электронный журнал
МОЛОДЕЖНЫЙ НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК
Издатель ФГБОУ ВПО "МГТУ им. Н.Э. Баумана". Эл No. ФС77-51038.

УДК 004.6, 004.8

Big data. Актуальность и перспективы использования

***Латышева А. М.**, студент*
Россия, 105005, г. Москва, МГТУ им. Н.Э. Баумана,
кафедра «Системы обработки информации и управления»

Научный руководитель: Гапанюк Ю.Е., к.т.н., доцент,
Россия, 105005, г. Москва, МГТУ им. Н.Э. Баумана
gapyu@bmstu.ru

1. Введение

Bid Data или Большие Данные уже давно у всех на слуху. Но не каждый точно знает, что же представляет собой это понятие. Чтобы внести ясность, рассмотрим пример из жизни.

Каждый, кому приходилось заранее планировать авиапутешествия, наверняка сталкивался с проблемой оптимизации расходов на билеты. Один и тот же маршрут завтра может стоить вдвое дороже, чем сегодня, а послезавтра – вчетверо дешевле. И сложно отгадать, когда выгоднее всего купить билет. У большинства людей ощущение экономического предательства растаяло бы очень быстро и ему на смену пришло бы смирение с неизбежностью. Но Орен Эциони – американский ученый в сфере компьютерных технологий, руководитель программы искусственного интеллекта в Вашингтонском университете, основатель множества компаний, занимающихся обработкой больших данных, еще до того, как термин «большие данные» приобрел известность – попавшись на финансовую удочку авиакомпаний, не захотел мириться с несправедливостью и стал искать способ, который помог бы определить выгодность той или иной цены в интернете.

Место в самолете – это товар. Все места на один рейс в целом одинаковы. А цены на них разительно отличаются в зависимости от множества факторов, полный список которых известен лишь самим авиакомпаниям. Эциони пришел к выводу, что не нужно учитывать все нюансы и причины разницы в цене. Нужно спрогнозировать вероятность того, что отображаемая цена возрастет или упадет. А это вполне осуществимо, причем без особого труда. Достаточно проанализировать все продажи билетов по заданному маршруту, а также соотношение цен и количества дней до вылета. Если средняя цена билета имела тенденцию к снижению, стоило подождать и купить билет позже. Если же

к увеличению — система рекомендовала сразу же приобрести билет по предложенной цене. Используя 12-тысячную выборку цен за 41 день, с трудом собранную на сайте путешествий, Эциони создал модель прогнозирования, которая обеспечивала его условным пассажирам неплохую экономию.

Результат не заставил себя долго ждать: новоиспеченный «прогнозист» – сайт Farecast.com – едва успел открыться в конце июля 2006 года для публичного бета-тестирования, как его накрыло потоком посетителей, а журнал «Time» включил его в свой престижный список Top 50. И это при том, что на первых порах сайт давал советы – купить билеты сейчас или подождать – только для одного маршрута между двумя американскими городами Бостоном и Сиэтлом. Уже к 2012 году система прогнозировала цены на авиабилеты для всех внутренних рейсов США, анализируя около триллиона записей. В 75% случаев система оказывалась права и позволяла путешественникам экономить на билете в среднем 50 долларов. В это время служба уже была выкуплена компанией Microsoft за 110 миллионов долларов США, после чего интегрирована в поисковую систему Bing.

Farecast — это воплощение компании, которая в своей работе использует большие данные; наглядный пример того, к чему идет мир. Еще пять или десять лет назад создать такую компанию было бы невозможно. Для большого объема данных необходимое количество вычислительных мощностей и хранилище обошлись бы слишком дорого. И хотя важнейшим фактором, сыгравшим на руку, стали изменения технологий, изменилось еще кое-что — едва уловимое, но более важное: само представление о том, как использовать данные.

2. Определение

Данные больше не рассматриваются как некая статичная или устаревшая величина, которая становится бесполезной по достижении определенной цели, например после приземления самолета. Скорее, они стали сырьевым материалом бизнеса, жизненно важным экономическим вкладом, используемым для создания новой экономической выгоды. Оказалось, что при правильном подходе их можно ловко использовать повторно, в качестве источника инноваций и новых услуг.

По сути, большие данные предназначены для прогнозирования. Обычно их описывают как часть компьютерной науки под названием «искусственный интеллект» (точнее, ее раздел «машинное обучение»). Рассматривается применение математических приемов к большому количеству данных для прогноза вероятностей, например таких, что электронное письмо является спамом; что вместо слова «коипя» предполагалось набрать

«копия»; что траектория и скорость движения человека, переходящего дорогу в неположенном месте, говорят о том, что он успеет перейти улицу вовремя и автомобилю нужно лишь немного снизить скорость. Но главное — эти системы работают эффективно благодаря поступлению большого количества данных, на основе которых они могут строить свои прогнозы. Более того, системы спроектированы таким образом, чтобы со временем улучшаться за счет отслеживания самых полезных сигналов и моделей по мере поступления новых данных.

В качестве определяющих характеристик для больших данных отмечают «три V»: объём (англ. volume, в смысле величины физического объёма), скорость (англ. velocity в смыслах как скорости прироста, так и необходимости высокоскоростной обработки и получения результатов), многообразие (англ. variety, в смысле возможности одновременной обработки различных типов структурированных и полуструктурированных данных).

3. Объем данных

К 2013 году количество хранящейся информации в мире составило 1,2 зеттабайта, из которых на нецифровую информацию приходится менее 2%.

Трудно представить себе такой объем данных. Если записать данные на компакт-диски и сложить их в пять стопок, то каждая из них будет высотой до Луны. В III веке до н. э. считалось, что весь интеллектуальный багаж человечества хранится в великой Александрийской библиотеке, поскольку египетский царь Птолемей II стремился сохранить копии всех письменных трудов. Сейчас же в мире накопилось столько цифровой информации, что на каждого живущего ее приходится в 320 раз больше, чем хранилось в Александрийской библиотеке.

Прогноз на 2014-2015 годы

- На технологии работы с Большими Данными в 2013 году в мире потрачено порядка 34 млрд долл., а к 2015 году в этом секторе будет создано 4,4 млн рабочих мест.
- В ближайшие 8 лет количество данных в мире достигнет 35 зеттабайт, по данным исследования IDC Digital Universe, опубликованного в декабре 2012 года. По прогнозам, количество данных на планете будет удваиваться каждые два года вплоть до 2020 года.

- Большую часть данных, которая будет произведена в период с 2012 по 2020 годы, сгенерируют не люди, а машины в ходе взаимодействия друг с другом и другими сетями данных. Сюда относятся, например, сенсоры и интеллектуальные устройства, которые могут взаимодействовать со сторонними девайсами.
- Количество серверов (виртуальных и физических) во всем мире вырастет десятикратно, в первую очередь за счет расширения и создания новых промышленных дата-центров, говорится в исследовании IDC. Тем не менее, количество обслуживающих их ИТ-специалистов увеличится не более чем в 1,5 раза. (Подробнее про рост вычислительной мощности см. рис.1.)
- Ожидается, что в будущем большая часть цифровой информации будет храниться в облаке. Однако в облаке будет производиться преимущественно обработка и процессинг данных, а непосредственно храниться в облаке будет только 15% информации.
- Инвестиции в управление, хранение, изучение битов в цифровой вселенной вырастут только на 40% в период с 2012 по 2020 году. В результате инвестиции на гигабайт в этот период снизятся с $2 до $0,2.
- Повышение вклада развивающихся ИТ рынков в наполнение цифровой вселенной новой информацией. Если в 2005 году, по данным IDC, 48% всех данных было сгенерировано в США и Западной Европе, а на развивающиеся страны в совокупности приходилось 20%, то в 2012 году доля развивающихся стран составила 36%, а к 2020 году достигнет 62%. Только на Китай будет приходиться 21% всей цифровой информации в мире.

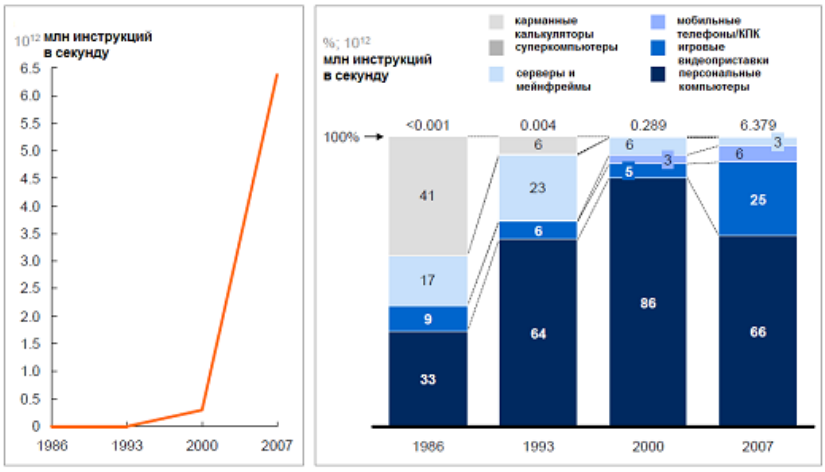

Рис. 1. Рост вычислительной мощности компьютерной техники (слева) на фоне трансформации парадигмы работы с данными (справа). Источник: Hilbert and López, `The world's technological capacity to store, communicate, and compute information,`Science, 2011Global

4. Источники

Источников больших данных в современном мире великое множество. В их качестве могут выступать непрерывно поступающие данные с измерительных устройств, события от радиочастотных идентификаторов, потоки сообщений из социальных сетей, метеорологические данные, данные дистанционного зондирования Земли, потоки данных о местонахождении абонентов сетей сотовой связи, устройств аудио- и видеорегистрации. Развитие и начало широкого использования этих источников послужило отправной точкой для проникновения технологий больших данных едва ли не во все сферы деятельности человека. В первую очередь в научно-исследовательскую деятельность, в коммерческий сектор и сферу государственного управления.

5. Методики анализа больших данных

Существует множество разнообразных методик анализа массивов данных, в основе которых лежит инструментарий, заимствованный из статистики и информатики (например, машинное обучение). Вот некоторые из них:

- *методы класса Data Mining*: обучение ассоциативным правилам (англ. *association rule learning*), классификация (методы категоризации новых данных на основе принципов, ранее применённых к уже наличествующим данным), кластерный анализ, регрессионный анализ;
- *краудсорсинг* — категоризация и обогащение данных силами широкого, неопределённого круга лиц, привлечённых на основании публичной оферты, без вступления в трудовые отношения;
- *смешение и интеграция данных* (англ. *data fusion and integration*) — набор техник, позволяющих интегрировать разнородные данные из разнообразных источников для возможности глубинного анализа, в качестве примеров таких техник, составляющих этот класс методов приводятся цифровая обработка сигналов и обработка естественного языка (включая тональный анализ);
- *машинное обучение*, включая обучение с учителем и без учителя, а также Ensemble learning (*англ.*) — использование моделей, построенных на базе статистического анализа или машинного обучения для получения комплексных прогнозов на основе базовых моделей (англ. *constituent models*, ср. состатистическим ансамблем в статистической механике);
- *искусственные нейронные сети*, *сетевой анализ*, *оптимизация*, в том числе генетические алгоритмы;
- *распознавание образов*;

- *прогнозная аналитика*;
- *имитационное моделирование*;
- *пространственный анализ* (англ. *Spatial analysis*) — класс методов, использующих топологическую, геометрическую и географическую информацию в данных;
- *статистический анализ*, в качестве примеров методов приводятся A/B-тестирование и анализ временных рядов;
- *визуализация аналитических данных* — представление информации в виде рисунков, диаграмм, с использованием интерактивных возможностей и анимации как для получения результатов, так и для использования в качестве исходных данных для дальнейшего анализа.

6. Аналитический инструментарий

Некоторые из перечисленных в предыдущем подразделе подходов или определенную их совокупность позволяют реализовать на практике аналитические движки для работы с большими данными. Из свободных или относительно недорогих открытых систем анализа Big Data можно порекомендовать:

- 1010data;
- Apache Chukwa;
- Apache Hadoop;
- Apache Hive;
- Apache Pig!;
- Jaspersoft;
- LexisNexis Risk Solutions HPCC Systems;
- MapReduce;
- Revolution Analytics (на базе языка R для мат.статистики).

Особый интерес в этом списке представляет Apache Hadoop – ПО с открытым кодом, которое за последние пять лет испытано в качестве анализатора данных большинством трекеров акций. В настоящее время практически все современные средства анализа больших данных предоставляют средства интеграции с Hadoop. Их разработчиками выступают как стартапы, так и общеизвестные мировые компании.

7. Адаптация технологий

- Каждый второй ИТ-директор готов потратиться на Big data

После нескольких лет экспериментов с технологиями Big data и первых внедрений в 2013 году адаптация подобных решений значительно возросла. Исследователи опросили ИТ-лидеров во всем мире и установили, что 42% опрошенных уже инвестировали в технологии Big data или планируют совершить такие инвестиции в течение ближайшего года (данные на март 2013 года).

Компании вынуждены потратиться на технологии обработки больших данных, поскольку информационный ландшафт стремительно меняется, требует новых подходов к обработке информации. Многие компании уже осознали, что большие массивы данных являются критически важными, причем работа с ними позволяет достичь выгод, не доступных при использовании традиционных источников информации и способов ее обработки. Кроме того, постоянная поддержка темы «больших данных» в СМИ подогревает интерес к соответствующим технологиям.

- Эксперты: Big Data провоцируют все больше «шума»

Все без исключения вендоры на рынке управления данными сегодня ведут разработку технологий для менеджмента Big Data. Этот новый технологический тренд также активно обсуждается профессиональным сообществом, как разработчиками, так и отраслевыми аналитиками и потенциальными потребителями таких решений.

8. Мировой рынок технологий Big Data

Ведущие игроки рынка

Интерес к инструментам сбора, обработки, управления и анализа больших данных проявляют едва ли не все ведущие ИТ-компании, что вполне закономерно. Во-первых, они непосредственно сталкиваются с этим в собственном бизнесе, во-вторых, большие данные открывают отличные возможности для освоения новых ниш рынка и привлечения новых заказчиков. Вот некоторые из таких компаний:

- Amazon
- Dell
- eBay
- EMC
- Facebook
- Fujitsu
- Google
- Hitachi Data Systems Corporation
- HP
- IBM
- LinkedIn

- Microsoft
- NetApp
- Oracle
- SAP
- SAS
- SGI (Silicon Graphics Inc)
- Teradata
- VMware
- Yahoo

Новая волна стартапов

В последнее время появляется множество стартапов, которые делают бизнес на обработке огромных массивов данных. Часть из них используют готовую облачную инфраструктуру, предоставляемую крупными игроками вроде Amazon.

Например, сайт http://www.ancestry.com/ пытается построить семейную историю всего человечества, основываясь на всех доступных на сегодняшний день типах данных: от рукописных записей во всевозможных книгах учета до ДНК-анализа. На сегодняшний день им удалось собрать уже около пяти миллиардов профилей людей, живших в самые разные исторические эпохи, и 45 миллионов генеалогических деревьев, описывающих связи внутри семей.

Главная сложность в этой работе заключается в том, что обрабатываемые данные страдают неполнотой, в них много неточностей, а идентифицировать людей нужно по отнюдь не уникальным именам, фамилиям, датам рождения, смерти и т.п. Стандартные алгоритмы не справляются с обработкой таких данных. Однако машинное обучение позволяет учитывать все эти неточности и с большой вероятностью выдавать правильные результаты.

Другой пример – проект http://www.eharmony.com/. Это сайт знакомств, на котором сейчас есть около 40 миллионов зарегистрированных пользователей. В анкетах можно указывать до 1000 различных признаков. Ежедневно система делает около 100 миллионов предположений о том, что два человека могут подходить друг другу. (См. рис.2.)

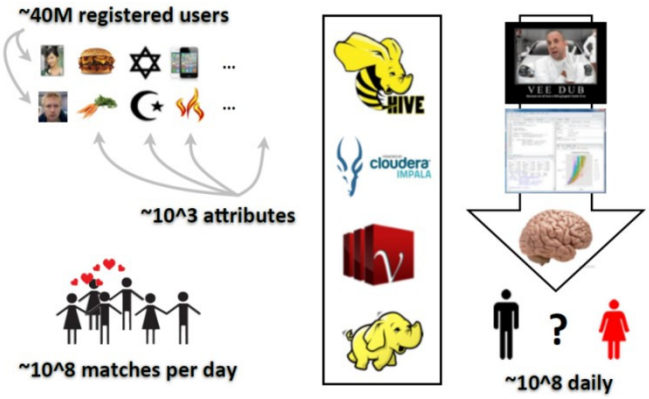

Рис. 2. Принцип работы проекта eHarmony

И предположения эти строятся не просто на банальном нахождении соответствий в указанных пользователями свойствах и пристрастиях. Например, выяснилось, что относительная площадь лица на фотографии в профиле может влиять на вероятность контакта между определенными людьми. Кроме того, оказалось, что люди с пристрастиями к определенным видам пищи могут обладать разной совместимостью друг с другом. Два вегетарианца с вероятностью в 44% найдут общий язык и начнут общение, в то время как два любителя гамбургеров с вероятностью 42% никаких отношений не заведут (подробнее см. рис.3.)

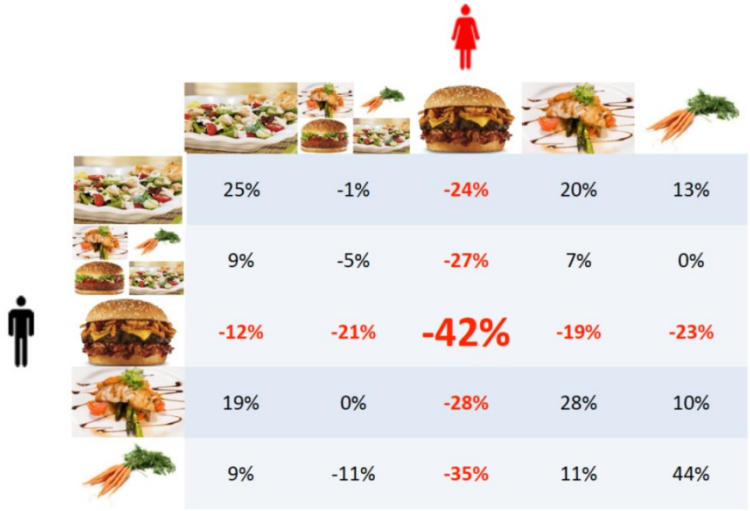

Рис. 3. Совместимость мужчин и женщин в зависимости от их питания

9. Заключение

Мир стоит на пороге эпохи больших данных, и люди сталкиваются с ними ежедневно. Онлайн магазины на основе сведений о покупках клиентов делают прогнозы по их предпочтениям, на основе которых им рассылаются всевозможные акции и скидки.

Сайты, способные построить семейное дерево любого человека, основываясь на любых доступных на сегодняшний день типах данных: от рукописных записей во всевозможных книгах учета до ДНК-анализа. Спам-фильтры разрабатываются с учетом автоматической адаптации к изменению типов нежелательных электронных писем. Сайты знакомств подбирают пары на основе корреляции многочисленных атрибутов с теми, кто ранее составил удачные пары. Функция автозамены в смартфонах отслеживает действия пользователя и добавляет новые вводимые слова в свой орфографический словарь. И это далеко не все примеры. От автомобилей, способных определять момент для поворота или торможения, до компьютеров IBM Watson, которые обыгрывают людей на игровом шоу Jeopardy, — этот подход во многом меняет наше представление о мире, в котором мы живем.

Список литературы

1. Майер-Шенбергер В., Кукьер К. Большие данные. Революция, которая изменит то, как мы живем, работаем и мыслим: пер. с англ./под ред. И. Гайдюк, М.: Манн, Иванов и Фербер, 2014. 221 с. [V. Mayer-Schönberger, K. Cukie Big Data: A Revolution that Will Transform How We Live. US: Houghton Mifflin Harcourt, 2013. 256 p.].
2. Лекция А. Себранта в Яндексе. Что такое на самом деле Big Data и чем они
3. прекрасны. Режим доступа: http://habrahabr.ru/company/yandex/blog/214217/ (дата обращения 14.04.14).
4. Большие данные (Big_Data)на новостном портале Tadviser. Режим доступа: http://www.tadviser.ru/index.php/Статья:Большие_данные_(Big_Data) (дата обращения 14.04.14).