востребовано. Особое внимание предлагается уделить следующим направлениям исследований:

– выделение и анализ основные причины затрат времени на этапе локализации бизнес-процесса;

– определение наборов ключевых факторов, влияющих на этап локализации;

– разработка подходов к решению задач обработки банков данных для последующей локализации бизнес-процесса клиента.

Список использованных источников:

1. Олифер В.Г., Олифер Н.А. Компьютерные сети: принципы, технологии, протоколы. 3-е изд. СПб: Издательство «Питер», 2008. 958 с.
2. Кульгин М.В. Практика построения компьютерных сетей. Для профессионалов. СПб.: Питер, 2001. 320 с.
3. Dehaghani, S. M. H., & Hajrahimi, N. “Which factors affect software projects mainten ance cost more? “. Acta Informatica Medica, 21(1), 63, March 2013.
4. Coen J. Burki, Dr. Harald H. Vogt “How to save on software maintenance costs. An Omnext white paper on software quality”, November 2014.

БОЛЬШИЕ ДАННЫЕ (BIG DATA): ОБЕЩАНИЯ И ПРОБЛЕМЫ

Белорусский государственный университет информатики и радиоэлектроники
г. Минск, Республика Беларусь

Перцев И.Ю.

Ярмолик В.Н. – д-р.техн. наук, профессор

Смартфоны с мегапиксельными камерами, карманные компьютеры, беспроводные сенсорные сети, вездесущие социальные сети, спутники на околоземной орбите, космические телескопы - все это создает больше данных, чем когда-либо прежде. Не будет преувеличением сказать, что за последние два года было произведено более 90% мировых данных, и этот объем данных продолжает расти с развитием новых технологий, переходом к гигабитным сетям и объемному Интернету вещей (Internet of Things, IoT).

Солнечная обсерватория НАСА (SDO) использует четыре телескопа, делающих восемь изображений Солнца каждые 12 секунд. В январе 2015 года SDO получило свое 100-миллионное изображение Солнца - всего лишь один пример того, как астрономы собирают все больше данных. В настоящее время примерно один Петабайт этих данных доступен онлайн, и этот объем растет со скоростью 0,5 петабайта в год. В Большом адронном коллайдере CERN 150 миллионов датчиков собирают данные о почти 600 миллионах столкновений в секунду.Работа, которая получила Нобелевскую премию в области химии в 2013 году, включала измерение и визуализацию поведения около 50 000 или более атомов в реакции в течение доли миллисекунды.

В сфере социальных сетей пользователи Facebook добавляют 300 миллионов новых фотографий в день. Более 300 миллионов пользователей Instagram ежедневно публикуют 60 миллионов фотографий. Каждую минуту на YouTube загружается более 100 часов видео.

Несмотря на то, что данные по бизнесу и социальным сетям редко рассматриваются более одного раза, и реже используются для детального анализа, существует теория, что это изменится в скором времени. В 2013 году только 22 процента такого рода данных было сочтено полезным, и менее 5 процентов от этой суммы было фактически проанализировано. К 2020 году более 35 процентов всех данных можно было бы считать полезными из-за увеличения производительности устройств IoT, а также потому, что они все чаще проектируются для достижения конкретных целей, таких как научное открытие или оптимизация процесса. Например, данные IoT от гигантских газовых турбин, генерирующих электроэнергию, имеют огромную ценность, поскольку это может оптимизировать производство электроэнергии и помочь в обслуживании и ремонте. Аналогичным образом, проект радиотелескопа Square Skil Array (SKA), который, как ожидается, будет запущен в 2020 году, позволит производить 2,8 Гбайт астрономических данных в секунду, что поможет создать самую большую когда-либо созданную карту Вселенной.

Большие данные можно определить как слишком большие и сложные данные для сбора, обработки и анализа с использованием современной вычислительной инфраструктуры. В настоящее время такого рода данные характеризуются пятью характеристиками:

1) объем данных измеряется в терабайтах (240) или даже в петабайтах (250). При этом объемы данных стретмительно растут, приближаясь к экзабайтам (260);

2) скорость получение данных очень высокая, и из-за этого огромного объема некоторым приложениям требуется возможность обработки данных в реальном времени, чтобы определить, хранить ли часть данных;

3) разнородные данные являются гетерогенными и могут быть высокоструктурированными, полуструктурированными или полностью неструктурированные;

4) достоверность — в связи с необходимостью промежуточной обработки, разнообразия источников данных и эволюции данных, особое внимание привлекают вопросы безопасности, конфиденциальности, доверия и подотчетности. Это создает, в свою очередь, необходимость проверки источника данных;

5) модели с предсказанием ценности, которые отвечают на запросы «что-если», анализ этих данных может дать противоречивые сведения и обучаемый искуственный интеллект.

Большие данные позволяют определить новые направления научных исследований, которые ранее

были ограничены объемом имеющихся данных. К примеру, многие проблемы, связанные с естественным языком и речью, плохо подходят для математически точных алгоритмических решений. Для лучшего решения этой проблемы часто используются статистические модели машинного обучения, которые требуют подготовки данных для построения и оценки [1]. Теперь точный выбор математической модели теряет свое значение, потому что есть достаточно большие данные, чтобы компенсировать их необходимость [2].

Способность эффективно обрабатывать массивные наборы данных стала неотъемлемой частью широкого круга научных и других академических дисциплин. Однако это не устраняет необходимости глубокого понимания теоретических основ проблемы. Большие данные позволяют ученым преодолевать проблемы, связанные с малыми выборками исходных данных таким образом, чтобы уменьшить влияние предположений на математическую модель. Это позволяет избежать проблемы с перестроением существующей модели для каждого нового набора входных данных, лучше обрабатывать зашумленные исходные данные.

В последние годы появилось множество новых систем для решения задач хранения и обработки больших массивов данных. В настоящее время более 220 таких систем относятся к так называемым NoSQL систем данных, и регулярно появляются новые решения (http://db-engines.com/en/). Принимая это во внимание, мы сталкиваемся с ограниченными или отсутствующими теоретическими основами для моделей данных и языков запросов, и нам не хватает четких стандартов, которые помогли бы избежать проблем при работе. Многие системы NoSQL предназначены для развертывания на компьютерах с распределенными кластерами, и предлагают широкий выбор спецификаций уровня согласованности данных, обеспечивают встроенную поддержку параллельной обработки с использованием структуры MapReduce. Однако не все проблемы с большими данными поддаются решению с применением MapReduce. Например, изменяющиеся во времени графики и динамические сети, требовательные к обработке в режиме реального времени, и масштабируемая обработка потока данных создают дополнительные проблемы. Однако при решении поставленной задачи нужно проявлять осторожность. Далеко не всегда нужно использование обработки больших объемов данных, им нужны правильные данные [4]. Часто данные могут быть противоречивыми, неполными, неточными, субъективными, избыточными, необъективными и зашумленными. Такие данные потенциально могут создать путаницу и дезинформацию, а не давать полезные сведения. Как описано в «Big Data or Right Data?» [4], нам необходимо задать правильный вопрос:

1) Как мы обрабатываем, фильтруем и моделируем исходные данные, чтобы получить нужные?
2) Как мы определяем достоверность таких данных?
3) Как мы различаем достоверные данные и ошибочные? Фильтрация ошибочных данных является нетривиальной проблемой и возможный источник проблем.
4) Как мы определяем и устраняем дубликаты?

Конфиденциальность данных также актуальна, поскольку она касается юридических и этических ограничений. О каких вопросах конфиденциальности следует позаботиться? Нужно ли анонимизировать данные? Это связано с тем, что отслеживание происхождения данных является основным требованием во многих крупных приложениях: информация о происхождении используется для преобразования данных, позволяет проводить аудит, моделировать подлинность, осуществлять контроль доступа для производных данных и оценивать качество и доверие к данным. Около 400 лет назад Галилей сказал, что «книга природы написана на языке математики». Сегодня это становится еще актуальнее, учитывая потенциал больших данных для инновационных открытий в науке и аналитике, основанной на данных. Большие данные могут стать следующим рубежом инноваций, конкуренции и производительности.

Списокиспользованныхисточников

1. V. Gudivada, D. Rao, and V. Raghavan, “Big Data–Driven Natural Language– Processing Research and Applications,” Big Data Analytics, V. Govindaraju, V. Raghavan, and C.R. Rao, eds., Elsevier, 2015 (in press).
2. Halevy, P. Norvig, and F. Pereira, “The Unreasonable Effectiveness of Data,” IEEE Intelligent Systems, vol. 24, no. 2, 2009, pp. 8–12.
3. V. Gudivada, D. Rao, and V. Raghavan, “Renaissance in Data Management Systems: SQL, NoSQL, and NewSQL,” Computer (in press).
4. R. Baeza-Yates. “Big Data or Right Data?” Proc. 7th Alberto Mendelzon Int’l Workshop on Foundations of Data Management (AMW 13), 2013, vol. 1087, paper 14; http://ceur-ws.org/Vol-1087/paper14.pdf.

ИССЛЕДОВАНИЕ И АНАЛИЗ ЗАЩИЩЕННОСТИ МОБИЛЬНЫХ ПРИЛОЖЕНИЙ НА ОСНОВЕ ПОСТРОЕНИЯ ТЕСТА НА ПРОНИКНОВЕНИЕ

Белорусский государственный университет информатики и радиоэлектроники
г. Минск, Республика Беларусь

Петухов А. В.

Медведев С. А. – канд. техн. наук, доцент

Огромный рост количества мобильных устройств за последнее время, открыл новые возможности