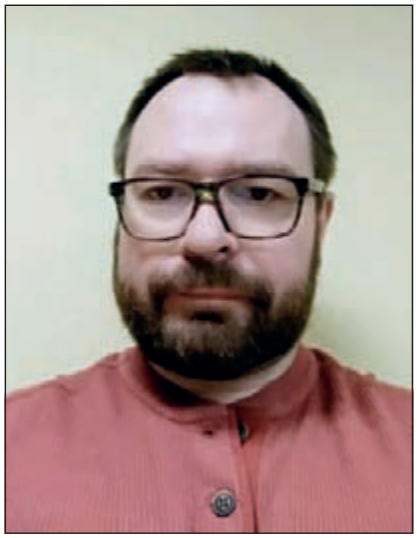
Темная сторона BigData

О том, что Искусственный интеллект в лице нейронных сетей, погрузившись в глубины BigData, сможет уже сегодня выявить всех мошенников, предотвратить сомнительные сделки и предсказать самые высокодоходные рынки, написано немало статей и выпущено немало новостных сообщений. Сама же по себе финансовая отрасль станет полностью автоматизированной под управлением мудрого искусственного интеллекта.

По мнению большинства футурологов, роботы вытеснят человека со многих позиций, в том числе и аналитических. Более того, общий вектор развития идет именно в этом направлении. Но, как говорил один известный восточный поэт: «Жизнь — цепь, а мелочи в ней — звенья. Нельзя звену не придавать значения».

И действительно, эти «звенья» порой преподносят множество сюрпризов. Например, такой факт: львиная доля всех данных, которые можно и нужно было бы использовать в задачах борьбы с мошенничеством, прогнозированием рынков, представляют собой неструктурированные текстовые данные. Количество ежедневно порождаемых письменных артефактов составляет миллиарды строк, анализ которых с помощью операторов практически бесполезен. Кто-то может поспорить, что все не так и большинство данных представляют собой обычные таблицы, которые хорошо обрабатываются статистическими методами. Выглядит вроде убедительно, что вроде бы подтверждается рапортами о широком использовании BigData в банках из TOP-30. Но если присмотреться повнимательнее, то даже в анализе структурированных данных, представленных в виде таблиц с цифрами, есть «нестыковки». Например, в отдельных столбцах названия товаров даны без SKU, а наименования организаций без указания каких-либо ИНН, присутствуют



Максим Ковалев
Генеральный директор IQSystems

фамилии и другие, скажем так, «неструктурированные данные».

По оценкам ряда западных источников, на «темные», или скрытые данные по разным странам приходится 50, а зачастую и более процентов[1]. В реальности же анализ проводится не более чем на 10–20 процентах информации.

Объединяющим все эти проблемы является тот факт, что никакими статистическими методами, будь то даже нейронные сети, решить эту задачу без применения поисково-аналитических систем семантического и семиотического анализа попросту невозможно. В качестве простого примера можно привести задачу борьбы с мошенничеством в сфере ипотечного кредитования или выдачу автокредита на покупку подержанного авто. Набор данных, которые хотелось бы получить, думается, понятен каждому: есть ли квартира или авто, под которые требуется выдать кредит в списках на продажу? А какова стоимость квадратного метра в этом же или соседнем доме, или цена аналогичного авто? А какова стоимость в пределах населенного пункта, а в пределах агломерации и т. д.? А как заемщик ведет себя в социальных сетях? Проблема же обработки товарных каталогов вообще является кошмаром любого аналитика...

Скачать данные с сайтов «как есть» на сегодняшний день не представляет собой технически сложной задачи. Получив такую базу, мы имеем миллионы записей с неструктурированной информаций и базу «категории» BigData во всей своей полноте.

В последнее время все больше разного рода государственных органов стали интересоваться темой семантического анализа данных. В качестве примера можно привести размещенный в мае 2017 года на сайте госзакупок электронный аукцион на разработку «аналитической подсистемы АИС ФНС»[2], в составе которой есть подсистема семантического анализа текстов.

К сожалению, за победными реляциями почему-то скрывается связан-

[1] По материалам https://www.veritas.com

[2] http://zakupki.gov.ru/epz/order/notice/ea44/view/documents.html?regNumber=0337100017717000138

ный с этим реальный пул проблем и упущенных возможностей. Попробуем разобраться хотя бы в некоторых из них.

Проблема «Трех V»

Под тремя V подразумевается цепочка: Объем — Скорость изменений — Разноформатность. Объем и скорость поступающих данных на сегодняшний день исключает возможность их обработки операторами. Следствием является острая необходимость в наличии на рынке продуктов, обеспечивающих решение задач Data Quality и Data Mining в автоматическом режиме с уровнем извлечения не ниже 90% при весьма высокой скорости обработки. При этом уровень ошибок не должен превышать одного-полутора процентов. Конечно, на это можно услышать возражение: дескать, есть различные распределенные решения, способные снять вопросы низкой производительности, которые присутствуют в линейках практически всех ведущих производителей — Oracle, IBM, Informatica и т. д.

Согласимся, но напомним, что такого рода процессы носят, как правило, циклический характер. А извлеченные данные должны сразу же пополнять справочники, поисковые индексы и т.д. Непересекающиеся в рамках одного потока данные могут неожиданно пересечься с данными из другого потока. Следовательно, количество вычислительных узлов должно быть сведено к максимальному минимуму, а производительность в рамках одного потока должна быть максимальна.

Проблема «темных» данных, или низкого процента использования данных

Основными причинами проблем являются их слабая структурированность на фоне низкого качества. Здесь сразу хочется пояснить, что проблема структурированности и низкого качества — это две совершенно разные проблемы. Неструктурированные данные, несмотря на то что они абсолютно достоверны и валидны, — трудно разложить на составляющие, чтобы построить какие-либо зависимости, проводить сравнительный анализ.

Невалидные, или данные с низким качеством, могут быть прекрасно структурированы, но не соответствовать объектам «реального» мира. Например, почтовый адрес может быть замечательно разложен по полям, но не существовать в природе.

Проблема обработки естественного языка

Назовем ее NLP-проблемой, или проблемой обработки естественного языка (от Natural Language Processing).

Безусловно, успехи в переводе данных компании Google достойны уважения. Но основаны они главным образом не на методиках семантического анализа, NLP-методах, а на Deep Learning. Иначе говоря, все на тех же статистических, пусть и продвинутых, моделях. В качестве аргумента можно привести пример с товарными каталогами. Какое количес-

тво элементов и сотен измерений характеристик на каждую единицу продукта создается ежедневно? А сколько вариантов написания одного и того же продукта? Ответ очевиден: равное количеству производителей, интернет-магазинов, логистических и дистрибуторских компаний, которые участвуют в цепочках поставок. Новые артефакты знаний и устойчивые или сленговые выражения рождаются у каждой, даже небольшой, группы людей каждый день. Да и смыслы, вкладываемые в одну и ту же фразу, могут отличаться до противоположных. А как обрабатывать ошибки, сокращения и аббревиатуры? Создавать базы вариантов написания, как предлагают большинство поставщиков решений?

Очевидно же, что это попросту невозможно.

Проблема языковой некомпетентности

Проблема отсутствия у западных систем компетенции в области семантики русского языка при выборе систем для работы с данными зачастую упускается сами аналитиками. Поставщики решений и системные интеграторы радушно обещают, что это легко решаемый вопрос, дескать, «наше решение уже присутствует во многих странах». К сожалению, тот факт, что это или международные организации, работающие на английском языке или это язык той же романской группы, или внедрение не полностью локализовано — попросту умалчивается. По нашему опыту, все известные на российском рынке попытки локализации задач семантического поиска не увенчались успехом, достигая уровня качества не выше 60-70 процентов от возможного.

Методологическо-эпистемологическая проблема

Что такое «наилучшая», или «истинная» структура знаний: справочников, атрибутов, документов? И каковы, например, связи между ними? К тому же, например, географическая локализация источников данных предопределит своеобразный «фильтр» трактовок и правил классификации каких-либо сущностей. В данном случае речь не идет о том, что в рамках информационного ландшафта существует несколько систем, а о том, что зачастую в рамках одной и той же системы одни и те же по своей сути объекты описаны и классифицированы по-разному. И причина не в невнимательности или нерадивости каких-то сотрудников. А — в контексте, условиях, национальных традициях, различном культурном коде, определяющих механизм подачи информации. Произвести однозначную регламентацию правил в этих условиях попросту невозможно.

Таким образом, задача использования больших данных, искусственного интеллекта и т. д. на самом деле требует более широко взгляда и комплексного подхода, который можно было бы выразить скорее термином Data Science, частью которого является процесс проектирования решений в формате BigData. А в процессе проектирования решений в области BigData следует уделять отдельное и не менее важное по сравнению с проблемами объемов и скоростями чтения записи значение вопросам очистки и извлечения данных. В противном случае мы будем по-прежнему копаться в массивах разнонаправленной и неструктурированной информации. Автоматизированный бардак — все равно бардак, как говорится в известной поговорке. Конструктивно порассуждать о возможных выходах из сложившейся ситуации сможем в продолжении темы в следующей статье. Так что продолжение следует.