Организатор

ОТКРЫТЫЕ СИСТЕМЫ
Open Systems Publications

21 октября
Москва, отель «Корстон»

конференция

БОЛЬШИЕ ДАННЫЕ
в национальной экономике

Тезисы докладов

В рамках деловой программы выставки softool

При поддержке

РФФИ РОССИЙСКИЙ ФОНД ФУНДАМЕНТАЛЬНЫХ ИССЛЕДОВАНИЙ

Российская Академия Наук

Вторая конференция «Большие Данные в национальной экономике»

Важнейшим условием успешного развития мировой экономики на современном этапе становится возможность фиксировать и анализировать огромные массивы и потоки информации. Существует точка зрения, что страны, которые овладеют наиболее эффективными методами работы с Большими Данными, ждет новая индустриальная революция. У России с ее колоссальным научным и образовательным потенциалом есть все шансы занять достойное место среди тех национальных экономик, где извлечение полезных знаний из больших объемов данных различной природы поставлено на службу индустриальному прогрессу.

Цель второй конференции «Большие Данные в национальной экономике» — предоставить экспертам в области работы с данными из научных организаций и индустрии площадку для обсуждения результатов своих проектов и перспектив их практического применения. Организаторы конференции также видят своей задачей популяризацию наиболее интересных работ, способствующих развитию Больших Данных как самостоятельного научного направления, а также расширению использования на практике методов работы с большими массивами информации в других областях научных исследований и в различных секторах российской экономики.

Оргкомитет конференции «Большие Данные в национальной экономике» выражает признательность за поддержку Российскому фонду фундаментальных исследований (грант 14—07—20305-г).

Тематика конференции «Большие Данные в национальной экономике»:
- Пленарная сессия. Перспективные методы анализа Больших Данных
- Секция. Большие Данные в научных исследованиях
- Секция. Прикладные аспекты Больших Данных

Тезисы докладов конференции «Большие Данные в национальной экономике» (Москва, 21 октября 2014 г.). /[Под ред. Дубовой Н.А.]. — М.: «Открытые системы», 2014. — 27 с.

В сборник включены доклады конференции «Большие Данные в национальной экономике», прошедшей 21 октября 2014 года в Москве в конференц-центре отеля «Корстон».

Целями конференции были обсуждение актуальных вопросов в области обработки и анализа больших объемов данных, представление результатов научно-практических исследований и консолидация наиболее значимых работ и коллективов, способных сделать вклад в формирование этого нового направления науки и индустрии, а также расширение использования на практике методов работы с большими массивами информации в других областях научных исследований и в различных секторах российской экономики. Материалы сборника предназначены для научных сотрудников, преподавателей, аспирантов и студентов, а также любых специалистов, интересующихся проблемами Больших Данных.

Подробную информацию о конференции «Большие Данные в национальной экономике» можно найти по адресу www.ospcon.ru.

Конференция «Большие данные в национальной экономике»
21 октября 2014 года, конференц-центр, отель «Корстон»

Организационный комитет

Гуляев Ю. В.	академик РАН, член президиума РАН, директор ИРЭ им. В. А. Котельникова РАН, председатель оргкомитета
Дубова Н. А.	научный редактор журнала «Открытые системы.СУБД», зам. председателя оргкомитета
Арлазаров В. Л.	чл.-корр. РАН, зав.отделом ИСА РАН, академик РАН, директор НИИСИ РАН
Будзко В. И.	д.т.н., зам.директора по научной работе ИПИ РАН
Волков Д. В.	с.н.с. ИПМ им.М.В.Келдыша РАН, гл. редактор журнала «Открытые системы.СУБД»
Гергель В. П.	д.т.н., декан факультета ВМК ННГУ им. Н. И. Лобачевского
Кузнецов Н. А.	академик РАН, президент Международного союза приборостроителей
Никитов С. А.	чл.-корр. РАН, зам. директора ИРЭ им. В.А.Котельникова РАН, зав. кафедрой МФТИ
Сухомлин В. А.	д.т.н., профессор, МГУ им. М. В. Ломоносова

Пленарная сессия. Перспективные методы анализа Больших Данных

Большие Данные: разделяй и властвуй

Сергей Кузнецов, д.т.н., главный научный сотрудник, **ИСП РАН**
Сравнительно установившейся идеей является то, что горизонтально масштабируемые системы обработки данных можно основывать только на подходе shared nothing (грубо говоря, на распределенных системах без наличия общих ресурсов между узлами). Легко видеть, что этому подходу в той или иной мере следуют как массивно-параллельные СУБД, так и решения NoSQL (в частности, map/reduce). Однако, как нетрудно заметить, подобные системы хорошо работают только в том случае, когда данные хорошо разделены по узлам системы (грубо говоря, разделение соответствует текущей рабочей нагрузке). И обещанное горизонтальное масштабирование реально возможно только в том случае, когда данные можно эффективно переразделить. Так что реально проблемой Больших Данных является проблема их разделения. Научимся разделять — проблемы почти и не будет.

Модели выбора для анализа Больших Данных

Фуад Алескеров, д.т.н., руководитель департамента математики факультета экономики, **НИУ ВШЭ**
Предлагаются модели выбора для задачи поиска в анализе Больших Данных. Эти модели включают процедуры, обобщающие модели, предложенные Г. Саймоном, и другие классические и новые модели выбора, такие как q-Паретовские правила, экстремизационные, экстремизационные с погрешностью, надпороговое, минимаксное правила и многие другие. Рассматриваются модели суперпозициии этих правил, и исследуется их сложность.
Показано, что предлагаемые модели работают эффективнее, чем многие известные правила, в частности метод опорных векторов.
Проведено сравнение эффективности различных процедур на данных компании Microsoft.

Методы и инфраструктуры интеграции разнородных Больших Данных

Алексей Вовченко, к.т.н., с.н.с., **Сергей Ступников**, к.т.н., с.н.с., **ИПИ РАН**
Работа выполнена при поддержке РФФИ (гранты 13—07—00579, 14—07—00548), Президиума РАН (программа фундаментальных исследований Президиума РАН № 16 ‹Фундаментальные проблемы системного программирования›), ИПИ РАН (тема 38.25 «*Спецификация и решение задач анализа данных в концептуальных терминах предметных областей с интенсивным использованием данных*» *государственного задания ИПИ РАН).*
Новая парадигма [1] в науке и информационных технологиях, доминирующая в последнее время, основывается на исследовании данных (data exploration). Роль данных особо подчеркивается и становится критической. Объемы данных фактически во всех областях деятельности растут со временем экспоненциально. Поэтому новая парадигма требует создания методов и средств оперирования данными, объемы которых выходят за рамки возможностей технологий баз данных, развивавшихся в последние десятилетия (преимущественно реляционных). Необходимы подходы, позволяющие справляться с разнообразием моделей данных (включая неструктурированные и слабоструктурированные данные), метаданных, семантики данных. Для поддержки новых моделей данных создаются системы управления данными, обладающие масштабируемостью, высокой доступностью, возможностью разбиения коллекций данных произвольным образом на разделы для параллельной обработки.

Данная работа относится к области конструирования средств поддержки систем с интенсивным использованием данных. Целью работы является разработка и реализация комбинированной виртуально-материализованной архитектуры среды интеграции неоднородных коллекций данных различного вида (структурированных, слабоструктурированных и неструктурированных). Такая среда должна поддерживать как виртуальную, так и материализованную интеграцию коллекций данных, представленных как в традиционных, так и нетрадиционных моделях данных.

Необходимость поддержки двух различных видов интеграции объясняется тем, что как виртуальный, так и материализованный подходы интеграции имеют свои достоинства и недостатки. *Виртуальная интеграция* осуществляется с использованием технологии предметных посредников [2], образующих промежуточный слой между пользователем (приложением) и неоднородными информационными ресурсами. При *материализованной интеграции* предполагается создание хранилища данных (warehouse), в которое загружаются коллекции данных, подлежащие интеграции. В процессе загрузки происходит преобразование данных из схемы коллекции

в общую схему хранилища.
Для материализованной интеграции ресурсов в комбинированной архитектуре необходима масштабируемая платформа манипулирования большими разноструктурированными данными (ПМБРД). В качестве такой платформы в данной работе выбрана связка системы Hadoop и системы организации реляционных хранилищ данных над Hadoop — WR-Hadoop (в качестве которой могут использоваться, например, системы Big SQL или Hive).

Платформа Apache Hadoop [3] была впервые представлена в 2005 году в составе проекта Apache Software Foundation и представляет собой набор программных средств распределенного хранения и обработки больших объемов данных. Платформы Hive [4] и Big SQL [5] представляют собой решения для организации реляционных хранилищ данных, разработанные на основе среды Hadoop. Фактически системы проецируют реляционную структуру на данные, хранящиеся в Hadoop, и предоставляют возможность исполнения SQL-подобных запросов на больших наборах данных путем компиляции их в программы MapReduce, исполняемые в среде Hadoop.
Таким образом, в комбинированной архитектуре обеспечивается возможность распределенного хранения, преобразования и интеграции больших разноструктурированных данных (при помощи Hadoop), а также унифицированный взгляд на материализованные данные через реляционную модель (при помощи Hive или Big SQL).

В комбинированной архитектуре ПМБРД рассматривается как еще один вид ресурсов, подлежащий виртуальной интеграции. Интеграция становится двухслойной: материализованная интеграция осуществляется внутри ПМБРД, виртуальная — на уровне предметных посредников.

Материализация в ПМБРД осуществляется путем помещения в Hadoop-кластер файлов, экспортированных из информационных ресурсов. Преобразование данных к реляционному виду для последующей интеграции производится при помощи программ на языке Jaql [6]. Jaql представляет собой язык запросов и сценариев, ориентированный на прозрачное применение модели программирования MapReduce: декларативно-императивные запросы Jaql переписываются в последовательность программ MapReduce, исполняемых в среде Hadoop. Сложные потоки обработки данных (очистки, устранения дублирования, слияния) и их интеграции реализуются с использованием комбинации языков Jaql и HIL. Декларативный язык HIL (High-level Integration Language) [7] разработан IBM для программирования сложных потоков обработки данных (ETL), агрегирующих факты из больших коллекций разноструктурированной информации в целевые коллекции унифицированных сущностей. Программы на HIL компилируются в Jaql, что позволяет использовать HIL для преобразования и интеграции данных в Hadoop.
Рассматриваемая в работе комбинированная среда интеграции нацелена, в частности, на исследование методов интеграции больших разноструктурированных данных. Этим и мотивирован выбор в качестве ПМБРД платформы Hadoop/WR-Hadoop и комбинации языков Jaql и HIL в качестве инструментов материализованной интеграции.

Литература

1. The Forth Paradigm: Data-Intensive Scientific Discovery. Eds. Tony Hey, Stewart Tansley, and Kristin Tolle. Redmond: Microsoft Research, 2009. — http://goo.gl/GqkDX1
2. Kalinichenko L.A., Briukhov D.O., Martynov D.O., Skvortsov N.A., Stupnikov S.A. Mediation Framework for Enterprise Information System Infrastructures. Proc. of the 9th International Conference on Enterprise Information Systems ICEIS 2007. — Funchal, 2007. — Volume Databases and Information Systems Integration. — P. 246-251.
3. Tom White. Hadoop: The Definitive Guide. O'Reilly Media; Third Edition edition. 2012.
4. Cynthia M. Saracco, Uttam Jain. What's the big deal about Big SQL? Introducing relational DBMS users to IBM's SQL technology for Hadoop. IBM DeveloperWorks, 2013. — http://www.ibm.com/developerworks/library/bd-bigsql/bd-bigsql-pdf.pdf
5. Edward Capriolo, Dean Wampler, Jason Rutherglen. Programming Hive Data Warehouse and Query Language for Hadoop. O'Reilly Media, 2012.
6. Kevin S. Beyer, Vuk Ercegovac, Rainer Gemulla, Andrey Balmin, Mohamed Eltabakh, Carl-Christian Kanne, Fatma Ozcan, Eugene J. Shekita. Jaql: A Scripting Language for Large Scale Semistructured Data Analysis. VLDB 2011.
7. Mauricio Hernández, Georgia Koutrika, Rajasekar Krishnamurthy, Lucian Popa, Ryan Wisnesky. HIL: a high-level scripting language for entity integration. Proceedings of the 16th International Conference on Extending Database Technology EDBT 2013. P. 549-560.

Предсказательная аналитика на основе потоков Больших Данных

Андрей Дмитриев, д.ф.-м.н., профессор, **НИУ ВШЭ**

Дается краткий обзор методов предсказательной аналитики на основе потоков больших массивов данных, генерируемых в режиме реального времени. Более подробно обсуждаются возможности и ограничения методов детектирования кризисов и предкризисных режимов кризисов в данных потоках. Предложена нелинейно-динамическая модель, генерирующая временные ряды рыночных цен спроса и предложения с одним фундаментальным управляющим параметром. Калибровкой данного параметра по историческим данным и выделением интервалов его постоянства при определенных условиях удалось детектировать предкризисные режимы. Данная модель апробирована на биржевых данных по ценам спроса и предложения на драгоценные металлы и стальные биллеты. Разработано приложение, позволяющее на основе данной динамической модели детектировать предкризисные режимы в потоках больших массивов данных, генерируемых в режиме реального времени.

Поиск похожих подпоследовательностей временных рядов на сопроцессорах Intel Xeon Phi

Михаил Цымблер, к.ф.-м.н., доцент, **Александр Мовчан**, **ЮУрГУ**

Временной ряд (time series) представляет собой совокупность вещественных значений, каждое из которых ассоциировано с последовательными отметками времени. Задача поиска похожих подпоследовательностей (subsequence matching) определяется следующим образом (см. рис. 1). Пусть дан временной ряд T, его подпоследовательности мы обозначаем как Tij, где $i<j$ — номера членов ряда; пусть задан запрос Q — временной ряд с длиной, не превышающей длину ряда T; имеется функция схожести $D(t1,t2)$, определяющая схожесть двух временных рядов. Необходимо найти подпоследовательности Tij, имеющие длину, равную длине запроса, для которых значение функции $D(Tij,Q)$ минимально.

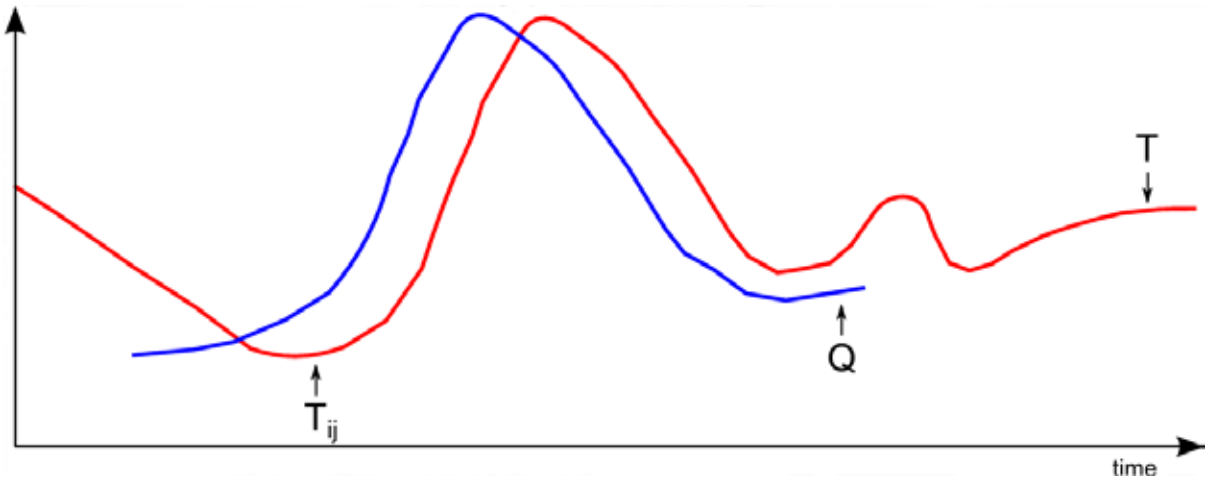


Рис. 1. *Поиск похожих подпоследовательностей*

Для определения схожести временных рядов можно использовать различные функции схожести. В настоящее время одной из наиболее популярных функций схожести временных рядов является *динамическая трансформация шкалы времени (Dynamic Time Warping, DTW)*, которая отличается от традиционной функции евклидова расстояния и вычислительно существенно более сложна. Преимуществом динамической трансформации шкалы времени является возможность сравнивать временные ряды, различающиеся скоростью изменения данных.

На сегодня алгоритм UCR-DTW, предложенный учеными Калифорнийского университета в Риверсайде, является, по-видимому, наиболее быстрым последовательным алгоритмом поиска похожих подпоследовательностей. Идея данного алгоритма заключается в применении каскада предварительных оценок, позволяющих отбросить непохожую подпоследовательность до выполнения вычислительно сложной динамической трансформации шкалы времени. Существуют реализации данного алгоритма для FPGA, а в нашем исследовании алгоритм UCR-DTW адаптируется для сопроцессора Intel Xeon Phi.

Intel Xeon Phi представляет собой сопроцессор, основанный на архитектуре Intel Many Integrated Core (Intel MIC). Intel Xeon Phi содержит 61 ядро, которые соединяются высокопроизводительной шиной. Каждое ядро сопроцессора имеет 4 потока за счет технологии Hyper-Threading и 512-разрядные векторные АЛУ,

обеспечивающие в одной инструкции до 16 операций над типом float или до 8 операций над типом double. В силу совместимости сопроцессора с архитектурой x86 при разработке приложения для Intel Xeon Phi имеется возможность использовать стандартные инструменты и технологии параллельного программирования, предназначенные для процессоров Intel Xeon.

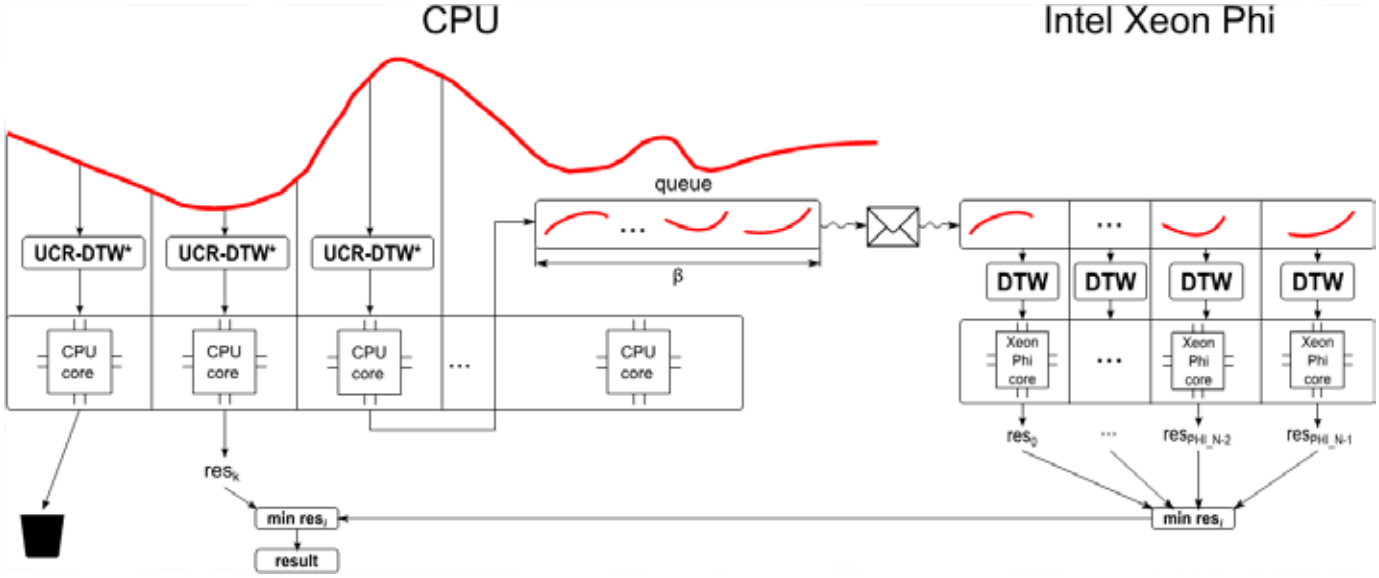


***Рис. 2**. Улучшенный параллельный алгоритм для сопроцессора*

Идея разработанного нами алгоритма (см. рис. 2) заключается в использовании на стороне процессора очереди подпоследовательностей, которые выгружаются на сопроцессор для вычисления динамической трансформации шкалы времени. Одна из нитей, выполняемых на ядрах процессора, объявляется мастером, остальные — рабочими. Мастер осуществляет выгрузку очереди на сопроцессор при ее заполнении. Рабочий вычисляет каскадные оценки и отбрасывает заведомо непохожую подпоследовательность либо добавляет эту подпоследовательность в очередь. Если очередь заполнена, то рабочий вычисляет динамическую трансформацию шкалы времени самостоятельно. По окончании выгрузки на процессор передается информация о найденных на сопроцессоре самых похожих подпоследовательностях. В итоге вычисляется самая похожая подпоследовательность среди найденных на процессоре и сопроцессоре.

Для исследования эффективности предложенного алгоритма нами проведена серия вычислительных экспериментов. В качестве аппаратной платформы экспериментов использовался вычислительный узел суперкомпьютера «Торнадо ЮУрГУ», характеристики которого приведены в табл. 1.

***Таблица 1.** Аппаратная платформа экспериментов*

Характеристики	Процессор	Сопроцессор
Модель	Intel Xeon X5680	Intel Xeon Phi SE10X
Количество ядер	6	61
Частота ядер,ГГц	3,33	1,1
Количество потоков на ядро	2	4
Производительность, TFLOPS	0,371	1,076

Эксперименты на синтетическом временном ряде, состоящем из 109 точек (см. рис. 3), показали преимущество улучшенной версии алгоритма.

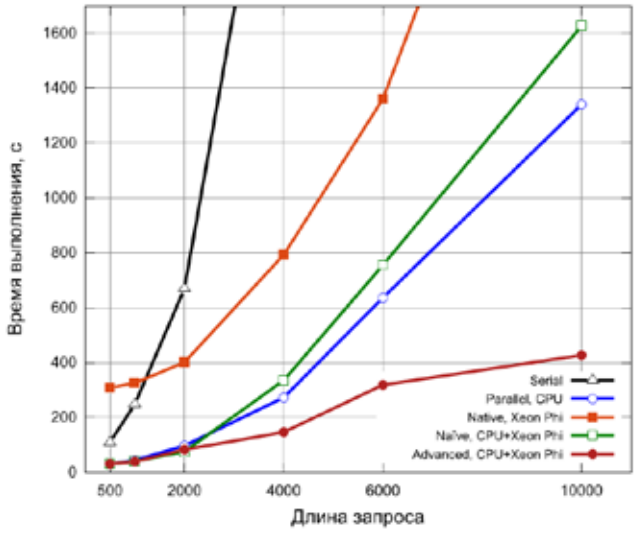


Рис. 3. *Производительность разработанного алгоритма на синтетических данных*

Эксперименты на реальных данных (см. рис. 4), в качестве которых использовались 2 · 108 точек данных ЭКГ (примерно 22 часа при частоте дискретизации 250 Гц), также показали преимущество улучшенной версии алгоритма.

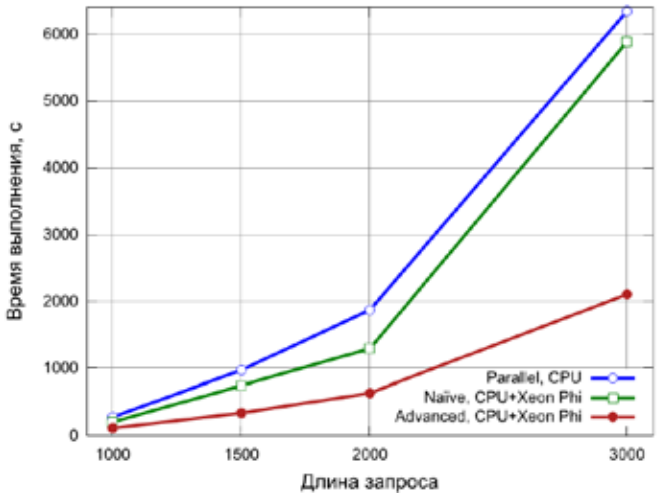


Рис. 4. *Производительность разработанного алгоритма на реальных данных*

Производительность разработанного алгоритма мы сравнили с аналогичными алгоритмами для GPU (NVIDIA Tesla C1060, 77,76 GFLOPS) и FPGA (Xilinx Virtex 5 LX—330, 65 GFLOPS) для запроса длиной 1024 точек данных, результаты показаны на рис. 5.

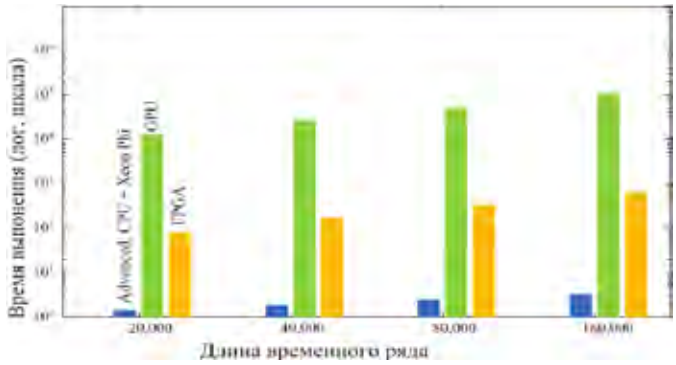


Рис. 5. *Сравнение разработанного алгоритма с алгоритмами для GPU и FPGA*

Источники Больших Данных и современные способы хранения данных

Михаил Комаров, к.т.н., доцент, **НИУ ВШЭ, Евгений Кучерявый,** PhD, руководитель лаборатории **ET4NBIC Lab, Технологический университет Тампере (Финляндия)**

Введение

Большие Данные на сегодняшний день являются достаточно известным трендом в области информационных технологий. Многие говорят о работе с Большими Данными, но немногие приводят конкретные примеры работы с ними, которые полностью соответствовали бы определению самих Больших Данных. Несмотря на это, нельзя оставить без внимания, все больший рост и развитие технологий работы с данными. Существует несколько важных направлений, которые активно развиваются в области информационных технологий и напрямую относятся к Большим Данным, но основным нужно выделить с точки зрения увеличения объемов данных — мобильные технологии.

Мобильные технологии, Интернет вещей и Большие Данные

Направления исследования в области носимых вычислительных систем не являются новыми, однако значительное продвижение данное направление получило с появлением смартфонов и развитием мобильных технологий. Если некоторое время назад ученые говорили о том, что технология Smart Dust для получения в режиме реального времени данных о том или ином объекте или территории будет внедряться десятки лет, то с появлением смартфонов и технологическим прогрессом в области мобильных технологий фактически данный подход уже внедряется. Каждый человек добровольно пользуется датчиком в виде мобильного устройства, которое снимает определенные данные и может их передавать для дальнейшей обработки.

Наличие мобильных устройств и развитие технологий передачи данных привело к тому, что на сегодняшний день много исследований ведется по теме Интернета вещей. В настоящее время количество устройств, подключенных к Интернету, постоянно растет, и уже в 2009 году их число превысило население планеты, то есть на каждого человека стало приходиться почти 2 подключенных устройства. Эти устройства являются не только смартфонами и ноутбуками, но и многими другими не столь традиционными предметами, подключенными к Интернету, — чайниками, холодильниками, автомобилями и другими. Если обратиться к истории, то еще в 1926 году Никола Тесла в интервью для журнала «Collier's» сказал, что в будущем радио будет преобразовано в «большой мозг», все вещи станут частью единого целого, а инструменты, благодаря которым это станет возможным, будут легко помещаться в кармане. А в 1990 году выпускник MIT Джон Ромки создал первую в мире интернет-вещь. Это был его тостер, подключенный к сети.

Нужно отметить, что Интернет вещей (то есть устройства, которые подключены к сети) является источником одного типа данных — более ориентированных на устройства. Однако есть еще и мобильные технологии, в рамках которых пользователи в повседневной деятельности используют мобильные приложения, а мобильные приложения являются источниками данных, более ориентированных на пользователя.

В настоящее время Интернет вещей состоит из разрозненных и практически не связанных между собой сетей, каждая из которых создана для решения своих конкретных задач и собирает, накапливает и анализирует данные с определенных устройств. Для объединения в единый работающий механизм необходима стандартизация методов работы этих устройств и передачи информации. Внедряя Интернет вещей в повседневную жизнь, мы сталкиваемся с его трехчастным построением (согласно Робу ван Краненбургу): Интернет вещей — это непрерывный поток данных, который начинается от нашего тела BAN (Body Area Network), домашней и рабочей обстановки LAN (Local Area Network), городской инфраструктуры WAN (Wide Area Network) и растворяется в глобальной информационной системе VWAN (Very Wide Area Network). Монетизация такой глобальной системы происходит в результате того, что конечные пользователи оплачивают предоставление доступа к данным, собранным в результате работы этого непрерывного потока, либо к услугам, которые предлагаются на основе полученных данных и их проведенного анализа.

Благодаря современному стандарту взаимодействия 6LoWPAN, позволяющему подключаться к Интернету маломощным устройствам, в настоящее время установить микрокомпьютер в любой предмет, начиная от браслета или зубной щетки, не представляет особой сложности. Но на текущий момент целесообразно говорить о множестве различных несвязанных сетей, которые решают отдельные задачи и соединяют отдельные устройства. Один из немногих путей решения проблемы разрозненности — это серьезное экономическое стимулирование взаимодействия производителей мобильных устройств либо провайдеров услуг, которые могут оказывать новые услуги в рамках объединенной сети.

Независимо от того, что до сих пор мы имеем раздробленные сети, на их основе уже выстроены различные бизнес-модели, по которым работают компании. Они занимаются мониторингом потребления ресурсов, экологической обстановки, слежением за здоровьем человека и животных, отслеживанием движения каждого

конкретного потребительского товара для оптимизации поставок и др. Интернет вещей может быть использован в том числе и страховыми компаниями для отслеживания поведения (перемещение, скорость автомобиля) клиентов и предоставления персонифицированных тарифов с учетом их наклонностей. И если говорить о перспективах, то одним из основных трендов во многих сферах выделена повсеместная персонификация, что позволит (совместно с развитием и распространением 3D-печати) изготавливать необходимые устройства прямо у себя дома и непосредственно под свои индивидуальные особенности. Данные об этих особенностях будут собираться с многочисленных датчиков, в том числе и мобильных телефонов. Кроме того, согласно Прогнозу научно-технологического развития Российской Федерации на период до 2030 года (Министерство образования и науки РФ), человек сможет отслеживать, какая деятельность ему полезна, а какая вредит. Таким образом, это увеличит как качество жизни, так и ее продолжительность. Конкуренция между различными компаниями будет усиливаться вследствие уравнивания их возможностей, поэтому основной упор будет делаться на эффективное использование открывающихся возможностей и, в результате получения больших объемов данных от подключенных устройств, на обнаружение новых возможностей для бизнеса. Получается, что сама информация, собираемая каким-либо продуктом по мере использования, становится активом наравне с физическими активами или трудом, пользование которым необходимо оплачивать. Это приводит к разрушению привычных бизнес-моделей, основанных на продаже продукта, и появлению новых, в которых монетизация идет через предоставление дополнительных услуг для приобретенного продукта и непосредственно для потребителя.

Необходимо также рассмотреть приоритетные проекты некоторых корпораций – лидеров в данной области: Cisco, HP и IBM. Проект Planetary Skin от Cisco и NASA (www.planetaryskin.org/) предполагает объединение спутниковой сети, беспилотных самолетов, а также наземных средств и датчиков для сбора информации о Земле, ее процессах и явлениях последующего контроля земных ресурсов и предоставления этой информации людям для повышения уровня жизни. Это и подтверждается миссией, приведенной на официальном сайте проекта. Среди основных направлений Planetary Skin — прогнозирование и оптимизация потребления энергии, прогнозирование стихийных бедствий, исследование и принятие решений в сфере управления водными ресурсами, поддержка сельского хозяйства и анализ рисков с использованием сенсорных сетей, программа мониторинга состояния лесов, а также мониторинг взаимосвязей потребления различных ресурсов.

Central Nervous System for Earth от HP заключается в повсеместном внедрении сенсоров, считывающих такие показатели, как давление, температура, освещенность, вибрация, влажность и некоторые другие. Также будут использоваться другие датчики, похожие на популярные RFID-метки, однако являющиеся еще и мощными акселерометрами. Сферы применения так же обширны, как и у Planetary Skin от Cisco. Датчики могут устанавливаться на мосты и строения, вдоль дорог для мониторинга загрузки. В дальнейшем возможны вхождение датчиков в бытовую электронику и в конечном счете, переход к «Интернету вещей» от изначальной сети, отвечающей за мониторинг состояния природы и инфраструктурных сооружений.

Большой интерес представляет проект IBM — Smart Planet (www.ibm.com/smarterplanet/us/en/). На официальном сайте отмечается, что многие компании собирают данных куда больше, чем могут позволить себе обработать. Однако на «Умной планете» наиболее крупные организации смогут обработать и превратить эти данные в ценную информацию о клиентах, бизнесе и окружающем мире в целом, и таким образом откроются новые возможности для оптимизации принимаемых решений. Это также всеобъемлющая сеть датчиков, осуществляющих мониторинг важнейших показателей окружающей среды. Наибольший интерес в данном случае представляет документ, затрагивающий сферу электронных устройств на «Умной планете». В нем говорится об открывающихся возможностях мониторинга жизненного цикла продукта, начиная от производства, и заканчивая тем, как его использует владелец. На основе полученных данных можно осуществлять поиск идей для новых услуг и сервисов. Также говорится о смене ориентации деятельности некоторых компаний, когда они получают основную прибыль, продавая не продукт, а услугу, с ним связанную. Одним из интересных направлений IBM является процесс превращения обычных городов в «умные». Это будет происходить через создание товаров и услуг для городских управлений. Такое направление получило название IBM Smarter Cities. Большие данные и Интернет вещей, являющийся одним из источников Больших Данных, также имеют одну общую проблему — достоверность получаемых данных и защиту от несанкционированного доступа к данным [1]. Таким образом, Большие Данные — это не только правило четырех V: Volume, Variety, Velocity и Value (объем, вариативность, скорость и ценность), но еще и Verification (подтвержденность, достоверность). В статье [1] достаточно подробно указаны угрозы нарушения прав человека, которые могут возникать в процессе обработки данных, более ориентированных на пользователя. Также приведены мнения специалистов из различных областей – бизнеса, науки, политики. Отмечается, что необходимы новые методы обработки Больших Данных, которые позволили бы исключить возможность идентификации пользователя по имеющемуся набору

данных, а также необходимо межгосударственное взаимодействие с точки зрения формирования базового законодательства из-за появления больших объемов данных о гражданах разных стран, которые хранятся в компаниях. Данный вопрос также затрагивает понятие транснациональных компаний и впоследствии, как это было уже упомянуто, платежи компаний пользователям за использование их данных в деятельности компаний в рамках формирования их информационных активов.

Заключение

В докладе основное внимание уделено Интернету вещей из-за стремительного технологического развития, поскольку на сегодняшний день речь идет уже о создании и внедрении сетей на наноуровне. Это означает, что потоки генерируемых данных, передаваемых от устройств наноуровня, будут в десятки и сотни раз больше, чем потоки данных, которые генерируются в рамках сетей Интернета вещей. В статьях [2, 3] отмечается, что устройства подобных размеров нуждаются в больших объемах встроенной памяти, для хранения и обработки базовых данных (на текущий момент известно, что на 10 μм можно получить лишь 7 Кбайт памяти для хранения данных). В таком случае пока речь идет лишь о базовом взаимодействии устройств на основе принципов молекулярного взаимодействия. Кроме разработки новых протоколов обмена информацией, ключевыми вопросами являются вопросы разработки новых методов обработки данных в режиме реального времени для минимизации требуемых объемов памяти совместно с новыми подходами организации хранения данных на наноуровне, что само по себе уже является революционным в условиях развития концепции Больших Данных.

Литература

1. P. Schaar, The Internet and Big Data — Incompatible with Data Protection? Mind — Multistakeholder Internet Dialog Vol. 7: Privacy and Internet Governance. Berlin : Internet & Society Collaboratory, P. 14–20, 2014.

2. Akyildiz, I. F., Jornet, J. M., and Pierobon, M. Nanonetworks: A New Frontier in Communications, Communications of the ACM, vol. 54, no. 11, pp. 84–89, November 2011.

3. Llatser I., Cabellos-Aparicio A. and E. Alarcon, Networking Challenges and Principles in Diffusion-based Molecular Communication, IEEE Wireless Communications, vol. 19, no. 5, pp. 36–41, October 2012.

Секция. Большие Данные в научных исследованиях

Научные вызовы Больших Данных

Евгений Павловский, к.ф.-м.н., старший преподаватель, **НГУ**, директор, **«Исследовательские системы»**
Аннотация. В докладе обсуждаются научные проблемы Больших Данных, приводятся новые постановки задач для научных исследований.

Проблема качества исходных данных
Проблема несоответствия цели сбора данных и цели их использования. Отсутствие должного внимания к качеству собираемых данных. Множество источников данных с неизвестной степенью истинности.

Проблема структурированности
Традиционно делят информацию на структурированную, полуструктурированную и неструктурированную. Однако следует понимать, что структурированность — это свойство воспринимающего информацию, так как понятие «информация» уже подразумевает приемник и передатчик. Если информацию воспринимает человек, скажем автор романа, то для него его произведение является хорошо структурированным текстом — есть главы, разделы, основные мысли и т. п. Однако для машины мы скажем, что такая информация (а точнее — такие данные) является неструктурированной. Это происходит потому, что эти данные машину еще не научили обрабатывать, в них неизвестная структура. Хотя если позвать специалиста-лингвиста, он с легкостью наведет грамматическую структуру на представленный текст. Итак, понятие структурированности — относительное. Если машина умеет разбирать представленные данные, то можно считать, что в них есть определенная структура. Можно сформулировать даже более сильное утверждение: структурированность — это мера приближенности к ожидаемой информативности данных. Если мы ожидаем от видеопотока возможностей распознавания лиц, и даже более — определения в них подозрительных лиц, но пока не имеем возможностей к этому, то мы говорим, что видеопоток является неструктурированными данными. Но как только машина научится самостоятельно выделять лица, присваивать им метки, классифицировать их на подозрительные и неподозрительные — мы тут же скажем, что видеопоток обогащен структурированной информацией, потому что мы сами задали определенный фрейм, в котором информация о видеопотоке для нас стала осмысленной. Если же мы вдруг увидим, что в видеопотоке на 1 час встречается 1 миллион человек, на каждого из них поставлены метки с определенной степенью уверенности, но нет никакой возможности сгруппировать эти метки по признакам, то мы вновь скажем, что информация о метках не является структурированной. Такова относительность структурированности.

Проблема смешения понятий «данные», «информация» и «знания»
Предыдущий тезис только подчеркивает необходимость наведения порядка в области понятийного аппарата. Мы немного пленились наличием цифровых данных, их растущими размерами и называем по привычке эти данные информацией, так как они представлены в виде массива нулей и единиц. Но на самом деле, как видно из проблемы структурированности, эти объемы для нас не всегда являются информацией. Они, скорее, — данные, то есть некоторая груда непонятных объектов, которые заведены в системы, способные их хранить и перерабатывать. Однако, если мы не знаем, что это за данные, не имеем некоторой метаинформации о них, то для нас это конечно же не информация. Если информацию понимать, как нечто, уменьшающее неопределенность, то можно навести порядок в терминах.

Вторая путаница происходит между терминами «знание» и «информация». При бытовом употреблении мы говорим, прочитав, скажем, электронное письмо, что узнали, когда состоится конференция. Однако, в строгих терминах, мы-скорее-получили информацию. Знания же-более глубокая сущность. Знание в контексте обработки Больших Данных — это-скорее-то, что позволяет предсказывать информацию в новых данных.

Отдельного рассмотрения заслуживают формализованные знания, представленные в виде данных.

Извлечение информации и знаний из текстов
При исследовании семантики текстов автоматизированными методами мы выполняем неестественные исследования, решая проблемы, навязанные представлением этих текстов в виде предложений, состоящих из слов, состоящих из букв. Между тем, чтобы понимать смысл текста, человек проходит долгий путь обучения. Сначала, в детстве, у него формируется сенсорный образ каждого слова, затем он идентифицирует его с фразами в рамках некоторой ситуации (контекста) и только потом переходит к словам, слогам и буквам. После этого синтезирует знание о буквах и слогах и снова как бы обучается письму и чтению. При работе

автоматизированными методами с текстами мы словно бы забываем про часть, которая до первого класса школы, а начинаем учить машину буквам, слогам, словам, морфологии, синтаксису, грамматике и, наконец, семантике. Причем, как выясняется, проблема обучения семантике самая сложная. Кроме того, представляя информацию для машины в виде битов, мы сразу навязываем себе следующие проблемы:

1) преобразование из двоичной информации в смысловую;
2) множественность смыслов;
3) наличие смысла в каком-либо отношении.

Осознание этого факта позволяет взглянуть на проблему семантики текстов по-новому.

Новые типы данных

Для того чтобы решать проблемы, возникающие в обработке Больших Данных, уместно провести ревизию существующих типов данных и предложить новые формы. Мы все привыкли представлять данные в двоичной форме, так как это удобно компьютеру. Прошел век кодирования всего в двоичную систему, и теперь мы имеем громоздкие массивы данных бинарного формата. Как новый примитивный тип данных можно рассмотреть понятие смысла, или контекста, то есть некоторой конкретной ситуации, в рамках которой множество терминов имеют однозначный смысл. Например, математической основой для таких типов данных также могут служить алгебраические системы, как и в случае с номинальными, абсолютными, порядковыми шкалами. Однако здесь можно добавить и новые возможности. Эти алгебраические системы могут быть нечеткими. Нечеткие, или булевозначные, модели больше подходят для описания ситуаций и текстов естественного языка [1].
Вторым перспективным направлением в области новых типов данных являются графы, позволяющие соединять методы линейной алгебры и теории графов.

Новые измерительные шкалы

В основе работы всех методов машинного обучения лежит понятие меры сходства. Эта мера, как правило, отображает пару объектов в пространство действительных неотрицательных чисел. Например, мера Хэмминга или метрика Евклида. Однако при сравнении объектов качественно мы часто хотим знать, насколько похож один объект на другой. Это невозможно точно определить, не имея некоторый третий объект, относительно которого производится сравнение. То есть можно ввести относительную меру сходства (функция конкурентного сходства, или сокращенно FRiS-функция) $F(x,y,z)$ $F(x,y,z)$, которая будет сообщать о похожести одного объекта x x на другой y y относительно некоторого третьего zz. Такая мера была выдвинута в [2], исследованы ее возможности в области машинного обучения. Результаты позволяют надеяться на прорыв в решении задачи выбора информативных признаков [3].
Отдельной задачей является переосмысление методов машинного обучения относительно уже не бинарных мер сходства, а относительно тернарных мер. Исследование измерительных шкал для бинарных мер проводилось полвека назад в [4].

Новые методы выборки данных

Для решения проблемы Больших Данных можно использовать как априорные методы (до исследования содержания данных), такие как случайная выборка (уменьшение количества объектов), так и апостериорные (после исследования данных), такие как методы выбора информативных признаков (уменьшение признакового пространства). Каждых из этих подходов имеет свои достоинства и недостатки [5]. Существует ли промежуточное решение, использующее достоинства обоих подходов?

Индуктивно-дедуктивные системы

Представляется, что соединение методов индуктивного формирования знаний (машинное обучение) и дедуктивный вывод (логический вывод) могут дать принципиально новые результаты в обработке Больших Данных, характеризующихся слабой структурированностью. Примером и прообразом может послужить семантический вероятностный вывод [6, с. 60–56].

Субквадратичные алгоритмы

Большие Данные требуют алгоритмов сложности не более $O(NlogN)$ $O(NlogN)$. Для задачи кластеризации решен вопрос о создании субквадратичного и даже линейного алгоритма с хорошим качеством [7].

Литература

1. Пальчунов Д. Е., Яхъяева Г. Э. Нечеткие алгебраические системы //Вестник Новосибирского государственного университета. Серия: математика, механика, информатика. — 2010. — Т. 10. — № 3. — С. 76–93.
2. A quantitative measure of compactness and similarity in a competitive space // N. G. Zagoruiko, I. A. Borisova, V. V.

Dyubanov and O. A. Kutnenko — Journal of Applied and Industrial Mathematics, 2011, Vol. 5, № 1, pp.144–154.
3. Использование алгоритма FRiS-GRAD для анализа активности генов при решении 9 медицинских задач // Загоруйко Н.Г., Борисова И.А., Дюбанов В.В., Кутненко О.А. — IV Международная конференция «Математическая биология и биоинформатика», Москва, 2012, с. 84–85.
4. Суппес П., Зинес Дж. Основы теории измерений // Психологические измерения. М.: Мир, 1967. С. 117–132.
5. National Research Council. 2013. Frontiers in Massive Data Analysis. Washington, D.C.: The National Academies Press.
6. Витяев Е. Е. Принципы работы мозга, содержащиеся в теории функциональных систем ПК Анохина и теории эмоций ПВ Симонова // Нейроинформатика (электронный рецензируемый журнал). — 2008. — Т. 3. — № 1. — С. 25–78.
7. Модификация алгоритма кластеризации FRiS-Tax для работы с большими данными / Зырянов А.О. // Академический форум корпорации EMC: сборник тезисов докладов участников академической секции. 23–28 сентября 2013 г., Ялта, АРК, Украина. — Симферополь: изд-во «Ариал». — 2013. — С.17–18.

Большие Данные и вычислительная наука: место и время встречи

Виктор Топорков, д.т.н., зав. кафедрой, **НИУ МЭИ**

Введение

Вопросы, связанные с решением больших задач, для решения которых не хватает вычислительных ресурсов даже суперкомпьютерного класса, привлекают к себе внимание уже не один десяток лет. Однако попытки использования масштабируемых сред заставляют многие проблемы переосмыслить и взглянуть на них по-новому.

Рис. 1

Масштабируемость подразумевает возможность наращивания количества процессоров, емкости памяти и независимость пропускной способности коммуникационной системы от числа процессорных узлов, участвующих в вычислениях. Распределенные среды, как известно, характеризуются потенциальной ненадежностью. Здесь хотелось бы привести высказывания двух известных специалистов об особенностях распределенных систем. Одно из них принадлежит Эндрю Таненбауму, другое – Лесли Лэмпорту (рис. 1).

Оба этих свойства характеризуют потенциальную ненадежность распределенных сред не с точки зрения возможного отказа или сбоя компьютеров. Здесь акцент делается на том, что у компьютеров нет никаких обязательств перед самой средой и в любой момент времени они могут отключиться от сети вне зависимости от того, завершена ли обработка выделенного задания. Тем не менее существуют, как известно, так называемые большие задачи, для решения которых не хватает вычислительных ресурсов даже суперкомпьютерного класса. Возникает вопрос: «как же в этом случае обеспечить согласование структуры задания с динамично изменяющимся составом, в общем случае, неоднородных ресурсов для эффективной организации вычислений?» Иными словами, как обеспечить масштабируемость вычислительной среды для выполнения задания при соответствующих значениях показателей качества обслуживания.

В таких средах поддержка образа общего ресурса сочетается с высокой степенью автономности процессорных узлов, не говоря уже о том, что в гридах в отдельных доменах могут применяться различные локальные планировщики и отсутствует единая политика администрирования вычислительных ресурсов. Все это заставляет задуматься над новыми принципами организации вычислений и распределения ресурсов, реализуемостью моделей распределенных программ и эффективным планированием процессов обработки.

Это и определяет круг вопросов, которые выносятся на обсуждение (рис. 2).
В работе основное внимание уделяется новым результатам, полученным в двух направлениях. Одно из них – построение моделей распределенной обработки, наиболее адекватных особенностям масштабируемых сред. Другое — выбор конфигурации ресурсов и планирование вычислений.

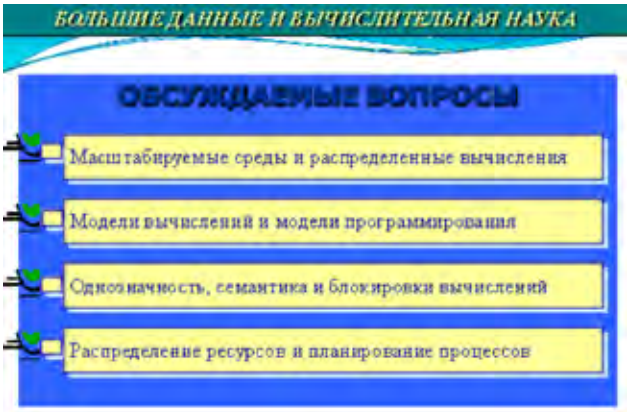


Рис. 2

Модели вычислений и парадигмы программирования
Модель вычислений служит связующим звеном между архитектурой среды и моделью программирования и в распределенных средах должна отражать взаимодействие процессов (рис. 3).

Это взаимодействие представляется явно в некоторых моделях программирования и может быть организовано с помощью стандартных коммуникационных библиотек, например MPI. Другие, более высокоуровневые модели освобождают программиста от выраженного управления параллельными процессами, примитивами передачи сообщений или синхронизации. Так, стандарт OpenMP может рассматриваться как модель программирования систем с разделяемой памятью и надстройка над библиотеками нитей.

Рис. 3

Модель вычислений должна помогать в поисках ответа на вопрос: «насколько эффективно реализуется программа на данной архитектуре?» Эффективность же реализации может интерпретироваться по-разному: как сбалансированность загрузки процессоров, согласованность степени распределенности вычислений и трафика обмена данными, длительность простоя процессоров из-за отсутствия данных и т. д.

Несмотря на большое разнообразие моделей выполнения программ, в масштабируемых средах чаще всего используются две из них — модель обмена сообщениями и модель общей памяти. Особенности моделей не предполагают их реализаций на архитектурах соответствующего типа. Так, первая может быть надстроена над любой архитектурой. Вторая чаще всего применяется в архитектурах с общей памятью, где программа рассматривается как система нитей, взаимодействующих через общие переменные и примитивы синхронизации. При этом программист имеет дело с привычным для него единым адресным пространством. Основная проблема – обеспечение согласованного состояния памяти.

Модель обмена сообщениями позволяет избавиться от проблем разделения памяти и может быть надстроена над архитектурами с общей памятью. Реализация же модели разделяемой памяти над архитектурами с передачей

сообщений значительно сложнее. Недетерминированный обмен неоднородными сообщениями адекватно представляет многие черты функционирования распределенных масштабируемых сред, такие как массовый параллелизм, взаимодействия типа «точка-точка», коммутация пакетов (рис. 4).

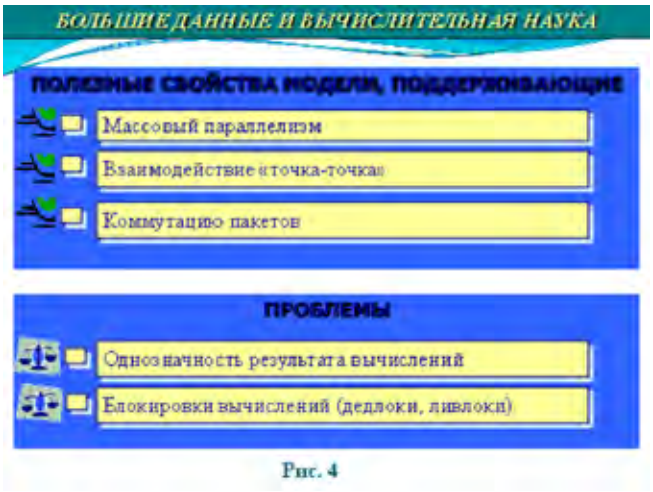


Рис. 4

Однако подобное сочетание свойств требует исследования реализуемости моделей (проблем однозначности результата и блокировки вычислений) и алгоритмической разрешимости задач их анализа.

Известно, что чем менее строгая модель, в смысле точного порядка выполнения инструкций, принимается программистом, тем эффективнее и быстрее работают его программы в распределенных системах. Например, разновидности модели общей памяти с ослабленными требованиями к согласованности данных являются более производительными, чем модель последовательной непротиворечивости. Но это одна сторона медали. Другая ее сторона — недетерминизм процессов вычислений: возможны различные истории выполнения программы на одних и тех же исходных данных. Естественным ограничением в этих условиях является требование однозначности результата.

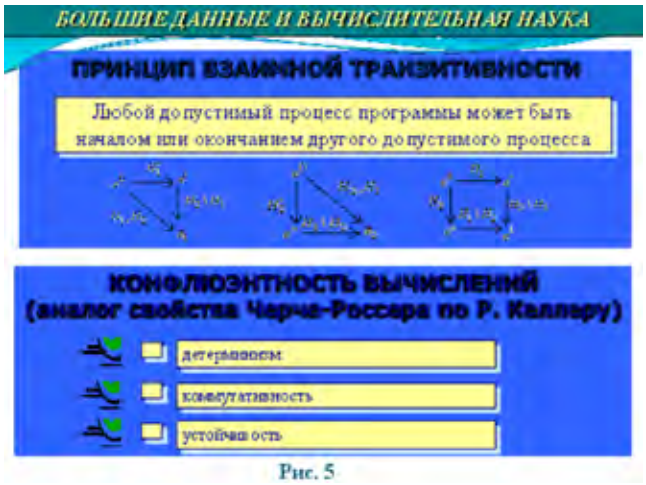


Рис. 5

Один из путей решения этой проблемы — реализация *принципа взаимной транзитивности:* любой допустимый процесс программы может быть началом или окончанием другого допустимого процесса (рис. 5). При этом оказывается, что в случае детерминированных вычислительных процессов конфлюэнтность, или аналог свойства Черча — Россера, в смысле, определенном Келлером, является частным проявлением свойства взаимной транзитивности.

Серьезнейшая проблема распределенных вычислений — взаимные блокировки процессов, разделяющих общий ресурс. При этом не все исчерпывается тупиками. В сложно организованных программах блокировки могут быть обусловлены и недетерминированным характером протекания процессов, и структурой задания, и структурой данных, которыми обмениваются процессы. Вопросы обнаружения и предотвращения таких ситуаций требуют специальных методов анализа свойств распределенных программ – в частности, использования потоковых моделей, представимых маркированными графами.

Алгоритмически разрешимый подкласс таких моделей представляют М-сети [1]. Проблема реализуемости распределенных вычислений сводится к достижимости стационарной разметки М-сетей. М-сети адекватно представляют связь проблем недетерминизма и блокировок в наиболее часто используемой буферной модели

обмена сообщениями. Этот аппарат строго формализован и обоснован. Он допускает исследование свойств как параллельных, так и распределенных программ. Эквивалентные структурные преобразования М-сетей позволяют избегать блокировок распределенных вычислений (рис. 6).

Рис. 6

Особенности планирования в распределенных средах

Давно и хорошо осознано, что организация вычислений в масштабируемых средах требует особых механизмов управления ресурсами и планирования процессов. Здесь уже не «работают» классические расписания: прикладную программу и аппаратную платформу можно рассматривать как программно-аппаратную среду, которая соответствующим образом настраивается, конфигурируется для ускорения вычислений, балансировки загрузки процессоров и т. д.

На практике типична следующая ситуация: имеют место ресурсный запрос пользователя и фактически складывающаяся динамика загрузки и доступности вычислительных ресурсов, которую необходимо учитывать для эффективной реализации программ.

Эта идея, кстати, хорошо согласуется и с парадигмой грид-технологий, согласно которой пользователь должен иметь полную иллюзию использования «чужих» ресурсов исключительно в интересах своего программного приложения.

Рис. 7

Один из возможных путей практической реализации этого подхода представлен проектами в духе AppLeS (Application Level Scheduling): агент-планировщик формирует и координирует выполнение расписания для повышения эффективности работы приложения с точки зрения конечного пользователя (рис. 7).

Для «настройки» ресурсов используются, как правило, эвристические приемы, дающие весьма приближенное решение. Подобный же подход реализуется и в системе X-Com (НИВЦ МГУ) [2]. Он рассчитан на использование доступных неоднородных распределенных ресурсов, когда не предполагается наличия какого-либо регламента в их предоставлении. Концептуально отличный подход, учитывающий коллективный характер функционирования гридов, ориентирован на образование виртуальных организаций. Здесь глобальная очередь заданий поддерживается метапланировщиком (рис. 8).

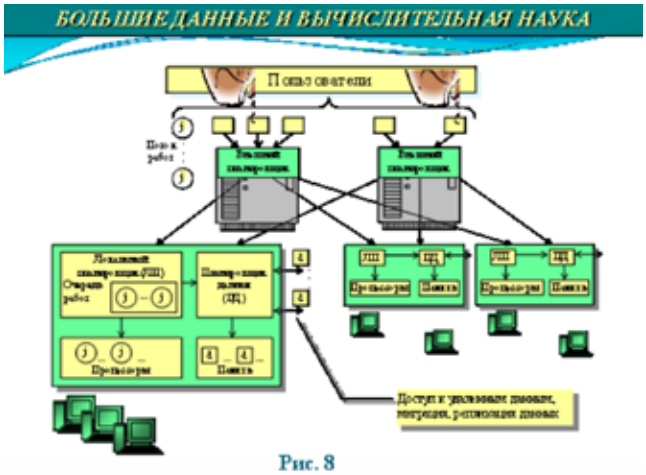


Рис. 8

Наконец, принципиально иной путь — совместное планирование процессов и согласованное выделение ресурсов, базирующиеся на так называемой метафоре масштабирования: программа и ресурс рассматриваются как единая настраиваемая среда вычислений.

Эта метафора основывается на специально разработанном математическом аппарате — методе критических работ [3]. Приведенный на рис. 9 пример иллюстрирует процедуру минимизации стоимости обработки при заданном ограничении на время выполнения задания. Стоимость определяется как функция от объемов и времени вычислений.

Рис. 9

Метафора масштабирования предполагает развитие с целью порождения стратегии (множества сценариев) вычислений на основе совокупности критериев, включающей, например, стоимость и загрузку процессоров, с учетом множества факторов [4].

Конкретный план выполнения задания может выбираться из стратегии в зависимости от временных параметров контрольных событий, наступление которых обусловлено динамикой загрузки процессорных узлов и коммуникаций, пространственно-временной конкуренцией различных программ и т. п. (рис. 10).

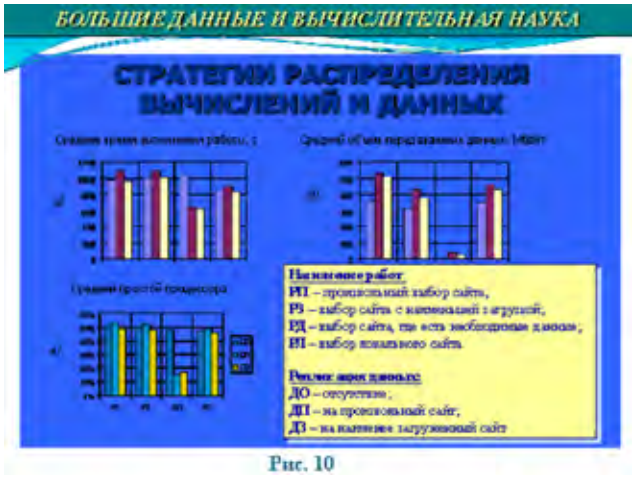


Рис. 10

Таким образом, несмотря на априорный характер формирования стратегии [3], управление ресурсами может осуществляться динамически [5].

Заключение

В одной статье вряд ли возможно охватить все проблемы, связанные с моделями вычислений и распределения ресурсов в масштабируемых средах. Также ясно, что нельзя предложить некой универсальной модели распределенной обработки. Вне рамок обсуждения остались вопросы надежности и защиты, технологии разработки распределенных приложений, файловых систем, систем распределенных документов и т. д.

Здесь сделана попытка обозначить те области в распределенных вычислениях, которые, на наш взгляд, требуют внимания в настоящее время и будут влиять на эффективность использования масштабируемых сред в будущем.

Литература

1. Топорков В.В. Проблема разрешимости задачи анализа потоковых моделей программ // Программирование. 2003.– № 3.–С. 3–14.

2. Воеводин Вл.В., Жолудев Ю.А., Соболев С.И., Стефанов К.С. Эволюция системы метакомпьютинга X-Com // Вестник Нижегородского университета им. Н.И. Лобачевского. 2009.– № 4. С.–157–164.

3. Топорков В.В. Опорные планы согласованного выделения ресурсов при организации распределенных вычислений на масштабируемых системах // Программирование. –2008. № 3.–С. 50–64.

4. Toporkov V., Tselishchev A., Yemelyanov D., Potekhin P. Metascheduling Strategies in Distributed Computing with Non-dedicated Resources // W. Zamojski and J. Sugier (eds.), Dependability Problems of Complex Information Systems, Advances in Intelligent Systems and Computing (AISC), vol. 307, pp. 129–148: Springer International Publishing Switzerland (2014)

5. Toporkov V., Toporkova A., Tselishchev A., Yemelyanov D. Slot Selection Algorithms in Distributed Computing // Journal of Supercomputing (2014), Vol. 69, No. 1, pp.53–60.

Распределенные вычисления и Большие Данные в ядерных исследованиях (на основе опыта России и ОИЯИ)

Владимир Кореньков, д.т.н., директор лаборатории информационных технологий, **ОИЯИ,** зав.кафедрой, **Международный университет «Дубна»,**

Алексей Климентов, к.ф.-м.н., **Брукхейвенская национальная лаборатория (США)**

В докладе представлены концепция и эволюция глобальной компьютерной инфраструктуры для хранения, обработки и анализа данных экспериментов на Большом адронном коллайдере в ЦЕРНе. Дается краткая информация об участии России в этом процессе. Представлен обзор проектов в области распределенных вычислений и Больших Данных, выполненных Лабораторией информационных технологий (ЛИТ ОИЯИ) в России, ЦЕРНе, США, Китае, странах — участницах ОИЯИ.

Особое внимание уделено созданию центра уровня Tier1 в России для хранения и обработки данных экспериментов на Большом адронном коллайдере, развитию облачной и гибридной инфраструктуры, а также модели компьютинга мегапроекта NICA в ОИЯИ. Представлены результаты и планы развития платформы для управления Большими Данными.

В настоящее время в мире информационных технологий интенсивно развиваются распределенные и параллельные вычисления, которые представлены широким спектром архитектурных решений. Активно развивается направление big data (Большие Данные), которое концентрирует усилия в организации хранения, обработки, анализа огромных массивов данных.

Грид-технологии играют важнейшую роль в развитии компьютинга для обработки данных экспериментов на Большом адронном коллайдере (LHC), для чего создана уникальная распределенная компьютерная инфраструктура WLCG (Worldwide LHC Computing Grid) [1]. Эта инфраструктура объединяет больше 200 компьютерных центров 60 стран мира, в которых производятся хранение, обработка, анализ и моделирование данных для получения знаний физиками крупнейших международных коллабораций. Глобальная грид инфраструктура используется не только для решения задач обработки данных экспериментов на Большом адронном коллайдере, но и для других задач в области биологии, медицины, наук о земле, высокотехнологичной промышленности и т. д.

На семинаре 4 июля 2012 года, посвященном наблюдению бозона Хиггса на экспериментальных установках CMS и ATLAS, директор ЦЕРНа Р. Хойер дал высокую оценку грид-технологиям и их значимости для мировой науки. Он выделил три составляющие, обеспечившие получение этого результата — ускорительный комплекс ЦЕРНа, экспериментальные установки и грид-инфраструктура БАКе. Грид-инфраструктура на БАКе позволила обрабатывать и хранить колоссальный объем данных, поступающих от экспериментов на коллайдере, и, следовательно, совершать научные открытия [5].

Российские центры, объединенные в консорциум RDIG (Russian Data Intensive GRID), участвуют в обработке

и анализе данных экспериментов на Большом адронном коллайдере. За 2014 год на ресурсах российской грид-инфраструктуры было выполнено около 17 млн заданий с суммарным процессорным временем около 650 млн нормализованных часов, что соответствует 4% от всей инфраструктуры WLCG [2, 3].
Большую роль в развитии грид-технологий играет Объединенный институт ядерных исследований (ОИЯИ). Сотрудники Лаборатории информационных технологий ОИЯИ активно участвуют во многих проектах по развитию систем глобального мониторинга, управления распределенными данными, модели компьютинга для мегапроектов [4].

Эти проекты требуют новых подходов и решений, так как между грид-сайтами возникают огромные потоки данных и задач.
Модель компьютинга для экспериментов на БАКе находится в постоянном развитии. Это выражается оптимизации потоков данных и задач, интеграции грид-технологий и облачных вычислений, использовании суперкомпьютеров для моделирования, развития технологий Больших Данных.

После модернизации Большого адронного коллайдера объем данных существенно увеличится, поэтому организация хранения данных и управления потоком задач играет первостепенную роль.

Для эксперимента ATLAS была разработана программная платформа «Панда», которая выполняет управление потоком заданий в разных грид-инфраструктурах. В настоящее время эта платформа активно развивается для работы с различными ресурсами (кластерами, облачными средами, суперкомпьютерами), что позволит существенно увеличить и разнообразить инфраструктуру распределенных вычислений для решения масштабных задач с использованием технологий Больших Данных[6]. В этом проекте активно участвуют сотрудники НИЦ «Курчатовский институт» и ОИЯИ.

В настоящее время на базе НИЦ КИ и ОИЯИ создан центр хранения данных для Большого адронного коллайдера уровня Tier-1. Это очень важный и престижный проект, так как обеспечивает надежное хранение и эффективный доступ к данным многим грид-сайтам мира. Обеспечение стопроцентной надежности и доступности требует колоссальных усилий и надежной работы всех систем.

Для обучения специалистов разных стран в ОИЯИ создана распределенная учебно-исследовательская инфраструктура, в которой представлены популярные современные технологии распределенных вычислений.

Литература

1. Портал проекта WLCG (Worldwide LHC Computing Grid; Всемирный грид для Большого адронного коллайдера)— http://wlcg.web.cern.ch/
2. В. А. Ильин, В. В. Кореньков, А. А. Солдатов: Российский сегмент глобальной инфраструктуры LCG // Открытые системы –2003–№1–с. 56–60.
3. В. Ильин, В. Кореньков: Компьютерная грид-инфраструктура коллаборации RDMS CMS // сб.: «*В глубь материи: физика XXI века глазами создателей экспериментального комплекса на Большом адронном коллайдере в Женеве*» *— М: Этерна –2009–с. 361–372.*
4. Портал по развитию грид-технологий в ОИЯИ — http://grid.jinr.ru/
5. А. Климентов, В. Кореньков: Распределенные вычислительные системы и их роль в открытии новой частицы // Суперкомпьютеры –2012–№3–(11)– с. 7–11.
6. А. Ваняшин, А. Климентов, В. Кореньков: За большими данными следит ПАНДА // Суперкомпьютеры –2013–№3 (15)–с. 56–61.

Воспроизводимость численных экспериментов

Андрей Устюжанин, к.ф.-м.н., руководитель проектов **CERN** и **«Яндекс»**, приглашенный исследователь, **Лондонский Имперский колледж**

Наше завтра во многом определяется уровнем развития науки и технологии сегодня. Эта зависимость отмечается многими общественными деятелями, которые все чаще подчеркивают важность подготовки технических кадров и развития инновационно-технологических инфраструктур [1]. Научные исследования становятся доступны все более широкому кругу лиц и становятся все более «медиализированными». Как следствие, повышается уровень ошибок в таких исследованиях, которые могут вводить в заблуждения представителей научных сообществ. Возникает необходимость в механизме проверки полученных и опубликованных результатов. Подавляющее большинство проводимых экспериментов опирается на численную обработку данных, полученных в ходе достаточно сложных измерений. Мы называем этот этап исследований *вычислительным*

экспериментом. Основные особенности численных экспериментов таковы:

- обработка и анализ могу происходить на вычислительных ресурсах, оторванных от экспериментальных установок, получающих данные (спутники, телескопы, секвенсоры и т. п.);
- методология обработки данных и получения содержательных выводов составляет самостоятельную область науки (во многих экспериментах правильная численная обработка данных представляет собой самую сложную часть анализа);
- методы и подходы обработки данных, накопленных в различных сферах науки/техники, могут быть очень близки.

Численные эксперименты можно рассматривать как современные микроскопы, позволяющие заглянуть вглубь наблюдаемых явлений и, лучше понять взаимосвязи различных аспектов окружающего нас мира.
Центральным звеном численных экспериментов являются алгоритмы [2]. Существует несколько возможностей для улучшения точности таких вычислительных «приборов»:

- использование более «продвинутых» алгоритмов (BDT, Deep Neural Networks, SVM, …);
- предобработка данных и выделение информативных признаков, основываясь на которых алгоритмы могут дать более точные результаты;
- использование более сложных сценариев обработки данных (k-folding, ensembling, blending, cascading, …).

Как правило, повышение точности обработки данных в ходе численных экспериментов позволяет снизить затраты (сократить время и стоимость) на проведение измерительной части эксперимента (работа большого адронного коллайдера, спутника или телескопа). Однако, повышение точности приводит к повышению сложности численных экспериментов. Источники сложности можно разделить на следующие группы:

- сложность понимания предметной области;
- сложность работы с источниками данных (форматы, версии);
- сложность стратегии численного эксперимента;
- сложность отдельных шагов эксперимента, алгоритмов;
- сложность выбора критериев оценки качества результата;
- сложность работы распределенной команды.

Каждый из указанных аспектов может оказаться важным как для получения результата, так и для его воспроизведения.

Одним из важных критериев успешности выполненного численного эксперимента является возможность повторного получения такого же результата другой группой исследователей. Численный эксперимент, для которого определена процедура повторного получения результата другими исследователями, мы будем называть воспроизводимым численным экспериментом. Для проведения таких экспериментов необходима специальная среда.

Reproducible Experiment Platform — программная среда для поддержки экосистемы совместной исследовательской работы над общими задачами, позволяющая:

- выполнять численные эксперименты над большими объемами данных;
- получать воспроизводимые результаты;
- использовать единообразные критерии качества.

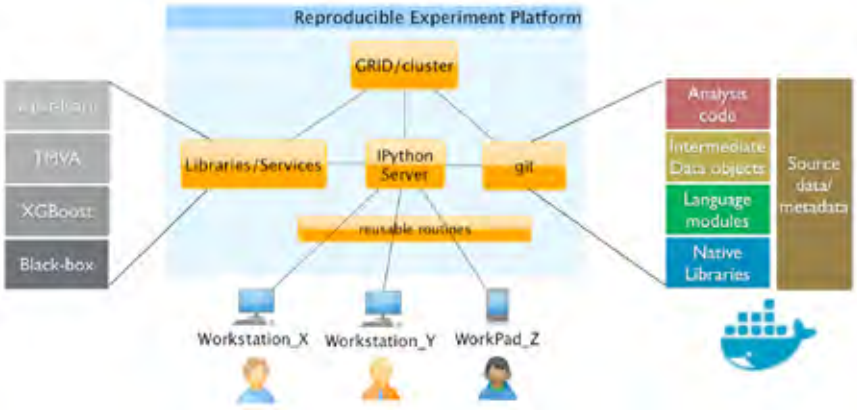


Рис 1. Диаграмма компонентов Reproducible Experiment Platform

К этой среде предъявляются следующие высокоуровневые требования:
- автоматизация и повторное использование разработанных компонент;
- автоматическая проверка полученных результатов;
- онлайн-доступ к вычислительным мощностям и алгоритмам;
- версионирование модулей, кода, окружения выполнения и данных;
- поддержка существующих реализаций алгоритмов обработки данных;
- масштабируемость по ресурсам;
- простота в изучении.

В настоящее время прототип этого продукта разрабатывается в ходе совместных работ «Яндекса» и ЦЕРНа для решения задач анализа данных физики частиц. На сегодняшний день он применялся для анализа распадов $Bs \to 4\mu$, $\tau \to 3\mu$ [3]. Однако его применение не ограничивается областью физики частиц, и он может быть востребован в таких областях, как:
- преподавание дисциплин инженерии данных (data science);
- исследования в различных областях науки таких как физика частиц, астрофизика, информационный поиск, и т. п.;
- междисциплинарные исследования, развитие методов анализа данных.

Примером междисциплинарного исследования может служить работа по построению нового метода классификации, который способен нивелировать корреляции между прогнозом классификатора и заранее указанной переменной [4].

Дальнейшие шаги развития системы предполагают:
- интеграцию с платформами распределенных вычислений (Hadoop, Hive, Spark, GRID);
- интеграцию с системами ведения проектов;
- развитие методологии ведения распределенных вычислительных экспериментов;
- поддержку публикации результатов исследований не в форме статей, а в форме контролируемо-исполняемых контейнеров.

Применение системы Reproducible Experiemnt Platform позволяет сделать исследовательскую работу намного более предсказуемой и увлекательной. Аналогично использованию методологий программирования (Scrum, Agile) и технологий контроля версий (SVN, git) применение данной системы позволит значительно расширить круг исследователей как в фундаментальных областях наук, так и в технологических сферах, что в свою очередь даст толчок развитию разнообразных инновационных направлений (народного хозяйства).

Литература

1. Felt, Ulrike et al (2013) 'Science in Society: Caring for our Futures in Turbulent Times', ESF Policy Briefing 50. European Science Foundation.– https://sts.univie.ac.at/en/people/ulrike-felt/
2. Breiman, Leo, «Statistical Modeling: The Two Cultures», 2001, Statistical Science, Volume 16, Issue 3 (2001), P. 199–231.
3. LHCb collaboration, Search for the lepton flavour violating decay $\tau^- \to \mu^- \mu^+ \mu^-$, http://arxiv.org/abs/1409.8548
4. Alex Rogozhnikov, Aleksandar Bukva, Vladimir Gligorov, Andrey Ustyuzhanin, Mike Williams. «New approaches for boosting to uniformity». – http://arxiv.org/abs/1410.4140v1

Облачные технологии в естественных и гуманитарных науках

Сергей Березин, к.ф.-м.н., доцент, руководитель совместного исследовательского центра **МГУ-Microsoft Research, МГУ им. М.В. Ломоносова**

В докладе будут рассмотрены облачные приложения FetchClimate и «Хронозум», совместно разработанные Microsoft Research и МГУ им. М.В. Ломоносова. FetchClimate ориентирован на специалистов в области вычислительного моделирования и предоставляет пользовательский интерфейс и веб-сервис для получения климатических данных. «Хронозум» — ориентированное на историков и представителей гуманитарных наук приложение для визуализации Универсальной истории. На примере FetchClimate и «Хронозум» будут рассмотрены вопросы создания, практического применения и возможного развития облачных приложений для работы с научными данными.

Секция. Прикладные аспекты Больших Данных

Влияние технологий Больших Данных на дизайн организации

Светлана Мальцева, д.т.н., профессор, **НИУ ВШЭ**

Исследуя влияние Больших Данных на дизайн организации, необходимо рассматривать несколько важных аспектов:

- область деятельности;
- внутренняя среда организации;
- зрелость технологии;
- ожидаемые результаты.

Область деятельности, связанная с конкретной социальной или экономической сферой, определяет особенности задач, решение которых технологии Больших Данных могут существенно улучшить по многим критериям, например, качества или скорости, а также новых задач, которые возникают вследствие возможностей, которые открывают эти технологии. В этом смысле технологии Больших Данных могут породить так называемые инновации ценности («value innovation»), которые сегодня обладают самым высоким потенциалом с точки зрения конкурентоспособности предприятия.

Среди направлений работы организации, в которых важность новых технологий существенна, можно выделить следующие:

- принятие решений;
- стратегическое планирование;
- маркетинг и коммуникации;
- клиентские сервисы и взаимодействие с потребителем;
- управление финансами;
- исследования и разработки.

Велико влияние технологий Больших Данных также на социальную сферу, в особенности на здравоохранение, образование, сферу услуг, жилищную сферу и другие. Усиливается влияние на сферу науки и культуры.

Внутренняя среда организации сегодня является критическим условием для внедрения Больших Данных. Обзор успешных историй внедрения показывает, что сегодня эти технологии внедрены в основном в крупных компаниях с развитой цифровой инфраструктурой и персоналом, обладающим развитыми цифровыми компетенциями, в особенности в области аналитики и социальных сетей. Важной особенностью этих компаний является также высокий инновационный потенциал организации, так как внедрение Больших Данных инициирует необходимость инноваций в продуктовой и в особенности в организационной сфере.

Зрелость технологий Больших Данных в организации определяется составом задач, для которых они используются, что, в свою очередь, сильно зависит от внутренней среды организации. В работе [1] предложено выделить пять уровней зрелости для бизнес-модели организации с точки зрения интегрированных групп задач, которые решаются на основе технологий больших данных:

- мониторинг бизнеса;
- анализ бизнеса;
- оптимизация бизнеса;
- монетизация данных;
- трансформация бизнеса.

Возможность проводить глубинный анализ данных, их интерпретацию на основе корреляции данных из разных источников позволяет не только отвечать на традиционные вопросы аналитики о том, что происходит и почему, но также на вопросы о том, что должно произойти (предсказательная аналитика) и какие действия следует предпринять. Предписывающая аналитика, которая используется в этом случае, направлена не только на поддержку решений, но и на автоматизацию принятия решений.

Предсказательная и предписывающая аналитика, особенно проводимая в режиме реального времени на потоках данных, позволяет переходить к задачам, возникающим на более высоких уровнях зрелости.

Уровень монетизации данных предполагает возникновение новых бизнес-моделей, составляющие которых возникают за счет возможностей технологий Больших Данных. Например, создание нового изделия со

встроенными средствами анализа Больших Данных; предоставление результатов исследования данных, как информационного продукта; сервисы, связанные с использованием предсказательной и/или моментальной аналитики.

Расширение использования технологий Больших Данных во всех сферах деятельности предприятия делает необходимым проведение существенных изменений в его организационной структуре.

Наиболее важные изменения связаны с изменением управления данными в организации; новыми требованиями к составу и компетенциям персонала; внедрением новых моделей бизнес-процессов, ориентированных на коллаборативное принятие решений; возникновением новых подразделений и должностей; пересмотром зон ответственности.

Менеджмент Больших Данных, учитывая их разнообразие с точки зрения структуры, содержания, характеристик источников, предполагаемого использования, должен опираться на использование существенно большего, чем при использовании традиционных хранилищ данных, спектра применяемых инструментов и методов, а также на многоплатформенность программных и аппаратных решений и тщательный подход к формированию персонала, ответственного за управление. Особенную сложность представляет управление потоковыми данными, поступающими в реальном времени.

Важной частью управления Большими Данными в организации является создание новых регламентов и стандартов, отличных от тех, которые используются для структурированной информации предприятия. Оно необходимо для обеспечения синергии существующего и нового подходов к аналитике.

Персонал является ключевой задачей предприятия, переходящего к технологиям Больших Данных. Сегодня отмечается явный недостаток специалистов в области исследования данных, который будет усиливаться в ближайшие годы [2].

Сегодня видение состава компетенций бизнес-аналитика в сфере Больших Данных только формируется. В работе [3] представлена модель, являющаяся развитием диаграммы Венна, предложенной Дрю Конвеем [4].

В предложенной модели интегрированные компетенции возникают на пересечении трех областей компетенций и навыков: в области бизнеса, информационных технологий и наук о данных. Для каждой из областей очерчен перечень ролей и обязанностей, которые она включает.

Бизнес-навыки предполагают умение правильно формулировать вопросы к аналитике для решения бизнес-задач, знание существующих ограничений, выбор критериев для оценки эффективности, принятие решений, обеспечение прозрачности применяемых механизмов и методов. ИТ-навыки включают знание технологий сбора, хранения и обработки данных. Важными составляющими являются компетенции в области высокопроизводительных вычислений и доставки данных. Компетенции и навыки в области исследования данных связаны с пониманием свойств данных, их моделей, соответствующих аналитических методов, способов представления и интерпретации данных. Область пересечения — это область интегрированных компетенций, позволяющих решать задачи бизнеса на основе аналитики Больших Данных.

Отсутствие на сегодняшний день специалистов с такими интегрированными компетенциями порождает как альтернативу использование многофункциональной команды для работы с большими данными. В организационной структуре возникают центры исследования данных. Для средних компаний привлекательным решением для использования аналитики Больших Данных является аутсорсинг.

Реализация высокого потенциала Больших Данных в организации связана также с общей информационной культурой компании с точки зрения доверия к новым аналитическим инструментам и понимания их положительного влияния на работу организации.

Литература

1. Bill Schmarzo. Big Data Business Model Maturity Chart. – https://infocus.emc.com/william_schmarzo/big-data-business-model-maturity-chart/

2. Philip Russom. Managing Big Data. – www.sas.com/content/dam/SAS/en_us/doc/whitepaper2/tdwi-managing-big-data—106702.pdf

3. Vincent Granville. The Data Science Venn Diagram Revisited. Data Science Central. – www.datasciencecentral.com/profiles/

blogs/the-data-science-venn-diagram-revisited
4. The Data Science Venn Diagram. – http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram
Mike Barlow. The Culture of Big Data. O'Reilly – 2013. – http://www.slideshare.net/sanghgautam_slide_share/culture-ofbigdata

Возможности извлечения маркетинговой информации в e-commerce

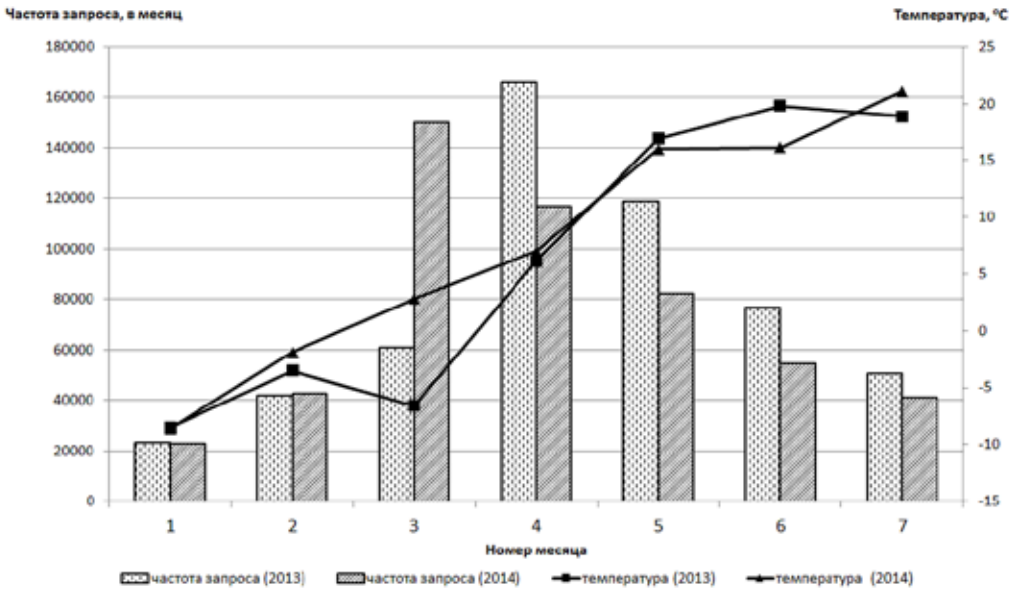
Михаил Сливинский, руководитель отдела маркетинговой и поисковой аналитики, **«Викимарт»**
Я бы хотел рассказать о возможностях, которые появились в маркетинге в последние годы благодаря Большим Данным. Мы не занимаемся фундаментальными исследованиями, наша задача — быстро создавать полезные для e-commerce решения на основе собственной статистики и внешних источников данных. Соответственно, исследовательская часть заведомо небезупречна, и мы знаем об этом. Однако надеюсь, наши идеи и подходы могут быть интересны.

Очевидный и важный источник Больших Данных — данные о действиях пользователей на сайте, поисковые запросы и транзакции пользователей и т. д. Эта информация легкодоступна владельцу сайта. Но есть и другие, менее популярные и очевидные источники, например:

- статистика запросов к поисковым системам;
- данные, собираемые браузерными плагинами (Alexa [1], Similarweb [2], Neiron [3] и т. д.);
- результаты поиска в поисковых системах;
- корпусы отзывов на товары и т. д.

Поговорим подробнее о практических способах извлечения ценной маркетинговой информации из этих данных.

Современные поисковые машины предоставляют доступ к статистике поисковых запросов (Yandex [4], Mail [5]). Благодаря этим инструментам стало возможно измерить динамику популярности поисковых запросов в срезах по регионам, полу и возрасту пользователей.
Доступность данных позволяет делать интересные наблюдения. Например, спрос на детские велосипеды очень зависим от температуры воздуха в весенние месяцы:
Другой извечный вопрос — что порекомендовать покупателю? Традиционно рекомендательные системы строят на коллаборативной фильтрации. Принцип ее прост: будем считать, что данного покупателя заинтересует то, к чему проявили интерес похожие по своему поведению покупатели. Модель хорошая, но есть два существенных недостатка: для обучения системы нужно много пользователей; обучаясь «с нуля», мы снижаем лояльность покупателей, демонстрируя им слаборелевантные предложения.



Мы разработали решение для подбора похожих товаров, основываясь на их сходстве по характеристикам, цене или бренду. Суть решения — предсказание вероятности изменения исходной потребности покупателя, основываясь на статистике действий пользователей. Взвесив эти вероятности между собой, мы можем оценить сходство любых двух товаров. Таким образом, мы научились отвечать на вопросы: «Если покупатель пришел за 40-дюймовым ЖК-телевизором Samsung, заинтересуют ли его: 40-дюймовые телевизоры Panasonic?

36-дюймовым и 42-дюймовым телевизоры Samsung? На 7% более дорогая модель?».
Другая наша работа — построение системы автоматического ранжирования товаров в категории. Многие покупатели начинают поиск товара и магазина, не сформулировав полностью свою потребность. Например, по запросу «холодильник» мы можем предложить сотни хороших товаров. Очевидно, пользователь не станет просматривать их все, поэтому у нас есть только несколько секунд, чтобы заинтересовать его действительно подходящим ассортиментом. Мы построили полином на нескольких важных признаках товаров. Наше ранжирование позволило увеличить на 40% выбранную целевую метрику. Подробнее мы рассказали об этом решении на конференции YaC/m 2013 [6].

Также весьма интересны корпусы отзывов и обсуждений товаров. Частотные словари по n-граммам позволяют выделить характерные коллокации, что проливает свет на потребности и задачи покупателя. Например: «регулятор крепости кофе», «долго держит заряд», «инструкция для сборки», «крепится к стеклу» и пр. Подробнее в докладе [7] нашего лингвиста Ирины Борисовой.

Получение трафика из поисковых систем базируется на связках ‹запрос-документ›. Для крупного проекта речь идет о сотнях тысяч и миллионах поисковых запросов. Мы решили задачу поиска большого числа поисковых запросов, характерных для заданной предметной области. Считая современные поисковые машины идеальными кластеризаторами, мы считали релевантными те запросы, в ответах на которые встретились заданные слова-маркеры. Применение этой идеи позволило найти множество неочевидных поисковых запросов. Например для тематического кластера «свадьба» нашлись такие запросы, как «диадема», «бонбоньерки», «какого размера рушник», «Вера Вонг» и т. д.
Другое возможное применение знания о популярности поисковых запросов — оценка востребованности различных типов контента (отзывы, видео, обзор и пр.). Так, относя популярность запроса, дополненного словом «отзывы» к собственной популярности запроса, можно оценить востребованность отзывов при поиске и выборе товаров в заданной категории.

Литература

1. http://www.alexa.com/
2. http://www.similarweb.com/
3. http://neiron.ru/
4. http://wordstat.yandex.ru/
5. http://webmaster.mail.ru/querystat
6. Аркадий Итенберг, Михаил Сливинский. Покупатель — лучший эксперт. Ранжируем товары умно! YaC/m. — 2013. — https://tech.yandex.ru/events/yac/m/talks/824/
7. Ирина Борисова. Прикладная лингвистика и искусственный интеллект — 2012. Лексическая статистика в оценке качества коммерческих текстов. Wikimart. — http://www.ashmanov.com/arc/aiconf2012/16-borisova-aiconf2012.pdf

СОДЕРЖАНИЕ