УДК 004.6–024.11

БОЛЬШИЕ ДАННЫЕ И ПРИНЦИПЫ РАЗРАБОТКИ АНАЛИТИЧЕСКИХ СИСТЕМ

С.М. Боровиков
Доцент кафедры проектирования информационно-компьютерных систем Белорусского государственного университета информатики и радиоэлектроники, кандидат технических наук

С.К. Дик
Первый проректор Белорусского государственного университета информатики и радиоэлектроники, кандидат физико-математических наук

С.С. Дик
Аспирант кафедры проектирования информационно-компьютерных систем Белорусского государственного университета информатики и радиоэлектроники, магистр технических наук

Белорусский государственный университет информатики и радиоэлектроники,
Республика Беларусь
E-mail: bsm@bsuir.by

С.М. Боровиков

Доцент кафедры проектирования информационно-компьютерных систем БГУИР. Основная область научных интересов: прикладные математические методы в проектировании изделий радиоэлектроники, включая алгоритмы статистического прогнозирования надёжности изделий электронной техники и оценку надёжности прикладного программного обеспечения на ранних этапах его разработки. Руководитель разработки программных комплексов по автоматизированному расчёту и обеспечению надёжности электронных устройств: система АРИОН (2008-2009 гг.), система АРИОН-плюс (2011-2015 гг.).

С.К. Дик

Декан факультета компьютерного проектирования (2007-2014 гг.), проректор по учебной и воспитательной работе (2014-2016 гг.), первый проректор Белорусского государственного университета информатики и радиоэлектроники (с 2016 г. по настоящее время), кандидат физико-математических наук. Область научных интересов: средства медицинской электроники, аппаратно-программная поддержка внедрения новых прогрессивных методов и технологий в диагностике и лечении людей.

С.С. Дик

Окончил Белорусский государственный университет информатики и радиоэлектроники (2016 г.), в настоящее время является аспирантом этого университета, магистр технических наук. Работает в ООО «Syber Genesis» в должности начальника отдела разработки программного обеспечения. Проводит научные исследования по созданию и внедрению информационно-компьютерных систем в процессы подготовки спортсменов.

Аннотация. В работе в простой и доступной форме даётся наиболее общее представление о больших данных (Big Data), технологии и принципах создания аналитических систем, использующих эти данные. Материал ориентирован на IT-специалистов, интересующихся тем, как сфера их профессиональной деятельности может быть расширена с использованием достижений в области больших данных и создания систем расширенной аналитики (Big Data and Advanced Analytics).

Ключевые слова: большие данные, технологии Big Data, аналитические системы, целевая переменная, прогнозирование.

История больших данных. Сам термин Big Data впервые был озвучен в 2008 году на страницах специального выпуска журнала Nature в статье главного редактора Клиффорда Линча. Этот номер издания был посвящён взрывному росту глобальных объёмов данных и их роли в науке. Многие специалисты утверждают, что большими данными допустимо называть любые потоки информации объёмом более 100 Гб в сутки. Сами по себе алгоритмы Big Data возникли при внедрении первых высокопроизводительных серверов (мэйнфреймов), обладающих достаточными ресурсами для оперативной обработки информации и пригодных для компьютерных вычислений и для дальнейшего анализа. Однако, в последние годы термин Big Data стал слишком популярным, его употребляют практически везде, где упоминаются потоки данных, и как следствие, он стал восприниматься слишком обобщённо и размыто.

Надо признать, что понятие «большие данные (Big Data)» являются довольно условным и относительным. На пользовательском уровне самое распространённое определение можно сформулировать как набор данных, по объёму превосходящих жёсткий диск одного персонального устройства, и обработка которых классическими методами и инструментами, применяемыми для меньших объёмов, невозможна, либо не может быть выполнена в требуемые сроки.

В настоящее время термином Big Data обозначают также технологии обработки структурированных и неструктурированных данных огромных объёмов для получения полезных и понятных человеку результатов. В бизнесе Big Data используется для поддержки принятия решений руководителем (например, на основании анализа финансовых показателей из учётной системы) или маркетологом (например, на основании анализа предпочтений клиентов из социальных сетей).

По мнению специалистов [1] Big Data – это не только данные больших объёмов, но и набор технологий, таких, как потоковая аналитика и аналитика неструктурированных данных, распределённые файловые системы и базы данных, массивно-параллельная обработка. Часть из них являются относительно новыми для массового применения в IT, а часть получила «второе дыхание» благодаря интересу к Big Data в силу расширения возможностей решения реальных задач. Big Data – это скорее новая возможность быстрой обработки больших и плохо структурированных данных, которую мы получили с развитием информационных технологий.

Одной из задач технологии обработки больших данных является создание различных аналитических отчётов и получение прогнозов, которые будут использованы компаниями и организациями (далее собирательно – компаниями) в своей деятельности. Эффективность деятельности компаний по производству продукции, оказанию услуг (коммерческих, образовательных, медицинских и др.) может быть описана целевой переменной (англоязычный вариант – target), одной или несколькими. Во многих случаях в задачах прогнозирования с использованием Big Data целевая переменная может предсказываться на основе набора признаков (англоязычный вариант – features). Причём этот набор должен быть исчерпывающим с точки зрения прогнозирования целевой переменной с высокой достоверностью [2].

Эксперты в области IT высказывают мнение, что расширение Big Data и ускорение темпа роста стало объективной реальностью. Ежесекундно гигантские объёмы контента генерируют такие источники, как социальные сети, информационные сайты, кассовые терминалы, файлообменники – и это лишь сотая часть поставщиков. Согласно исследованию IDC Digital Universe, к 2020 году объём данных на планете вырастет до 40 зеттабайт, то есть на каждого живущего на Земле человека будет приходиться объём данных более 5 тысяч Гб [3].

По данным специалистов примерно 20-25 % цифровых данных содержат «полезные» сведения. Однако, только примерно 0,5 % данных в мире в действительности анализируется

[4], что подчёркивает важность технологии и талант, чтобы извлечь скрытые закономерности и знания из всех этих данных.

Принципы проектирования аналитических систем. Для успешного проектирования аналитических систем, использующих большие данные, аналитики должны иметь чёткое представление о прогнозировании как предметной области научного предсказания целевой переменной (target), описывающей процессы в рассматриваемой сфере человеческой деятельности (бизнес, образование, медицина, государственное управление и т.д.). В работе [5] авторы в систематизированной форме попытались пояснить суть некоторых методов прогнозирования и указать возможные области применения методов, которые могут быть востребованы в аналитических системах, использующих Big Data (таблица 1).

Таблица 1

Возможное применение методов прогнозирования в технологии Big Data [5]

Вид, метод прогнозирования	Возможное применение
1. Эвристическое	1. Оценка важности (ценности) больших данных, полученных из разных источников. 2. Оценка достоверности используемых больших данных, полученных из разных источников.
2. Математическое	Оценка прогнозных значений целевой переменной, описывающей деятельность компаний и организаций бизнеса, образования, транспорта, медицины и т.д.
3. Групповое (прогнозирование для группы однотипных объектов)	В случаях, когда интересуются эффектом деятельности корпораций, сообществ в целом (сети магазинов, производственные объединения, однотипные учреждения образования и т.д.)
4. Экстраполяцией целевой переменной (анализ временных рядов)	В случаях непосредственной оценки на основе Big Data значений целевой переменной для временных точек в прошлые и настоящий моменты времени
5. Экстраполяцией целевой переменной (обратное прогнозирование)	В задачах, когда необходимо перестроить деятельность компаний и организаций, не доводя их работу до возникновения потерь (экономических, технических, социальных и т.п.)
6. Прогнозирование целевой переменной по признакам	Оценка значения целевой переменной, одной или нескольких, описывающей деятельность компаний и организаций в разных областях хозяйственной, общественной и социальной сферах. Это наиболее типичные задачи с использованием Big Data

Подбор данных для обработки и выбор алгоритма анализа может стать не меньшей проблемой, так как отсутствует понимание, какие данные следует собирать и хранить, а какие можно игнорировать. Технологии Big Data – это не волшебная палочка, которая принимает на вход «простыню» (ещё один термин, который приходится слышать во многих компаниях), что-то делает внутри и выдаёт на выходе то, «что душе угодно». В прогнозных задачах целевая переменная (target) предсказывается, как правило, на основе набора признаков (features), который должен быть исчерпывающим в плане влияния на целевую переменную.

Один из возможных подходов к прогнозированию целевой переменной был рассмотрен в [6]. Согласно этому подходу, после определения целевой переменной, важнейшей для рассматриваемой сферы деятельности, рекомендуется определить ценность и достоверность данных (Big Data) для конкретной решаемой задачи; а также получить или уточнить характеристики степени важности или влиятельности (Value) и степени достоверности (Veracity) данных, которые будут использоваться для уточнения признаков и формирования прогнозной оценки целевой переменной. При выполнении этого этапа во многих случаях

имеет смысл воспользоваться приёмами эвристического прогнозирования или общепринятыми сведениями, которым можно доверять. Например, очевидным является то, что данные, получаемые от кассовых терминалов являются более ценными и достоверными, нежели данные форумов в социальных сетях. Для обработки неструктурированных данных, полученных из разных источников и принятых во внимание при получении характеристик важности и достоверности данных может быть использована программная модель Google MapReduce, позволяющая решать задачи сортировки и группировки данных. С её помощью, например, удобно организовать счётчик появления искомых слов в большом файле (построение Term-вектора) или счётчик частоты обращений к заданному адресу, вычислить объём всех веб-страниц со всех URL конкретного хост-узла или же создать список всех адресов, содержащих необходимые данные.

Для выделения переменных, рассматриваемых как признаки, и выявления из них тех, которые оказывают заметное влияние на целевую переменную, пригодны хорошо разработанные методы Data Mining, включающие в том числе статистические методы (корреляционный и регрессионный анализ, факторный анализ, дисперсионный анализ, компонентный анализ, дискриминантный анализ, анализ временных рядов и др.). Методы Data Mining позволяют обнаружить в данных ранее неизвестные, нетривиальные, практически полезные и доступные для интерпретации закономерности, необходимые для принятия решений о прогнозной оценке целевой переменной в рассматриваемой сфере человеческой деятельности (бизнес, медицина, образование и т.д.).

Получение прогнозной оценки целевой переменой (forecasting) является одной из задач Data Mining и одновременно одним из ключевых моментов при принятии решений.

Прогнозирование является распространённой и востребованной задачей во многих областях человеческой деятельности (бизнес, государственное управление, медицина, образование). В результате прогнозирования уменьшается риск принятия неверных, необоснованных или субъективных решений. Важно оценить достоверность решения, рекомендуемого к принятию (исполнению) и возможный при этом риск.

Согласно [6], в предположении независимости достоверности данных, полученных из m разных источников, можно воспользоваться формулой

$$R_{\text{пр}} = 1-(1-r_1)(1-r_2)\ \ldots\ (1-r_m),$$

где $R_{\text{пр}}$ – достоверность оценки целевой переменной, выраженная вероятностью принятия правильного (удачного) решения; r_j – достоверность данных, полученных из i-го источника, выраженная вероятностью (i = 1, 2, …, m); m – число источников, принятых во внимание аналитической системой при прогнозировании целевой переменной.

Расчёт риска, обусловленного принятием решения на основе прогноза целевой переменной (как результата работы аналитической системы), использующей Big Data, следует выполнять с учётом вероятности недоверия прогнозу, сделанному для целевой переменной, и размера возможных потерь в случае, если прогнозное решение не подтвердится. В качестве вероятности недоверия прогнозу можно использовать разность (1 – $R_{\text{пр}}$).

Заключение. Авторы попытались в простой доступной форме дать наиболее общее представление о больших данных (Big Data), технологии и принципах создания аналитических систем, использующих эти данные. Материал ориентирован на IT-специалистов, интересующихся тем, как сфера их профессиональной деятельности может быть расширена с использованием достижений в области Big Data and Advanced Analytics. При этом акцент сделан на рассмотрении принципов проектирования аналитических систем принятия решений, обеспечивающих возможность прогнозирования (на основе обработки больших данных) целевой переменной (target), являющейся важнейшей для описания процесса в рассматриваемой сфере человеческой деятельности: бизнес, образование, медицина, государственное управление.

Литература

[1]С. Кузнецов: Под термином Big Data скрываются самые разные вещи [Электронный ресурс]. – Режим доступа : http://www.iksmedia.ru/articles/5033748-SKuzneczov-Pod-terminom-Big-Data.html (дата обращения: 22.01.2019).

[2]Фрэнкс, Б. Укрощение больших данных. Как извлекать знания из массивов / Б. Фрэнкс ; пер. с англ. – М. : Изд-во «Технологии развития ООО», 2014. – 352 с.

[3]Что такое Big Data (большие данные) в маркетинге: проблемы, алгоритмы, методы анализа. [Электронный ресурс]. – Режим доступа: http://lpgenerator.ru/blog/2015/11/17/chto-takoe-big-data-bolshie-dannye-v-marketinge-problemy-algoritmy-metody-analiza/#ixzz48TK3zzF3 (дата обращения: 23.01.2019).

[4]Работа с Big Data: основные области и возможности [Электронный ресурс]. – Режим доступа: marketing.spb.ru›lib-research/methods/Big_Data.htm (дата обращения: 23.01.2019).

[5]Borovikov, S. Prediction in Big Data Technology / Borovikov, S. [and oth.] // BIG DATA and Advanced Analytics. Использование BIG DATA для оптимизации бизнеса и информационных технологий : сб. материалов II Междунар. науч.-практ. конф. (Минск, Республика Беларусь, 15–17 июня 2016 года). – Минск : БГУИР, 2016. – С. 98-101.

[6]Batura, M. Big Data Volumes and Some Approaches to the Creation of Corporate Analytical Systems / M. Batura [and oth.] // BIG DATA and Advanced Analytics. Использование BIG DATA для оптимизации бизнеса и информационных технологий: сб. материалов II Междунар. науч.-практ. конф. (Минск, Республика Беларусь, 15–17 июня 2016 года). – Минск : БГУИР, 2016. – С. 74-80.

BIG DATA AND PRINCIPLES OF DEVELOPMENT OF ANALYTICAL SYSTEMS

S.M. BOROVIKOV
PhD, associate professor of the department of Information and Computer Systems Design of the Belarusian State University of Informatics and Radioelectronics

S.K. DZICK
Ph.D., First Vice-Rector of the Belarusian State University of Informatics and Radioelectronics, associate professor

S.S. DZICK
Master of engineering, PG student of the Belarusian state university of informatics and Radioelectronics

Belarusian State University of Informatics and Radioelectronics, Republic of Belarus
E-mail: bsm@bsuir.by

Abstract. In the work in a simple and accessible form, the most general idea of big data (Big Data), technologies and principles of creating analytical systems using this data is given. The material was prepared for IT- professionals who are interested in how their professional field can be expanded using the achievements in the field of big data and the creation of advanced analytical systems.

Keywords: big data, Big Data technology, analytical systems, target variable, forecasting.