава и его коллеги с необычайной легкостью демонстрируют лучшие практики организации аналитических проектов, активно используя PolyAnalyst как в исследовательской, так и в образовательной деятельности, продвигают актуальные технологии в образовательном сообществе, формируя новое качество научного и практического знания.

Всем читателям я хотел бы пожелать продуктивной работы с материалом. В свою очередь будем рады получить обратную связь, что поможет сделать рынок отечественных цифровых решений еще более передовым.

Павел Голосов,
директор Центра цифровых решений
и искусственного интеллекта,
директор Института общественных наук,
Президентская академия

ВВЕДЕНИЕ

Жизнь в эпоху тотальной цифровизации всего и вся диктует, естественно, инновационный подход к анализу данных. И хотя само словосочетание «тотальная цифровизация» уже порядком набило оскомину, однако и в условиях новой реальности важнейшим ключом к достижению успеха в цифровом мире остается своевременное получение полезной и достоверной информации.

Легендарный афоризм гласит: «Кто владеет информацией, тот владеет миром». При этом велик риск того, что человек может попросту потеряться в «большой помойке» экспоненциально растущих массивов накопленных данных, которые необходимо проанализировать для извлечения информации, актуальной для решения именно его конкретной задачи. Поэтому в цифровую эпоху крайне важно уметь извлекать в агрегированном виде полезную информацию, т. е. обладать навыками использования современных методов анализа данных.

Данные о клиентах, поставщиках, продуктах и продажах, потоки данных с разнообразных датчиков, текстовые отчеты компаний, научные материалы, договоры, счета, накладные, технологическая документация, архивы, открытые данные различных организаций, данные компаний, Интернет, социальные сети — данных накопилось так много, что их уже невозможно «переварить». Доступ к данным в наши дни зачастую не является проблемой. Проблема — что с ними делать: как преобразовывать горы данных в знания, позволяющие принимать эффективные решения.

В XXI веке не только резко вырос объем создаваемых в мире данных. Они стали доступнее, поскольку стоимость их сбора и хранения кардинально упала. При этом скорость передачи данных в сетях связи также выросла — при одновременном падении тарифов. Многие даже не задумываются о том, что тариф подключения к Интернету за последние 15–20 лет почти не изменился, но при этом скорость подключения выросла на несколько порядков.

Рост объема и доступности данных в совокупности с ростом вычислительной мощности привел к развитию новых математических алгоритмов и средств анализа, ориентированных на сверхбольшие объемы разнородных данных. Произошел «квантовый скачок».

Например, нейросети как концепция и математический объект были изучены и описаны еще 60–70 лет назад. Но лишь с появлением суперкомпьютеров и кластерных вычислений появилась возможность рассчитывать за разумное время значения сотен миллиардов параметров нейросетей в процессе

их обучения. Это и стало ключевым фактором взрывного роста нейросетей типа ChatGPT и ее многочисленных аналогов. Наличие вычислительных мощностей — ключевой фактор конкуренции в технологической гонке за лидерство между странами.

Анализ данных, в особенности текстовых, требует применения технологий, основанных на алгоритмах машинного обучения и алгоритмах обработки естественного языка. Здесь важно помнить, что выбор методов анализа зависит от формата представления информации. Набор методов анализа структурированных данных, имеющих четкую структуру (например, таблицы с числами), отличается от методов анализа неструктурированных данных, когда информация представлена в произвольной форме (например, текстовые документы, сообщения в социальных сетях, новости из Интернета).

Развитие цифровой экономики и цифровизация промышленности привели к росту спроса на бизнес-аналитиков — специалистов, которые умеют структурировать окружающий нас океан информации, часто избыточной и «зашумленной», использовать данные в интересах бизнеса и государства, обрабатывать, интерпретировать результаты анализа, чтобы принимать на их основе взвешенные решения.

Как аналитики решают такие задачи?

Во-первых, информацию надо найти и собрать.

Во-вторых, следует провести анализ информации. Анализ делается кем-то — аналитиком (это человек) или компьютером.

В-третьих, нужно использовать результаты анализа. Анализ проводится для кого-то.

Таким образом, всегда есть потребитель (заказчик) — сотрудник компании, руководитель, чиновник, гражданин и т. д. Потребители, как правило, не могут сами провести сложный анализ данных, но именно они должны принимать управленческие и бизнес-решения на основе полученных результатов. Им нужен простой для понимания и удобный в использовании «образ результата» — такой, как «генеральский слайд» или аналитическая панель.

С точки зрения авторов, значительная часть практических аналитических задач в окружающем нас мире может быть решена на основе уже имеющейся алгоритмической базы. Такой «инженерный» подход прекрасно сочетается с концепцией LowCode, в которой механизм построения сценариев анализа данных и представления результатов реализуется с помощью визуального программирования в интуитивно понятном графическом интерфейсе.

Аналитик — это «художник», у которого есть краски (алгоритмы) и кисти (его знания). Художник создает свои шедевры на полотне. Математик — специалист по Data Science — разрабатывает свои сценарии анализа в вычислительной среде. Раньше его «полотном» был персональный компьютер — сейчас все чаще на смену ПК приходит распределенный виртуальный кластер.

Современные алгоритмы и модели, основанные, например, на нейросетях, часто типизированы, оптимизированы и отлажены. Есть множество готовых библиотек для использования, т. е. алгоритмы и модели можно рассматривать как элементы конструктора, из которых остается лишь собирать все что угодно. Знать бы только как...

Ключевая компетенция современного аналитика — это понимание того, какие алгоритмы, в каком случае и как применять, умение создавать из них сценарий анализа, ориентированный на получение конечного результата для заказчика. Аналитик должен знать, где и какие данные взять, как их очистить, агрегировать и обогатить, какая у них точность, какие модели анализа к ним можно применить с учетом различных ограничений, как интерпретировать результат и донести его до заказчика. Для самостоятельного погружения в тему анализа данных (а это целая вселенная!) читатель может обратиться к изданиям [1–8].

В традиционной парадигме компании покупают для своих аналитиков несколько ИТ-продуктов — программное обеспечение. Например, отдельно приобретаются средства ETL (Extract, Transformation, Load — загрузка, агрегация, очистка и преобразование данных), библиотеки алгоритмов Data Mining и Text Mining, отдельно — пакет программного обеспечения для визуализации результатов BI (Business Intelligence). Для того чтобы построить готовое промышленное решение или сервис, специалистам необходимо интегрировать эти программные продукты друг с другом, что отнимает много времени и усилий.

Есть и другой подход. В компании «Мегапьютер» мы называем это «демократизацией» работы с данными — когда все этапы работы с данными (в английском языке для этого есть термин Data Processing Pipeline) реализуются в единой среде разработки. Аналитик может преобразовать исходные данные, обогатить, агрегировать их, применить модель анализа, получить результат, визуализировать его, оценить, интерпретировать, «поиграть» данными или параметрами модели, пересчитать и т. д. Все это «не выходя из машины».

Именно такой подход положен в основу российской системы анализа данных PolyAnalyst (регистрация № 4414 от 16.04.2018 в Реестре российского ПО Минцифры РФ). Это набор функциональных узлов OCR (распознавание текстов), ETL, Data Mining, Text Mining (технологии обработки естественного языка NLP — Natural Language Processing), а также узлов BI (визуализация результатов анализа). Таким образом реализуется комбинация методов анализа текстов и данных, основанных на машинном обучении, включая нейросети и подход Rule-Based. Эти решающие правила реализованы в разработанном для PolyAnalyst языке XPDL (eXtended Pattern Definition Language).

Создание сценариев обработки данных в PolyAnalyst осуществляется как «сборка» конечного решения в интересах заказчика из деталей «конструктора», в котором элементами выступают алгоритмы, оформленные как

отдельные функциональные узлы. Каждый узел имеет вход, внутренние настройки и выход. Использование широкого набора инструментов позволяет реализовать концепцию комплексной аналитики, т. е. формирования в единой среде многошаговых аналитических сценариев анализа данных: их загрузки и преобразования, проведения исследований, интерактивного тестирования и доработки моделей, а также визуализации результатов анализа средствами блока BI на основе интерактивных графических объектов и многих других форматов представления.

PolyAnalyst имеет низкий «порог входа» при освоении (курс обучения занимает один-два дня) и позволяет на основе концепции LowCode в десятки раз сократить время решения задач по обработке больших объемов числовой и текстовой информации в интересах бизнеса и государственных заказчиков. Использование PolyAnalyst упрощает процесс разработки моделей и обеспечивает сотрудникам любой организации (непосредственным «владельцам» данных) возможность самостоятельно, не обращаясь к высокооплачиваемым специалистам (или с привлечением их ограниченного количества), проводить анализ данных.

Удобный интерфейс, представление информации в наглядном и для аналитика, и для руководителя виде, высокая производительность, способность обработки больших объемов данных, анализ любых предметных областей — все это помогает решать широкий спектр задач. Не удивительно, что на основе системы PolyAnalyst реализован ряд комплексных аналитических проектов в интересах компаний из различных отраслей промышленности и государственных органов.

Платформа PolyAnalyst постоянно развивается, соответственно и объем сопроводительной документации к ней растет. Сейчас платформа содержит более 100 узлов, а их описание занимает более 2000 страниц. Безусловно, в документации детально описаны «внутренности» системы, что для продвинутых специалистов чрезвычайно важно и необходимо. Тем не менее, многие наши пользователи, особенно представители вузов, в которых готовят специалистов по анализу данных, в том числе в рамках проекта «Цифровые кафедры» программы Минобрнауки России «Приоритет 2030», а также государственные органы и корпоративные заказчики давно просили нас написать краткий «курс молодого бойца» по PolyAnalyst.

Результат вы видите перед собой.

Надо особо отметить, что данное пособие появилось во многом благодаря поддержке и обратной связи от вузов — участников Ассоциации «Университетский консорциум исследователей больших данных». Компания «Мегапьютер» изначально была одним из его индустриальных партнеров. Благодаря инициативе, вдохновению и усилиям генерального директора Ассоциации Вячеслава Гойко, а также команды Томского государственного университета на базе этого вуза был создан центр коллективного пользования PolyAnalyst. Это позволило аналитикам высших учебных заведений со всей страны в режиме дистанционного доступа проводить свои исследования и

расчеты на больших объемах данных на суперкомпьютере СКИФ Cyberia ТГУ, используя функциональные возможности платформы PolyAnalyst. Результаты таких исследований вы можете найти в списке рекомендованной литературы [19-30].

Авторы хотят выразить особую благодарность за бесценную обратную связь по функционалу платформы PolyAnalyst, советы и рекомендации по форматам обучения методам анализа данных и изложению материалов пособия Екатерине Митягиной — доктору социологических наук, профессору, проректору по развитию на основе анализа данных Вятского государственного университета; Павлу Голосову — кандидату технических наук, директору Центра цифровых решений и искусственного интеллекта, директору Института общественных наук Президентской академии; Наталье Ястреб — доктору философских наук, доценту, директору Института социальных и гуманитарных наук Вологодского государственного университета; Ивану Пикалову — кандидату педагогических наук, доценту, руководителю Научно-методического центра разработки информационных систем и анализа данных Курского государственного университета; Константину Воронцову — доктору физико-математических наук, профессору РАН, заведующему лабораторией машинного обучения и семантического анализа Института Искусственного Интеллекта МГУ им. М.В. Ломоносова, и.о. заведующего кафедрой математических методов прогнозирования ВМК МГУ; Петру Иванову — проректору по цифровому развитию СВФУ им. М.К. Аммосова; Михаилу Мягкову — научному руководителю Центра прикладного анализа больших данных ТГУ, председателю совета Университетского консорциума исследователей больших данных; Давиду Калхиташвили — ассистенту кафедры системного анализа и анализа данных Института экономики, математики и информационных технологий Президентской академии; Борису Игольникову — кандидату экономических наук, доценту, руководителю образовательной программы «ИТ-сервисы и технологии обработки данных на транспорте» Российского университета транспорта (МИИТ).

Мы очень многому научились у наших клиентов и партнеров. С учетом их настойчивых требований и рекомендаций был значительно доработан функционал PolyAnalyst. За это говорим отдельное спасибо Михаилу Петрову — директору Департамента цифровой трансформации Счетной палаты РФ; Евгению Нечепоренко — исполнительному директору по исследованию данных ПАО «Сбербанк»; Алексею Гололобову — руководителю Управления информационных технологий Ситуационного центра Центра национальных проектов Аналитического центра при Правительстве Российской Федерации; Алексею Лохматову — первому заместителю начальника Центра моделирования бизнес-процессов ОАО «РЖД» и Виктору Положишникову — кандидату технических наук, главному эксперту этого подразделения.

Отрадно, что многие из наших клиентов и партнеров стали соратниками, единомышленниками и друзьями.

Компания «Мегапьютер» является активным членом Ассоциации разработчиков программных продуктов «Отечественный софт» — крупнейше-

го объединения российских производителей тиражируемого программного обеспечения. Ассоциация ведет огромную работу по консолидации игроков рынка для совместной работы над ключевыми вопросами развития российской ИТ-отрасли. Пользуясь случаем, мы хотели бы выразить благодарность Ренату Лашину, Исполнительному директору АРПП «Отечественный софт», за его вклад в развитие отечественного ИТ-рынка, готовность к диалогу, бесконечную энергию и неоценимую помощь всем членам ассоциации.

PolyAnalyst продолжает стремительно развиваться. Значительно расширен функционал по работе с нейросетями, средства ML. Обеспечена возможность подключения внешних библиотек и алгоритмов, включения в сценарий анализа готового кода на Python и R. Выпущена версия под Linux. При поддержке Российского фонда развития информационных технологий (РФРИТ) разработана кластерная версия PolyAnalyst GRID, которая поддерживает распределенную архитектуру вычислений и умеет распараллеливать алгоритмы, что критически необходимо при работе со сверхбольшими потоками и объемами данных.

Но это уже следующая история.

Мы очень надеемся, что данное издание послужит читателям ориентиром в огромном многообразии инструментов анализа данных, предлагаемых PolyAnalyst, поможет красиво, эффективно и элегантно решать прикладные задачи, а значит — профессионально расти.

Авторы будут признательны вам, уважаемые читатели, за рекомендации и пожелания по доработке данного учебного пособия.