## V МЕЖДУНАРОДНАЯ СОЦИОЛОГИЧЕСКАЯ ГРУШИНСКАЯ КОНФЕРЕНЦИЯ «БОЛЬШАЯ СОЦИОЛОГИЯ: РАСШИРЕНИЕ ПРОСТРАНСТВА ДАННЫХ»

А.Б. Мосягин

# ИСПОЛЬЗОВАНИЕ МЕТОДОЛОГИИ DATA MINING ПРИ РЕШЕНИИ ЗАДАЧ ОБРАБОТКИ СОЦИАЛЬНЫХ ДАННЫХ

*МОСЯГИН Александр Борисович – доцент кафедры прикладных информационных технологий Института общественных наук РАНХиГС, кандидат технических наук, доцент. E-mail: albor99@mail.ru*

Сегодня в прикладных социологических исследованиях происходит настоящая революция, связанная с появлением принципиально новых источников данных, прежде всего основанных на так называемой объективной регистрации реального поведения людей. На основе новых информационных технологий различные субъекты (госорганы и бизнес-структуры) собирают огромные массивы данных (Big Data), которые используются в социальной диагностике и прикладных исследованиях. Радикально настроенные аналитики даже предрекают смерть традиционным методам социологических исследований, в большей мере основанным на субъективной информации, получаемой в ходе разного рода опросов.

Существует хорошее высказывание, что «за последние годы, когда, стремясь к повышению эффективности и прибыльности бизнеса, при создании БД все стали пользоваться средствами обработки цифровой информации, появился и побочный продукт этой активности – горы собранных данных. И все больше распространяется идея о том, что эти горы полны золота». В прошлом процесс добычи золота в горной промышленности состоял из выбора участка земли и многократного дальнейшего ее просеивания.

Термин Data Mining часто переводится как добыча данных, извлечение информации, раскопка данных, интеллектуальный анализ данных, средства поиска закономерностей, извлечение знаний, анализ шаблонов, раскопка знаний в базах данных. Понятие «обнаружение знаний в базах данных» (Knowledge Discovery in Databases, KDD) можно считать синонимом Data Mining [Encyclopedia….,2009].

Понятие Data Mining, появившееся в 1978 г., приобрело высокую популярность в современной трактовке примерно с первой половины 1990-х гг. До этого времени обработка и анализ данных осуществлялсись в рамках прикладной статистики, при этом в основном решались задачи обработки небольших баз данных.

Таким образом, методология Data Mining – это мультидисциплинарная область, возникшая и развивающаяся на базе таких наук, как прикладная статистика, распознавание образов, искусственный интеллект, теория баз данных и др. (рис. 1).

Возникновение и развитие Data Mining обусловлено различными факторами, и основные среди них [Vercellis, 2009]:

- совершенствование аппаратного и программного обеспечения;
- совершенствование технологий хранения и записи *данных*;
- накопление большого количества ретроспективных *данных*;
- совершенствование алгоритмов обработки информации.

Data Mining – это процесс поддержки принятия решений, основанный на поиске в данных скрытых закономерностей (шаблонов информации) [Паклин, 2013], т.е. это процесс обнаружения в сырых данных ранее неизвестных, нетривиальных, практически полезных и доступных интерпретаций знаний, необходимых для принятия решений в различных сферах человеческой деятельности.

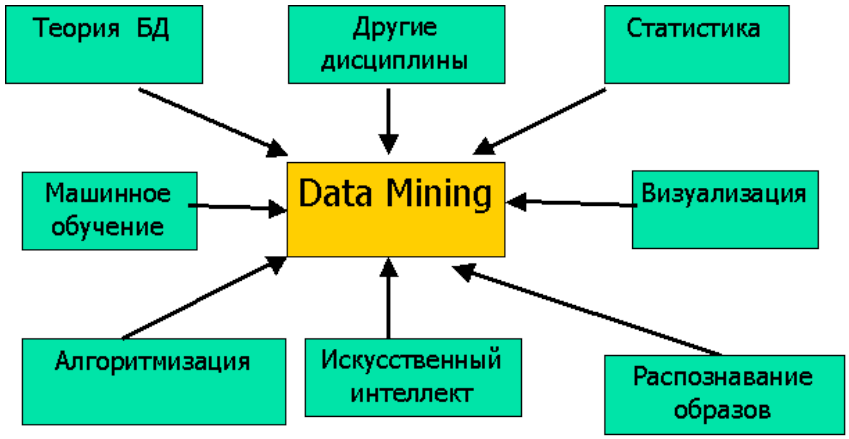


**Рисунок 1** - Data Mining как мультидисциплинарная область

Суть и цель технологии Data Mining заключаются в извлечении из больших объемов данных неочевидных, объективных и полезных на практике закономерностей. Это значит, что найденные закономерности не обнаруживаются стандартными методами обработки информации или экспертным путем, они будут полностью соответствовать действительности и им можно найти практическое применение в социологии.

В основу технологии Data Mining положена концепция шаблонов (patterns), которые представляют собой закономерности, свойственные выборкам данных, которые могут быть выражены в форме, понятной человеку. Цель поиска закономерностей – представление данных в виде, отражающем искомые процессы. Построение моделей прогнозирования также является целью поиска таких закономерностей. Чтобы максимально использовать мощность масштабируемых инструментов Data Mining, в социологических исследованиях необходимо выбрать, очистить и преобразовать данные, иногда интегрировать информацию, добытую из внешних источников, и установить специальную среду для работы Data Mining алгоритмов.

Результаты Data Mining в большой мере зависят от уровня подготовки данных, а не от чудесных возможностей некоего алгоритма или набора алгоритмов. Около 75% работы над Data Mining состоит в сборе данных, который совершается еще до того, как запускаются сами инструменты. Прежде чем использовать технологию Data Mining, необходимо тщательно проанализировать ее проблемы, ограничения и критические вопросы, с ней связанные, а также понять, что эта технология не может. Например, Data Mining не может заменить аналитика, она всего лишь дает ему мощный инструмент для облегчения и улучшения его работы.

Необходимы тщательный выбор модели и интерпретация обнаруженных зависимостей или шаблонов. Поэтому работа с такими средствами требует тесного сотрудничества между экспертом в предметной области и специалистом по инструментам Data Mining. Построенные модели должны быть грамотно интегрированы в бизнес-процессы для возможности оценки и обновления моделей. В последнее время системы Data Mining поставляются как часть технологии хранилищ данных.

В отличие от статистических средства Data Mining теоретически не требуют наличия строго определенного количества ретроспективных данных. Эта особенность может стать причиной обнаружения недостоверных, ложных моделей и, как результат, принятия на их основе неверных решений. Также необходимо контролировать статистическую значимость обнаруженных знаний.

Традиционные методы анализа данных (статистические методы) и OLAP в основном ориентированы на проверку заранее сформулированных гипотез (verification-driven data mining) и на грубый разведочный анализ, составляющий основу оперативной аналитической обработки данных (OnLine Analytical Processing, OLAP), в то время как одно из основных положений Data Mining – поиск неочевидных закономерностей. Инструменты Data Mining могут находить такие закономерности самостоятельно и также самостоятельно строить гипотезы о взаимосвязях. Поскольку именно формулировка гипотезы относительно зависимостей является самой сложной задачей, преимущество Data Mining по сравнению с другими методами анализа очевидно [Современные…].

Исследования отмечают, что существуют как успешные решения, использующие Data Mining, так и неудачный опыт применения этой технологии [Data Mining…]. Области, где применения технологии Data Mining, скорее всего будут успешными, имеют такие особенности:

- требуют решений, основанных на знаниях;
- имеют изменяющуюся окружающую среду;
- имеют доступные, достаточные и значимые данные;
- обеспечивают высокие дивиденды от правильных решений.

И все эти характеристики присущи социологии.

Таким образом, технология Data Mining постоянно развивается, привлекает к себе все больший интерес как со стороны научного мира, так и со стороны применения достижений технологии в бизнесе, социологических исследованиях. С сентября 2014 г. в Институте общественных наук создана и успешно функционирует кафедра прикладных информационных технологий, состоящая из математиков-информатиков, обладающих большим опытом использования, разработки и внедрения информационных технологий в различные прикладные области. В том числе есть специалисты, способные обучать и передавать знания методологии Data Mining, особенности использования алгоритмов и инструментов программных приложений для обработки и анализа структурированных данных.

Литература


1 Паклин Н.Б., Орешков В.И. Бизнес-аналитика: от данных к знаниям. СПб.: Питер, 2013.

2 Современные технологии использования Data Mining в прикладных областях, особенности применения инструментария технологии URL: http://www.kdnuggets.co

3 Data Mining and Knowledge Discovery. URL: http://www.knowldiscovery.com/

4 Encyclopedia of Data Warehousing and Mining. Idea Group Inc., 2006.

5 Vercellis C. Business Intelligence: Data Mining and Optimization for Decision Making. Wiley Publishing, Inc., 2009.

УДК: 330.111.4:004.738.5

Ю.М. БОЛЬШАКОВА

# КОНЦЕПЦИЯ БАЗ ДАННЫХ И ОБЛАЧНЫЕ ТЕХНОЛОГИИ В СТРАТЕГИИ ПРОДВИЖЕНИЯ ИНТЕГРИРОВАННЫХ КОММУНИКАЦИЙ БИЗНЕСА

*БОЛЬШАКОВА Ю.М. – к. полит. н., доцент Санкт-Петербургского государственного экономического университета*

Процесс масштабирования ИТ-инфраструктуры бизнеса, расположенной на площадке предприятия, может быть медленным, и организациям часто не удается достичь оптимального уровня использования ИТ-инфраструктуры.

Облачные технологии – это смена парадигмы, которая обеспечивает поддержку вычислений с использованием Интернета. Сервис облачных вычислений состоит из высокооптимизированных и виртуализированных центров обработки данных, обеспечивающих предоставление различных программных, аппаратных и информационных ресурсов, когда их использование оказывается необходимым [Handler et al., 2012].

Бизнес может подключаться к вычислительному облаку, чтобы использовать доступные ресурсы и оплачивать только фактически потребляемый объем услуг. Это помогает компаниям избегать капитальных затрат на установку дополнительных элементов инфраструктуры на своих площадках, а также мгновенно увеличивать или уменьшать объем используемых вычислительных ресурсов согласно своим бизнес-требованиям [Research..., 2013]. Гибридная облачная среда представляет собой сочетание частной и общедоступной моделей интегрированных коммуникаций. В гибридной облачной среде отдельные ресурсы выполняются или используются в общедоступной облачной среде, а другие выполняются или используются на площадке заказчика, в частной облачной среде. Это обеспечивает повышение эффективности.

Нефтяная отрасль, индустрия путешествий являются достаточно изменчивыми отраслями экономики, зачастую сталкивающимися с глобальными вызовами, бизнес трактует это как «вся эта затея – игра с очень высокими ставками», «отличительная особенность нашего бизнеса – это технологии, которые мы используем», «нам нужен способ продвижения в этом направлении и быстрого развертывания подобной системы».

В каждом из этих случаев залогом успеха интегрированных коммуникаций было использование облачных вычислений: общедоступной облачной среды, новой парадигмы разработки приложений и обработки больших объемов данных.

Рассмотрим данные влияния облачных услуг на IT-бюджет фирмы (см. рис. 1). По результатам социологического опроса «The hidden truth about cloud spending», опубликованного в 2011 г. [Handler et al., 2012], 55% респондентов отметили, что затраты выросли на виртуализацию данных, 51% на обеспечение ИТ-безопасности, 50% – на соответствие стандартам и управление, сетевое оборудование, 44% – на использования серверов, 43% опрошенных подчеркнули, что затраты выросли на сетевое программное обеспечение (ПО), хранение данных, серверное обслуживание ПО.

32% респондентов отметили, что затраты уменьшились за счет внедрения облачных услуг в направлении серверного обслуживания, 20% отметили снижение затрат за счет использования серверного ПО, 19% за счет использования коммерческого ПО, 18% респондентов отметили, что затраты уменьшились за счет внедрения программ повышения производительности ПК.

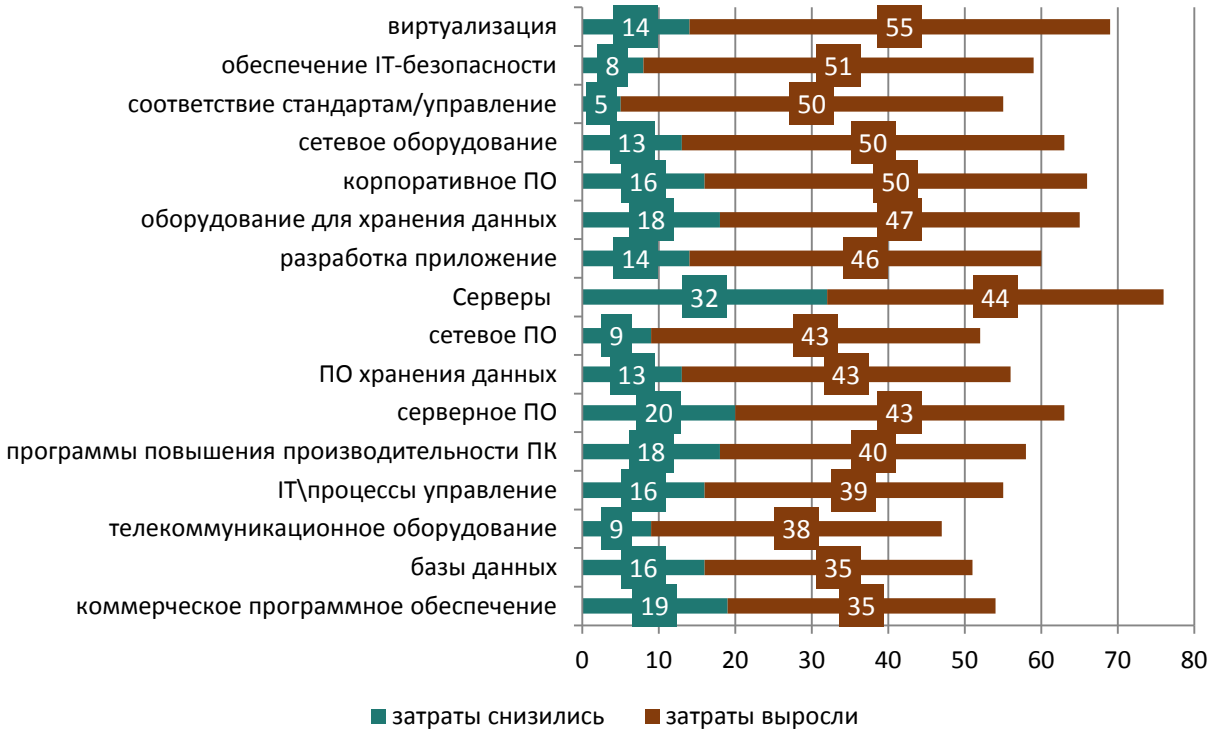


**Рисунок 1** - Доля респондентов отметивших, что затраты увеличились (+), уменьшились (–) за счет внедрения облачных услуг в 2010 г.

Существует несколько серьезных стратегических установок для переноса бизнеса любой компании в облачную среду.

Первая причина – масштабирование, когда бизнес может свободно масштабировать используемые ресурсы.

Вторая причина – эластичность, организация получает возможность использовать невероятный набор инструментов для выполнения анализа данных, обработки графики, потокового вещания видео и решения любых других задач.

И третья – доступ бизнеса к базам данных и расчетам в реальном времени. На вашем устройстве все происходит в реальном времени, что позволяет вам принимать решения. Последняя причина носит принципиальный характер – речь идет о переходе от приобретения ИТ-оборудования для проектов с целью последующего использования на своих площадках к модели, в которой вы приобретаете услугу по мере ее использования. Пользователь имеет доступ по мере необходимости, и это обеспечивает большую гибкость с точки зрения прибыли и убытков – такова модель операционной деятельности многих компаний во всем мире.

Рассмотрим практику интегрированных коммуникаций для использования цифровых услуг в современном мире (рис. 2). Так, в 2010 г. бизнес в основном использовал необлачные технологии в объеме 22 зетабайт.


**Рисунок 2** - Анализ использования цифровых услуг в современном мире, 2014 г. [1]

Рассматривая расходы современного бизнеса на ИТ-безопасность следует отметить, что с 2008 по 2012 г. расходы увеличились на 74%, по данным компании «SG Cross Assert Research», основные статьи расходов на безопасность – это обеспечение защиты на уровне пользователя (5,4 млрд долл США в 2012 г.), корпоративная безопасность (4,4 млрд долл США в 2012 г.), резервное хранение и восстановление данных (4,4 млрд долл США в 2012 г.), подробные данные представлены на рис. 3.


**Рисунок 3** - Расходы на обеспечение информационной безопасности (в млрд долл США) [4]

Самое последнее новшество – это социальные сети. Можно утверждать со всей определенностью, что социальные сети с течением времени могут стать крупнейшим или одним из крупнейших трендов в интегрированных коммуникациях, поскольку они дают возможность людям объединяться в группы, обмениваться информацией и совместно работать над решением различных задач. Обеспечение более эффективных возможностей для обмена данными между людьми и их совместной работы всегда служило фактором эволюции технологий во всем мире, а социальные сети – это нечто совсем новое. Облачная среда предоставляет множество возможностей и преимуществ для решения проблемы злоумышленников. Компании Microsoft, Google, Salesforce и другие поставщики услуг доступа к облачным средам могут реализовать множество сервисов и технологий безопасности, сконцентрировать для этой цели усилия лучших сотрудников на всей планете. Компании, подробные Google, Microsoft и Salesforce, способны предоставлять услуги защиты данных своим клиентам, которые используют G-mail и документы Google. Важнее всего то, что эти компании имеют мотивацию обеспечивать более высокий уровень открытости при появлении проблем, которые они могут обсудить друг с другом и клиентами, чтобы понять суть возникших проблем и возможные способы их решения.

Литература


1 Handler D.P., Barbier J., Schottmiller P. SMB Public Cloud Adoption. $51 Billion of Enterprise Disruption. Cisco. Internet Business Solutions Group (IBSG). 2012.
2 Research in action White Paper: the hidden costs of managing applications in the cloud. Research In Action GmbH. 2103.
3 Kepes B. Cloudonomics: the Economics of Cloud Computing. Diversity Limited, 2011.
4 Technology Hardware. This special SG Global Report. «SG Cross Assert Research». 2014.

УДК: 331.522:316.334.22

# П.А. ЛЕБЕДЕВ, Д.А. ШУРЫГИНА
# АНАЛИТИКА ДАННЫХ РЕЗЮМЕ И ВАКАНСИЙ

*ЛЕБЕДЕВ Павел Андреевич – к. соц. н., руководитель направления исследований портала Superjob.ru (г. Москва). E-mail: p.lebedev@superjob.ru*
*ШУРЫГИНА Дарья Алексеевна – старший аналитик портала Superjob.ru (г. Москва. E-mail: d.shurygina@superjob.ru*

Исследовательский центр рекрутингового портала SuperJob занимается анализом тенденций на рынке труда с 2005 г. Основой для аналитических выкладок служит большой, постоянно пополняемый архив резюме (более 15 млн на январь 2015 г.) и вакансий (около 300 тысяч активных вакансий ежедневно).

Традиционно большая часть аналитики строится на основе четырех базовых показателей:

- предложение (резюме);
- спрос (вакансии);
- зарплатные предложения в вакансиях;
- зарплатные ожидания в резюме.

Остальные показатели чаще всего являются или производными (например, конкуренция на рынке труда – соотношение числа резюме к количеству вакансий), или динамическими метриками.

Усредненные данные на макроуровне мало интересуют внутренних и внешних заказчиков. Более интересен региональный и/или отраслевой разрез. Эти параметры задаются пользователем в процессе создания резюме/вакансии – их необходимо привязать к городу и отрасли/сфере деятельности. На этом уровне мы имеем дело с данными, большими только по объему, а не по сложности структуры или скорости прироста.

Однако соискатель и работодатель мыслят категорией позиции/должности, и задача аналитиков усложняется. В контексте внутренних бизнес-задач наиболее значимым оказывается вопрос правильной классификации/группировки должностей, для того чтобы реализовывать релевантную функцию поиска и всевозможных подсказок, а также для адекватного SEO-продвижения.

Для внешней аналитики разделение на должности необходимо при подготовке:

- обзоров заработных плат по определенной позиции, в том числе в зависимости от навыков и опыта;
- оперативных внутриотраслевых и межрегиональных обзоров по рынку труда;
- анализа карьерных траекторий.

Только на первый взгляд кажется, что название должности/позиции единообразно заполняется большинством пользователей. Синонимы и различные варианты написания приводят к тому, что массив данных становится слабоструктурированным, а необходимость постоянной актуализации и пересчета данных в итоге определяют две других характеристики больших данных.

В рамках доклада на конкретных примерах будет показано, каким образом аналитические задачи решались вне парадигмы Big Data и как они начинают решаться в процессе внедрения парадигмы Big Data.

| Задача | До парадигмы Big Data | Парадигма Big Data |
|---|---|---|
| Классификация должностей | Ограниченное количество должностей (около 100). Словари синонимов. Редкое обновление | Все возможные варианты написания должностей. Текстовая кластеризация (около 5000 позиций) |
| Анализ заработных плат по позиции | Выборочный анализ, основанный на экспертной позиции. Стандартный набор метрик (минимум, максимум, медиана, персентили) | Анализ всего массива резюме и вакансий. Синтетические метрики, группировка по отраслевым сегментам |
| Анализ должностных обязанностей | Качественный подход, генерализация | Количественный анализ. Выявление навыков и умений, анализ из взаимосвязей для разных позиций и должностей |
| Анализ карьерных траекторий | В контексте расширения функционала и/или освоения менеджерских функций. Простые модели | Сложные модели, учитывающие разнообразные факторы |

На данный момент мы говорим о постепенном переходе Superjob от одной аналитической парадигмы к другой. Некоторые задачи уже решаются в парадигме Big Data, другие находятся в стадии тестирования и пилотных проектов, третьи – только оформляются как идеи. Говоря об этом переходе, нельзя обойти стороной сложности и барьеры на пути внедрения. Большая часть этих сложностей традиционна для любого процесса внедрения инноваций на производстве.

УДК: 316.477

# Е.С. МИТРОФАНОВА, А.В. АРТАМОНОВА
# ОСОБЕННОСТИ ПОДГОТОВКИ ДАННЫХ О СОБЫТИЯХ ЖИЗНЕННОГО ПУТИ К АНАЛИЗУ ПРОДВИНУТЫМИ СТАТИСТИЧЕСКИМИ МЕТОДАМИ

*Е.С. МИТРОФАНОВА – магистр социальных наук, младший научный сотрудник Института демографии НИУ ВШЭ, научный сотрудник Института социального анализа и прогнозирования РАНХиГС. E-mail: mitrofanovy@yandex.ru*

*А.В. АРТАМОНОВА – студентка НИУ ВШЭ, участник Научно-учебной группы НИУ ВШЭ «Изучение рождаемости, формирования, развития и распада семей на данных выборочных обследований». E-mail: alyona89152694371@yandex.ru*

Большинство современных исследователей работает на стыке дисциплин, так как каждая научная область разрабатывает свой инструментарий, элементы которого могут представлять интерес для коллег из других сфер науки и практики. Так, анализ жизненных событий невозможно выполнить в рамках одной предметной области, ведь сама концепция жизненного пути является междисциплинарной с момента ее основания: зародившись в психологии, она стала важной теоретической и методологической рамкой для социологов, антропологов, демографов [Кон, 1999; Levy, 2005; Levy, Ghisletta, 2005].

Методы, которые используются для анализа жизненных событий, включают наработки из медицины, социологии, демографии, статистики, эконометрики, программирования. В данной работе мы уделим внимание наиболее продвинутым способам анализа событий жизненного пути: Event history analysis (анализ наступления событий) [Бурдяк, 2007; Blossfeld, Rohwer, 2002; Mills, 2011] и Sequence analysis (анализ последовательностей) [ Abbott, 1995; Billari, Fürnkranz, Prskawetz, 2006]. Фокусом нашего доклада будет не детальное описание непосредственной работы данными методами, которые можно найти в научных статьях, на форумах, в справочниках, а особенности подготовки данных, о которых большинство исследователей умалчивает.

Одна из основных причин того, почему о подготовке данных к исследованию методами событийного и последовательного анализа говорится так мало, заключается в том, что этот процесс очень трудоемок, требует нестандартных, творческих решений, хорошей подготовки в области программирования. Мы не претендуем на универсальность разработанных нами решений, но предлагаем их в качестве ориентира для тех коллег, которые в той или иной мере исследуют события жизненного пути. К событиям жизненного пути можно отнести демографические события (вступление в союзы, расставания, рождение детей, переезды), социоэкономические (трудоустройство, получение образования, отделение от родителей, взятие кредита) и другие.

Для того чтобы анализировать события жизненного пути, нужны биографии индивидов, а значит, либо лонгитюдные, либо ретроспективные данные. Таким условиям удовлетворяет лишь несколько обследований, проведенных в России. Наиболее известные из них – панельное и ретроспективное обследование «Родители и дети, мужчины и женщины», проведенное в 2004, 2007 и 2011 гг.[1] и ретроспективное обследование «Человек, семья, общество», проведенное в 2013 г. [ Малева, 2014].

[1] Подробнее: http://www.socpol.ru/gender/RIDMIZ.shtml.

После того как панельные данные гармонизированы, а все расхождения между датами наступления событий в биографиях респондентов устранены, можно приступать к подготовке данных к анализу продвинутыми методами. Эта подготовка различается в зависимости от метода, который мы будем использовать. Для начала рассмотрим случай работы с Event history analysis.

*Event history analysis (EHA)*

Первое решение, которое предстоит принять, – в каком статистическом пакете и с какой степенью детализации мы будем анализировать события. Если исследователь располагает ретроспективными данными или панельными данными за небольшой период, работать можно в SPSS. SPSS подойдет и в том случае, если не ставится цель очень специфицированной настройки моделей. Если же исследователь располагает панельными данными за большой промежуток времени или ему нужен дополнительный функционал при анализе событий, больше подойдут такие статпакеты, как Stata или R.

В случае, если для работы выбран SPSS, никакой особой подготовки, кроме создания специальной временной переменной и переменной-индикатора наступления искомого события, не требуется. Если выбраны Stata или R, появляется возможность углубить анализ за счет использования данных неиндивидуального уровня (Person-Level Data), а формата «индивид–период» (Person–Period Data). Различие между этими форматами состоит в том, что в первом каждому респонденту соответствует одна строка, где содержится вся информация о нем (рис. 1), а во втором – у одного респондента может быть столько строк, сколько было волн опроса (рис. 2). Главное преимущество формата «индивид–период» заключается в том, что в каждой новой строке мы можем фиксировать состояние индивида на момент каждой волны опроса. Это могут быть сопутствующие жизненные события, финансовое положение, ценностные установки и т.д.

| | ID | дата_рождения | работа1 | брак1 | ребенок1 | брак_обязателен |
|---|---|---|---|---|---|---|
| 1 | 1 | 10.09.67 | 04.03.84 | 12.12.85 | 05.09.87 | 5 |
| 2 | 2 | 04.03.90 | 21.03.08 | . | . | 1 |
| 3 | 3 | 11.01.81 | 02.02.05 | 01.03.07 | 06.01.11 | 5 |
| 4 | 4 | 21.07.54 | 09.09.70 | 17.11.72 | 28.05.73 | 5 |

**Рисунок 1** - Данные индивидуального уровня (Person-Level Data)

| | ID | дата_рождения | работа1 | брак1 | ребенок1 | брак_обязателен |
|---|---|---|---|---|---|---|
| 1 | 1 | 10.09.67 | 04.03.84 | 12.12.85 | 05.09.87 | 4 |
| 2 | 1 | 10.09.67 | 04.03.84 | 12.12.85 | 05.09.87 | 4 |
| 3 | 1 | 10.09.67 | 04.03.84 | 12.12.85 | 05.09.87 | 5 |
| 4 | 2 | 04.03.89 | . | . | . | 2 |
| 5 | 2 | 04.03.89 | . | . | . | 2 |
| 6 | 2 | 04.03.90 | 21.03.08 | . | . | 1 |
| 7 | 3 | 11.01.81 | . | . | . | 2 |
| 8 | 3 | 11.01.81 | 02.02.05 | 01.03.07 | . | 5 |
| 9 | 3 | 11.01.81 | 02.02.05 | 01.03.07 | 06.01.11 | 5 |
| 10 | 4 | 21.07.54 | 09.09.70 | 17.11.72 | 28.05.73 | 5 |
| 11 | 4 | 21.07.54 | 09.09.70 | 17.11.72 | 28.05.73 | 5 |
| 12 | 4 | 21.07.54 | 09.09.70 | 17.11.72 | 28.05.73 | 5 |

**Рисунок 2 -** Данные в формате «индивид–период» (Person–Period Data)

Еще одно существенное отличие Stata и R от SPSS в том, что в последней реализована только регрессия Кокса, которая является полупараметрическим методом, т.е. она не позволяет делать предположения о распределении данных, кроме уже заложенного в данную модель – о пропорциональности рисков наступления событий. Stata и R позволяют работать со всеми типами методов, в том числе комбинировать разные распределения в рамках одной модели.

### *Sequence analysis (SA)*

В отличие от EHA, очень распространенного в зарубежной социологии и демографии, а также стремительно завоевывающего свою аудиторию в России, SA ввиду трудоемкости использования все еще набирает популярность в России и мире. Поэтому если об особенностях работы методом EHA довольно легко найти информацию даже на русском языке, то про SA написано еще очень мало, в основном в медицине и биологии. Количество программ, в которых реализован SA, также невелико. Наиболее знакомый из них социологам и демографам – R. Для R разработан специальный пакет TraMineR[2], содержащий готовые функции для анализа последовательностей. Чтобы воспользоваться этим пакетом, работа с которым хорошо описана в англоязычном справочнике [ Gabadinho, 2009], необходимо осуществить целый ряд шагов по преобразованию данных.

В первую очередь необходимо учесть следующую особенность: для построения цепочки событий нужно, чтобы для всех событий, про которые упомянул респондент, у нас имелась дата. Если на вопрос «Было ли у вас событие N?» респондент ответил утвердительно, но на следующий вопрос «Когда у вас наступило данное событие?» не дал ответа, данный респондент должен быть исключен из наблюдения. Ведь фактически данное событие произошло, но мы не можем разместить его на временной оси и сопоставить с остальными событиями жизни данного индивида, что внесет искажение в общий анализ последовательностей.

Для того чтобы построить фильтр для отсечения данных респондентов, нужно сопоставить ответы на оба вопроса – про факт наступления события и про дату. Особое внимание следует уделить неоднократным событиям (браки, рождения детей, образования разных уровней) и ситуации, когда приходится собирать фактологическую переменную сквозь несколько волн.

После того как искомые респонденты удалены, наступает этап подготовки данных непосредственно к SA. Необходимо перейти от формата событий к формату статусов, т.е. вместо даты наступления каждого события (левый прямоугольник на рис. 3) должны быть указаны статусы респондентов в каждый момент времени (правый прямоугольник на рис. 3). В качестве интервала можно взять месяцы или годы, а в качестве стартовой точки для анализа, например, взросления, 15-летие. Далее нужно определиться с «алфавитом» – набором буквенных сокращений для каждого события. Например, работа – Р, зарегистрированный союз – С, дети – Д. И противоположные состояния: безработица – Б, холост/не замужем – Х, нет детей – Н.

| | ID | дата_рождения | работа1 | брак1 | ребенок1 | возр_15 | возр_16 | возр_17 | возр_18 | возр_19 | возр_20 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 10.09.67 | 04.03.84 | 12.12.85 | 05.09.87 | БХН | БХН | РХН | РСН | РСН | РСД |
| 2 | 2 | 04.03.90 | 21.03.08 | . | . | БХН | БХН | БХН | РХН | РХН | РХН |
| 3 | 3 | 11.01.81 | 02.02.05 | 01.03.07 | 06.01.11 | БХН | РХН | РХН | РХН | РХН | РХН |
| 4 | 4 | 21.07.54 | 09.09.70 | 17.11.72 | 28.05.73 | БХН | РХН | РХН | РСН | РСД | РСД |

[2] Подробнее: http://mephisto.unige.ch/traminer.

Рисунок 3 - Переход от формата событий к формату статусов

Перейти от формата событий к формату статусов не так просто. Сначала надо закодировать событие каждого типа, указав факт его наступления или ненаступления, стартовую и конечную точку, а затем все эти события сгруппировать в статусы для каждого временного интервала и разместить по ячейкам. Если мы анализируем взросление и исследуем временное пространство между 15- и 30-летием с точностью до месяца, нам понадобится 180 ячеек для каждого респондента.

Существует много нюансов при подготовке данных к анализу событий жизненного пути продвинутыми методами. Мы постарались упомянуть самые важные шаги, которые необходимо сделать при подготовке к работе методами событийного и последовательного анализа.

Литература


1 Бурдяк А.Я. Применение метода «Анализ наступления события (Event history analysis)» с помощью пакета SPSS // SPERO. 2007. Т. 6. С. 189–202.Кон И.С. Социологическая психология. Воронеж: МОДЭК, 1999. С. 254–270.

2 Малева Т.М. и др. Разработка методологии и проведение первой пилотной волны регулярного общенационального репрезентативного обследования населения по изучению демографического, социального и экономического поведения, включая пенсионное поведение. М.: ФГБОУ ВПО Российская академия народного хозяйства и государственной службы при Президенте РФ, 2014.

3 Abbott A. Sequence Analysis: New Methods for Old Ideas //Sociological Methods Research. 1995. Vol. 21, N 1. P. 93.

4 Billari F., Fürnkranz J., Prskawetz A. Timing, Sequencing, and Quantum of Life Course Events: A Machine Learning Approach // Eur. J. Popul. 2006. Vol. 22, N 1. P. 37–65Blossfeld H.-P., Rohwer G. Techniques of event history modeling: new approaches to causal analysis. New York: Lawrence Erlbaum Associates, 2002.

5 Gabadinho A. et. al. Mining sequence data in R with the TraMineR package: A users guide for version 1.2. – Geneva: Geneva Univ., 2009.

6 Levy R. et. al. Incitations for Interdisciplinarity in Life Course Research // Adv. Life Course Res. 2005. Vol. 10. P. 361–391.Levy R., Ghisletta P. Towards an interdisciplinary perspective on the Life course, 2005

7 Mills M. Introducing survival and event history analysis. – SAGE Publications, 2011.

УДК: 303.432

# А.С. ДМИТРИЕВ
# BIG DATA, 4V: VOLUME, VELOCITY, VARIETY, VALUE

*ДМИТРИЕВ Александр Станиславович – ведущий системный архитектор IBM, MBA, Warwick University (Warwick Business School). E-mail: admitriev@ru.ibm.com*

Так называемые решения в области больших данных (Big Data) являются качественным скачком в возможностях работы с данными. В первую очередь появилась возможность получения огромного количества разнообразных данных об окружающей нас реальности в режиме реального или близкого к реальному времени и извлечения из этого потока неожиданно полезной информации. Однако эти новые возможности ставят целый ряд достаточно сложных задач, связанных как с техническими особенностями проведения исследований на основании сложных компьютеризированных систем, так и с этическими аспектами.

Основные рекомендации для успешного решения выдвигаемых задач – тесное сотрудничество с техническими специалистами по сложным системам и сочетания так называемых инновационных и традиционных методов социологии для контроля полученных результатов.

## Что такое Big Data и основные сферы применения

«Big Data» – новый, не устоявшийся еще термин в сфере информационных технологий.

Мнемонически принято считать, что Big Data обладают четырьмя особенностями, на английском – 4V: Volume, Velocity, Variety, Value, т.е. большими объемами, скоростью их возникновения, разнообразием и внутренним ценным смыслом.

Тема Big Data напрямую связана с появлением таких объемов данных, для обработки которых в настоящее время нет адекватного инструментария.

Вторым важным моментом является не то, что такого инструментария нет в принципе, важно, что для той или иной отрасли, фирмы, конкретного случая такой инструментарий может отсутствовать либо на его приобретение/создание нужны недоступные в конкретном случае ресурсы.

Таким образом, задача по решению проблемы Big Data сводится к следующим положениям:

- обработка больших массивов данных в режиме, соответствующем скорости бизнес-процесса;
- полученные результаты должны иметь практическое для бизнеса значение и давать при принятии решений конкретные конкурентные преимущества.

Итак, речь идет о создании новых рынков сбыта и новой индустрии производства.

В целом Big Data должны обеспечить:

- Прозрачность и скорость использования информации о потенциальном объекте бизнеса для быстрого принятия правильных решений.

- Превращать любую информацию в строгую цифровую отчетность, позволяющую моделирование или контролируемые эксперименты для принятия правильных решений при многих возможных вариантах.
- Создавать более четкую категоризацию в интересующей области (сегментация клиентов, продуктов, сервисов и т.д.).
- Поддержать процесс принятия решений.
- На основании полученной реакции от рынка, клиентов, других объектов бизнеса формировать следующие, усовершенствованные на основании обработки и анализа полученных данных, продукты и услуги.

### Решения для рынка и возможности в них для социологии

Рассмотрим 3 важных направления:

1 маркетинг (формирование предложений на основе учета вкусов, предпочтений и персональной истории клиента/групп клиентов);
2 финансовый сектор (отслеживание кредитной истории, обоснование выделений финансовых ресурсов, борьба с мошенничеством);
3 социальный сектор (отслеживание предпочтений тех или иных групп населения, отслеживание формирования новых социальных образований, возможность влияния на эти группы).

Что касается маркетинга, это задача связана с отслеживанием предпочтений потенциальных клиентов и их персональной истории.

Остаются открытыми следующие вопросы:

1 выборка;
2 валидность/верификация полученных данных;
3 интерпретация полученных данных.

Есть и более сложные проблемы, в первую очередь речь идет о правильной постановке задачи техническим исполнителям, создателям инструментария для обследования. Классическому социологу необходимо разбираться не только в статистических методах, но и в получении, обработке и анализе данных, равно как и в механизмах ИТ-индустрии вообще. Наиболее важным, видимо, будет понимание модели облачных решений как одной из перспективных форм предоставления сервисов, в ходе реализации которых могут собираться данные о респондентах.

Проблемой является также искажение результатов (или проблемы интерпретации). Методологически определение поправки на интернет-искажение (назовем это так) – весьма сложная задача, которая тоже пока не решена.

### Общая архитектура решения Big Data

При работе с традиционными массивами данных заранее известно, в каких единицах измеряются наши данные, что они представляют, что мы хотим из них извлечь. При работе с Big Data заранее неизвестно, какие зависимости будут обнаружены в многоформатном,

постоянно поступающем в реальном времени массиве данных. Это различие в подходах в литературе обычно называется Information Architecture Paradigm Shift.

Вторым принципиальным отличием является то, что в традиционных массивах данные как бы загружались в аналитический инструментарий для их анализа, а в Big Data аналитический инструментарий приближается к данным, непосредственно встраивается в механизмы получения и обработки информации таким образом, чтобы обрабатывать их на лету, максимально быстро.

Типы данных для обработки при работе с Big Data подразделяются на целый ряд отличающихся категорий, в которых необходимо разбираться.

**Необходимые технологические элементы: вычислительные ресурсы, «датчики» для сбора данных, каналы передачи информации, обрабатывающие программы, инструменты передачи обратной связи (порталы, целевая рассылка и пр.)**

В этом разделе дается описание элементов компьютерной системы, работающей с большими данными, собираемыми в реальном режиме времени.

**Поддерживающие механизмы (облачная инфраструктура, необходимый сервис)**

В этом разделе рассматриваются варианты облачных решений для понимания того, с кем и когда придется сотрудничать социологам при работе в подобных проектах.

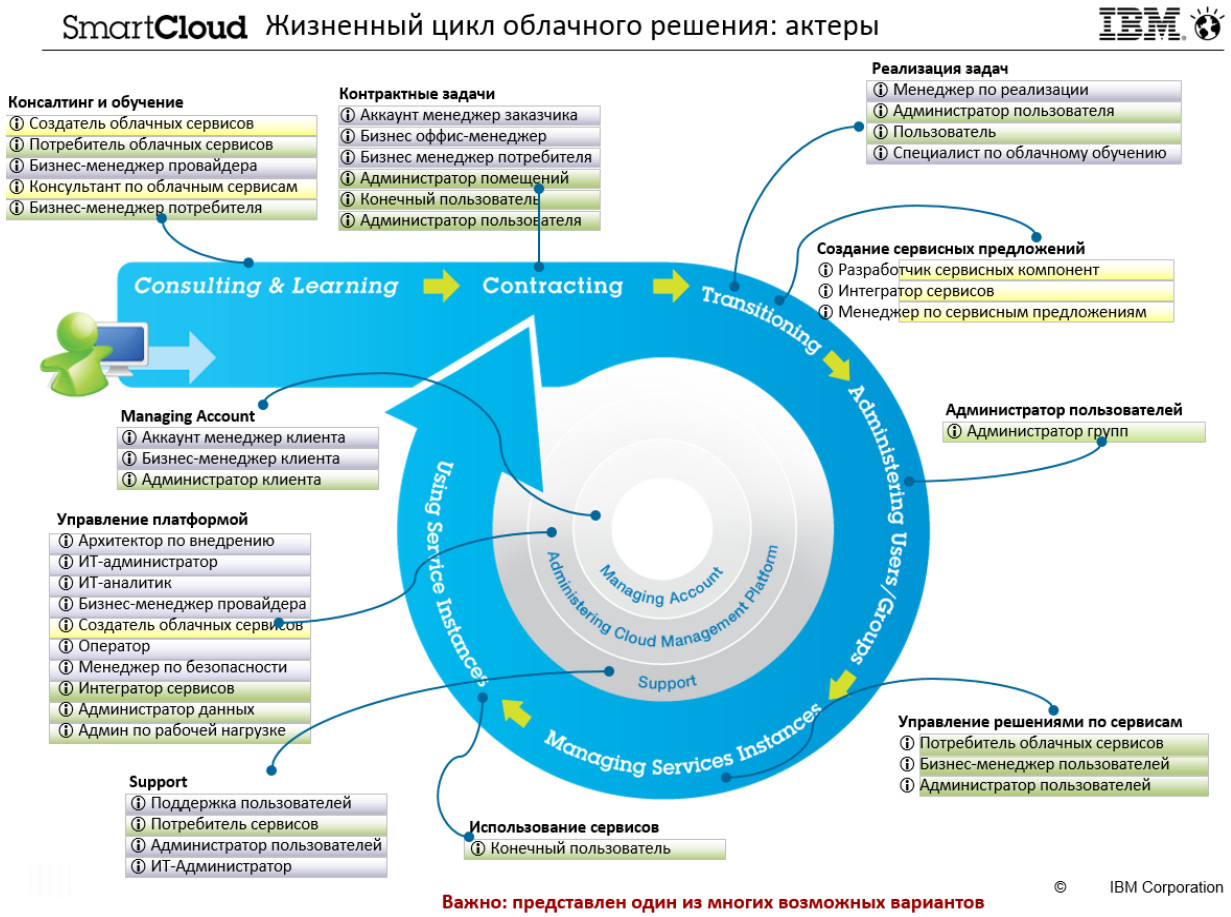


**Рисунок 2** - Роли в поддержке облачного решения. Жизненный цикл решения. Материалы IBM

На разных этапах создания облачного решения им занимаются разные либо одни и те же люди, выступающие в разных ролях.

При составлении опросников либо вариантов получения обратной связи в другой форме необходимо будет ставить задачи разработчикам решения, тесно общаться с теми, кто будет, собственно, получать эти данные. Не понимая суть процесса, получить положительный результат затруднительно.

Между источником данных и социологом-рецепиентом, если можно так выразиться, пролегает огромная технологическая структура. Принципы ее организации и работы желательно понимать, чтобы не получить искаженных результатов.

### Дополнительные необходимые аспекты. Обеспечение очистки данных и соблюдение единства форматов

Понимание форматов данных, существующих при работе с Big Data, а также типов программного обеспечения для работы с ними может существенно облегчить задачу их очистки, трансформации, анализа. В этом разделе даются общепринятые типы/форматы данных и программного обеспечения, работающего с ними.

Категоризация и понимание форматов данных является важнейшим условиям обеспечения правильного методологического подхода при социологических исследованиях.

### Обеспечение защиты (надежность, непрерывность и т.д.), в том числе шифрации данных, этические аспекты

Вопросы обеспечения защиты данных, которыми обычно занимаются ИТ-специалисты в области безопасности, а также специалисты по построению защищенных от различных воздействий (катастроф, сбоев электропитания, человеческих ошибок и т.д.), касаются и социологов при работе с большими данными.

Одна из основных проблем – защита полученных данных, а также результатов анализа. Соблюдение режима анонимности при определенных исследованиях обязательно для социологов, особенно если это касается возможных негативных последствий для респондента. Последствия же неверного решения могут быть весьма печальными.

Литература


1 An Oracle enterprise architecture white paper – an enterprise architect’s guide to big data, 2015, whitepaper. URL: http://www.oracle.com/technetwork/topics/entarch/oea-framework-133702.pdf

2 Big data: The next frontier for innovation, competition, and productivity. – McKinsey Global Institute, 2011.

3 Внутренние материалы IBM