Математические структуры и моделирование
2020. № 1(53). С. 139–143

УДК 004.6
DOI 10.24147/2222-8772.2020.1.139-143

О ПРИМЕНИМОСТИ СУЩЕСТВУЮЩИХ АЛГОРИТМОВ ОБРАБОТКИ ДАННЫХ К BIG DATA

А.И. Горев
к.ю.н., доцент, e-mail: gorev_a@inbox.ru
Е.Г. Горева
к.ф.-м.н., доцент, e-mail: gorev_a@inbox.ru

Омский государственный университет им. Ф.М.Достоевского, г. Омск, Россия

Аннотация. Определены процессы получения BIG DATA при функционировании компьютерных технологий, целью которых не является создание массивов данных. Однако использование полученных данных может принести большой положительный эффект.

Ключевые слова: big data, автоматизированные информационные системы, структурированные данные.

Современное состояние информационного общества характеризуется повсеместным использованием компьютерных технологий сбора, фиксации и обработки информации. При этом надо обратить внимание, что указанные операции могут быть «побочным продуктом» функционирования автоматизированных информационных систем (далее — АИС). Так, сотовые операторы, предоставляя услуги телефонной связи и передачи данных, формируют биллинговые базы данных (далее — БД), банки при предоставлении услуг расчётных и кредитных карт, ведут БД расходов абонентов в формате кассового чека и пр.

Применение компьютерных технологий во всех сферах жизни общества привело к появлению доселе невозможного процесса формирования Big Data. И если ранее сбор данных для исследования общественного поведения ограничивался тысячами, то современные технологии предоставляют возможность сбора данных о миллионах, что недавно было просто физически невозможно. Причём эти данные, являясь «побочным» продуктом технологий, часто не востребованы. Более того, всё чаще выдвигается утверждение о том, что сначала надо собрать как можно больше данных, на основе которых потом можно будет строить теорию [1].

Складывается парадоксальная ситуация: практики с пренебрежением относятся к научным теориям, выдвигаемым учёными, а учёные, получив возможности Data Science, увлеклись сбором данных, оставив исследования «на потом». Для оправдания бездействия выдвигается проблема плохой структурированности данных («на то они и большие данные, что плохо структурированы») и невозможности их восприятия. При этом единственным способом восприятия определяется визуализация данных [2]. Однако визуализация данных применима только для последующей ручной обработки, что существенно ограничивает

возможное применение результатов исследований. А ведь именно своевременное исследование Big Data может позволить государству и обществу произвести процесс принятия решений в режиме реального времени и в правильном направлении.

Ссылка на плохую структурированность данных также не выдерживает критики. Современные компьютерные системы сбора информации хранят собранные данные в базах данных, что однозначно предопределяет их структурированность. Обязательными параметрами являются дата / время и идентификатор пользователя / клиента системы. Остальные фиксируемые факторы зависят от направленности действия АИС. Более того, структуры данных могут быть изменены (перекодированы) по желанию исследователя без потери самих данных.

На практике возникают нелепые ситуации: не понимая особенностей компьютерного формата хранения данных и возможностей их перекодировки в зависимости от потребностей заказчика, специалисты в области социологии, юриспруденции, экономики и пр. без консультаций с IT-специалистами выбирают те программы обработки данных, которые работают со знакомыми им форматами. Так, при выборе для тестирования с последующей закупкой программно-аппаратных комплексов контроля автотрасс заказчик руководствовался не качеством распознавания объектов, а применяемыми форматами хранения изображений, совпадавшими с накопленной базой данных.

Таким образом, задачи обработки Big Data сводятся к пониманию целей их обработки и построению алгоритма. Конечно, реализация программы тоже является непростой задачей, однако можно отметить, что задачи подобного класса уже были неоднократно реализованы в интересах различных заказчиков.

Примером могут служить специализированные операторские системы класса Fraud Management System (далее — FMS), созданные в качестве мер технической защиты от фрода (от англ. Fraud, — вид мошенничества в области информационных технологий, в частности несанкционированные действия и неправомерное пользование ресурсами и услугами), способные обнаруживать проявления мошенничества в телекоммуникационных системах в режиме реального времени. Работа FMS заключается в мониторинге событий, происходящих в сети, т. е. в обнаружении активности по заранее заданным признакам и последующем анализе. Критерии, по которым система вычленяет неоднозначные действия, разнообразны. Учитываются не только типичные параметры вроде длительности вызова или географии вызова / приёма, но и, например, возраст абонента или период, в течение которого он пользуется услугами компании. Наблюдения используются для формирования групп по аналогии. Для получения максимально широкого информационного охвата FMS интегрируют с другими операторскими системами [3]. Существуют как зарубежные, так и отечественные программы данного класса.

Аналогичные системы давно и успешно функционируют в банковской сфере, проводя анализ активности клиентов банков по тем же критериям географии, времени и типа услуги. Банковские системы ориентированы на выявление противоправных действий с платёжными картами, счетов «салями» и иных мошенничеств. Изучив профиль клиента (регулярность поступления и величину

денежных сумм, кредитную историю, операции по счету и пр.), система способна оценить его платежеспособность при кредитовании. Указанные АИС не являются системами искусственного интеллекта, хотя и способны к самообучению.

Однако указанные системы, начало разработки которых относится к 90-м гг. XX века, действуют в интересах специализированных предприятий и организаций и направлены, как правило, на выявление нестандартных действий с противоправной направленностью. Тем не менее, именно эти алгоритмы могут быть взяты за основу при обработке Big Data. Сдерживающим фактором обработки Big Data является ограниченное распространение пакетов прикладных программ для анализа миллионов записей. Однако рассмотренные выше системы типа FMS, собирая миллионы записей, строили статистически значимый портрет для групп. Такой подход применим и в настоящее время. Имея на входе миллионы данных, программа обработки производит их сортировку по группам характеризующих признаков. Дальнейшая обработка ведётся по выделенным группам. Такой подход, с одной стороны, обезличивает возможные персональные данные, с другой — значительно сокращает объём обработки. После этапа сортировки обработка групп может вестись традиционными методами. Несомненно, могут возникать ошибки, однако при непрерывном анализе и группировке входящих данных с последующим изменением характеризующих признаков групп они будут на уровне статистической погрешности.

Сфера сбора данных с каждым днем становится шире. Примером такой нерешённой задачи являются городские пассажиропотоки, не совпадающие с городской транспортной схемой. Повышение эффективности городского транспорта решает не только вопрос рентабельности, но и экологическую проблему любого мегаполиса, поскольку именно городской транспорт является основным источником загрязнений. В 80-е гг. XX века изучение пассажиропотоков проводилось в ручном режиме статистами в салонах городского транспорта. Такое трудоёмкое решение всё равно не позволяло точно оценить пассажиропотоки из-за ограниченного времени наблюдения и «человеческого фактора», связанного со статистами. В современном городе с переходом на безналичные системы оплаты проезда данная задача перешла в разряд вычислительных. Конечно, построение такой модели является непростой задачей, однако она выполнима и может быть реализована.

В качестве элементов наблюдения и анализа данная система может рассматривать транспортные и банковские карты оплаты, предъявляемые пассажирами. Используемые карты оплаты изначально обезличены и не составляют проблему оборота персональных данных. Фиксация их в кассовом аппарате с указанием времени позволяет с высокой точностью определить место посадки клиента. При возвращении пассажира в вечернее время в БД будет зафиксировано место обратной посадки. Ежедневные сотни тысяч записей позволяют с высокой достоверностью собрать статистические данные о пассажиропотоках в течение дня в разные дни недели в любой период года.

Транспортная схема, разработанная с учётом пассажиропотоков, может подстраиваться под изменения интенсивности и направленности движения. Посто-

янное обновление БД позволит системе реагировать на изменения и перенаправлять транспорт. Ещё больший эффект может быть получен при взаимодействии данной системы с картой дорожной обстановки, пробок и местоположения подвижного состава.

В отдельности многие из названных задач уже существуют: карта дорожной обстановки, пробки, местоположение подвижного состава, БД оплаты проезда. Но максимальный эффект может быть получен только при построении интегрированной системы. И таких задач с Big Data с каждым днем становится всё больше. Необходимо решать парадоксальную ситуацию, в которой оказалось современное информационное общество — много современной техники, ещё больше фиксируемых данных и полное отсутствие тенденций улучшения в прикладных сферах жизни общества.

Литература

1. Сафронов П. Методы науки о данных в социальных исследованиях. URL: `//https://postnauka.ru/talks/82202`(дата обращения: 15.09.2019).
2. Новиков А. Проблема больших данных в городских исследованиях. URL: `https://postnauka.ru/video/83423` (дата обращения: 15.09.2019).
3. Семенов Г.В., Бирюков П.Н. Ответственность за «мошенничество» в сетях сотовой связи: учебное пособие. Воронеж: Изд-во Воронежского гос. ун-та, 2002. С. 41.

Дата поступления в редакцию: 30.11.2019

ABOUT THE APPLICABILITY OF EXISTING DATA PROCESSING ALGORITHMS TO BIG DATA

E.G. Gorev
Ph.D. (Phys-Math.), Associate Professor, e-mail: gorev_a@inbox.ru

Abstract. The processes of obtaining BIG DATA during operation of computer technologies, the purpose of which is not to create data arrays, are defined. However, the use of the obtained data can have a great positive effect.

Keywords: big data, structured data, automated information systems.

References

1. Takie mysli vydvinul v 2008 g. Kris Anderson, redaktor Wired, v stat'e ≪The End of Theory≫. Tsit. po Safronov P. Metody nauki o dannykh v sotsial'nykh issledovaniyakh Safronov P. Metody nauki o dannykh v sotsial'nykh issledovaniyakh. URL: `//https://postnauka.ru/talks/82202`(15.09.2019). (in Russian)
2. Novikov A. Problema bol'shikh dannykh v gorodskikh issledovaniyakh. URL: `https://postnauka.ru/video/83423` (15.09.2019). (in Russian)

3. Semenov G.V., Biryukov P.N. Otvetstvennost' za «moshennichestvo» v setyakh sotovoi svyazi: uchebnoe posobie. Voronezh: Izd-vo Voronezhskogo gos. un-ta, 2002. pp. 41. (in Russian)

Дата поступления в редакцию: 30.11.2019