Открытые системы

SofTool
www.softtool.ru

ОТКРЫТЫЕ СИСТЕМЫ
Open Systems Publications

Конференция

БОЛЬШИЕ ДАННЫЕ
в национальной экономике

Тезисы докладов

БОЛЬШИЕ ДАННЫЕ

22 октября 2013 года, ЦВК «Экспоцентр»

Конференция «Большие Данные в национальной экономике»

Ключевым фактором успеха для многих отраслей российской экономики становится возможность эффективно обрабатывать огромные массивы и потоки информации. Поиск оптимальных способов обработки больших объемов данных приобретает важнейшее значение для развития перспективных направлений науки и индустрии, таких как биоинформатика, энергетика нового поколения, экономическое моделирование, социология и др.

Сегодня уже очевидно, что Большие Данные — одно из ключевых направлений компьютерной науки, открывающее новые перспективы для исследований в сфере представления, анализа и извлечения полезных знаний из больших объемов данных различной природы.

Цель конференции «Большие данные в национальной экономике» — собрать отечественных экспертов в области анализа данных, с тем чтобы оценить общее состояние дел в этой области, выделить и консолидировать наиболее значимые работы и коллективы, способные сделать вклад в формирование нового направления науки и индустрии и тем самым способствовать дальнейшему развитию национальной экономики России.

РФФИ
РОССИЙСКИЙ ФОНД ФУНДАМЕНТАЛЬНЫХ ИССЛЕДОВАНИЙ
Российская Академия Наук

Оргкомитет конференции «Большие Данные в национальной экономике» выражает признательность за поддержку Российскому фонду фундаментальных исследований (грант 13-07-06055-г), Институту проблем информатики РАН, компаниям «ЕС-Лизинг» и «Информ-Консалтинг».

Тематика конференции «Большие Данные в национальной экономике»:

Пленарная сессия. Платформы и методы обработки Больших Данных

Секция. Большие Данные в науке и индустрии

Секция. Большие Данные и общество

Тезисы докладов конференции «Большие Данные в национальной экономике» (Москва, 22 октября 2013 г.). /[Под ред. Дубовой Н.А.]. — М.: «Открытые системы», 2013. — 53 с.

В сборник трудов включены доклады конференции «Большие Данные в национальной экономике», прошедшей 22 октября 2013 года в Москве в рамках XXIV выставки Softool-2013.

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 оценка перспектив создания новых учебных программ для подготовки специалистов соответствующей квалификации (data scientists). 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» 22 октября 2013 года, ЦВК «Экспоцентр»

Организационный комитет

Гуляев Юрий Васильевич	**академик РАН, член президиума РАН, директор ИРЭ им. В. А. Котельникова РАН, председатель оргкомитета**
Дубова Наталия Аркадиевна	**научный редактор журнала «Открытые системы. СУБД», зам. председателя оргкомитета**
Арлазаров Владимир Львович	**чл.-корр. РАН, зав. отделом ИСА РАН**
Волков Дмитрий Владимирович	**с.н.с. ИПМ им. М. В. Келдыша РАН, гл. редактор журнала «Открытые системы. СУБД»**
Гергель Виктор Павлович	**д.т.н., декан факультета ВМК ННГУ им. Н. И. Лобачевского**
Кузнецов Николай Александрович	**академик РАН, президент Международного союза приборостроителей**
Никитов Сергей Аполлонович	**чл.-корр. РАН, зав. кафедрой МФТИ**
Сухомлин Владимир Александрович	**д.т.н., профессор, ВМК МГУ им. М. В. Ломоносова**

Пленарная сессия. Платформы и методы обработки Больших Данных

Перспективы работы с большими объемами данных

Черняк Л.С. ***(cherniak@osp.ru)*** **— научный редактор, журнал «Открытые системы»**

Кто-то удачно сравнил происходящее сейчас с окончанием каменного века. Тогда человек научился извлекать металл из руды, а сейчас — информацию из данных. Интернет, сенсорная революция, бесчисленные медийные устройства привели к тому, что порождается невероятное, немыслимое еще десять лет назад количество новой руды современной культуры. Именно это не совсем удачно назвали Большими Данными.

В 1998 году термин Big Data впервые использовал Джон Мэши, главный ученый Silicon Graphics. Тогда термин не получил широкого распространения, поскольку Мэши предсказывал будущий рост данных и его значение, адресуясь к узкому кругу коллег. Свою нынешнюю популярность словосочетание Big Data обрело после известных публикаций в журнале Nature в 2008 году, где обсуждались проблемы, вызванные ростом объемов данных, получаемых в процессе проведения современных научных экспериментов, и, как следствие, в связи с появлением нового поколения науки, называемого электронной наукой (e-science). За прошедшие с момента выхода статьи пять лет этот термин получил более широкое распространение и был хорошо освоен маркетингом. Он вошел в обиход бизнес-компьютинга, причем стал использоваться с такой невероятной интенсивностью, что, еще не будучи достаточно понят, стал вызывать негативную эмоциональную реакцию у некоторых специалистов.

Как бы ни были сложны и важны технологии работы с данными, они остаются инструментами. Есть следующий уровень, ради которого эти инструменты создаются. Прежде всего тектонический сдвиг, отличающий существующий компьютинг от компьютинга будущего, — это переход от программируемого компьютинга к когнитивному компьютингу. Во вторую очередь надо назвать возвращение искусственного интеллекта. Кроме того, предпринимаются первые шаги к формированию экономики нового типа — экономики обратной связи (feedback economy).

Почему в этом перечне нет Data Science? Казалось бы, об этом сейчас так много говорят и пишут. Прежде всего потому, что это понятие еще недостаточно определено, что в ряде случаев освобождает употребляющих этот термин от необходимости глубже погрузиться в суть происходящего. Кроме того, очевидно, что Data Science нельзя переводить буквально как «наука о данных», поскольку в английском science — не только «наука», но еще и «мастерство», «искусство» и «умение». Следовательно, Data Science точнее было бы интерпретировать еще и как умение, а в некоторых случаях — и искусство, работать с данными. Такая интерпретация точнее отражает специфику деятельности data scientists, не занимающихся изучением данных, а использующих свой комплекс знаний и навыков для получения требуемых результатов при анализе данных.

Раньше других непривычный пока термин «когнитивный компьютинг» в широкий оборот ввела корпорация IBM. Там считают, что когнитивный компьютинг — это качественно новое явление, которое знаменует собой наступление объявленной IBM третьей эры в компьютинге. Первой была эра табуляторов, которая началась с дифференциальной машины Чарльза Бэб-

биджа и достигла своего расцвета усилиями Германа Холлерита, создавшего производительные электромеханические табуляторы и основавшего компанию Tabulating Machine Company, позже преобразованную в IBM. Следующая, вторая эра — нынешняя, она ассоциируется с программируемой схемой Джона фон Неймана, из которой следует обработка данных по заранее заданной программе. Что же касается когнитивного компьютинга в его нынешнем виде, то это не полный отказ от схемы фон Неймана — скорее некоторый паллиатив, который предполагает отказ от схемы на верхнем уровне при сохранении традиционных процессоров.

В полном смысле когнитивный компьютер, то есть не содержащий в себе следов неймановского наследия, еще не существует. Работа над его созданием ведется, например, в IBM по программе SyNAPSE (Systems of Neuromorphic Adaptive Plastic Scalable Electronics) по заказу агентства DARPA. Предполагается, что он будет нейроморфным, то есть имитирующим деятельность мозга. Близкие по содержанию работы идут в целом ряде лабораторий и университетов США и Европы.

Нынешний подъем интереса к искусственному интеллекту начался с отказа от непродуктивных попыток повторить человеческий мозг и эмулировать человеческую логику. Непредвзятому наблюдателю в годы первой волны работ в области искусственного интеллекта было непонятно, зачем заставлять делать машину то, что и без того с успехом делает человек, — например, отвечать на вопросы так, как это делает человек, писать музыку или сочинять стихи, когда есть такие сферы, где машины могут действовать успешнее человека.

Есть важное отличие прежних работ по искусственному интеллекту от исследований эпохи Больших Данных. В этой области происходит переход, аналогичный переходу от детерминированной ньютоновской к релятивистской физике. Если видеть мир во всей его полноте, а именно эту возможность предоставляют Большие Данные, то приходится признать, что не существует абсолютно детерминированной реальности, что есть вероятность и непредсказуемость событий. Вместе с осознанием такой картины мира дисциплина искусственного интеллекта переходит от классического детерминированного подхода к пониманию сложности окружающего мира. На смену расчету, занимавшему монопольное положение, приходят анализ в самых разных формах и майнинг данных.

До сих пор рынок остается единственным способом реализации обратной связи. Внедрению альтернативных регуляторов, основанных на обратной связи, до последнего времени мешало то, что не было возможности собирать необходимые данные и справляться с огромными объемами сведений о реальном состоянии экономики. Любые, даже самые сложные технические системы, которые создавались прежде, будь то атомный или химический реактор, самолет или энергоблок, порождают на порядки меньше данных, чем экономика. Как ни парадоксально, но этого не понимали считавшие себя кибернетиками создатели Киберсина в Чили при власти Сальвадора Альенде и Общегосударственной автоматизированной системы управления производством (ОГАС) в СССР. Они по наивности или по каким-то иным соображениям полагали, что, не имея достаточно мощной петли обратной связи, они смогут управлять государством. Невозможность или, скорее, нежелание оценивать реальное состояние дел с неизбежностью приводит к экономическому диктату. И только сейчас, когда создается мощная информационная инфраструктура, появилась возможность дополнить стихийно сложившиеся регуляторы дополнительными, созданными искусственно, при этом ни в коем случае не претендуя на ту глобальность замыслов, на которую рассчитывали Стаффорд Бир и академик В. М. Глушков.

Интеграция параллелизма в СУБД с открытым кодом

Пан К. С., Соколинский Л. Б. (*sokolinsky@gmail.com*), Цымблер М. Л. — ЮУрГУ (Челябинск)

В настоящее время СУБД с открытым исходным кодом (например, PostgreSQL, MySQL, SQLite и др.) являются надежной альтернативой коммерческим СУБД. PostgreSQL представляет собой одну из наиболее популярных СУБД с открытым кодом. Проект PostgreSQL был начат в 1995 году как ответвление от проекта POSTGRES М. Стоунбрейкера и до сих пор разрабатывается группой энтузиастов. Сначала PostgreSQL отличался от своего предка только наличием SQL-синтаксиса в запросах. На сегодня PostgreSQL представляет собой полноценную объектно-реляционную СУБД с открытым кодом для практически всех популярных операционных систем. PostgreSQL поддерживает стандарт SQL:2011, ACID-транзакции и хранимые процедуры на различных языках высокого уровня. Максимальный размер таблицы в PostgreSQL равен 32 Тбайт, максимальный размер поля таблицы — 1 Гбайт. PostgreSQL имеет хорошо документированный код и применяется в многочисленных коммерческих организациях, госструктурах и университетах (например, Apple, Sun, Cisco, Fujitsu, Red Hat, U.S. State Department, United Nations Industrial Development Organization и др.).

Проект PargreSQL посвящен разработке параллельной СУБД PargreSQL путем внедрения фрагментного параллелизма в СУБД PostgreSQL.

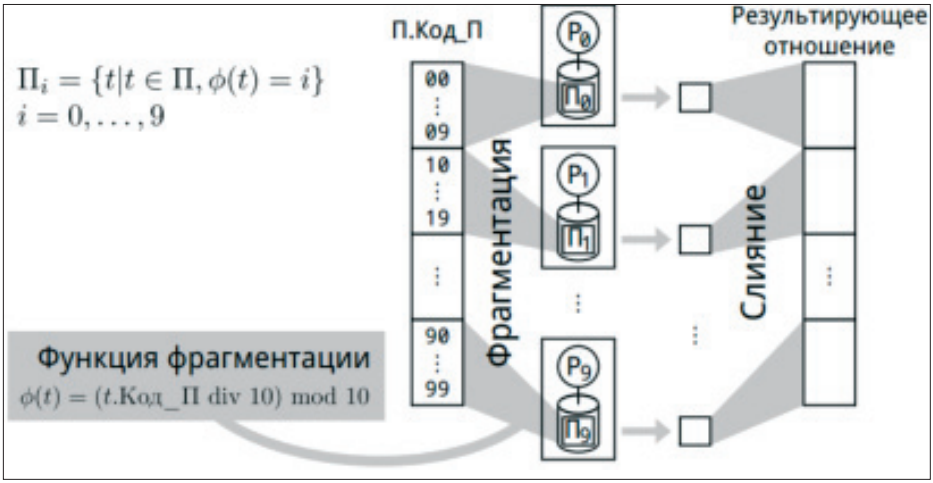


Рис. 1. Фрагментный параллелизм

Фрагментный параллелизм (см. рис. 1) подразумевает горизонтальную фрагментацию каждой таблицы базы данных по дискам кластерной системы. Способ фрагментации определяется функцией фрагментации, которая получает значение некоторой колонки таблицы и выдает номер диска, где хранится данная запись. На каждом узле запускается параллельный агент, представляющий собой модифицированный экземпляр СУБД PostgreSQL, который обрабатывает свои фрагменты, и затем частичные результаты сливаются в результирующую таблицу.

Технология внедрения параллелизма кратко может быть описана следующим образом. При обработке запроса мы добавляем к стандартным шагам (разбор запроса, разрешение представлений, построение плана запроса, выполнение плана запроса) еще два: построение параллельного плана запроса из последовательного плана и балансировка нагрузки узлов кластера при выполнении запроса (см. рис. 2).

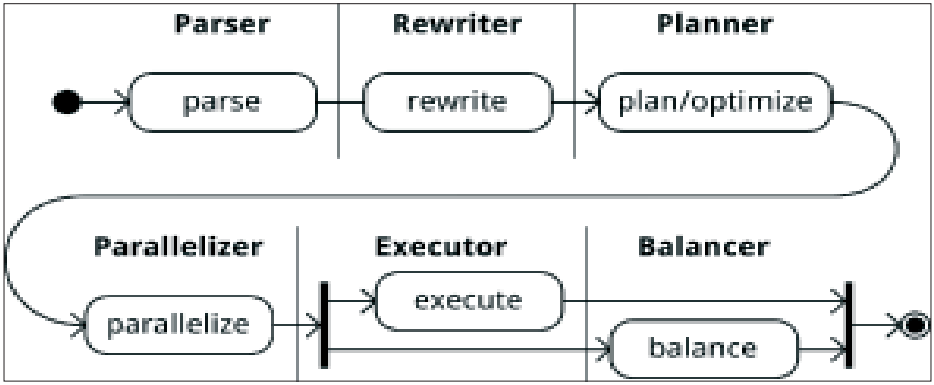
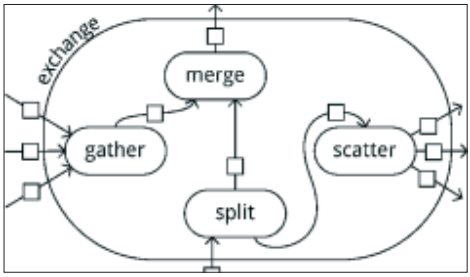


Рис. 2. Этапы обработки запроса в СУБД PargreSQL

Для реализации фрагментации в синтаксис команды PostgreSQL создания таблиц нами добавлен дополнительный параметр, специфицируя который программист при создании таблицы указывает имя целочисленного поля, используемого для фрагментации. В качестве функции фрагментации берется остаток от деления поля фрагментации на количество вычислительных узлов в кластере.

Нами разработан набор макросов, который подменяет вызовы оригинальных функций PostgreSQL на их PargreSQL-копии и обеспечивает прозрачный переход последовательных приложений PostgreSQL на параллельные приложения PargreSQL.



Наше ноу-хау представляет оператор обмена EXCHANGE. Оператор обмена EXCHANGE инкапсулирует в себе все механизмы, необходимые для реализации внутриоперационного параллелизма, существенно облегчает реализацию параллельных алгоритмов в системах баз данных и делает их более надежными.

Рис. 3. Оператор обмена EXCHANGE

Оператор EXCHANGE (см. рис. 3) состоит из четырех операторов: Split, Scatter, Gather и Merge. Split разделяет записи на «свои» (они должны быть обработаны на текущем узле кластера) и «чужие» (их необходимо передать на другой узел). Scatter отправляет «чужие» записи на соответствующие им узлы. Gather принимает «свои» записи от других узлов. Merge попеременно выдает результаты Gather и Split. Оператор обмена EXCHANGE встав-

ляется в нужные места последовательного плана запроса и реализует пересылки данных между параллельными агентами, необходимые для обеспечения корректности результата запроса.

Нами проведены эксперименты по исследованию ускорения и расширяемости СУБД PargreSQL, а также ее эффективности на задачах класса OLTP (обработки транзакций). Эксперименты проводились на суперкомпьютере «Торнадо ЮУрГУ», который занял 249-е место в 41-й редакции рейтинга TOP500 (июнь 2013 года).

В экспериментах на исследование ускорения выполнялся запрос, предполагающий соединение двух таблиц размерами $3x10^8$ и $7{,}5x10^5$ записей соответственно. Нами получено ускорение, близкое к линейному (см. рис. 4).

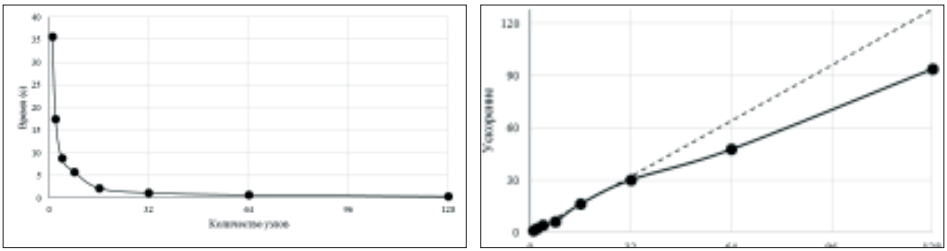


Рис. 4. Результаты экспериментов по исследованию ускорения

На том же запросе нами была исследована расширяемость СУБД PargreSQL с использованием от 1 до 128 узлов, когда одновременно с увеличением количества узлов равно увеличивается объем данных (от $1{,}2x10^6$ и $3x10^5$ записей до $1{,}5x10^8$ и $3{,}8x10^6$ записей соответственно). Эксперименты показали (см. рис. 5), что расширяемость близка к линейной.

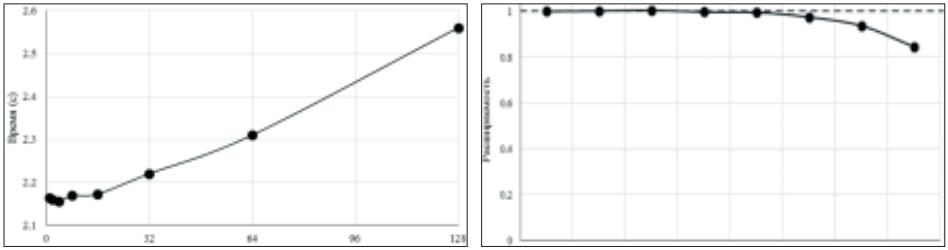


Рис. 5. Результаты экспериментов по исследованию расширяемости

Эффективность СУБД PargreSQL на задачах OLTP измерялась на стандартном тесте TPC C консорциума TPC (Transaction Processing Council), в котором моделируется складской учет. На конфигурации «12 складов, 30 клиентов» с использованием 12 узлов кластера PargreSQL показал производительность 2,2 млн запросов в минуту.

Рейтинг-лист теста TPC C среди параллельных СУБД, реализованных на кластерах

Rank	Company	System	Performance (tpmC)	DBMS	OS
1	Oracle	SPARC SuperCluster with T3-4 Servers	30 249 688	Oracle Database 11g R2 Enterprise Edition w/RAC w/ Partitioning	Oracle Solaris 10 09/10
2	IBM	IBM Power 780 Server Model 9179-MHB	10 366 254	IBM DB2 9.7	AIX Version 6.1
3	Oracle	Sun SPARC Enterprise T5440 Server Cluster	7 646 486	Oracle Database 11g Enterprise Edition w/RAC w/ Partitioning	Sun Solaris 10 10/09
	SUSU	Tornado	2 202 531	PargreSQL	Linux CentOS 6.2
4	HP	GHz-64p	1 184 893	Enterprise Edition	Linux AS 3

Это позволяет говорить о том, что PargreSQL входит в пятерку лидеров рейтинга TPC C (см. таблицу) среди параллельных СУБД, реализованных на кластерных вычислительных системах.

Подводя итоги, можно заключить следующее. Нами реализована параллельная СУБД PargreSQL путем внедрения концепции фрагментного параллелизма в свободную СУБД PostgreSQL. Нами проведены эксперименты, показавшие хорошую масштабируемость PargreSQL в задачах обработки сверхбольших данных. Предложенная технология внедрения параллелизма может быть применена к другим свободным СУБД с открытым кодом, например MySQL.

Кластер-анализ как средство анализа и интерпретации данных

Миркин Б. Г. (*bmirkin@hse.ru*) — НИУ ВШЭ (Москва)

Поддержано Программой фундаментальных исследований НИУ ВШЭ через грант «Учитель-ученики» 2011–2012, НУГ «Методы визуализации и анализа текстов» 2013 и Научную лабораторию ЛАВР (Москва) 2010–2013.

Для успеха обработки данных необходима автоматизация выработки заключений. Кластер-анализ — одно из средств такой автоматизации, пока, к сожалению, далеко не совершенное. Цель доклада — обзор возможностей использования кластер-анализа.

Кластер — это совокупность элементов, которые являются однородными или похожими в данной системе признаков.

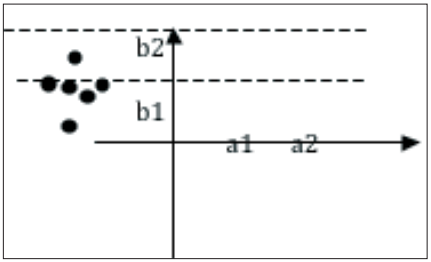
Цели кластер-анализа [1]: а) структуризация (представление общей структуры данных); б) описание кластеров в терминах тех или иных признаков; в) установление взаимосвязи между различными аспектами явлений; г) формирование обобщающих утверждений о свойствах данных и явлений; д) визуализация данных в процессах принятия решений.

Типичные данные объект-признак — таблица типа реляционной базы данных (фрагмент):

Town	Pop	PS	D	Ho	Ba	Sst	Pet	DIY	Swi	Po	CAB	FM
Mullion	2040	1	0	0	2	0	1	0	0	1	0	0
So Brent	2087	1	1	0	1	1	0	0	0	1	0	0
St Just	2092	1	0	0	2	1	1	0	0	1	0	0
St Colum	2119	1	0	0	2	1	1	0	0	1	1	0
Nanpean	2230	2	1	0	0	0	0	0	0	2	0	0
Gunnisla	2236	2	1	0	1	0	1	0	0	3	0	0
Mevagiss	2272	1	1	0	1	0	0	0	0	1	0	0
Ipplepen	2275	1	1	0	0	0	1	0	0	1	0	0
Alston	2362	1	0	0	1	1	0	0	0	1	0	0
Lostwith	2452	2	1	0	2	0	1	0	0	1	0	1
StColumb	2458	1	0	0	0	1	3	0	0	2	0	0
Padstow	2460	1	0	0	3	0	0	0	0	1	1	0
Perranpo	2611	1	1	0	1	1	2	0	0	2	0	0
Kingsbri	5291	1	1	0	5	3	1	0	1	1	1	0
Wadebrid	5676	2	0	0	4	4	1	0	0	2	1	1
Dartmout	6466	4	1	0	8	4	4	0	1	3	1	0
Launcest	6929	2	1	1	7	2	1	0	1	4	0	1

Технически кластер — это скопление объектов как точек многомерного пространства (рис.1).

Есть кластеры



Нет кластеров

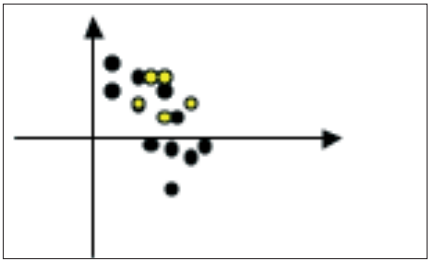


Рис. 1. Кластер представляет собой скопление объектов как точек многомерного пространства

Ниже представлены примеры применения методов кластер-анализа для продвижения в вышеуказанных целях.

(a) Структуризация (представление общей структуры данных)

a1. Структуризация деятельности (совместно с С. Насименто, Лиссабон)

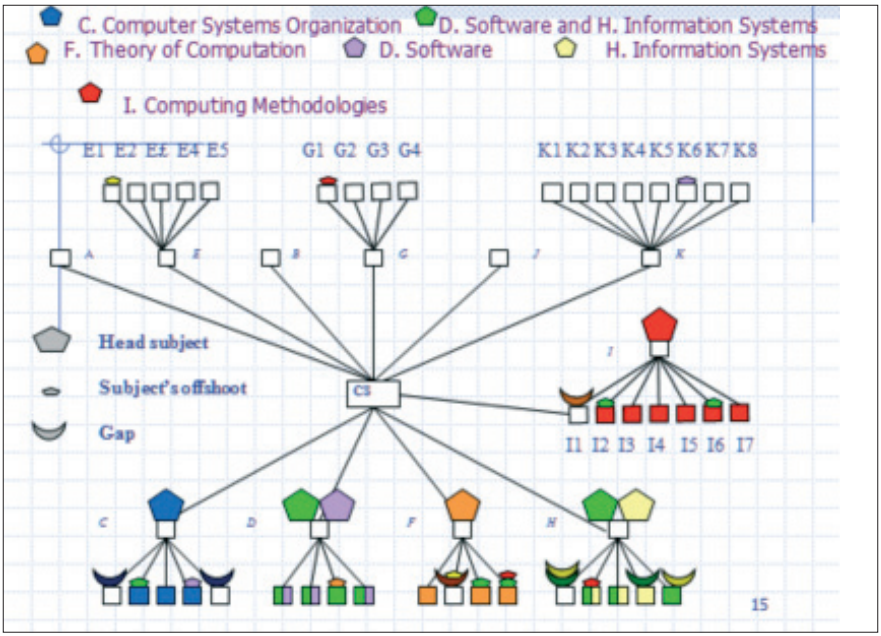


Рис. 2. Структура научной тематики работ ЦЕНТРИА, Лиссабон (в таксономии АВМ)

Структура научной тематики работ ЦЕНТРИА может быть представлена шестью кластерами тем из классификации компьютерной тематики, разработанной Ассоциацией вычислительных машин (рис. 2, раскрашены в разный цвет). Все кластеры попадают в свои гнезда классификации, кроме одного — зеленого. Он отображается сразу в две головные темы: «Программное обеспечение» и «Информационные системы». Это, по идее, обозначает какое-то исследование, ломающее границы между направлениями и, значит, несущее инновации. В данном случае, действительно, одна из крупных тем исследования — новое направление, называемое сейчас Software Engineering, не отраженное в классификации.

a2. Анализ жалоб жителей (совместно с Э. Бабкиным, Нижний Новгород)

Сформированы кластеры писем жителей в администрацию города. Они отображены на таксономию городских служб и проблем, решаемых ими (рис. 3).

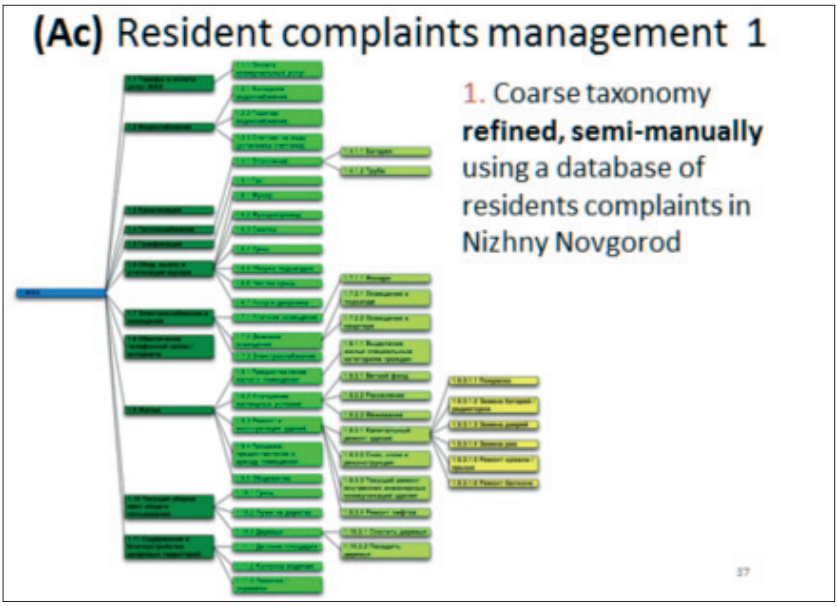


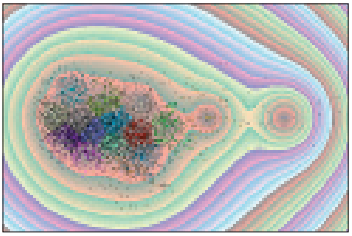
Рис. 3. Таксономия потребностей горожан и соответствующих служб города (по письмам населения)

Оказалось, что кластеры писем не «ложатся» в гнезда таксономии, а скорее идут поперек. Вывод: кластеры потребностей **не вписываются** в структуру городского хозяйства. Например, при протечке крыши не только нужна помощь плотников, но и требуется ремонт стен, электропроводки и т. п. Следовательно, нужны комплексные центры услуг для горожанина, которые могли бы оперативно формировать бригады для помощи жителям. В Москве подобные центры начали создаваться (и без нас).

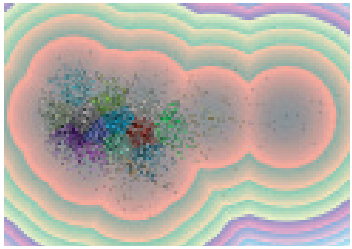
(б) Описание кластеров в терминах признаков

Проблема определения принадлежности нового химического соединения (совместно с Е. Колосовым и Р. Станфортом)

Чтобы применить прогностическую формулу активности того или иного соединения, нужно сначала определить, к какой группе оно относится. Пример — совокупность 14 тыс. химических соединений (признаки структуры) (рис. 4).



а)



б)

Рис. 4. Карта химических соединений, автоматически сгруппированных в кластеры (41): а — нашим методом; б — популярным методом нечетких к-средних

(в) Формирование обобщающих утверждений о свойствах явлений

Ростовцев, Миркин, Шанин (1981): исследование заболеваний органов дыхания и их факторов риска в Академгородке г. Новосибирска. Обработано 50 тыс. анкет, получено 14 иерархических кластеров (рис. 5).

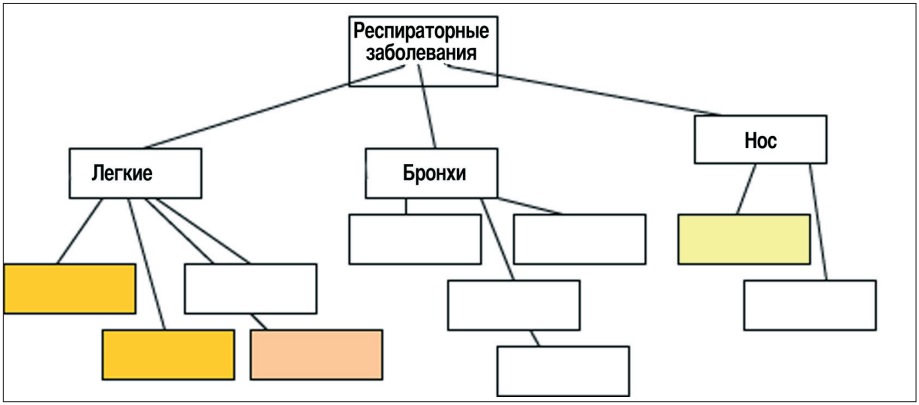


Рис. 5. Кластер-анализ заболеваний органов дыхания и их факторов риска

Предполагалось, что основными факторами риска являются курение и алкоголь. По полученным нами кластерам факторами риска оказались наличие заболевания в семье и плохие жилищные условия, а курение и алкоголь оказались никак не связаны с респираторными заболеваниями.

К сожалению, наши выводы, ставшие сейчас общим местом, тогда были отвергнуты (1981 год), так как противоречили всем устоявшимся представлениям, — нередкий случай, когда анализ данных оказывается бессилен.

(г) Визуализация данных в процессах принятия решений

г1. Кластерный анализ данных по 47 районам Московской области за 1979 и 1988 годы позволил нам увидеть разрушение структуры факторов прироста населения, произошедшее за эти 9 лет (Панфилова, Миркин, 1990).

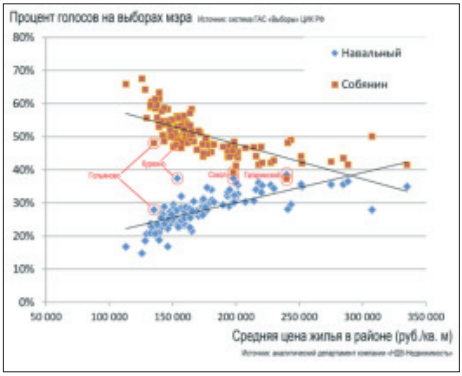


Рис. 6. Два кластера в пространстве «стоимость жилья × % голосов за кандидата»

г2. Выборы в Москве (Гурьянов В. Закон Бершидского: стоимость квадратного метра определила результаты выборов мэра // Квадратъ. 16 сентября 2013. № 44) (рис. 6): два кластера.

Заключение: «кластеры» являются важной частью автоматизации анализа данных, которая еще ждет своей более полной разработки и интеграции в системы анализа данных.

Литература

1. B. Mirkin. Mathematical classification and clustering. Kluwer AP, Dordrecht, 1996.

Стратегические угрозы XXI века в области ИТ: компьютеры против людей

Шмид А. В. — «ЕС-Лизинг» (Москва)

Согласно стратегическому прогнозу развития информатики на период 2005–2015 годов, разработанному в IBM, к 2015 году новое поколение обучаемых компьютеров (экспертных систем, ЭС) будет превосходить людей по качеству принимаемых решений сначала в некоторых, а затем и в большинстве областей человеческой деятельности.

Для подтверждения состоятельности этого прогноза достаточно отметить, что ЭС Watson фирмы IBM в 2013 году на общих основаниях сдала экзамены и получила диплом врача, приобретя юридическое право лечить людей. И в области онкологии уже демонстрирует блестящие результаты. Коммерчески доступны также индустриальные и банковские приложения ЭС. На этом фоне показательно заявление генерального директора IBM Вирджинии Рометти [1]: «...в течение следующих 5 лет все фирмы разделятся на победителей и побежденных в зависимости от качества принимаемых корпоративных решений (с применением ЭС!!!)».

В ближайшем будущем качество (решений) уже не может и не будет опираться на опыт и интуицию: конкурентные преимущества будут достигаться с учетом прогнозирования последствий принимаемых решений (predictive analytics). А технологическая «гонка вооружений» в области ИТ будет идти и уже идет за достижение превосходства по основным характеристикам применяемых ЭС: информированности и интеллектуальности [2].

Согласно прогнозам McKinsey [3], эта новая область ИТ в недалеком будущем станет и новой областью экономики, превосходящей по своей значимости нефтегазовый сектор, с тем отличием, что «сырьем» для переработки здесь будут не нефть и газ, а огромные и быстрорастущие мировые данные с необходимостью создания «заводов» по переработке «сырья» — ЭС.

Налицо поэтому две угрозы для национальной экономики:

А. Потеря конкурентоспособности предприятий РФ, для которых будут недоступны современные и будущие технологии принятия конкурентоспособных решений.

Проблема в том, что современные ЭС уже относятся к классу обучаемых систем, в которых персонал постоянно совершенствует приданные ему средства автоматизации в процессе эксплуатации (дообучает ЭС). И конкурентные свойства ЭС определяются совокупностью качеств персонала и ЭС. И если можно себе представить, что на рынке будут доступны из-

начально обученные ЭС, то остается открытым вопрос доступности персонала, способного дообучать ЭС в ходе эксплуатации на конкурентном уровне.

Представить себе наем зарубежных сотрудников в массовом порядке (десятки тысяч предприятий) уже невозможно, необходимо организовать соответствующее обучение в РФ.

Вместе с тем, по мнению Рометти [1], в США этому пока не учат, но собираются учить (predictive analytics).

Проблема обучения персонала для создания, эксплуатации и развития ЭС является международной.

В. Снижение конкурентоспособности экономики РФ В ЦЕЛОМ в случае выпадения из мирового разделения труда в новом секторе инновационной экономики — «машиностроении» ЭС («заводов» по переработке «сырья» — Больших Данных). Несмотря на новизну и необычность сектора создания нового класса нематериальных активов (обучаемых ЭС), базовые законы экономики никто не отменял.

Объемы производства и в этом секторе, как и ранее, будут определяться как числом работающих, так и их производительностью труда. Производительность труда при создании ЭС (как и ранее в программировании) будет определяться наличием средств автоматизации разработки заключительного продукта и готовых крупных строительных блоков программ (в программировании — языки высоко уровня, СУБД, мониторы транзакций и т. д.).

Если же рассматривать создание ЭС на основе коммерчески доступной платформы, то, например, основные компоненты платформы IBM BIG DATA (более 600) позиционируются фирмой-изготовителем именно как акселераторы (ускорители) разработки ЭС. То есть как средства радикального повышения производительности труда в новом «машиностроении» — производстве ЭС.

Итак, для удовлетворения потребностей инновационного развития экономики РФ, по экспертным оценкам, в ближайшие годы потребуются десятки тысяч специалистов в области создания и развития ЭС — нового поколения обучаемых компьютеров.

Необходимым условием (но не достаточным!) самой возможности организации обучения такого рода специалистов является наличие доступа при обучении к современным средствам автоматизации проектирования ЭС: платформам BIG DATA.

С целью удовлетворения потребности доступа к такого рода технологиям фирмой «ЕС-Лизинг» (ЕСЛ) совместно с IBM в конце 2012 года создан первый в РФ Центр компетенции по технологиям платформы IBM BIG DATA — базовым технологиям создания ЭС Watson. В этом центре на основе ВЦ ЕСЛ (вся линейка оборудования IBM, включая IBM z и Netezza) развернуты в полном объеме продукты платформы IBM BIG DATA, а также организованы основные университетские лабораторные работы IBM по начальному обучению этим продуктам.

Само обучение специалистов в настоящее время осуществляется на основе базовой кафедры ЕСЛ «Информационно-аналитические системы», созданной в МИЭМ НИУ ВШЭ в 2013 году, с предоставлением облачного доступа преподавателям и студентам к возможностям центра.

Особенностью обучения на базовой кафедре ЕСЛ является непременное участие студентов под руководством преподавателей в практическом проектировании систем, реализуемых для заказчиков ЕСЛ.

Центр компетенции BID DATA IBM-ЕСЛ является материально-технической основой для проведения НИР и ОКР в области создания современных информационно-аналитических

систем, в том числе ЭС. Для заинтересованных организаций Центром компетенции IBM-ЕСЛ оказывается целый спектр услуг по оказанию помощи в освоении технологий BIG DATA и организации обучения персонала.

Выводы:

1. В конкурентной борьбе как за качество принимаемых корпоративных решений, так и за долю рынка в новой экономике ключевую роль играет производительность труда проектировщиков ЭС. Целевая производительность труда может быть достигнута только с применением платформ BIG DATA — крупноблочных конструкторов для строительства ЭС.

2. Применение технологий BIG DATA приводит к появлению обучаемых ЭС, отличающихся от традиционных ЭС схемой принятия решений и наличием возможности обучения в ходе эксплуатации, с соответствующим изменением требований к ролям и квалификации персонала.

3. Подготовка специалистов для крупноблочного проектирования и последующего обучения ЭС требует доступа обучаемых к тренажерам ЭС, обеспечивающим формирование необходимых знаний и навыков.

Литература

1. http://www.cfr.org/technology-and-science/conversation-ginni-rometty/p30181

2. http://4cio.activetextbook.com/active_textbooks/34#page642

3. McKinsey Global Institute. Disruptive technologies: Advances that will transform life, business, and the global economy, May 2013.

Системы высокой доступности и Большие Данные

Будзко В. И. (*VBudzko@ipiran.ru*) — ИПИ РАН (Москва)

По мере увеличения степени встраиваемости средств электронной информационной технологии (ЭИТ) в различные направления современного общества возрастают требования к таким их характеристикам, как живучесть, адаптируемость, масштабируемость, поэтому сформировалось и постоянно развивается научно-техническое направление, связанное с созданием систем высокой доступности (СВД). Понятие «высокая доступность» введено в рекомендациях международного комитета TPC (Transaction Processing Performance Council) [1] и предполагает, что доступ к системе и получение соответствующего обслуживания выполняется, когда это необходимо и с приемлемой производительностью. При этом обслуживание системой не прерывается при любых условиях (например, при ее масштабировании — добавлении дополнительных мощностей без остановки работы, расширении — вводе в промышленную эксплуатацию новых прикладных систем на той же системотехнической базе, изменении версии системного программного обеспечения, отказе оборудования, катастрофе и пр.).

От таких систем требуется повышенная готовность осуществлять информационное обслуживание пользователей или управляемых объектов. А это означает, что предъявляются и повышенные требования ко всем обеспечивающим средствам [2, 3].

Большие Данные (Big Data, BD) — общий термин, используемый для описания огромного количества неструктурированных и частично структурированных данных, которые создает

компания. Это данные, хранение которых в реляционной базе данных для анализа заняло бы слишком много времени и стоило бы слишком много денег [4]. Хотя Большие Данные не относятся к какой-либо конкретной величине, когда о них говорят, часто используют термины «петабайты» и «экзабайты данных».

В настоящее время мы являемся свидетелями эволюции коммерческих предложений в области инструментария создания аналитических систем предприятий: от частных решений, например по визуализации аналитических результатов, к представлению интегрированных платформ с полной функциональностью для проведения аналитики по любым типам данных в ранее недоступных объемах. Например, платформы IBM Big Data или HP IDOL10. Это эволюция от ЭИТ-станков к ЭИТ-заводам.

Говоря об СВД, которая обладает свойством BD, мы не должны забывать о втором присущем ей обязательном свойстве — доступности всех необходимых данных для своевременной выработки адекватного информационного продукта, на основе которого может быть принято оптимальное решение [5]. От полноты и точности доступных данных, подлежащих обработке, зависит качество получаемого «информационного продукта» (ИП) и, соответственно, качество информационной поддержки процесса принятия решений пользователем. Высокая доступность применительно к автоматизированным информационным системам (АИС) предполагает не только своевременность выработки информации, но и высокое качество последней. Таким образом, АИС ВД должна включать средства своевременного сбора точных и полных данных и средства их своевременной обработки для получения информации, обеспечивающей своевременное принятие эффективного решения. Поэтому уместно называть такие системы не просто СВД, а системами высокой доступности данных (СВДД).

Технологии выбора наилучших из возможных решений отработаны и не менялись на протяжении веков. При их реализации возникает известная проблема распределения усилий между четырьмя «D»: выявление (discovery), отбор (discrimination), переработка (distillation), доведение информации в нужном представлении (delivery/dissemination) [6–8].

В чем состоит своевременность получения ИП? Введем простое условие. Если время Т, прошедшее с момента возникновения релевантных исходных данных (ИД) до момента завершения выполнения действия на основании выработанного ИП, достаточно для своевременного и, соответственно, эффективного воздействия на происходящие процессы, то Т удовлетворяет требованиям непрерывности бизнес-процесса конкретной предметной области (организации, предприятия и пр.). СВДД должна своевременно выявлять ИД, доставлять ИД, обрабатывать ИД, производить ИП и доводить ИП до руководителя. Обозначим как ТРЕГ максимально допустимое значение времени Т, превышение которого нарушит непрерывность бизнес-процесса. Определим составляющие Т.

Т = ТВ + ТП + ТО + ТИП + ТР + ТД, где

ТВ — время, прошедшее с момента появления ИД до момента их отбора поисковыми средствами;

ТП — время, затрачиваемое на передачу ИД обработчику;

ТО — время, затрачиваемое обработчиком на подготовку данных для аналитика;

ТИП — время, затрачиваемое аналитиком на подготовку ИП;

ТР — время, затрачиваемое руководителем на принятие решения;

ТД — время, затрачиваемое на выполнение действия по решению руководителя.

Тогда в системе высокой доступности данных должно обеспечиваться условие: Т < ТРЕГ.

Высокая доступность требуемых для своевременного принятия решения сведений из ЭИ определяется не только временем доступа, но и своевременным выявлением их наличия.

Системы, которые накапливают и обрабатывают полностью структурированные данные, имеют результат поиска в конкретной БД по конкретному поисковому предписанию, полностью соответствующий состоянию БД на данный момент, обеспечивается стопроцентная полнота и стопроцентная точность. Аналитический ИП — всегда результат обработки структурированных данных. Поэтому если для решения аналитической задачи требуется привлечь неструктурированные или слабо структурированные данные, то необходимо иметь средство их преобразования в структуру.

Система должна быть ориентирована на обеспечение пользователей с учетом специфики их деятельности и соблюдение необходимых мер безопасности. При этом отдельные пользователи могут выступать как в качестве потребителей специально подобранной для них информации, так и в качестве ее поставщиков по заданиям или инициативно. Проблемы, которые при этом приходится решать, имеют универсальный характер. Прежде всего они связаны с созданием рациональной «человекомашинной» системы использования электронных источников, включающей информационные, технологические и организационные составляющие.

Потенциальные опасности, которые могут сбить с толку при проявлении инициативы перехода на аналитику BD; — это отсутствие внутри организации аналитиков с необходимыми навыками и высокая стоимость найма опытных профессиональных аналитиков, а также проблемы интеграции новых технологий и действующих хранилищ данных, хотя продавцы начинают предлагать специальное программное обеспечение для такого рода соединений.

Маркетологи вендоров аппаратного и программного обеспечения начали перемаркировывать каждый продукт и решение на BD; реляционные и другие традиционные подходы обработки бросались в «общий котел». Подход можно было бы считать лицемерным, но в действительности это подчеркивает мысль, высказанную ранее: BD — это вся цифровая информация, которая существует и когда-либо собиралась и производилась, включая традиционные транзакционные, основные и информационные данные, которые при помощи ЭИТ собраны, произведены и управлялись с незапамятных времен.

Традиционные данные составляют меньше 10% цифровой информации, которой управляет бизнес. Доля традиционной реляционной технологии уменьшается в ИТ-бюджете и составляет 15–25%, но в большинстве случаев все еще превышает расходы на программное обеспечение NoSQL. При этом реляционная технология существенно развилась в направлении работы с большими объемами данных и достижения более высоких скоростей обработки. Massively Parallel Processing (MPP), колоночные и расположенные в памяти базы данных позволили реляционной технологии поддержать существенно большую нагрузку и более высокие скорости. [9, 10]

Литература

1. A Recommendation for High-Availability Options in TPC Bench - break marks by Dean Brock. Data General 2001 NPC. http://www.tpc.org/information/other/articles/ha.asp

2. Будзко В.И., Соколов И.А., Синицин И.Н. Построение информационно-телекоммуникационных систем высокой доступности // Системы высокой доступности. 2005. № 1, т. 1. С. 6-14.

3. Будзко В.И., Шмид А.В. О создании отказоустойчивых и катастрофоустойчивых центров коллективной обработки банковской информации // 6-я Всероссийская конференция «Информационная безопасность России в условиях глобального информационного общества»: сборник материалов. Москва, 2004. С. 145-158

4. Preimesberger, Chris. Hadoop, Yahoo, 'Big Data' Brighten BI Future (англ.). EWeek (15 August 2011).

5. Будзко В.И., Леонов Д.В., Николаев В.С., Оныкий Б.Н., Соколина К.А. Мультиагентные информационно-аналитические системы (МИАС) по научно-техническим направлениям // Системы высокой доступности. 2011. № 4, т. 7. С. 5-16

6. Steel Robert D. ON INTELLIGENCE: Spies and Secrecy in an Open World. AFCIA International Press (AIP), USA, 2000.

7. Budzko V. The Russian Viewpoint on Electronic Open Source Technologies // Proceedings of conference «Open Source Solutions 21», Washington, D.C., USA, 15-17 May 2000, pp. 61-69.

8. Budzko V. Electronic open sources. Technology of application // Proceedings of conference «Open Source Solutions 21», Washington, D.C., USA, 15-17 May 2000, pp. 70-76.

9. Zikopoulos P. C., Eaton C., deRoos D, Deutsch T., Lapis G. Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data. The McGraw-Hill Companies, 2012, 141pp.

10. Zikopoulos P. C., deRoos D., Parasuraman K., Deutsch T., Corrigan D., Giles J. Harness the Power of Big Data The IBM Big Data Platform. The McGraw-Hill Companies, 2013, 242 pp.

YT — эволюция системы распределенных вычислений

Бабенко М. А., Пузыревский И. В. (*sandello@yandex-team.ru*) — «Яндекс» (Москва)

В течение последних трех лет мы разработали, реализовали и внедрили YT — новую платформу для хранения и обработки больших объемов статистических и аналитических данных. Платформа задумывалась как замена существующей в «Яндексе» с 2008 года MapReduce-подобной системе обработки данных с улучшенными показателями эффективности, доступности и масштабируемости. В данном докладе мы хотели бы дать краткий обзор развитию технологии распределенных вычислений, поделиться опытом, полученным в процессе разработки и эксплуатации новой системы.

Введение

В современных компаниях, занимающихся информационными технологиями, все чаще возникают задачи, связанные с обработкой больших массивов данных. К примеру, в компании «Яндекс» каждый день обрабатываются петабайты информации для решения широкого спектра вопросов: анализ статистики работы сервисов, поведения пользователей, обработка текстов интернет-документов, картинок, временных рядов. Многие из вышеперечисленных задач допускают эффективное и масштабируемое решение в модели вычислений MapReduce.

Модель вычислений MapReduce была представлена в 2004 году [1]; первая внутренняя система в «Яндексе», поддерживающая вычисления в MR-модели, появилась в 2006 году и

развивается и эксплуатируется до сих пор (далее по тексту — YAMR). Однако к 2011 году стали очевидными некоторые проблемы:

- единая точка отказа в виде мастер-сервера;
- совмещение точки координации (мастер-сервера) и точки планирования (планировщика) ограничивает возможность масштабирования;
- слабая поддержка метаданных и возможностей интроспекции состояния системы;
- плохая модульность системы, усложняющая ее развитие и поддержку.

Платформа YT появилась как замена существующей системе с улучшенными показателями эффективности, доступности и масштабируемости. Далее мы рассмотрим ключевые отличия от предшественника, причины изменений, а также опишем планы развития платформы в ближайшем будущем.

Реплицированное состояние

Как уже было сказано, в архитектуре системы YAMR присутствует единая точка отказа — мастер-сервер, хранящий метаданные о расположении блоков информации и обеспечивающий фоновые процессы их репликации и балансировки. Сбой в работе данной компоненты ведет к полной недоступности кластера и его длительному перезапуску.

В системе YT сервис с метаданными был реализован как реплицированный конечный автомат (Replicated Finite-State Machine). Используемая техника репликации подтверждает внесение изменений в (мета-) состояние только при наличии активного кворума на изменение, что позволяет обеспечивать работоспособность сервиса в случае недоступности меньшей части машин-копий.

Стоит отметить, что для реализации надежного транспорта событий на кворум машин используется алгоритм, основанный на ведении журнала изменений [2, 3], а не на разрешении задачи консенсуса [4]. Это позволило сделать поведение системы простым и доступным для понимания.

С технической стороны реализованная нами модель репликации позволяет обрабатывать порядка 100 тыс. событий в секунду. Также наличие отказоустойчивости к единичным сбоям позволяет производить «мягкие» обновления системы без глобального перезапуска путем поочередного обновления машин, что оказалось полезным на практике.

Планировщик

В системе YAMR планировщик был совмещен с мастер-сервером, что ограничивало масштабируемость. В системе YT эти два сервиса разделены. Отказоустойчивость планировщика достигается за счет периодического создания слепков состояния и поддержания теневых копий сервиса, которые перехватывают планирование в случае отказа основной копии сервиса.

Практическое использование кластера также показывает, что модель планирования с одним ресурсом не позволяет максимально использовать ресурсы кластера, так как запускаемые пользователями операции весьма существенно варьируются в потреблении памяти и процессорного времени. Для более эффективной утилизации ресурсов была внедрена мультикритериальная система планирования. В настоящий момент при планировании исполнения учитываются потребности в памяти, вычислительных ядрах и сетевой полосе.

Хранение данных

В качестве модели данных и в системе YAMR, и в системе YT используются несхематизированные таблицы — наборы записей, состоящих из пар «ключ-значение». Данная модель зарекомендовала себя как успешная, потому что дает возможность пользователю системы не заботиться о вопросах хранения данных и целиком сфокусироваться на решаемой задаче.

В системах распределенного хранения данных важным аспектом является отказоустойчивость. Стандартным средством ее обеспечения является репликация. К примеру, сохранение данных в трех копиях позволяет системе работать при отказе двух машин. Именно такой подход и использовался в системе YAMR для обеспечения гарантий сохранности данных.

Однако с ростом объемов данных возникла потребность в более эффективной схеме хранения информации. В системе YT реализована поддержка кодов Рида — Соломона, позволяющая обеспечивать гарантии по устойчивости к тому же количеству отказов, что и при репликации данных, но с меньшими накладными расходами. К примеру, трехкратная репликация требует 300% объема для хранения данных, в то время как кодирование по схеме RS (6,3) требует 150% объема при тех же гарантиях.

С кодами Рида — Соломона тем не менее связана другая проблема хранения: для восстановления данных после сбоя нужно прочитывать большой объем данных. К примеру, для схемы RS (12,3) нужно прочитать 12-кратный объем данных для восстановления после сбоя. Для решения данной проблемы были внедрены так называемые локально-реконструируемые коды [5], которые позволяют вдвое снизить требования по просматриваемому объему данных.

* **

В настоящий момент система YT успешно и эффективно эксплуатируется в компании «Яндекс» в разных отделах. Будущие направления развития системы включают масштабирование на новые объемы данных и размеров кластеров, а также привнесение интерактивной функциональности, позволяющей строить процессы обработки данных с субсекундной латентностью.

Литература:

1. Jeffrey Dean, Sanjay Ghemawat. MapReduce: Simplified Data Processing on Large Clusters. OSDI'04: Sixth Symposium on Operating System Design and Implementation, San Francisco, 2004.

2. Brian M. Oki, Barbara H. Liskov. Viewstamped replication: A new primary copy method to support highly-available distributed systems. PODC'88: Proceedings of the seventh annual ACM Symposium on Principles of distributed computing, New York, 1988.

3. Diego Ongaro, John Ousterhout. In Search of an Understandable Consensus Algorithm.

In draft, Stanford University, 2013.

4. Tushar Chandra, Robert Griesemer, Joshua Redstone. Paxos made live: an engineering perspective. PODC'07: Proceedings of the twenty-sixth annual ACM symposium on Principles of distributed computing, New York, 2007.

5. Cheng Huang, Huseyin Simitci, Yikang Xu, Aaron Ogus, Brad Calder, Parikshit Gopalan, Jin Li, and Sergey Yekhanin. Erasure coding in Windows Azure storage. USENIX, ATC, 2012.

Территория Больших Данных: Terra Incognita?

Агамирзян И. Р. — РВК (Москва)

Я собираюсь говорить совершенно не о технических деталях Больших Данных, а скорее о методических вещах и позиционировании этого рынка и его перспективности с точки зрения инвестиционной привлекательности и того места, которое он несомненно должен занять в экономике.

Прежде всего хотелось бы подчеркнуть, что человечество обрабатывает большие данные уже давно. Мне довелось несколько лет проработать в корпорации EMC, занимаясь программным обеспечением систем хранения данных. Очень похожие задачи — обеспечение отказоустойчивости, эффективности хранения данных, систем резервирования — мы уже тогда успешно решали, хотя термин Big Data еще не был по-настоящему известен.

В течение последних двух десятилетий происходит экспоненциальный рост по очень многим характеристикам индустрии, связанной с обработкой больших объемов данных. Понятно, что он не может продолжаться бесконечно, поэтому уже довольно многие из тех, кто занимается научно-техническим и экономическим прогнозированием, ведут разговор о сингулярности. Модель развития должна измениться — по разным прогнозам, до этого осталось от 10 до 20 лет.

Традиционные правила ведения бизнеса в текущей ситуации просто перестают работать. Поэтому сегодня нам не просто нужно принимать правильные решения, а необходимо принимать их быстро. Неточные прогнозы приводят к крушениям бизнеса. Если обратиться к истории компьютерного бизнеса, то можно наблюдать непрерывную череду самоубийств лидирующих компаний, связанную с принятием неправильных решений и выбором неверных стратегий. Поэтому в сегодняшнем мире, в пока ускоряющемся развитии экономики и технологий необходимо вырабатывать реальный инструментарий для своевременного принятия правильных решений.

Большие Данные предлагают для этого новый и перспективный подход, связанный с выявлением закономерностей в структурированных и неструктурированных данных. Тем не менее очень важно понимать, что так же, как и в других популярных направлениях, в области Больших Данных всегда присутствует смесь маркетингового, технологического и экономического подходов.

То же самое относится к облачным вычислениям — маркетинговому обозначению для целой совокупности технологий, которые позволяют нам построить новую экономическую модель. Набор технологий для облачных вычислений известен по крайней мере последние 15 лет. Конкретная экономическая модель перевода капитальных затрат в оперативные доходы и расходы разработана, привлекательна и просчитываема, и сегодня мы наблюдаем бум интереса к облачным вычислениям.

В значительной мере это относится и к Большим Данным. Большие Данные были всегда. Возможно, сегодня их объемы действительно становятся огромными, но, имея опыт работы в корпорации EMC, которая на протяжении 25 лет занимается хранением и обработкой больших объемов данных в чрезвычайно ответственных отраслях (90% финансовых транзакций

во всем мире производится на оборудовании EMC), я прекрасно понимаю, что почти все наработки в этой области существуют достаточно давно. За исключением, возможно, определенных технологий — например, Map Reduce, разработанной Google около 10 лет назад.

Тем не менее, как о явлении, о Больших Данных стали говорить около пяти лет назад. Сегодня вокруг этого направления развился большой маркетинг и реально стал выделяться соответствующий рынок, который растет очень высокими темпами. Практически все представители агентств и корпораций предсказывают не только значительные темпы роста, но и называют абсолютные объемы этого рынка, уже сегодня достигшие больших значений и миллиардных оборотов.

Последние несколько лет шли массовые эксперименты, результаты которых крупными компаниями типа Google и «Яндекс» реализовывались в практических продуктах. Но сегодня происходит процесс выхода технологий, связанных с Большими Данными, на уровень корпоративного использования. Почти половина руководителей во всем мире, ставших участниками опроса Gartner, либо уже сделали вложения, либо собираются инвестировать в решения Больших Данных.

Сегодняшний мир — это мир, генерирующий данные. Трудно назвать область деятельности, которая не была бы связана с огромными потоками данных. Самолет Dreamliner за время трансатлантического перелета собирает несколько десятков гигабайт телеметрических данных. Правда, пока не совсем понятно, что со всем этим делать. В истории технологий есть совершенно потрясающие факты — за 60 с лишним лет развития космонавтики из всего объема телеметрии, которая была собрана и хранится до сих пор на самых разных носителях, реально было использовано не более полутора процентов. Особенностью технологий Больших Данных как раз и является то, что они, пожалуй, впервые позволяют обрабатывать данные из большого числа самых разнообразных источников, сопоставляя ранее несопоставимые наблюдения, структурированную и неструктурированную информацию, и извлекать из них новые знания.

Рынок формируется, и ключевыми игроками на нем становятся не только крупные транснациональные корпорации, но и государства, которые стремятся все зарегулировать (в ряде стран уже принимаются национальные программы, связанные не только с образованием и подготовкой специалистов, но и, например, с медициной), и интеграторы, стремящиеся заработать на предоставлении и продвижении соответствующих решений. А также наиболее интересная для меня отрасль — стартапы и венчурная инфраструктура.

Рост венчурной индустрии Больших Данных вполне заметен. В 2012 году значительное число венчурных компаний приобрели новых инвесторов или привлекли инвестиции в ходе размещения акций. Капитализация лидеров в этой отрасли уже измеряется миллиардами долларов.

Сегодня, как нам представляется, для компаний есть определенное «окно возможностей» для вступления на новый рынок, когда ценой относительно небольших, но своевременных усилий можно вырасти в глобального лидера. Такое «окно возможностей» обычно очень короткое и быстро закрывается. После этого возникает конкурентная среда с большим количеством слияний и поглощений, в ходе которых более удачливые участники рынка скупают менее удачливых. Это происходило и происходит во всех индустриях: автомобильной, авиационной, компьютерной...

«Окно возможностей» в Интернете закрылось в конце 90-х годов. В сфере приложений для социальных сетей это, скорее всего, произошло лет 5–6 назад. «Окно возможностей» в

области облачных вычислений и технологий Больших Данных пока еще открыто. Но его закрытие — вопрос ближайших нескольких лет, если не месяцев.

Для России, ее инновационной экономики направление Больших Данных имеет высокий потенциал в силу ряда причин. Одной из них является то, что у нас в стране все еще сохраняются сильные научные и математические школы. Технологии Больших Данных связаны со сложной аналитикой и, по крайней мере потенциально, могут предоставить обширное поле деятельности для российских ученых и разработчиков.

Кроме того, сегодня вокруг тематики Больших Данных складывается собственная экосистема, которая наследует многие элементы инфраструктуры информационных технологий и в первую очередь ориентацию на открытое программное обеспечение, что делает доступным самый современный инструментарий и значительно понижает порог вступления на территорию Больших Данных.

Современные тенденции Больших Данных. Взгляд технолога

Позин Б. А. ***(bpozin@ec-leasing.ru)*** **— «ЕС-лизинг» (Москва)**

Технология Больших Данных (БД) — это прежде всего возможность значительно повысить способность компаний и органов управления страны в области анализа данных и подготовки принятия решений в самых разных областях на основе оперативной и достоверной информации об управляемом объекте. Для проведения анализа разных видов информации необходимы математические, экономические и другие методы, и их развитие весьма важно.

Роль технологии БД в национальной экономике будет повышаться по мере того и в той степени, как и в которой технология БД будет использоваться для решения практических задач компаний.

Объемы производства и в этом секторе, как и ранее, будут определяться как числом работающих, так и производительностью их труда. Под работающими здесь и далее имеются в виду не специалисты в области ИТ или математики, а специалисты-предметники: врачи, управленцы среднего и высокого уровня, работающие на заводах, фабриках, торговых предприятиях оптовой и розничной торговли, и широкий спектр специалистов самых разных специальностей. Они должны быть оснащены такими инструментами анализа, с которыми смогут общаться на языке своей предметной области и получать ответы, интерпретируемые в понятиях их предметной области.

Сегодня таких специалистов не готовят нигде в мире. Вместе с тем, чтобы использовать преимущества технологии БД, важно как можно быстрее начать их использовать. И это возможно уже сегодня. Надо начинать работать на готовых платформах БД, чтобы обучить специалистов по информационно-аналитическим системам, создать работающие в компаниях технологии, а затем совершенствовать как технологии, так и используемые в них методы.

Применение технологии БД при решении задач управления компаниями и корпорациями

В настоящее время компаниями на основе наработанных структурированных данных о ходе оперативной деятельности проводится анализ степени соответствия состояния компании ее среднесрочной стратегии.

Использование только внутренних данных не позволяет выявить новые внешние факторы, которые могут повлиять не только на тактику, но и на стратегию компании на среднесрочном интервале. Для подготовки принятия подобных решений необходимо привлекать данные, внешние для компании, то есть не вырабатываемые ею: сведения о стратегии и тактике государства, состоянии и тенденциях рынка, данные о конкурентах, перспективных и реальных клиентах, их планах и действиях. Эти данные далеко не всегда являются структурированными, их необходимо добывать, исследовать и преобразовывать в информацию новыми методами, которые при анализе структурированных данных ранее не применялись. Эти методы относятся в настоящее время к технологии БД.

Кроме того, в связи с появлением новых высокоэффективных методов обработки данных возникают новые возможности в обработке данных «на проходе», с выделением наиболее интересных из них по критериям, вырабатываемым конечными пользователями. Это данные, которые могут использоваться как в оперативном, так и в тактическом управлении компанией — например, обработка сотен тысяч и миллионов единиц данных о функционирующих технических средствах компании (в том числе и географически распределенных) для снижения издержек на их эксплуатацию и предотвращения нештатных ситуаций, анализ свойств потоков покупателей, в частности для определения устойчивого спроса на товары в крупных супермаркетах, и т. п.

Возможность быстро реагировать на внутренние события компании и на внешние воздействия, быстрое моделирование и прогнозирование развития ситуации для поддержки принятия управленческих решений позволят компаниям превратить использование обеспечивающей эти возможности технологии БД в конкурентное преимущество.

Две тенденции в развитии технологии БД

Научно-технические и популярные издания активно рекламируют очевидные преимущества технологии БД. Однако свидетельств об успешно решенных с использованием этой технологии бизнес-задачах пока не так много.

Опыт разработки и внедрения информационно-аналитических систем (ИАС), анализирующих структурированные данные, показывает, что конечный пользователь начинает осваивать технологию использования средств автоматизации анализа и постановки соответствующих задач по мере создания для него «полной» ИАС, подключенной к реальным источникам данных. Именно она является «толчком» к постановке новых прикладных аналитических задач.

С этой точки зрения в области БД в настоящее время существуют две тенденции:

- освоение и формирование новых методик решения практически значимых задач на основе доступных комплексов инструментальных средств, или платформ;
- развитие методов и технологий БД и их имплементация в состав доступных платформ.

Обе тенденции требуют интенсивной подготовки кадров, владеющих соответствующими знаниями, методиками и инструментами.

Креативные команды для решения задач в междисциплинарной области и требования к платформам БД

Задачами исследования данных для поддержки принятия решений обычно занимается креативная команда, составленная из аналитиков, прикладников и специалистов по работе с данными и их обработке конкретными, в том числе и математическими, методами. Для обе-

спечения креативной команды соответствующими методами, методиками и инструментами существует комплекс требований к платформе Больших Данных:

- платформа должна быть ориентирована на использование командой различных специалистов-предметников и ИТ-специалистов при поиске, создании и использовании механизмов решения новых задач в комплексной, междисциплинарной области;
- промежуточные и финальные результаты, получаемые разными членами команды, должны интегрироваться на средствах визуализации для принятия решений по результатам работы команды в целом;
- инструменты, входящие в состав платформы, должны содержать средства постановки задачи на языке предметников, а не ИТ-специалистов; обеспечивать решение основных задач в предметной области с использованием структурированных, неструктурированных и потоковых данных; быть проинтегрированы по управлению и по данным (метаданным), иметь открытые механизмы (правила и средства) подключения новых инструментов и встраивания новых методов/алгоритмов.

В настоящее время этому комплексу требований в наибольшей степени удовлетворяет платформа IBM Big Data, состав и структура которой приведены в докладе. Эта платформа содержит более 600 реализованных методов и алгоритмов обработки и анализа данных разных типов.

Обеспечение жизненного цикла ИАС, построенных на базе платформы БД

Опыт построения ИАС на базе элементов платформы IBM Big Data показывает, что важнейшим вопросом при ее использовании является создание не только собственно «полной» ИАС, но и определение того, как должны работать использующие ее люди, как она должна сопровождаться и развиваться в процессе ее жизненного цикла — 10–15 лет и, возможно, более. Рассмотрены основные цели, задачи и опыт решения задач обеспечения жизненного цикла ИАС.

Большие данные и экономические модели организации вычислений в распределенных средах

Топорков В. В. (*toporkovvv@mpei.ru*) — НИУ МЭИ (Москва)

Работа выполнена при частичном содействии Совета по грантам Президента Российской Федерации для поддержки ведущих научных школ (шифр НШ-316.2012.9), Российского фонда фундаментальных исследований (грант № 12-07-00042), Министерства образования и науки Российской Федерации в рамках федеральной целевой программы «Научные и научно-педагогические кадры инновационной России» на 2009–2013 годы (государственный контракт № 16.740.11.0516).

Введение

Большие задачи, например обработка данных физических экспериментов на LHC (ЦЕРН), зачастую требуют привлечения распределенных вычислительных ресурсов, часть из которых используется совместно с их владельцами [1–4]. Среди различных подходов к планированию вычислений в распределенных средах можно выявить следующие тенденции. Одна из них основывается на использовании доступных ресурсов и планировании вычислений на

уровне приложений. Зачастую при этом не предполагается наличия какого-либо регламента в предоставлении ресурсов [5]. Роль посредников между пользователями и вычислительными узлами выполняют агенты приложений [6–11] — брокеры ресурсов. Другая тенденция связана с образованием виртуальных организаций (ВО) пользователей и предполагает планирование на уровне потоков заданий [12–15]. Наличие определенных правил предоставления и потребления ресурсов в ВО, основанных, в частности, на экономических моделях [1–4, 15–18], позволяет повысить эффективность планирования и распределения ресурсов на уровне потоков заданий.

Идея «конвергенции» вышеупомянутых подходов была декларирована довольно давно [14, 19–21]. Однако попытки ее реализации обладают рядом серьезных ограничений. В одних из известных моделей отыскивается лишь подходящий набор ресурсов [22–24] и не поддерживаются оптимизационные механизмы планирования заданий. В других моделях [14, 16, 17] не представлены аспекты, связанные с динамикой изменения загрузки узлов, конкуренцией независимых пользователей, а также глобальных и локальных потоков заданий.

Предлагается модель справедливого разделения ресурсов для управления выполнением независимых заданий на основе экономических принципов [25–27].

Справедливое разделение ресурсов

В [26, 27] предложена циклическая схема планирования (ЦСП) на основе динамично обновляемых расписаний выполнения заданий в локальных процессорных узлах (рис. 1). Среди основных ограничений ЦСП можно выделить следующие. Отсутствует возможность влиять на ход выполнения отдельного пользовательского задания: поиск отдельных альтернатив осуществляется по принципу «первая подходящая», а выбор их оптимальной комбинации отражает интересы всей ВО. Таким образом, не учитываются предпочтения отдельных пользователей, что не позволяет говорить о справедливом разделении ресурсов. Планирование выполнения пакета заданий основывается на пользовательской, зачастую весьма неточной, оценке времени выполнения конкретного приложения. В случае несостоятельности оценки преждевременно высвобожденные ресурсы могут простаивать, тем самым понижая уровень загрузки среды. Для планирования пакета заданий необходимо выделение нескольких «непересекающихся» (по занятым ресурсам и времени) альтернатив, а для выполнения задания выбирается одна альтернатива. Это может приводить к фрагментации ресурсов, в преодолении которой может помочь использование бэкфиллинга [28].

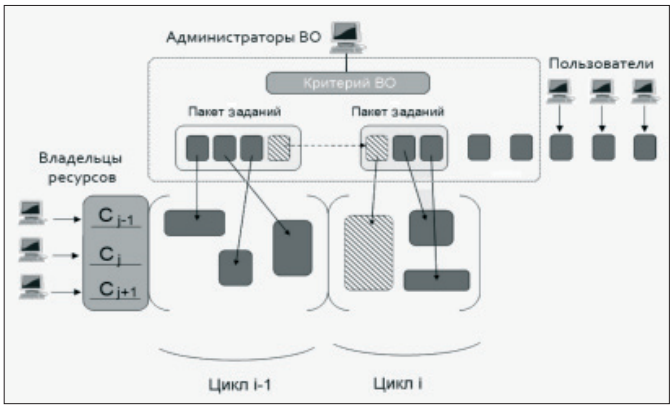


Рис. 1. Цикличное планирование потока заданий

В модифицированной схеме для учета интересов пользователей в формат ресурсного запроса дополнительно вводится критерий оптимизации (рис. 2). Для выбора альтернатив, оптимальных по заданному критерию, используется алгоритм AEP (Algorithm searching for Extreme Performance), подробно описанный в [29].

Рис. 2. Пользовательские критерии оптимизации выполнения заданий

Другое отличие МПСР от ЦСП состоит в алгоритме формирования системы заданий. МПСР предполагает разделение исходного пакета заданий на множество подпакетов и планирование каждого подпакета в отдельности на заданном интервале планирования (рис. 3). Необходимость процедуры «разрезания» пакета в МПСР особенно ярко проявляется при высоком уровне загрузки ресурсов среды.

МПСР оптимизирует выполнение потока заданий в соответствии с интересами участников ВО при условии, что найдены альтернативные наборы слотов для выполнения заданий. Бэкфиллинг реагирует на досрочное освобождение ресурсов и позволяет производить перепланирование «на лету», что очень важно в условиях, когда оценка времени выполнения и действительное время выполнения задания могут существенно отличаться. Предлагается комбинированный подход. В каждом цикле планирования из исходного пакета выделяется подмножество приоритетных заданий — например, самых «дорогих» (по стоимости выполнения) либо самых требовательных к ресурсу (по времени выполнения). Эти задания группируются в отдельный подпакет и планируются, возможно, и без соблюдения дисциплины очереди. Далее планирование этого подпакета производится с помощью МПСР на основе актуального расписания загрузки среды на рассматриваемом интервале планирования. Планирование остальных, возможно менее требовательных к ресурсам, заданий осуществляется с помощью бэкфиллинга на основе динамично обновляющейся информации о реальной загрузке узлов.

Эксперименты

Табл. 1 демонстрирует результаты планирования отдельных заданий в зависимости от заданного пользователем критерия оптимизации: времени старта и завершения задания, времени и стоимости выполнения задания в относительных единицах симулятора [27]. AEP минимизирует значение заданного критерия.

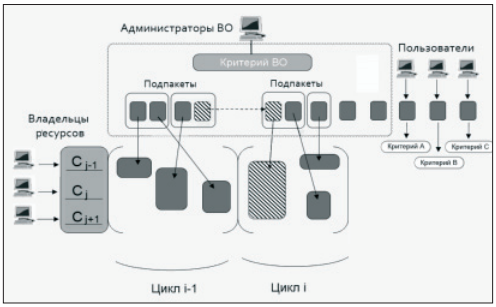


Рис. 3. Планирование потока заданий с разделением на подпакеты

Таблица 1. Учет интересов пользователей ВО

Критерий	*Na*	Время старта	Время выполнения	Время завершения	Стоимость
Время старта	12,8	171,7	56,1	227,8	1281,1
Время выполнения	10,6	214,5	39,3	253,9	1278,5
Время завершения	12,2	169,6	45	205,5	1283,2
Стоимость	12,9	262,6	55,5	318	1098,3
ЦСП	12,1	222	50,3	272,3	1248,4

Использование критерия оптимизации в МПСР при выполнении отдельных заданий позволяет более чем на 23% уменьшить время старта завершения заданий, на 21% уменьшить общее время выполнения и на 12% снизить стоимость выполнения по сравнению с ЦСП. Среднее количество Na альтернатив выполнения в табл. 1, найденных для заданий на одном цикле планирования, практически не зависит от выбора критерия оптимизации.

Следующий эксперимент посвящен сравнительному анализу результатов планирования при делении исходного пакета заданий в МПСР на различное число подпакетов при разных уровнях загрузки среды. При выборе оптимальной комбинации альтернатив выполнения решалась задача минимизации среднего процессорного времени выполнения заданий $T_{проц}$. Процессорное время для альтернативы вычисляется как сумма длительностей слотов, входящих в сформированное «окно». На рис. 4 представлены значения $T_{проц}$ в зависимости от числа k подпакетов, на которое делится исходный пакет заданий, и уровня загрузки среды. Уровень загрузки среды при проведении серии экспериментов определяется относительным числом Y неудач — циклов планирования, в ходе которых не удается найти план выполнения для всех заданий пакета. Эксперименты проводились при высоком (Y = 0,3), среднем (Y = 0,03) и низком уровнях загрузки (Y < 0,0002) (на рис. 4 — ВЗС, СЗС и НЗС соответственно).

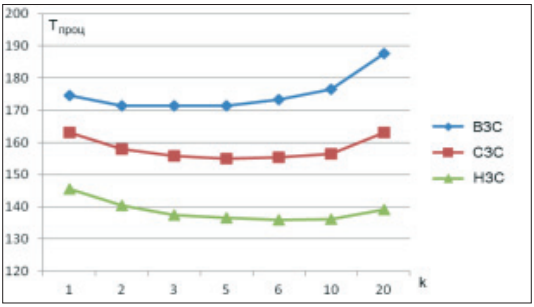


Рис. 4. Зависимость среднего времени $T_{проц}$ выполнения заданий пакета от числа k подпакетов

Анализ результатов показывает, что увеличение числа формируемых подпакетов обуславливает рост числа альтернатив для выполнения отдельного задания, уменьшение суммарной стоимости прохождения пакета заданий и увеличение относительного числа неудач Y.

Табл. 2 демонстрирует результаты планирования в МПСР с позиций учета интересов владельцев ресурсов в зависимости от удельной стоимости c, назначенной на интервале планирования T = 600. В табл. 2: Lc — суммарное процессорное время загрузки узла на интервале планирования; U — средняя относительная величина загрузки ресурса на рассматриваемом интервале планирования, P — средняя прибыль, полученная владельцем ресурса, Y — относительное количество неудачных попыток планирования.

Из табл. 2 видно, что владельцы ресурсов могут управлять собственной прибылью P и уровнем загрузки U вычислительных узлов на интервале планирования T путем выставления удельной стоимости c использования узла. Экстремум прибыли достигается при выставлении стоимости, близкой к «среднерыночной», то есть средней стоимости за экземпляр ресурса аналогичной производительности, выставленной другими владельцами ресурсов.

Таблица. 2. Учет интересов владельцев ресурсов ВО

c	*Lc*	*U*	*P*	*Y*
2	256,6	0,44	527,1	0
4	234,9	0,39	939,6	0,001
6	185,4	0,31	1112,3	0,013
8	109,8	0,18	878,7	0,024
10	71	0,12	710,3	0,025

На рис. 5 представлены среднее время выполнения $T_{проц}$ и среднее время старта $T_{старт}$ заданий пакета в зависимости от соотношения $N_{цсп}/N$, в котором происходит разделение N заданий на подпакеты, $N_{цсп}$ — число заданий в первом подпакете, планирование которых осуществляется на основе ЦСП. Из рис. 5 видно, что, если планирование большей части задания осуществлять с помощью МПСР, достигается лучшее значения целевого критерия ВО — времени выполнения $T_{проц}$, при этом среднее время старта $T_{старт}$ заданий отодвигается.

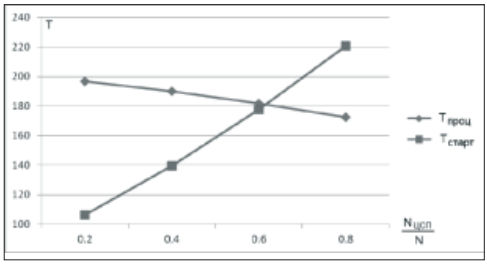


Рис. 5. Среднее время выполнения $T_{проц}$ и старта $T_{старт}$ заданий в МПБ

Для исследования и сравнения эффективности планов, построенных ЦСП, МПСР и бэкфиллингом, было проведено имитационное моделирование, в котором реальное время выполнения заданий существенно отличается от времени предварительного резервирования ресурсов. Фактическое время выполнения заданий задавалось случайной величиной, подчиняющейся равномерному распределению на интервале $[0{,}2 * T_{рез}, T_{рез}]$, где $T_{рез}$ — время резервирования ресурсов, отведенное для выполнения задания.

В табл. 3 приведены значения среднего времени выполнения (критерия оптимизации) и среднего времени старта заданий, полученные на этапе предварительного планирования на основе оценки времени выполнения заданий $T_{рез}$ (колонка «Запланированное») и в результате моделирования выполнения составленного плана с учетом фактического времени выполнения заданий (колонка «Фактическое»). Из табл. 3 видно, что даже при значительной разнице между временем резервирования ресурсов и фактическим временем выполнения заданий преимущество МПСР перед бэкфиллингом по целевому критерию оптимизации в ВО не только сохраняется, но и увеличивается.

Таблица 3. Фактическое и запланированное время выполнения заданий

Алгоритм	Время выполнения		Среднее время старта заданий	
	Запланированное	Фактическое	Запланированное	Фактическое
Бэкфиллинг	187,7	115,1	69	37,3
ЦСП	150,1	90,4	281,2	281,2
МПСР	138,6	83,5	223,8	223,8
Преимущество МПСР перед бэкфиллингом	26,2%	27,5%	-69%	-83%

Выводы

В докладе рассмотрены стратегии планирования с различными целевыми критериями, основанные на модели справедливого разделения ресурсов, учитывающей предпочтения всех участников ВО. Предложен подход к решению задачи справедливого разделения ресурсов между участниками ВО. Показана состоятельность планов, составленных МПСР, в условиях, когда фактическое время выполнения заданий может существенно отличаться от пользовательской оценки.

Литература

1. Garg, S.K., Buyya, R., Siegel, H.J.: Scheduling Parallel Applications on Utility Grids: Time and Cost Trade-off Management. In: 32nd Australasian Computer Science Conference (ACSC 2009), pp. 151-159. Wellington, New Zealand (2009).

2. Degabriele, J.P., Pym, D.: Economic Aspects of a Utility Computing Service, Trusted Systems Laboratory HP Laboratories Bristol HPL-2007-101. Technical Report, July 3, pp. 1-23 (2007).

3. Garg ,S.K., Yeo, C.S., Anandasivam, A., Buyya ,R.: Environment-conscious Scheduling of HPC Applications on Distributed Cloud-oriented Data Centers. J. Parallel and Distributed Computing 71, 6, 732-749 (2011).

4. Tesauro, G., Bredin, J.L.: Strategic Sequential Bidding in Auctions Using Dynamic Programming. In: 1st International Joint Conference on Autonomous Agents and Multiagent Systems, part 2, pp. 591-598. ACM, New York (2002).

5. 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.

6. Berman, F.: High-performance Schedulers. In: Foster, I., Kesselman, C. (eds.). The Grid: Blueprint for a New Computing Infrastructure. Morgan Kaufmann, San Francisco, pp. 279-309 (1999).

7. Yang, Y., Raadt, K., Casanova, H.: Multiround Algorithms for Scheduling Divisible Loads. IEEE Trans. Parallel and Distributed Systems 16, 8, 1092-1102 (2005).

8. Natrajan, A., Humphrey, M.A., Grimshaw, A.S.: Grid Resource Management in Legion. In: Nabrzyski, J., Schopf, J.M., Weglarz J. (eds.). Grid resource management. State of the art and future trends. Kluwer Academic Publishers, Boston, pp. 145-160 (2003).

9. Beiriger, J., Johnson, W., Bivens, H.: Constructing the ASCI Grid. In: 9th IEEE Symposium on High Performance Distributed Computing, pp. 193-200. IEEE Press, New York (2000).

10. Frey, J., Foster, I., Livny, M.: Condor-G: a Computation Management Agent for Multi-institutional Grids. In: 10th International Symposium on High-Performance Distributed Computing, pp. 55-66. IEEE Press, New York (2001).

11. Abramson, D., Giddy J., Kotler L.: High Performance Parametric Modeling with Nimrod/G: Killer Application for the Global Grid? In: International Parallel and Distributed Processing Symposium, pp. 520-528. IEEE Press, New York (2000).

12. Foster, I., Kesselman C., Tuecke S.: The Anatomy of the Grid: Enabling Scalable Virtual Organizations. Int. J. of High Performance Computing Applications 15, 3, 200-222 (2001).

13. Ranganathan, K., Foster, I.: Decoupling Computation and Data Scheduling in Distributed Data-intensive Applications. In: 11th IEEE International Symposium on High Performance Distributed Computing, pp. 376-381. IEEE Press, New York (2002).

14. Kurowski, K., Nabrzyski, J., Oleksiak, A., Weglarz, J.: Multicriteria Aspects of Grid Resource Management. In: Nabrzyski, J., Schopf, J.M., Weglarz J. (eds.). Grid resource management. State of the art and future trends. Kluwer Academic Publishers, Boston, pp. 271-293 (2003).

15. Garg S.K., Konugurthi P., Buyya R.: A Linear Programming-driven Genetic Algorithm for Meta-scheduling on Utility Grids. J. Par., Emergent and Distr. Systems 26, 493-517 (2011).

16. Buyya, R., Abramson, D., Giddy, J.: Economic Models for Resource Management and Scheduling in Grid Computing. J. Concurrency and Computation 14, 5, 1507-1542 (2002).

17. Ernemann, C., Hamscher, V., Yahyapour, R.: Economic Scheduling in Grid Computing. In: Feitelson, D.G., Rudolph, L., Schwiegelshohn, U. (eds.) JSSPP 2002. LNCS, vol. 2537, pp. 128-152. Springer, Heidelberg (2002).

18. Lee, Y.C., Wang, C., Zomaya, A.Y., Zhou, B.B.: Profit-driven Scheduling for Cloud Services with Data Access Awareness. J. Par. and Distr. Computing 72, 4, 591-602 (2012).

19. Toporkov, V.V.: Job and Application-Level Scheduling in Distributed Computing. Ubiquitous Computing and Communication J. Applied Computing 4, 3, 559-570 (2009).

20. Toporkov, V.V., Toporkova, A., Tselishchev, A., Yemelyanov, D.: Job and Application-Level Scheduling: an Integrated Approach for Achieving Quality of Service in Distributed Computing. In: 4th International Conference on Dependability of Computer Systems, pp. 202-209. IEEE CS Press, Los Alamitos (2009).

21. Toporkov, V.: Application-Level and Job-Flow Scheduling: an Approach for Achieving Quality of Service in Distributed Computing. In: Malyshkin, V. (ed.) PaCT 2009. LNCS, vol. 5968, pp. 350-359. Springer, Heidelberg (2009).

22. Aida, K., Casanova, H.: Scheduling Mixed-parallel Applications with Advance Reservations. In: 17th IEEE Int. Symposium on HPDC, pp. 65-74. IEEE CS Press, New York (2008).

23. Ando, S., Aida, K.: Evaluation of Scheduling Algorithms for Advance Reservations. Information Processing Society of Japan SIG Notes. HPC-113, 37-42 (2007).

24. Elmroth, E., Tordsson, J.: A Standards-based Grid Resource Brokering Service Supporting Advance Reservations, Coallocation and Cross-Grid Interoperability. J. of Concurrency and Computation 25, 18, 2298-2335 (2009).

25. Toporkov, V., Toporkova, A., Bobchenkov, A., Yemelyanov, D.: Resource Selection Algorithms for Economic Scheduling in Distributed Systems. Procedia Computer Science. Elsevier 4, 2267-2276 (2011).

26. Топорков В.В. Пакетная обработка заданий в распределенных вычислительных средах с неотчуждаемыми ресурсами // АиТ. 2012. № 10. С. 52-70.

27. Toporkov, V., Tselishchev, A., Yemelyanov, D., Bobchenkov, A.: Dependable Strategies for Job-flows Dispatching and Scheduling in Virtual Organizations of Distributed Computing Environments. In: Complex Systems and Dependability. AISC, vol. 170, pp. 240-255. Springer, Heidelberg (2012).

28. Moab Adaptive Computing Suite, http://www.adaptivecomputing.com/products/moab-adaptive-computing-suite.php.

29. Toporkov, V., Toporkova, A., Tselishchev, A., Yemelyanov, D.: Slot Selection Algorithms for Economic Scheduling in Distributed Computing with High QoS Rates. In: New Results in Complex Syst.

Машинное обучение как инструмент современного ученого

Устюжанин А. Е. (*anaderi@yandex-team.ru*) — «Яндекс» (Москва)

Научные открытия сегодня требуют тщательной экспериментальной проверки большого количества гипотез. Для таких проверок необходимо уметь эффективно обрабатывать огромные объемы данных, извлекать из них закономерности. Для этих задач одних методов статистики бывает недостаточно. Следующий виток развития технологий анализа данных принято называть общим термином «машинное обучение».

Он не является чем-то кардинально новым. Методы и подходы машинного обучения разрабатываются со времен, когда люди начали извлекать информацию из сырых данных.

В настоящее время происходит лишь смещение акцентов благодаря развитию технологий хранения и виртуализации данных. Известным примером технологий машинного обучения являются нейросети.

Одним из важных отличий машинного обучения является то, что в этом подходе требуется построение оперативного прогноза по имеющимся данным и его не может построить человек (в силу различных обстоятельств: большие объемы информации, высокая скорость принятия решений). Такой прогноз может сделать машина. Простой пример — предсказание возраста, пола и уровня дохода человека по паттернам его поведения в Интернете (сервис «Яндекс.Крипта»). Или предсказание интереса этого человека к тому или иному рекламному объявлению. Построить такие прогнозы вручную вряд ли кому-то под силу. Но написать программу, которая строит модель, принимающую во внимание определенные признаки и предсказывающую ту или иную характеристику наблюдаемого объекта, оказывается возможно. Дальше на вопросы отвечает не человек, а модель. Удивительным оказывается то, что законы обучения таких моделей гораздо менее многочисленны, чем задачи реального мира. Например, один и тот же алгоритм может быть использован для предсказания интереса пользователя к выбранной тематике и для выборки очень редких событий, собранных Большим адронным коллайдером, или предсказания автомобильных пробок на дорогах. Библиотеки таких алгоритмов становятся все более и более доступны, если раньше для их использования требовалось изучение специальных языков и сред, то сейчас эти алгоритмы реализованы и на языках общего назначения типа Python (http://scikit-learn.org, http://beta.graphlab.com, http://nltk.org).

Можно выделить несколько важных качеств построенных моделей, которые необходимо иметь в виду при работе с ними: эффективность (в терминах выбранной задачи), устойчивость (насколько меняется качество с изменением обучающей выборки), скорость работы и скорость обучения.

Поэтому в арсенале современного ученого должны быть не только инструменты для анализа данных, построения моделей, но и средства оценки их эффективности, устойчивости и т. д.

Инструменты по проверке моделей/гипотез существуют сами по себе, но чтобы работа по их отладке/настройке и применению была максимальна продуктивна, необходимо нечто большее, нежели набор инструментов. Необходима интерактивная среда, в которой можно было бы описывать типовые алгоритмы работы с данными, обучения моделей и проверки их эффективности. Такая среда должна поддерживать различные форматы файлов и распределенное выполнение заданий, работу с предварительными выборками данных для быстрой оценки и проверки гипотез.

Также она должна удовлетворять таким свойствам, как:

- читаемость/прозрачность записанных алгоритмов,
- модульность,
- совместная работа,
- измеримость результатов,
- воспроизводимость результатов.

Здесь уместна аналогия создания промышленного конвейера, который Генри Форд применил для сборки автомобилей. Чтобы заметно продвинуться в решении задач, необходимо уметь эффективно разбивать задачи на составные части и собирать полученные результаты.

Многие интернет-компании используют собственные наработки для реализации такого процесса. Мы видим, что процессы машинного обучения не являются чем-то уникальным и могут найти применение в самых разных областях: науке, медицине, бизнесе. На базе наработок «Яндекса», которые в основном используются для решения задач интернет-поиска и других внутренних продуктов, мы разрабатываем среду, доступную специалистам любых отраслей и обладающую описанными свойствами. Эта среда получила название «Фабрика данных». Нам представляется, что она может оказаться полезной как аналитикам, работающим на современные компании, так и физикам, астрономам, биологам, ищущим ответы на загадки нашей Вселенной.

Решетки формальных понятий в современных методах анализа данных и знаний

Кузнецов С. О. (*skuznetsov@yandex.ru*), Игнатов Д. И. (*dmitrii.ignatov@gmail.com*) — НИУ ВШЭ (Москва)

В докладе рассказывается об анализе формальных понятий (АФП) и его роли в разработке моделей майнинга данных (data mining) и открытия знаний, приводится краткий обзор основных приложений АФП в области анализа больших объемов данных сложной природы.

Автоматизация подготовки и ведения нормативно-справочной информации на основе ретроспективного анализа Больших Данных

Путин Р. В. (*RPutin@incon.ru*) — «Информ-Консалтинг» (Пермь)

Глобальная производственно-распределительная сеть оперирует, по всей видимости, десятками миллионов позиций товарной номенклатуры . Это множество не статично, его обновление происходит со все возрастающей скоростью. В процессе функционирования системы генерируются огромные объемы структурированных и неструктурированных данных. Каждое предприятие, взаимодействуя с этой стихией, описывает ее в своих информационных системах в виде «справочников материалов», рано или поздно приходя к мысли создать «единый справочник материалов». Наша компания с момента основания помогает своим клиентам решать эту задачу.

По нашему мнению, основные сложности и риски в решении задачи создания «единого справочника материалов» проистекают из игнорирования двух факторов: во-первых, многообразие номенклатуры, используемой предприятием, значительно меньше потенциально возможного (речь можно вести о различиях как минимум на 2–3 порядка), во-вторых, доля повторно закупаемой номенклатуры в общем количестве незначительна (для производственного предприятия не более 10% от года к году). В этом контексте становится очевидно, что идее качества данных как соответствия формальным требованиям необходимо противопоставить идею качества данных как их операционной ценности при приоритете последней.

Методология

В основе методологии подготовки справочника материалов лежит метод классификации, результатами его применения являются:

- система классификации, включающая иерархическую структуру классов, с разработанным для каждого класса перечнем признаков (фасет) и наборов их значений;
- шаблоны наименований, связанные с системой классификации;
- исходный массив данных, прошедший процедуру классификации и обработки.

В качестве исходного массива данных используются исторические справочники материалов, в которых отражается специфика материального потребления предприятия. Анализ данных, прошедших процедуру классификации, позволяет вскрыть эту специфику: какие группы товарной номенклатуры (классы), какие технические и конструкционные характеристики (фасеты), какие «типоразмеры» (значения фасет и комбинации значений фасет) используются на предприятии. Результатом анализа должны стать решения относительно того, какая должна быть номенклатура в справочнике и с какой детализацией описания она должна быть представлена. Использование в качестве эталона предлагаемых производителями и нормативно-технической документацией детализированных описаний товарной номенклатуры не является универсальным решением. На практике мы сталкиваемся с тем, что:

- одна и та же товарная номенклатура для различных предприятий имеет разную значимость и характеристики потребления (регулярность, разнообразие, объемы);
- на каждом предприятии имеются специфические традиции, свой словарь общения, существуют особенности в архитектуре информационной системы, системы управления имеют разную степень формализации.

Игнорирование этих фактов приводит к нарушению коммуникативных функций справочника материалов, в том числе из-за роста в геометрической прогрессии его объема, что с определенного момента может создать значительные, ничем не оправданные трудности. Чтобы их избежать, необходимо определить разумные пределы «глубины описания» и зафиксировать их для каждого класса. Обобщение в разумных пределах не только не мешает, но, напротив, позволяет системе материального обеспечения предприятия функционировать более эффективно. Поиск компромисса выполняется в ходе диалога с представителями функциональных служб предприятия, анализа накопленных данных по движению товарно-материальных ценностей. В дальнейшем достигнутый компромисс должен динамически поддерживаться. Результат в каждый момент времени фиксируется в системе классификации и шаблонах, в которых представлены наборы обязательных фасет, порядок следования значений фасет и разделителей. Разработанные шаблоны позволяют унифицировать исходный массив данных, устранить дублирование данных, выявить проблемные позиции.

Классы, фасеты, шаблоны как более устойчивые информационные объекты (по отношению к изменяющемуся перечню номенклатурных позиций) позволяют обеспечить управляемость справочника материалов, выстроить вокруг них процедуру ведения, повысить качество и аналитичность данных в информационных системах. При этом надо понимать, что устойчивость системы классификации относительна, она также нуждается в развитии, а все достигнутые компромиссы должны динамически балансироваться.

Программное и информационное обеспечение

Инструментальную поддержку работ, согласно предложенной методологии, эффективно обеспечивает специализированное программное обеспечение — Автоматизированная система классификации (АСК), разработанная и развиваемая нашей компанией. Сервисы АСК обеспечивают поддержку всего процесса работ по подготовке нормативно-справочной информации и последующему ведению на этапе промышленной эксплуатации.

На всех этапах работ применяются различные методы автоматизации операций по обработке данных. Методы автоматизации строятся на возможности использовать массив данных прошлых проектов, выступающий в качестве «обучающей выборки». Для повышения эффективности автоматических операций наряду с совершенствованием алгоритмов ведется работа по повышению качества «базы знаний» — перечни значений фасет, допустимых сочетаний значений фасет регулярно пополняются и верифицируются.

На текущий момент база данных АСК включает:

- 7 млн позиций товарной номенклатуры, полученных из исторических справочников материалов заказчиков;
- 5 тыс. фасет и 1,4 млн значений фасет;
- 3 млн допустимых сочетаний значений фасет.

При наличии такой базы данных и соответствующих сервисов АСК позволяет:

- автоматизировать иерархическую и фасетную классификацию исходного массива данных (нечеткий поиск и автоматическое распознавание);
- на основе ряда заданных для номенклатурной позиции значений фасет автоматически выводить недостающие значения фасет, а также выполнять контроль непротиворечивости описания номенклатурной позиции по сочетанию значений фасет (вывод на базе знаний).

Интернет-сервисы для подписчиков научных изданий

Павел Самарский *(pauls@osp.ru)*, Павел Христов *(pkh@osp.ru)* — издательство «Открытые системы» (Москва)

Классические механизмы привлечения подписчиков, основанные на периодическом издании подписных каталогов специализированными подписными агентствами, ранее использовавшиеся газетами и журналами, утратили свою эффективность.
Специализированным, нишевым изданиям, особенно научным, необходима альтернатива классическим подписным каталогам — их услуги слишком дороги для таких изданий, кроме того, они не достигают фокусной аудитории.

Разработанный в издательстве «Открытые системы» онлайн-сервис, позволяющий создавать виджеты с информацией об изданиях и условиях подписки для внедрения на сайты изданий, а также административной части.
Виджет подписки и личный кабинет позволяют подписчику самостоятельно оформлять и редактировать новые заказы на подписку, продлевать подписку, редактировать собственный профиль, хранить ранее приобретенные цифровые продукты издателя, получать информацию и персональные предложения, связанные с новыми продуктами и издателя, и т. д.

Использование сервиса доказало свою особую эффективность в случае, когда издание адресовано узкоспециализированной аудитории.

Большие Данные в информационной безопасности

Касперская Н. И. (*Natalya.Kaspersky@infowatch.com*) — InfoWatch (Москва)

Большие Данные — очень популярная тема, о которой много говорят в различных областях за исключением сферы информационной безопасности. Чтобы рассмотреть вопрос именно с этой точки зрения, для начала определимся с самим понятием. Большие Данные — это вид данных, которые обладают следующими свойствами:

- Большие Данные слишком велики для ручного просмотра и позволяют делать выводы только на основании всей их совокупности, а не на отдельной выборке;
- Большие Данные в основном касаются людей и собраны из разнородных источников (интернет-трафик, электронная почта, мобильный телефон, видеопоток и др.);
- природа этих данных может быть совершенно различной — фото, видео, звук, электронный файл и т. д.;
- Большие Данные имеют временную компоненту, то есть зависят от времени, могут рассматриваться ретроспективно либо для прогнозирования возможных событий в будущем.

Большие Данные сегодня анализируются с различными целями в области маркетинга, в розничной торговле, интернет-рекламе, поисковых системах, социальных сетях, мобильной связи, но не в сфере корпоративной безопасности. Это в особенности удивительно, поскольку компания имеет существенно лучшие условия для сбора и анализа Больших Данных, чем, например, социальные сети. Ведь в компании можно отследить всю переписку сотрудников, все выходы в Интернет, вся информация на компьютерах пользователей корпоративной сети принадлежит организации, на сотрудников собраны кадровые данные, известны используемые персоналом мобильные устройства и т. д. Но, к сожалению, все эти Большие Данные в совокупности пока не используются компаниями для анализа в области информационной безопасности. И это при том, что ценность данных в сфере ИБ для компаний гораздо выше, чем, например, в области маркетинга. Задача маркетологов — продать массовой аудитории свой продукт, и ценность одного потребителя в маркетинговой кампании невысока. Для компании же ценность одного сотрудника с точки зрения информационной безопасности может быть очень высокой, в особенности если речь идет о топ-менеджере или о специалисте, владеющем огромным количеством конфиденциальных корпоративных сведений. Если такой сотрудник начнет вести себя недружественно по отношению к компании, его поведение подвергнет эту организацию серьезным рискам.

В этой связи мне кажется, что будущее внутренней информационной безопасности находится в области управления рисками с применением анализа Больших Данных. В центре такого анализа, на мой взгляд, должен стоять человек, поскольку именно он является слабым звеном любой системы безопасности и источником различных рисков. При этом наличие в корпоративной среде большого числа каналов и источников данных позволяет нам делать такие выводы о поведении человека, которые другими способами сделать невозможно. Например, проанализировав переписку сотрудников по различным каналам, мы выяснили, что в организации есть человек, который позволяет себе резко негативные высказывания по отношению к компании. У нас возникли опасения, что его поведение может негативно

сказаться на имидже компании во внешней среде и снизить лояльность других работников. В результате с данным сотрудником была проведена профилактическая беседа, по итогам которой человек скорректировал свое поведение в отношении компании.

Конечно, не всегда нелояльного сотрудника можно вычислить по переписке, люди часто бывают скрытными. Однако некие отклонения в поведении сотрудника выявить можно, поскольку у любого работника есть некая стандартная модель поведения и отклонения от нее можно заметить, изучив информацию о человеке из различных источников.

Современные системы защиты ориентированы лишь на определенный вид рисков: антивирусы борются с вирусами, системы защиты от уязвимостей, соответственно, с уязвимостями и так далее. Эту разрозненность средств защиты необходимо ликвидировать, объединив системы защиты для сбора и анализа Больших Данных, чтобы делать выводы на будущее и предотвращать возможные инциденты.

Мне кажется, что компании рынка ИБ уже начинают думать в этом направлении, однако результатов пока не видно, и тому есть несколько причин. Во-первых, для создания систем прогнозирования рисков нужны технологии искусственного интеллекта и машинного обучения — очень сложная интеллектуальная разработка. Большинство же разработчиков систем ИБ сосредоточены сугубо на технологиях защиты, то есть имеют иную квалификацию, даже если речь идет о крупных компаниях с несколькими направлениями разработки. Кроме того, с чисто прагматической точки зрения разработка системы прогнозирования рисков является дорогим удовольствием, ее создание требует значительных инвестиций. При этом рынок данных систем пока не развит, их востребованность на современном этапе невысока, поэтому нельзя ожидать, что такие системы появятся уже завтра, но будущее мне видится за ними.

Резюмируя, отмечу, что, на мой взгляд, системы комплексного анализа рисков и угроз предприятия появятся на рынке в ближайшие 3–5 лет. Такие системы будут использовать анализ Больших Данных как в отношении персонала, так и инфраструктуры предприятия. Такой подход позволит предприятиям перейти от парадигмы реагирования на инциденты в области информационной безопасности к обнаружению угроз на ранней стадии и оценке рисков.

Большие Данные: от науки до практики

Исаев К. Ш. (*Kamil.Isaev@emc.com*) — Центр исследований и разработок EMC (Москва)

Большие Данные являются основой для создания принципиально новых преимуществ для бизнеса. Они улучшают эффективность, качество и уровень персонализации продуктов и услуг, тем самым повышая степень удовлетворенности заказчиков и качество их обслуживания. В презентации речь пойдет о перспективных направлениях исследований в области Больших Данных, которые проводит компания EMC.

Проблема Больших Данных в развитии национальной медицины и здравоохранения

Турлапов В. Е. (*vadim.turlapov@cs.vmk.unn.ru*), Сапрыкин В. А., Гаврилов Н. И. — ННГУ им. Н. И. Лобачевского (Нижний Новгород)

Работа выполнена при поддержке гранта Президента РФ № НШ-1960.2012.9 и гранта Правительства РФ № 11.G34.31.0066.

Охрана здоровья как задача национальной экономики. По данным госстатистики Израиля, за период 2010–2012 годов уровень смертности (на 1 тысячу человек за 1 год) составляет: в Израиле — 5,33 умерших; в США — 6,42; в Бельгии — 6,35; в Норвегии — 5,49; в Италии — 5,116; в России — 14,3; на Украине — 15,2. Вероятность умереть в возрасте 15–60 лет в РФ почти в два раза выше, чем в среднем по Европе.

На рис. 1 показано изменение уровня смертности в России в 1990–2010 годах согласно данным Росстата.

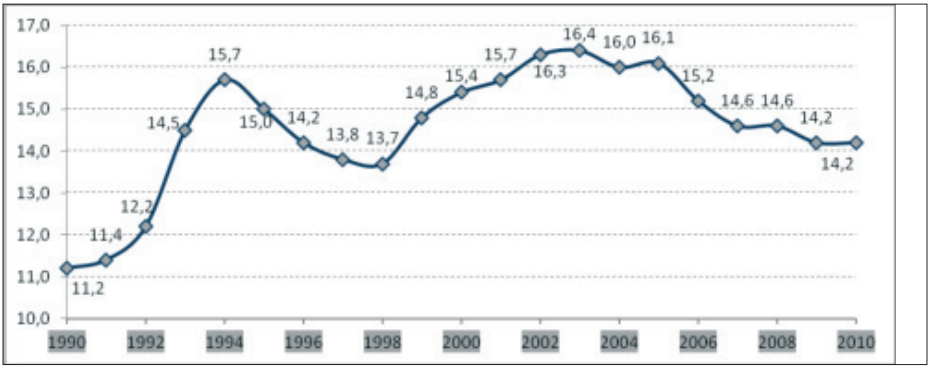


Рис. 1. Уровень смертности в РФ на 1000 чел населения в 2009–2010 (по данным Росстата 1997, 2001, 2006, 2009, 2011а)

По данным федеральной статистики (http://fedstat.ru), за период с 2005 по 2012 год число станций скорой помощи сократилось с 3277 до 2841, а число учреждений, оказывающих медицинскую помощь населению, — с 16 531 до 7986. Заболеваемость с впервые в жизни установленным диагнозом злокачественного новообразования на 100 тыс. человек населения за период с 2005 по 2012 год выросла с 330,5 до 367,6 (более чем на 11%).

Заболеваемость раком растет и во всем мире, а прогнозируемый прирост до 2030 года в странах со средним и низким уровнем дохода, по прогнозу CMAJ (журнал канадской медицинской ассоциации), может составить соответственно 70 и 80%. Россия в 2012 году формально относилась к странам с высоким уровнем дохода (чуть выше нижней границы), однако по расходам на здравоохранение, согласно данным Всемирного банка, опубликованным в 2012 году, Россия с долей 3,2% занимает 106-е место из 187 стран мира (http://gtmarket.ru/ratings/expenditure-on-health/info) и находится по этому показателю в окружении стран с низким уровнем дохода.

Согласно анализу предотвратимой смертности, выполненному для отчета Международной рабочей группы по расширению доступа к лечению рака и контролю над его распро-

странением в развивающихся странах [1], ежегодно регистрируется от 2,4 до 3,7 млн случаев смерти от рака, которые можно было предотвратить. Примерно 80% таких предотвратимых случаев смерти приходится на долю стран с низким и средним уровнем доходов (СНСУД), причем доля случаев смерти от рака, которые можно предотвратить путем профилактики и лечения, составляет в СНСУД 50–60%. Тяжелые социально значимые заболевания связаны с большими экономическими потерями для страны. Согласно тому же отчету, величина потерь, понесенных людьми из-за рака вследствие неполученного дохода, расходов на медицину за свой счет, оценивается в 2,5 трлн долларов США (2010 год) или более 4% мирового ВВП. Констатирована прямая связь роста заболеваемости раком с ростом хронических форм инфекционных заболеваний.

Авторы отчета [1] предлагают «диагональный подход» к разрешению создавшейся ситуации, заключающийся в создании инновационной системы здравоохранения, расширяющей доступ к лечению рака, обеспечивающей контроль над его распространением, укрепляющей систему здравоохранения путем одновременного учета общих задач здравоохранения. Борьба с раком может служить индикатором эффективности по каждому компоненту обновляемой системы здравоохранения. Инновации в системе здравоохранения должны охватывать шесть пересекающихся компонентов континуума лечения рака: первичную профилактику, раннее выявление, диагностику, лечение, выживаемость, а также долгосрочное наблюдение и паллиативный уход.

Создание инновационной системы здравоохранения России на основе трехмерного томографического скрининга населения и хранилищ больших медицинских данных. Мировая медицина сегодня интенсивно наполняется новыми средствами диагностики и лечения, информационными технологиями. По всему миру функционируют центры вычислительной медицины и вычислительной биологии.

Совмещение анализа результатов разных типов томографии (и не только) в 3D (КТ и ПЭТ, КТ и ОФЭКТ, КТ и МРТ, МРТ и УЗИ и др.) стало эффективнейшим способом диагностики и наблюдения хода лечения онкологических заболеваний. Стало актуальным использование 3D-модели пациента как основы его полного исследования современными методами. Компания IBM в 2008–2009 годах запустила и активно поддерживает проекты по использованию 3D-модели пациента как основы его медицинской карты [2]. В 2012 году IBM начала проект по моделированию действия лекарств на 3D-аватаре конкретного человека [3]. Отмечается существенный рост числа томографов и интенсивности их применения в системах здравоохранения всех стран. Так, по данным CDC (Center for Disease Control and Prevention, США) 2009 года, в США на каждый миллион населения применялось 60 томографов (26 МРТ и 34 КТ), которыми выполнялись 91 тыс. МРТ- и 228 тыс. КТ-обследований в год. При такой интенсивности в 2009 году было обследовано около 100 млн человек в год, то есть около трети всего населения США.

Современная многосрезовая КТ (МСКТ) фактически стала методикой объемного исследования всего тела человека, так как полученные аксиальные томограммы составляют трехмерный массив данных, позволяющий выполнить любые реконструкции изображений, в том числе многоплоскостные реформации, объемную (и при необходимости — стерео-) визуализацию, виртуальные эндоскопии. Трехмерная реконструкция томограммы может быть также эффективным инструментом обучения в медицинском образовании. На настоящий момент в мире, и в России в частности [4], достигнута производительность 3D-реконструкции томограмм, достаточная для создания простейших виртуальных анатомических столов. Первый

такой анатомический стол, произведенный фирмой Anatomage, с мая 2011 года находится в экспериментальной эксплуатации в Стэнфордском университете.

Все указанное говорит о возможности создания в России инновационной системы здравоохранения, реально охватывающей все шесть упомянутых выше компонентов континуума лечения рака. Предлагается создать инновационную систему, обеспечивающую принципиально более высокое качество здравоохранения в России, на основе трехмерного томографического скрининга населения и хранилищ больших медицинских данных.

Создание такой инновационной системы здравоохранения в течение ближайших 5–10 лет предполагает систематическую работу по следующим направлениям:

1) Обеспечение скрининга необходимым числом томографов, в то же время доступным для бюджета. Подготовка и переподготовка персонала

Число томографов определяется из условия прохождения обследования на томографе всем населением 1 раз в 2 года при суточной нагрузке на томограф — 15 обследований. Таким образом, на РФ с населением около 144 млн человек потребуется как минимум 16 тыс. штук (144 млн чел. / 600 дн. / 15 чел/день). А для города с населением 1 млн — около 110 томографов. Компании Philips, General Electric, Siemens и Toshiba по данным 2009 года ежегодно продают в России около 100–150 томографов. Предполагается, что отечественные томографы будут на порядок дешевле зарубежных (http://unitom.ru/ru/press) и для скрининга будут закупаться преимущественно низкопольные (0,15 Т – 0,4 Т) томографы отечественного производства (http://unitom.ru, http://az-mri.com/ms). Это позволит закупать порядка 1,5 тыс. томографов в год, не отказываясь в необходимых случаях от закупок зарубежных томографов. Необходима также масштабная подготовка и переподготовка кадров, способных обслуживать данную технику и задачи здравоохранения. Такая подготовка может быть выстроена по иерархическому принципу в ходе закупки томографов.

2) Создание, наполнение, классификация, обслуживание хранилищ больших медицинских данных на вычислительных кластерах

Отдельным вопросом является обеспечение условий и возможностей превращения больших медицинских данных в технологический и стратегический ресурс, работающий на развитие страны и здоровья нации. Это постоянная работа по: а) классификации хранимых данных для целей автоматизации диагностики; б) созданию программного обеспечения медицинских сервисов на Больших Данных (как задач многопотоковой массивно-параллельной обработки Больших Данных); в) обеспечению высокопроизводительных коммуникаций для массового доступа к данным.

Необходимо наращивание существующих центров обработки данных и вычислительных кластеров хранилищами Больших Данных с резервированием. Приближение хранения к центрам обработки значительно снизит издержки на передачу данных и ускорит их обработку, в том числе облачную. Объем хранилищ может быть определен в зависимости от численности населения территории, которую будет обслуживать кластер.

Необходимый ресурс хранения для РФ, заполняемый за 2 года с 16 тыс. томографов: 144×10^6 чел. $\times 10^9$ Байт/чел = 144 Пбайт. Такой объем ресурсов хранения может быть накоплен только к концу периода закупки 16 тыс. томографов. Соответственно, для города с миллионным населением и при наличии 110 томографов это будет около 1 Пбайт. При сегодняшнем количестве томографов в РФ в первые 2 года работы такой системы здравоохранения в РФ будет заполнено не более 10 Пбайт. Ресурс хранения на персональном уровне, необходимый

в течение 10 лет из расчета 1 Гбайт на обследование 1 раз в 2 года, составит всего 5 Гбайт, что в три раза меньше ресурса Google, выделяемого каждому пользователю под электронную почту, и может храниться пользователем на компактном носителе.

При этом объем данных, поступающих в хранилище Больших Данных с одного томографа в год из расчета 1 Гбайт на каждое обследование, составит: 1 Гбайт/чел × 15 чел/день × 300 рабочих дней = 4,5 Тбайт, что сопоставимо с объемом одного современного жесткого диска.

Объемы данных, связанные с современными методами диагностики, прежде всего с методами медицинской томографии, и методами лечения (КТ + ПЭТ, КТ + ОФЭКТ и др.), будут ежегодно пополняться, но с разработкой специальных методов сжатия данных томограмм, как и в связи с процессом персонализации медицины. Скорее всего, в итоге потребуется существенно меньший объем централизованного хранения.

3) Создание и развитие для скрининга населения технологий автоматической диагностики на больших медицинских данных, веб- и облачных сервисов на этой основе

Это прежде всего обеспечение визуальной диагностики по трехмерным (и, конечно, всем другим) данным, доступной в повседневной практике любому врачу-клиницисту, 3D-визуализация томограммы человека в полный рост в реальном времени. Обеспечение для скрининга предварительной автоматической диагностики, телеконсультирования, виртуальных анатомических столов, эмуляторов и образовательных 3D-пособий (в том числе стерео) для медицинского образования и переподготовки. А также такие новые методы и технологии как:

a. геометрическая 3D-реконструкция, морфометрия органов, систем и тела в целом, 3D-моделирование и печать (в том числе имплантов) для хирургии и травматологии;

b. слияние в 3D-модели человека всех методов исследования (прежде всего УЗИ, томографии, акустотермометрии), гладкая сшивка 3D-моделей с томограммами;

c. трехмерное моделирование и планирование облучения в онкологии, предварительное моделирование оперативных вмешательств, в том числе с использованием биоимиджинга.

Опыт создания хранилищ больших медицинских данных и медицинских сервисов для инновационной системы здравоохранения России. Необходимый опыт создания хранилищ больших медицинских данных и медицинских сервисов для инновационной системы здравоохранения России накоплен в Нижегородском государственном университете им. Н. И. Лобачевского. Для создания хранилища томограмм использован ресурс хранения вычислительного гетерогенного кластера объемом 36 Тбайт на основной территории ННГУ им. Н. И. Лобачевского в Нижнем Новгороде и СУБД с открытым кодом PostgreSQL. Выбор PostgreSQL был обусловлен как ее доступностью, так и полным соответствием ее характеристик требованиям организации экспериментального хранилища больших медицинских данных (прежде всего томограмм): максимальный размер таблицы в PostgreSQL составляет сегодня 32 Тбайт, а максимальный размер поля — 1 Гбайт. На сегодняшний день хранилище медицинских данных содержит около 10 тыс. томограмм. Данные хранятся в DICOM-формате в обезличенном виде в режиме долговременного хранения. Организация взаимодействия хранилища с локальной вычислительной сетью (PACS) медучреждения, хранящего персональные данные, показана на рис. 2.

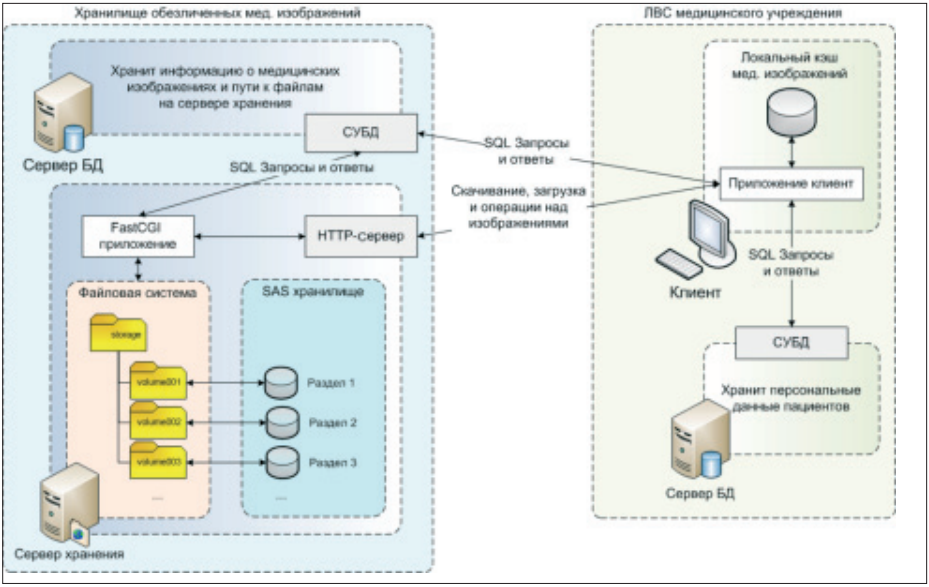


Рис. 2. Схема отношений хранилища больших медицинских данных с локальной вычислительной сетью (PACS) медучреждения

В лаборатории компьютерной графики ННГУ разработано также программное обеспечение медицинского сервиса с широкими возможностями для 2D-3D-визуализации томограммы человека в полный рост в реальном времени, а также для анализа данных томограмм, общедоступное в повседневной практике любому врачу-клиницисту (см. рис. 3 и рис. 4).

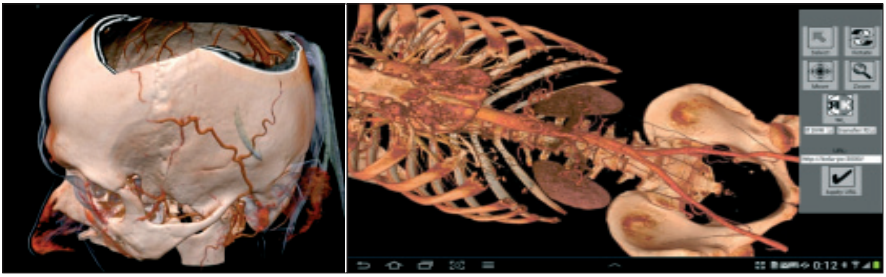


Рис. 3. Пример качества визуализации и работы облачной 3D-визуализации томограмм на мобильном устройстве

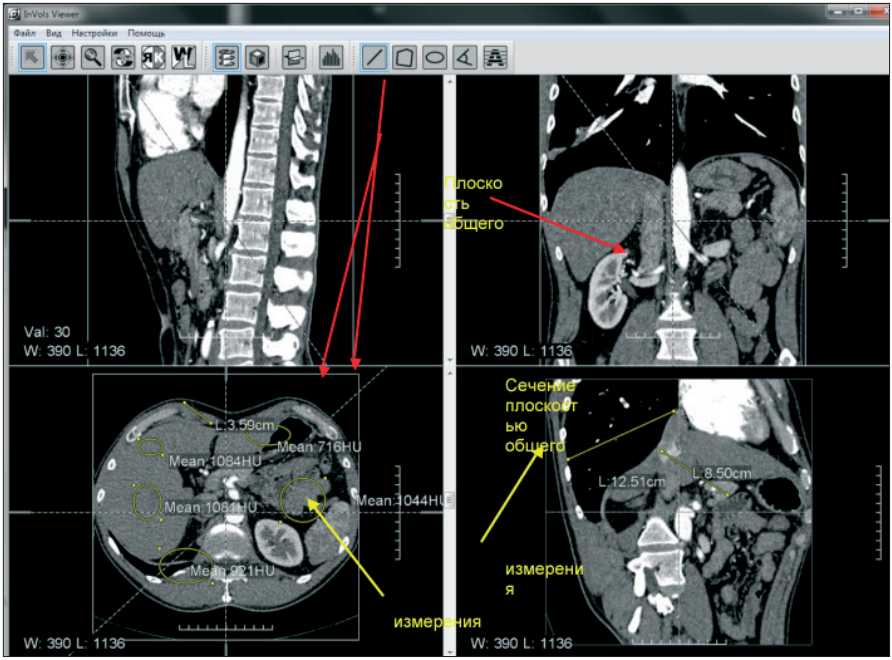


Рис. 4. Пример интерфейса и функций разработанной системы визуализации и анализа томограмм

Система использует графические процессоры независимо от производителя, отличается высоким качеством сгенерированного изображения при сохранении высокой производительности. На основе запуска данной системы на сервере как FastCGI-приложения, организованного через веб-интерфейс с использованием современных возможностей HTML5, построен облачный сервис 3D-визуализации томограмм. Сервис доступен как через проводной Интернет, так и на мобильных устройствах с мультикасательным интерфейсом (см. рис. 3).

Заключение. Рассмотрена проблема охраны здоровья населения России как задача национальной экономики в контексте состояния и подходов к решению этой проблемы в мире. Предлагается создание инновационной системы здравоохранения России на основе трехмерного скрининга населения и хранилищ медицинских данных, использующих технологии Больших Данных.

Сформулированы задачи создания новой системы здравоохранения, сделаны количественные оценки необходимых ресурсов хранения, технического и программного обеспечения на фоне существующих закупок.

Показан пример создания хранилища больших медицинских данных и базового медицинского облачного сервиса по медицинской 2D-3D-визуализации, работающего на нем, который мог бы стать основой для реорганизации системы здравоохранения на основе технологий Больших Данных.

Не секрет, что в России разработаны несколько медицинских информационных систем для поддержки персонализированной медицины и персональной электронной медицинской карты.

Все это показывает, что в России есть условия для создания инновационной системы, качественно повышающей уровень здравоохранения. Необходимо лишь принятие достаточной бюджетной государственной программы для ее реализации.

Литература

1. Преодоление неравенства в борьбе против рака: программа по расширению доступа к лечению в странах с низким и средним уровнем доходов/ Гарвардская глобальная инициатива по обеспечению справедливости, Бостон, Массачусетс, ноябрь 2011. 238 с. (http://gtfccc.harvard.edu).

2. Charette R.N. Visualizing Electronic Health Records With «Google-Earth for the Body»/ IEEE Spectrum. Jan.2008 (http://spectrum.ieee.org/jan08/5854).

3. Charette R.N. Using Avatars to Understand Adverse Drug Reactions / March 06, 2012 (http://spectrum.ieee.org/riskfactor/biomedical/diagnostics/using-avatars-to-understand-adverse-drug-reactions/).

4. Гаврилов Н.И., Турлапов В.Е.. Подходы к оптимизации GPU-алгоритма volume raycasting для применения в составе виртуального анатомического стола // Научная визуализация. 2012. Кв. 2. Т. 4. № 2. С. 21-56 (http://sv-journal.com/2012-2/index.php?lang=ru).

Большие Данные — онтология и применение в маркетинге

Аршавский А. В. (*andzhey@mac.com*) — «Лаборатория Цифрового Общества» (Москва)

В докладе будет еще раз поднята тема определения термина Большие Данные и будет предложена альтернативная концепция его интерпретации. Также подробно рассматривается ошеломляюще быстро растущее направление применения технологий Больших Данных в маркетинге и продажах на рынках B2C и B2B. Это явление рассматривается с точки зрения технологий и данных, а также представляется в качестве яркого примера масштабного воплощения в повседневный бизнес еще не так давно мало понятного технического направления Big Data.

Большие Данные в образовании: новые возможности и новые вызовы

Мальцева С. В. (*smaltseva@hse.ru*) — НИУ ВШЭ (Москва)

Образование является одной из сфер, в которой новые технологии Больших Данных являются очень актуальными, обеспечивают возможность перехода к новым, более эффективным образовательным моделям.

Широкое использование информационных технологий в современном образовательном процессе позволяет собирать, хранить и обрабатывать большие объемы данных, в том числе и в оперативном режиме. Аналитика этих данных позволяет решать новые задачи, обеспечивающие поддержку таких важных образовательных трендов, как использование среды открытых знаний, внедрение адаптивных систем обучения, реализация коллаборативного обучения.

Приход в образовательную сферу технологий Больших Данных означает существенное усиление роли в ней информационных технологий, делает их двигателем развития образовательных институтов и структур. В первую очередь этому способствуют возможности нового аналитического инструментария. Развитие концепции образовательной аналитики предполагает создание систем, функционал которых охватывает задачи извлечения данных из разнородных источников, моделирование образовательных процессов, накопление баз педагогических практик.

Доступность больших объемов данных и инструментов по извлечению информации из этих данных существенно влияет также на развитие научного и инновационного потенциала университетов.

Включение в образовательный процесс курсов, формирующих знания и умения использовать технологии обработки Больших Данных, внедрение инструментария, обеспечивающего возможность аналитики больших объемов информации, являются важным условием формирования у студентов компетенций, соответствующих современному уровню знаний в различных областях деятельности.

Данные, которые накапливались в образовательной системе в течение многих лет, носили преимущественно структурированный характер и были представлены, как правило, в форме отчетов и статистики. В большей степени они относились к так называемой академической аналитике, ориентированной на интегральную оценку образовательных учреждений.

Трансформация образовательной сферы под влиянием ИТ привела к внедрению новых, интерактивных форм обучения, которые обеспечивают поток неструктурированных данных, порождаемых интеллектуальными системами обучения, моделирующими программами и обучающими играми. Новые неструктурированные данные несут возможности изучения реальных результатов обучения, позволяют оценить компетенции студентов. Отмечается необходимость учета иерархической структуры этих данных, их зависимость от ситуации, в которой они возникают, а также их временные характеристики, которые являются важными для интерпретации данных. В целом использование аналитики этих данных позволяет реализовать идеи персонализированного, адаптивного обучения.

Создание адаптивных систем обучения связано с моделированием как самих образовательных процессов, так и поведения студентов. Основное назначение этих систем — формирование индивидуального подхода к обучению, который позволил бы с учетом индивидуальных особенностей обучаемого и конкретных задач, заложенных в программе обучения, обеспечить синтез необходимых знаний и навыков на основе получаемой информации. Важную роль здесь играет активное вовлечение обучаемого в процесс формирования его компетенций. Новые возможности образовательной аналитики связаны со сбором данных о поведении обучаемого на всех этапах его жизненного цикла, связанных с различными видами и программами обучения, и предоставлением этих данных по запросам образовательных учреждений (при условия соблюдения всех правовых и этических норм).

Образовательная аналитика нацелена на сбор и исследование данных на уровне образовательной программы, отдельных занятий. Ее возможности обеспечивают повышение качества обучения за счет постоянного мониторинга обучения и глубинного анализа причин, которые приводят к плохим результатам обучения, демонстрируемым на итоговом тестировании или других видах аттестации. Этот анализ позволяет обеспечить своевременную коррекцию процессов, причем персонализированно по отношению к отдельному студенту. Он также позволяет сформулировать рекомендации как для преподавателя, так и для студента.

Технологии Больших Данных позволяют активно использовать данные, собираемые в рамках задач, решаемых образовательной аналитикой, для анализа деятельности университета в целом — как на уровне внутренней администрации, так и на уровне внешних организаций.

Задачи анализа таких данных для целей внутреннего администрирования направлены на оценку образовательных программ, преподавательского состава, образовательной среды университета. На внешнем уровне данные используются для сравнительного анализа университетов, вычисления рейтингов и т. д.

Одной из критических задач внедрения образовательной аналитики является сбор данных. Необходимыми элементами здесь являются некоторая стандартизация и категоризация данных, а также наличие у образовательных учреждений регламентов и инструментария для организации сбора таких данных и необходимость решения этических проблем — например, получение согласия обучаемого и преподавателя на постоянный мониторинг их деятельности. Важным условием является понимание участниками процесса полезности для них результатов анализа собираемых данных, а также знакомство с технологиями и аналитическим инструментарием, примерами полезного использования.

Для обоснования значительных затрат на внедрение технологий и инструментария Больших Данных в образовательных учреждениях необходимо создание новых метрик качества и эффективности образовательного процесса, которые были бы связаны с показателями, используемыми в интегральной академической аналитике, внедрение этих метрик для оценки образовательных учреждений, в том числе и на уровне государственных структур.

Необходимо также учитывать, что внедрение образовательной аналитики в полном объеме существенно повлияет на изменения самого процесса образования, а также повысит значение образовательных бизнес-сетей как инструмента обучения и обмена знаниями.

Профессия Data Scientist

Жуков Л. Е. (*leonid.e.zhukov@gmail.com*) — НИУ ВШЭ (Москва)

Harvard Business Review в 2012 году назвал профессию data scientist «the sexiest job of the 21st century» [1]. Sexiest не только в смысле модной и желанной профессии, а еще и такой, когда спрос на людей этой профессии намного превышает предложение. McKinsey оценивает возможную нехватку специалистов с глубокими аналитическим знаниями к 2018 году в 140–190 тыс. человек [2].

Эта нехватка ощущается уже сегодня, когда все крупнейшие ИТ-компании, такие как Apple, IBM, EMC, Google, Facebook, ищут себе сотрудников. В результате средняя заработная плата data scientist на сегодняшний день в Силиконовой долине составляет около 140 тыс. долл. в год, что существенно превышает зарплаты и аналитиков, и программистов.

В чем же секрет такой привлекательности data scientists? Как правило, их навыки могут быть эффективно использованы для решения конкретных задач бизнеса и создания новых продуктов, основанных на собираемых бизнесом данных (data driven). Они способны не только анализировать данные, но также проверять гипотезы и строить предсказательные модели на очень больших объемах данных, которые на сегодняшний день нельзя обработать готовыми инструментами и программами [3].

Откуда берутся data scientists? В первую очередь это люди с отличными практическими навыками, глубокими теоретическими знаниями и большим кругозором. Они должны обладать исследовательским типом ума и в то же время уметь самостоятельно, своими руками решать практические задачи. Как правило, они эксперты в статистике, анализе данных и методах машинного обучения. На сегодняшний день образование в computer science дает наилучшую подготовку, однако много кандидатов приходят из других точных наук, таких как физика, астрономия, инженерные дисциплины. Ключевым навыком является умение работать с Большими Данными, и он часто вырабатывается у людей, занятых обработкой измерений и экспериментальных данных. Как правило, на позицию data scientist претендуют кандидаты наук (PhDs) и иногда выпускники магистратуры.

Какими же инструментами и навыками необходимо владеть на сегодняшний день data scientist? Большинство современных технологий обработки Больших Данных построено на свободном ПО (open source), поэтому в первую очередь необходимо знать операционную систему Linux и ее стандартные инструменты. Самыми популярными на сегодняшний день системами обработки Больших Данных являются платформа Hadoop и парадигма распределенных вычислений MapReduce. За последние годы было разработано много программных продуктов, облегчающих работу с Большими Данными, в том числе такие как Hive, Pig и другие. Конечно, необходимы базовые знания SQL, но также и опыт работы с современными базами данных NoSQL (HBase, Cassandra, MongoDB, Neo4j). Основными языками программирования для data scientists являются Pyhon, Java и становящийся популярным язык функционального программирования Scala. Для алгоритмов машинного обучения и статистического анализа часто используется язык R. Также применяются различные библиотеки на Python (NumPy, SciPy, Nltk, SkiKit) и Java (Mahoot, Weka).

Как правило, задачи data scientist получает из различных подразделений компании — маркетинга, управления продуктом или разработки. В любом случае самый первый шаг заключается в обсуждении с «заказчиком» и проработке постановки задачи. После этого начинается практическая работа по получению, подготовке, форматированию и очистке данных. Далее — исследование и поиск закономерностей в данных. В зависимости от поставленной задачи возможно проведение классификации или кластеризации данных (обучение с учителем и без учителя), а также построение предсказательных моделей. Как правило, бывает очень полезно визуализировать полученные результаты. Процесс заканчивается обсуждением полученных результатов и возможного продолжения проекта с «заказчиком». Каждый шаг процесса может занимать от нескольких часов до дней и недель, в зависимости от объемов и сложности задачи. Как правило, в разработке проекта участвует целая команда data scientist.

В чем же отличие data scientist от традиционной профессии аналитика? Аналитики, как правило, работают с данными, измеряющими поведение бизнеса компании, проводят сравнительный анализ, находят тренды и сегменты рынка, готовят бизнес-отчеты. Обычно такие данные уже хранятся в структурированном виде, готовом к SQL-запросам. Типичным инструментом для работы является Excel, хотя в последнее время появляется много новых систем, позволяющих аналитикам работать с Большими Данными (Tableau, Datameer, Karmasphere, Platfora). Ключевым отличием data scientist является работа с такими типами и размерами данных, для которых нет готовых коммерческих решений, часто это неструктурированная информация терабайтных и петабайтных масштабов, требующая распределенной обработки с использованием технологий Hadoop или схожих с ними. Часть рабочего времени data scientist уходит на написание специализированных программ для решения конкретных задач, часть — на исследование и анализ данных с использованием написанных программ.

Задачи приходят как со стороны бизнеса, маркетинга, так и от инженерных и продуктовых подразделений и часто включают создание предсказательных моделей и data driven продуктов и систем.

Data scientists решают разнообразные прикладные задачи — например, в маркетинге это могут быть сегментация рынка, моделирование приобретения и оттока клиентов, анализ социальных медиа. В финансовых и страховых компаниях задачами являются выявление мошенничества, детектирование аномального поведения, анализ кредитных рисков, страховое моделирование, оптимизация портфолио. В таких информационно емких областях, как здравоохранение и фармакология, к прикладным задачам для data scientist относятся генетический анализ, анализ клинических испытаний, клинические системы принятия решений.

Компании часто создают отдельные команды и целые подразделения data scientist. В таких командах data scientists могут играть разные роли в зависимости от своих навыков, опыта и экспертизы [4]. В частности, в результатах опроса [5] среди data scientists было выделено четыре типа: «data business person», «data creative», «data developer» и «data researcher», каждый из которых обладает разными уровнями навыков в бизнесе, методах машинного обучения, математике, статистике и программировании. Полноценная и оптимальная команда data science должна состоять из сотрудников с комплиментарными навыками.

Из-за значительной нехватки специалистов компании, специализирующиеся на разработке систем обработки Больших Данных, стали создавать подготовительные программы, курсы и сертификации для data scientists. К известным программам относятся Cloudera University и EMC Education Services.

Ряд крупных университетов США уже создали программы обучения — в частности, New York University, Columbia University, University of Washington, University of California Berkeley, University of Southern California. Многие другие университеты также готовят свои программы. В России отделение прикладной математики и информатики Высшей школы экономики с 2014 года запускает новую магистерскую программу «Наука о данных». В преподавании будут участвовать более 20 профессоров и преподавателей кафедры.

Основные предметы обучения, необходимые для data scientist, включают в себя программирование, алгоритмы и структуры данных, базы данных, статистику, анализ данных, машинное обучение, компьютерную обработку текста, распределенные системы, инструменты Больших Данных, визуализацию данных.

Отличным источником знаний является серия книг издательства O'Reily, посвященная различным методам и инструментам работы с Большими Данными.

Последние тренды data science и технологии Больших Данных бывают широко представлены на многочисленных специализированных конференциях и форумах. Среди известных конференций стоит выделить O'Reilly Strata Conference Making Data Work, Hadoop World, Big Data Techcon. Последние алгоритмические и научные достижения обсуждаются на KDD Knowledge Discovery and Data Mining, SIGIR Information Retrieval, ICML International Conference on Machine Learning, ICDM International Conference on Data Mining.

Как и в любой новой науке и профессии, есть ряд вопросов, на которые еще предстоит ответить data science. Насколько важно быть экспертом в предметной области решаемой задачи (domain expertise)? Что более важно в профессии data scientist — образование или практический опыт? Какие перспективы у профессии data scientist, будут ли они в будущем замещены готовыми программными решениями?

Литература

1. Thomas Davenport, D.J. Patil. Data Scientist: The Sexiest Job of the 21st Century. Harvard Business Review, October 2012.

2. Big data: The next frontier for innovation, competition, and productivity. McKinesy Global Institute report, June 2011.

3. DJ Patil. Building Data Science Teams. O'Reilly Strata, 2011.

4. Rachel Schutt, Cathy O'Neil. Doing Data Science: Straight Talk from the Frontline. O'Reilly Media, 2013.

5. Harlan Harris, Sean Murphy, and Marck Vaisman. Analyzing the Analyzers. O'Reilly Strata, 2012.

СОДЕРЖАНИЕ

Секция. Большие Данные и общество