ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ «ОРЕНБУРГСКИЙ ГОСУДАРСТВЕННЫЙ АГРАРНЫЙ УНИВЕРСИТЕТ»

МЕТОДИЧЕСКИЕ МАТЕРИАЛЫ ДЛЯ ОБУЧАЮЩИХСЯ ПО ОСВОЕНИЮ ДИСЦИПЛИНЫ

Б1.В.ДВ.06.01 Аналитика больших массивов данных

Направление подготовки (специальность) 27.03.04 Управление в технических системах

Профиль подготовки (специализация) Интеллектуальные системы обработки информации и управления

Квалификация выпускника бакалавр

Форма обучения очная

СОДЕРЖАНИЕ

1. КОНСПЕКТ ЛЕКЦИЙ

1.1. Лекция № 1, 2 (4 часа)

Тема: «Определение больших данных. Технологии хранения больших данных»

1.1.1. Вопросы лекции:

1. Большие данные (big data) в информационных технологиях.

2. Совокупность подходов, инструментов и методов обработки структурированных и неструктурированных данных огромных объёмов и значительного многообразия.

3. Средства массово-параллельной обработки неопределённо структурированных данных решениями категории NoSQL, алгоритмами MapReduce, программными каркасами и библиотеками проекта Hadoop.

4. Определяющие характеристики для больших данных: объём, скорость, многообразие.

1.1.2. Краткое содержание вопросов:

Для «больших данных» нет строгого определения. Изначально идея состояла в том, что объем информации настолько вырос, что рассматриваемое количество уже фактически не помещалось в памяти компьютера, используемой для обработки, поэтому инженерам потребовалось модернизировать инструменты для анализа всех данных. Так появились новые технологии обработки, например модель MapReduce компании Google и ее аналог с открытым исходным кодом – Hadoop от компании Yahoo. Они дали возможность управлять намного большим количеством данных, чем прежде. При этом важно, что их не нужно было выстраивать в аккуратные ряды или классические таблицы баз данных. На горизонте также появились другие технологии обработки данных, которые обходились без прежней жесткой иерархии и однородности. В то же время интернет-компании, имеющие возможность собирать огромные массивы данных и острый финансовый стимул для их анализа, стали ведущими пользователями новейших технологий обработки, вытесняя компании, которые порой имели на десятки лет больше опыта, но работали автономно.

Согласно одному из подходов к этому вопросу (который мы рассматриваем в этой книге), понятие «большие данные» относится к операциям, которые можно выполнять исключительно в большом масштабе. Это порождает новые идеи и позволяет создавать новые формы стоимости, тем самым изменяя рынки, организации, отношения между гражданами и правительствами, а также многое другое.

И это только начало. Эпоха больших данных ставит под вопрос наш образ жизни и способ взаимодействия с миром. Поразительнее всего то, что обществу придется отказаться от понимания причинности в пользу простых корреляций: променять знание почему на что именно. Это переворачивает веками установленный порядок вещей и ставит под сомнение наши фундаментальные знания о том, как принимать решения и постигать действительность.

Большие данные знаменуют начало глубоких изменений. Подобно тому как телескоп дал нам возможность постичь Вселенную, а микроскоп – получить представление о микробах, новые методы сбора и анализа огромного массива данных помогут разобраться в окружающем мире с использованием способов, ценность которых мы только начинаем осознавать. Но настоящая революция заключается не в компьютерах, которые вычисляют данные, а в самих данных и в том, как мы их используем.

Чтобы понять, на каком этапе находится информационная революция, рассмотрим существующие тенденции. Наша цифровая Вселенная постоянно расширяется. Возьмем астрономию.

Когда в 2000 году стартовал проект «Слоуновский цифровой обзор неба», его телескоп в Нью-Мексико за первые несколько недель собрал больше данных, чем

накопилось за всю историю астрономии. К 2010 году его архив был забит грандиозным количеством информации: 140 терабайт. А его преемник, телескоп Large Synoptic Survey Telescope, который введут в эксплуатацию в Чили в 2016 году, будет получать такое количество данных каждые пять дней.

За подобными астрономическими цифрами не обязательно далеко ходить. В 2003 году впервые в мире расшифровали геном человека, после чего еще десять лет интенсивной работы ушло на построение последовательности из трех миллиардов основных пар. Прошел почти десяток лет – и то же количество ДНК анализируется каждые 15 минут с помощью геномных машин по всему миру. В 2012 году стоимость определения последовательности генома человека упала ниже одной тысячи долларов. Эта процедура стала доступной широким массам. Что касается области финансов, через фондовые рынки США каждый день проходит около семи миллиардов обменных операций, из них около двух третей торгов решаются с помощью компьютерных алгоритмов на основе математических моделей, которые обрабатывают горы данных, чтобы спрогнозировать прибыль, снижая при этом по возможности риски.

Перегруженность в особенности коснулась интернет-компаний. Google обрабатывает более петабайта данных в день – это примерно в 100 раз больше всех печатных материалов Библиотеки Конгресса США. Facebook – компания, которой не было в помине десятилетие назад, – может похвастать более чем 10 миллионами загрузок новых фотографий ежечасно. Люди нажимают кнопку «Нравится» или пишут комментарии почти три миллиарда раз в день, оставляя за собой цифровой след, с помощью которого компания изучает предпочтения пользователей. А 800 миллионов ежемесячных пользователей службы YouTube компании Google каждую секунду загружают видео длительностью более часа. Количество сообщений в Twitter увеличивается приблизительно на 200 % в год и к 2012 году превысило 400 миллионов твитов в день.

От науки до здравоохранения, от банковского дела до интернета… Сферы могут быть разными, но итог один: объем данных в мире быстро растет, опережая не только наши вычислительные машины, но и воображение.

Немало людей пыталось оценить реальный объем окружающей нас информации и рассчитать темп ее роста. Они достигли разного успеха, поскольку измеряли разные вещи. Одно из наиболее полных исследований провел Мартин Гилберт из школы коммуникаций им. Анненберга при Университете Южной Калифорнии. Он стремился сосчитать все, что производилось, хранилось и передавалось. Это не только книги, картины, электронные письма, фотографии, музыка и видео (аналоговые и цифровые), но и видеоигры, телефонные звонки и даже автомобильные навигационные системы, а также письма, отправленные по почте. Он также брал в расчет вещательные СМИ, телевидение и радио, учитывая охват аудитории.

По его расчетам, в 2007 году хранилось или отправлялось примерно 2,25 зеттабайта данных. Это примерно в пять раз больше, чем 20 лет назад (около 435 экзабайт). Чтобы представить это наглядно, возьмем полнометражный художественный фильм. В цифровом виде его можно сжать до файла размером в один гигабайт. Экзабайт состоит из миллиарда гигабайт. Зеттабайт – примерно в тысячу раз больше. Проще говоря, немыслимо много.

Если рассматривать только хранящуюся информацию, не включая вещательные СМИ, проявляются интересные тенденции. В 2007 году насчитывалось примерно 300 экзабайт сохраненных данных, из которых около 7 % были представлены в аналоговом формате (бумажные документы, книги, фотоснимки и т. д.), а остальные – в цифровом. Однако совсем недавно наблюдалась иная картина. Хотя идея «информационного века» и «цифровой деревни» родилась еще в 1960-х годах, это действительно довольно новое явление, учитывая некоторые показатели. Еще в 2000 году количество информации, хранящейся в цифровом формате, составляло всего одну четверть общего количества информации в мире. А остальные три четверти содержались

в бумажных документах, на пленке, виниловых грампластинках, магнитных кассетах и подобных носителях.

В то время цифровой информации насчитывалось не так много – шокирующий факт для тех, кто уже продолжительное время пользуется интернетом и покупает книги онлайн. (В 1986 году около 40 % вычислительной мощности общего назначения в мире приходилось на карманные калькуляторы, вычислительная мощность которых была больше, чем у всех персональных компьютеров того времени.) Из-за быстрого роста цифровых данных (которые, согласно Гилберту, удваивались каждые три с лишним года) ситуация стремительно менялась. Количество аналоговой информации, напротив, практически не увеличивалось.

Таким образом, к 2013 году количество хранящейся информации в мире составило 1,2 зеттабайта, из которых на нецифровую информацию приходится менее 2 %[13 - По оценкам за 2013 год, объем сохраненной информации равен 1,2 зеттабайта, из которых нецифровая информация составляет менее 2 % (из интервью Гилберта Кукьеру).].

Трудно представить себе такой объем данных. Если записать данные в книгах, ими можно было бы покрыть всю поверхность Соединенных Штатов в 52 слоя. Если записать данные на компакт-диски и сложить их в пять стопок, то каждая из них будет высотой до Луны. В III веке до н. э. считалось, что весь интеллектуальный багаж человечества хранится в великой Александрийской библиотеке, поскольку египетский царь Птолемей II стремился сохранить копии всех письменных трудов. Сейчас же в мире накопилось столько цифровой информации, что на каждого живущего ее приходится в 320 раз больше, чем хранилось в Александрийской библиотеке.

Процессы действительно ускоряются. Объем хранящейся информации растет в четыре раза быстрее, чем мировая экономика, в то время как вычислительная мощность компьютеров увеличивается в девять раз быстрее. Неудивительно, что люди жалуются на информационную перегрузку. Всех буквально захлестнула волна изменений.

Рассмотрим перспективы, сравнив текущий поток данных с более ранней информационной революцией. Она была связана с изобретением ручного типографского станка Гутенберга около 1450 года. По данным историка Элизабет Эйзенштейн, за 50 лет – с 1453 по 1503 год – напечатано около восьми миллионов книг. Это больше, чем все книжники Европы произвели с момента основания Константинополя примерно 1650 годами ранее. Другими словами, потребовалось 50 лет, чтобы приблизительно вдвое увеличить информационный фонд всей Европы (в то время, вероятно, она представляла львиную долю всего мирового запаса слов). Для сравнения: сегодня это происходит каждые три дня.

Что означает это увеличение? Питер Норвиг, эксперт по искусственному интеллекту в компании Google, прежде работавший в Лаборатории реактивного движения НАСА, любит в этом случае проводить аналогию с изображениями. Для начала он предлагает взглянуть на наскальные изображения лошади в пещере Ласко во Франции, которые относятся к эпохе палеолита (17 тысяч лет назад). Затем – на фотографию лошади или, еще лучше, работы кисти Пабло Пикассо, которые по виду не слишком отличаются от наскальных рисунков. Между прочим, когда Пикассо показали изображения Ласко, он саркастически заметил: «[С тех пор] мы ничего не изобрели».

Он был прав, но лишь отчасти. Вернемся к фотографии лошади. Если раньше, чтобы нарисовать лошадь, приходилось потратить много времени, теперь ее можно запечатлеть гораздо быстрее. В этом и состоит изменение. Хотя оно может показаться не столь важным, поскольку результат по большому счету одинаков: изображение лошади. А теперь представьте, как делается снимок лошади, и ускорьте его до 24 кадров в секунду. Теперь количественное изменение переросло в качественное. Фильм коренным образом отличается от стоп-кадра. То же самое и с большими данными: изменяя количество, мы меняем суть.

Из курса физики и биологии нам известно, что изменение масштаба иногда приводит к изменению состояния. Обратимся к другой аналогии, на сей раз из области нанотехнологий, где речь идет об уменьшении объектов, а не их увеличении. Принцип, лежащий в основе нанотехнологий, заключается в том, что на молекулярном уровне физические свойства меняются. Появляется возможность придать материалам характеристики, недоступные ранее. Например, медь, которая в обычном состоянии проводит электричество, на наноуровне обнаруживает сопротивление в присутствии магнитного поля, а серебро имеет более выраженные антибактериальные свойства. Гибкие металлы и эластичная керамика тоже возможны на наноуровне. Подобным образом при увеличении масштаба обрабатываемых данных появляются новые возможности, недоступные при обработке меньших объемов.

Иногда ограничения, которые мы воспринимаем как должное и считаем всеобщими, на самом деле имеют место только в масштабе нашей деятельности. Рассмотрим третью аналогию, и на сей раз из области науки. Для людей важнейшим физическим законом является гравитация: она распространяется на все сферы нашей деятельности. Но для мелких насекомых гравитация несущественна. Ограничение, действующее в их физической вселенной, – поверхностное натяжение, позволяющее им, например, ходить по воде. Но людям, как правило, до этого нет дела.

То же самое с информацией: размер имеет значение. Так, поисковая система Google определяет распространение гриппа не хуже, чем официальная статистика, основанная на реальных визитах пациентов к врачу. Для этого системе нужно произвести тщательный анализ сотен миллиардов условий поиска, в результате чего она дает ответ в режиме реального времени, то есть намного быстрее, чем официальные источники. Таким же образом система Farecast прогнозирует колебания цен на авиабилеты, вручая потребителям эффективный экономический инструмент. Однако обе системы достигают этого лишь путем анализа сотен миллиардов точек данных.

Эти два примера, с одной стороны, демонстрируют научное и общественное значение больших данных, а с другой – показывают, что с их помощью можно извлечь экономическую выгоду. Они знаменуют два способа, которыми мир больших данных готов радикально изменить все: от бизнеса и естественных наук до здравоохранения, государственного управления, образования, экономики, гуманитарных наук и других аспектов жизни общества.

Мы стоим на пороге эпохи больших данных, однако полагаемся на них ежедневно. Спам-фильтры разрабатываются с учетом автоматической адаптации к изменению типов нежелательных электронных писем, ведь программное обеспечение нельзя запрограммировать таким образом, чтобы блокировать слово «виагра» или бесконечное количество его вариантов. Сайты знакомств подбирают пары на основе корреляции многочисленных атрибутов с теми, кто ранее составил удачные пары. Функция автозамены в смартфонах отслеживает действия пользователя и добавляет новые вводимые слова в свой орфографический словарь. И это только начало. От автомобилей, способных определять момент для поворота или торможения, до компьютеров IBM Watson, которые обыгрывают людей на игровом шоу Jeopardy, – этот подход во многом изменит наше представление о мире, в котором мы живем.

По сути, большие данные предназначены для прогнозирования. Обычно их описывают как часть компьютерной науки под названием «искусственный интеллект» (точнее, ее раздел «машинное обучение»). Такая характеристика вводит в заблуждение, поскольку речь идет не о попытке «научить» компьютер «думать», как люди. Вместо этого рассматривается применение математических приемов к большому количеству данных для прогноза вероятностей, например таких: что электронное письмо является спамом; что вместо слова «коипя» предполагалось набрать «копия»; что траектория и скорость движения человека, переходящего дорогу в неположенном месте, говорят о том, что он успеет перейти улицу вовремя и автомобилю нужно лишь немного снизить

скорость. Но главное – эти системы работают эффективно благодаря поступлению большого количества данных, на основе которых они могут строить свои прогнозы. Более того, системы спроектированы таким образом, чтобы со временем улучшаться за счет отслеживания самых полезных сигналов и моделей по мере поступления новых данных.

1.2. Лекция № 3, 4, 5 (6 часов)

Тема: «Технологии обработки больших данных. Технологии анализа больших данных. Научные проблемы в области больших данных»

1.2.1. Вопросы лекции:

1. Методы и техники анализа, применимые к большим данным.
2. Методы класса Data Mining: обучение ассоциативным правилам.
3. Классификация, кластерный анализ, регрессионный анализ.
4. Краудсорсинг, смешение и интеграция данных.
5. Машинное обучение, сетевой анализ, оптимизация.

1.2.2. Краткое содержание вопросов:

Объем данных в организациях настолько возрос, что привел к увеличению массива знаний, который выходит за рамки экономической ценности и практической применимости. Это дало толчок к развитию информационных технологий, появлению интеллектуальных технологий анализа деловых данных, аналитических систем и систем интеллектуальной поддержки принятия решений на их базе. Новые ИТ позволили найти нетривиальные подходы к автоматизации управленческого труда и отказаться от старых методов управления.

Технологии интеллектуального анализа данных обеспечивают формирование аналитических данных посредством выполнения операции очищения данных локальных баз организации, применения статистических методов и других сложных алгоритмов. Появлению аналитических систем способствовало осознание руководящим звеном предприятий факта, что в базах данных содержится не только информация, но и знания (скрытые закономерности). Последние позволяют охарактеризовать процесс управления предприятием и дать интеллектуальную информацию для более обоснованного принятия решений.

Можно выделить следующие технологии интеллектуального анализа данных:

- оперативный анализ данных посредством OLAP-систем;
- поиск и интеллектуальный выбор данных Data Mining;
- деловые интеллектуальные технологии BIS;
- интеллектуальный анализ текстовой информации.

Аналитические системы OLAP(On-Line Analytical Processing) предназначены для анализа больших объемов информации в интерактивном режиме для создания интеллектуального капитала (аналитических данных), позволяющего руководителю принять обоснованное решение. Они обеспечивают:

- агрегирование и детализацию данных по запросу.
- выдачу данных в терминах предметной области.
- анализ деловой информации по множеству параметров (например, поставщик, его местоположение, поставляемый товар, цены, сроки поставки и т.д.).
- многопроходный анализ информации, который позволяет выявить не всегда очевидные тенденции в исследуемой предметной области.
- произвольные срезы данных по наименованию, выбираемых из разных внутренних и внешних источников (например, по наименованию товара).
- выполнение аналитических операций с использованием статистических и других методов.

- согласование данных во времени для использования в прогнозах, трендах, сравнениях (например, согласование курса рубля).

Аналитические системы позволяют использовать данные новым образом. Вместо поиска отдельных фактов они позволяют получать результаты не через экспериментирование, теоретизирование или моделирование, а посредством информационных операций (установление корреляций, тенденций, других статистических методов). Появилась еще одна форма информационного процесса - наблюдение за текущей информацией.

Концепция технологии OLAP была сформулирована Эдгаром Коддом в 1993 году. Она стала ключевым компонентом организации данных в информационных хранилищах и их применении. Эта технология основана на построении многомерных наборов данных - OLAP-кубов. Целью использования технологий OLAP является анализ данных и представление этого анализа в виде, удобном для восприятия управленческим персоналом и принятия на их основе решений.

Основные требования, предъявляемые к приложениям для многомерного анализа:

- предоставление пользователю результатов анализа за приемлемое время (не более 5 сек.);
- осуществление логического и статистического анализа, его сохранение и отображение в доступном для пользователя виде;
- многопользовательский доступ к данным;
- многомерное представление данных;
- возможность обращаться к любой информации независимо от места ее хранения и объема.

Аналитические данные содержат факты и агрегатные данные.

Факт- это число, значение. Над фактами производятся различные операции: суммирование, группировка, вычисление средних, максимальных, минимальных значений для получения агрегатных данных.

Агрегатное данное - суммарное, среднее, минимальное, максимальное и другое значение, полученное посредством статистических операций над фактами. Операции над фактами выполняются вдоль определенных измерений.

Под измерением понимается один из ключей данных, в разрезе которого можно выполнять разные операции: получать, фильтровать, группировать и отражать информацию о фактах. Примеры измерений: страна, клиент, товар, поставщик. Измерения могут иметь иерархическую структуру. Например, в стране может быть несколько городов, в городе - несколько клиентов, их могут обслуживать различные поставщики из тех же или других городов и стран. Для отображения иерархии измерений используются раз-личные модели иерархий. Модели иерархий служат основой построения многомерных баз данных и метаданных в информационных хранилищах.

Разработано несколько способов хранения аналитических данных Наибольший эффект достигается при использовании многомерных кубов. Следует отметить, что OLAP-функциональность может быть реализована различными способами, начиная с простейших средств анализа данных в офисных приложениях и заканчивая распределенными аналитическими системами, основанными на серверных продуктах. Т.е. OLAP — это не технология, а идеология.

Прежде чем говорить о различных реализациях OLAP, давайте подробнее рассмотрим, что же представляют собой кубы с логической точки зрения.

Мы будем использовать для иллюстрации принципов OLAP базу данных Northwind, входящую в комплекты поставки Microsoft SQL Server и представляющую собой типичную базу данных, хранящую сведения о торговых операциях компании, занимающейся оптовыми поставками продовольствия. К таким данным относятся сведения о поставщиках, клиентах, список поставляемых товаров и их категорий, данные о заказах и заказанных товарах, список сотрудников компании.

Возьмем для примера таблицу Invoices1, которая содержит заказы фирмы. Поля в данной таблице будут следующие:

- Дата Заказа
- Страна
- Город
- Название заказчика
- Компания-доставщик
- Название товара
- Количество товара
- Сумма заказа.

Какие агрегатные данные мы можем получить на основе этого представления? Обычно это ответы на вопросы типа:

- Какова суммарная стоимость заказов, сделанных клиентами из определенной страны?
- Какова суммарная стоимость заказов, сделанных клиентами из определенной страны и доставленных определенной компанией?
- Какова суммарная стоимость заказов, сделанных клиентами из определенной страны в заданном году и доставленных определенной компанией?

Все эти данные можно получить из этой таблицы вполне очевидными SQL-запросами с группировкой.

Результатом этого запроса всегда будет столбец чисел и список атрибутов его описывающих (например, страна) – это одномерный набор данных или, говоря математическим языком, – вектор.

Представим себе, что нам надо получить информацию по суммарной стоимости заказов из всех стран и их распределение по компаниям доставщиков – мы получим уже таблицу (матрицу) из чисел, где в заголовках колонок будут перечислены доставщики, в заголовках строк – страны, а в ячейках будет сумма заказов. Это – двумерный массив данных. Такой набор данных называется сводной таблицей (pivot table) или кросс-таблицей.

Если же нам захочется получить те же данные, но еще в разрезе годов, тогда появится еще одно изменение, т.е. набор данных станет трехмерным (условным тензором 3-го порядка или 3-х мерным «кубом»).

Очевидно, что максимальное количество измерений – это количество всех атрибутов (Дата, Страна, Заказчик и т.д.), описывающих наши агрегируемые данные (сумму заказов, количество товаров и т.п).

Так мы приходим к понятию многомерности и его воплощению – многомерному кубу. Такая таблица будет у нас называться «таблицей фактов». Измерения или Оси куба (dimensions) – это атрибуты, координаты которых – выражаются индивидуальными значениями этих атрибутов, присутствующих в таблице фактов. Т.е. например, если информация о заказах велась в системе с 2003 по 2014 год, то эта ось годов будет состоять из 12 соответствующих точек. Если заказы приходят из трех стран, то ось стран будет содержать 3 точки и т.д. Независимо от того, сколько стран заложено в справочнике Стран. Точки на оси называются ее «членами» (Members).

Сами агрегируемые данные в данном случае буду назваться «мерами» (Measure). Чтобы избежать путаницы с «измерениями», последние предпочтительней называть «осями». Набор мер образует еще одну ось «Меры» (Measures). В ней столько членов (точек), сколько мер (агрегируемых столбцов) в таблице фактов.

Члены измерений или осей могут быть объединены одной или несколькими иерархиями (hierarchy). Что такое иерархия, поясним на примере: города из заказов могут быть объединены в районы, районы в области, области страны, страны в континенты или

другие образования. Т.е. налицо иерархическая структура – континент-страна-область-район-город – 5 уровней (Level). Для района данные агрегируются по всем городам, которые в него входят. Для области по всем районам, которые содержат все города и т.п. Зачем нужно несколько иерархий? Например, по оси с датой заказа мы можем хотеть группировать точки (т.е. дни) по иерархии Год-Месяц-День или по Год-Неделя-День: в обоих случаях по три уровня. Очевидно, что Неделя и Месяц по-разному группируют дни. Бывают также иерархии, количество уровней в которых не детерминировано и зависит от данных. Например, папки на компьютерном диске.

Агрегация данных может происходить с использованием нескольких стандартных функций: сумма, минимум, максимум, среднее, количество.

Как исходные, так и агрегатные данные могут храниться либо в реляционных, либо в многомерных базах данных MDD (MultiDimensional Data). В настоящее время применяются три способа хранения многомерных баз данных:

- Системы оперативной аналитической обработки многомерных баз данных MOLAP (Multidimensional OLAP) - исходные и агрегатные данные хранятся в многомер-ной базе данных. Многомерные базы данных представляют собой гиперкубы или поли-кубы. В гиперкубах все измерения имеют одинаковую размерность. В поликубе каждое измерение имеет свою размерность. Многомерная база данных оказывается избыточной, так как она полностью содержит исходные данные реляционных баз.
- Системы оперативной аналитической обработки реляционных баз данных ROLAP (Relational OLAP) - исходные данные остаются в реляционной базе, агрегатные данные размещаются в кэш той же базы.
- Гибридные системы оперативной аналитической обработки данных HOLAP (Hybrid OLAP) - исходные данные остаются в реляционной базе, а агрегатные данные хранятся в многомерной базе данных (MDD).

Многомерный анализ данных может быть произведен посредством клиентских приложений и серверных OLAP-систем.

Клиентские приложения, содержащие OLAP-средства, позволяют вычислять агрегатные данные. Агрегатные данные размещаются в кэш внутри адресного пространства такого OLAP-средства. Кэш - быстродействующий буфер большой емкости, работающий по специальному алгоритму. При этом, если исходные данные находятся в реляционной базе, вычисления производятся OLAP-средствами клиентского приложения. Если исходные данные размещаются на сервере баз данных, то OLAP-средства приложений посылают SQL-запросы на сервер баз данных и получают агрегатные данные, вычисленные сервером.

Примерами клиентских приложений, содержащими OLAP-средства, являются приложения статистической обработки данных SEWSS (Statistic Enterprise - Wide SPS System) фирмы StatSoft и MS Excel 2003, 2007, 2010, 2013. Excel позволяет создать и сохранить небольшой локальный многомерный OLAP-куб и отобразить его двух- или трехмерные сечения (разреза).

Многие средства проектирования позволяют создавать простейшие OLAP-средства. Например, Borland Delphi и Borland C++ Builder.

Отметим, что клиентские приложения применяются при малом числе измерений (не более шести) и небольшом разнообразии значений этих измерений.

Серверные OLAP-системы развили идею сохранения кэш с агрегатными данными.

В них сохранение и изменение агрегатных данных, поддержка содержащего их хранилища осуществляется отдельным приложением (процессом), называемым OLAP-сервером. Клиентские приложения делают запросы к OLAP-серверу и получают требуемые агрегатные данные. Серверные OLAP-системы рассчитаны на любое количество измерений.

Применение OLAP-серверов сокращает трафик сети, время обслуживания запросов, сокращает требования к ресурсам клиентских приложений.

В масштабе предприятия обычно используются OLAP-серверы типа Oracle Express Server, MS SQL Server 2000 Analysis Services и др.

Заметим, что MS Excel позволяет делать запросы к OLAP-серверам.

Серверные OLAP-системы на базе информационных хранилищ поддерживают все три способы хранения данных.

Аналитическая система обеспечивает выдачу агрегатных данных по запросам клиентов. Сложность аналитических систем вызвана реализацией сложных интеллектуальных запросов. Интеллектуальные запросы осуществляют поиск по условию или алгоритму вычисления ответа. Например, выбрать для выпуска изделия, приносящие максимальную прибыль. Само условие может доопределяться в ходе формирования ответа, что усложняет алгоритм формирования ответа. Данные для формирования ответа могут находиться в разных внутренних и внешних базах. Существующий язык запросов SQL рас-ширяется возможностью построения интеллектуальных запросов. Пример такого запроса - сравнить данные о продажах в конкретные месяцы, но разные годы. Для таких запросов используются непроцедурные языки обращения к многомерным базам данных. Приме-ром такого языка запросов является язык MDX (Multidimensional Expressions). Он позволяет формировать запрос и описывать алгоритм вычислений. Язык SQL используется для извлечения данных из локальных баз. Язык MDX служит для извлечения данных из многомерных баз и информационных хранилищ.

Аналитические данные используются в системах поддержки принятия решений

В идеале работа аналитиков и руководителей различных уровней должна быть организована так, чтобы они могли иметь доступ ко всей интересующей их информации и пользоваться удобными и простыми средствами представления и работы с этой информацией. Именно на достижение этих целей и направлены информационные технологии, объединяющиеся под общим названием хранилищ данных и бизнес-анализа.

В соответствии с определением Gartner, бизнес-анализ (BI, Business Intelligence) - это категория приложений и технологий для сбора, хранения, анализа и публикации данных, позволяющая корпоративным пользователям принимать лучшие решения. В русскоязычной терминологии подобные системы называются также системами поддержки принятия решений (СППР).

Сбор и хранение информации, а также решение задач информационно-поискового запроса эффективно реализуются средствами систем управления базами данных (СУБД). В OLTP (Online Transaction Processing)-подсистемах реализуется транзакционная обработка данных. Непосредственно OLTP-системы не подходят для полноценного анализа информации в силу противоречивости требований, предъявляемых к OLTP-системам и СППР.

Для предоставления необходимой для принятия решений информации обычно приходится собирать данные из нескольких транзакционных баз данных различной структуры и содержания. Основная проблема при этом состоит в несогласованности и противоречивости этих баз-источников, отсутствии единого логического взгляда на корпоративные данные.

Поэтому для объединения в одной системе OLTP и СППР для реализации подсистемы хранения используются концепция хранилищ данных (ХД). В основе концепции ХД лежит идея разделения данных, используемых для оперативной обработки и для решения задач анализа, что позволяет оптимизировать структуры хранения. ХД позволяет интегрировать ранее разъединенные детализированные данные, содержащиеся в исторических архивах, накапливаемых в традиционных OLTP-системах, поступающих из внешних источников, в единую базу данных, осуществляя их предварительное согласование и, возможно, агрегацию.

Подсистема анализа может быть построена на основе:

- подсистемы информационно-поискового анализа на базе реляционных СУБД и статических запросов с использованием языка SQL;
- подсистемы оперативного анализа. Для реализации таких подсистем применяется технология оперативной аналитической обработки данных OLAP, использующая концепцию многомерного представления данных;
- подсистемы интеллектуального анализа, реализующие методы и алгоритмы Data Mining.

Примером OLAP-систем является Brio Query Enterprise корпорации Brio Technology. OLAP-средства включают в свои системы фирмы 1С, Парус и др.

Технологии Data Mining (добыча данных)разработаны для поиска и выявления в данных скрытых связей и взаимозависимостей с целью предоставления их руководителю в процессе принятия решений. Для этого используются статистические методы корреляции, оптимизации и методы, позволяющие находить эти зависимости и синтезировать дедуктивную (обобщающую) информацию. Технологии Data Mining обеспечивают:

- Поиск зависимых данных (реализацию интеллектуальных запросов);
- Выявление устойчивых бизнес-групп (выявление групп объектов, близких по заданным критериям);
- Ранжирование важности измерений при классификации объектов для проведения анализа (страна, город, район, поставщик);
- Прогнозирование бизнес-показателей (например, ожидаемые продажи, спрос);
- Оценка влияния принимаемых решений на достижение успеха предприятия;
- Поиск аномалий и т.д.

Технологии Data Mining позволяют наблюдать за текущей информациейс целью поиска отклонений, тенденций без вникания в смысл самих данных. Их используют, на-пример, для оценки поведения покупателей, чтобы внести изменения рекламную тактику, для корректировки выпуска продукции, изменения ценовой политики и т.д.

Интеллектуальные деловые технологии BIS(Business Intelligence Services) преобразуют информацию из внутренних и внешних баз в интеллектуальный капитал (аналитические данные). Главными задачами систем интеллектуального выбора данных является поиск функциональных и логических закономерностей в накопленных данных для подсказки обоснованных управленческих решений. Они основаны на применении технологий информационного хранилища и алгоритмов автоматизации деловых процессов (Workflow). Аналитические данные предоставляются руководству всех уровней и работникам аналитических служб организации по запросам в удобном виде.

Для интеллектуального анализа текстовой информации разработаны струк-турные аналитические технологии (CAT).Они ориентированы на углубленную обра-ботку неструктурированной информации. Реализуют уникальную способность человека интерпретировать (толковать) содержание текстовой информации и устанавливать связи между фрагментами текста. CAT реализованы на базе гипертекстовой технологии, лингвистических процессоров, семантических сетей.

Структурные аналитические технологии предназначены для решения разнообразных задач аналитического характера на основе структуризации предварительно отобранной текстовой информации. Являются инструментом создания аналитических док-ладов, отчетов, статей, заметок для использования в информационно-аналитических службах организаций, отраслей, государственного управления, СМИ и т.д.

1.3. Лекция № 6, 7, 8 (6 часов)

Тема: «Статистические методы анализа больших данных»

1.3.1. Вопросы лекции:

1. Понятие прогноза и предвидения. Отличие прогнозирования от предвидения.

2. Закон распределения случайной величины. Статистические оценки параметров. Доверительные области. Теория моментов.

3. Корреляционный анализ. Модель множественной линейной регрессии. Доверительные интервалы для зависимой переменной. Сглаживание временных рядов.

4. Многомерный анализ. Регрессия. Классификация. Кластеризация.

5. Адаптивные и мультипликативные методы прогнозирования. Экспоненциальное сглаживание. Авторегрессионные модели. Модели скользящего среднего.

1.3.2. Краткое содержание вопросов:

Сегодня для анализа Больших Данных пытаются использовать подходы и методы, разработанные еще при создании технологий информационных хранилищ, и хотя при этом делаются поправки с учетом количественных показателей, в них не учитываются принципы сбора, обработки и анализа данных. Вместе с тем некоторые особенности традиционных операций могут противоречить специфике обработки Больших Данных.

Существенное различие задач оперативной и аналитической обработки данных стало проявляться еще на заре развития технологий баз данных. Термин хранилища данных (Data Warehouse) был предложен Биллом Инмоном еще в 70-х годах, однако всплеск интереса к этим технологиям произошел лишь 20 лет спустя, когда, во-первых, возникла реальная потребность в подобных системах и, во-вторых, стали доступны необходимые вычислительные мощности.

Цикл обработки данных в информационном хранилище включает в себя сбор, очистку, загрузку, анализ и, наконец, представление результатов анализа. Здесь нет смысла подробно останавливаться на этих этапах, но необходимо четко обозначить главный тезис — если пытаться применить технологии информационных хранилищ для анализа Больших Данных, то следует обратить внимание не только на алгоритмы анализа, но и на все этапы работы с данными.

Сбор

В русле информационных хранилищ предполагается, что информация извлекается из оперативных баз данных, преобразуется к необходимому виду, проверяется и только затем загружается в систему. Перечисленные операции выполняются с некоторой периодичностью, и здесь сразу же возникает вопрос: всегда ли возможна такая «периодичность» при работе с Большими Данными, которые потоком поступают на вход и должны быть доступны для анализа как можно раньше? Может оказаться, что промежуток времени между появлением информации и ее доступностью для анализа меньше, чем время, необходимое для выполнения операций по построению информационного хранилища. Примером такой задачи является мониторинг социальных сетей на предмет выявления негативных высказываний или фактов утечек конфиденциальной информации — все эти события должны быть выявлены и нейтрализованы как можно раньше. Однако здесь мы имеем дело с неформализованным представлением данных, для переработки которых требуются алгоритмы интеллектуального анализа текста, изначально не отличающиеся пока высоким быстродействием.

Мониторинг социальных сетей

На вход этой задачи подаются высказывания пользователей, комментарии, выставленные оценки, фотографии и т. д. Цели мониторинга могут быть самые разные: выявление негативных высказываний, пресечение утечек информации, определение быстро распространяющейся информации и ее источников, определение цепочек влияния и авторитетных пользователей. Очевидно, что следует осуществлять мониторинг популярных социальных сетей с большим количеством пользователей. Также очевидно,

что из-за большого числа пользователей и их высокой активности решение данной задачи подразумевает сбор и обработку больших объемов данных, представленных в неформализованном виде.

В информационных хранилищах неявно подразумевается, что при предварительной обработке данных (например, при поиске несоответствий) может использоваться ранее накопленное содержимое хранилища, что трудновыполнимо при работе с Большими Данными. Проблема заключается прежде всего в том, что они всегда распределены, причем не так, как было бы удобно для анализа, а так, как было удобно их собирать — например, если речь идет о телекоммуникационных системах, то данные «складываются» на региональных серверах. С точки зрения анализа более удобно распределять данные не по территориальному, а по временному признаку (каждый сервер отвечает за конкретный промежуток времени) и т. п. Однако опять же информация станет доступной для анализа только после того, как будет перенесена на необходимые серверы.

Итак, в традиционных информационных хранилищах все данные всегда проходят через единый логический блок, отвечающий за их конвертацию, проверку, очистку, загрузку, и время выполнения этих операций редко бывает критичным для всей остальной системы. Однако при обработке Больших Данных такого единого блока быть не может. Справедливости ради стоит отметить, что пока существует еще не так много задач с интенсивным входным потоком данных и еще возможно реализовать блок сбора, очистки, преобразования и загрузки в виде распределенной, но логически единой системы.

Мониторинг массовых рассылок

Одним из способов борьбы с нежелательными рассылками является определение факта массовых рассылок одного и того же или немного откорректированного письма. Эта задача выполняется обычно почтовым сервером, однако для ее точного решения требуется еще учитывать статистику с других серверов. Возникает задача анализа больших потоков входящих писем, сбора необходимой статистики, обмена статистикой с другими серверами и принятия решения о том, является ли каждое полученное письмо массовым. Допустимое время на принятие решения по каждому письму составляет примерно одну-две минуты — в противном случае полезные письма будут доходить до адресата с задержкой. Если исходить из того, что на почтовый сервер приходит примерно 500 писем в секунду, то получается, что в оперативной обработке сервера будет находиться примерно 60 тыс. писем (500 писем x 60 секунд x 2 минуты). Если средний размер одного письма составляет 5 Кбайт, то потребуется примерно 300 Мбайт памяти. Вместе с тем одно и то же спам-сообщение может рассылаться многократно в течение продолжительного периода времени (например, в течение 20 дней). В этом случае серверу необходимо хранить статистику о 850 млн писем (500 писем * 60 секунд x 60 минут x 24 часа x 20 дней). Если по каждому письму хранить хотя бы 1 Кбайт данных, то вместе с индексами получается уже больше 1 Тбайт постоянно обновляющихся данных на каждом сервере. А если учитывать общее количество почтовых серверов, то итоговая сумма получается внушительной.

Анализ

Традиционно информационные хранилища предоставляют примерно одинаковый набор инструментов анализа данных: многомерный анализ (OLAP), регрессия, классификация, кластеризация и поиск закономерностей. Сегодня появились и продукты — например, SAP HANA, Greenplum Chorus, Aster Data nCluster, — позволяющие запускать перечисленные методы и на Больших Данных.

Для понимания потенциальных возможностей таких решений необходимо рассмотреть лежащие в их основе алгоритмы, а также проанализировать пути для их возможного распараллеливания — ключа к обработке Больших Данных. При этом важно не привязываться к конкретным технологиям распределенной обработки данных (например, MapReduce), а лишь учитывать ключевые параметры, характерные для Больших Данных (интенсивность сетевого взаимодействия и объемы).

Важным фактором, влияющим на скорость работы любой СУБД, является количество операций ввода/вывода и эффективность построенных индексов. Для работы с Большими Данными применимы все уже существующие методы, начиная от классических B-деревьев и заканчивая сложными структурами для оперирования многомерной информацией.

Многомерный анализ

Суть метода заключается в построении многомерного куба и получении его различных срезов. Результатом анализа, как правило, является таблица, в ячейках которой содержатся агрегированные показатели (количество, среднее, минимальное или максимальное значение и так далее). В зависимости от реализации, системы многомерного анализа делятся на MOLAP, ROLAP и HOLAP. Среди них ROLAP-системы являются наиболее прозрачными и изученными, поскольку основываются на широко распространенных реляционных СУБД, в то время как внутреннее устройство MOLAP и HOLAP обычно более закрыто и относится к области «ноу-хау» конкретных коммерческих продуктов.

MOLAP представляет информацию в виде «честной» многомерной модели, но внутри используются те же подходы, что и в ROLAP: схемы «звезда» и «снежинка». С точки зрения СУБД база данных ROLAP — это обыкновенная реляционная база, и для нее необходимо поддерживать весь перечень операций. Однако это не позволяет, во-первых, жестко контролировать этапы ввода данных. Во-вторых, собирать статистику и подбирать оптимальные структуры для хранения индексов. В-третьих, оптимизировать размещение данных на диске для обеспечения высокой скорости ввода/вывода. В-четвертых, нет возможности для кэширования промежуточных агрегатных значений. В-пятых, при выполнении аналитических запросов из-за высоких требований к быстродействию нет возможности произвести глубокий статистический анализ и выработать оптимальный план выполнения. В ROLAP используются «родные» реляционные оптимизаторы запроса, которые никак не учитывают «многомерность» базы данных. Технологии MOLAP лишены перечисленных недостатков и благодаря этому позволяют добиться большей скорости анализа.

Выбор технологии MOLAP/ROLAP/HOLAP при анализе Больших Данных зависит от частоты обновления базы данных. С точки зрения распараллеливания обработки, на первый взгляд, все просто — любой многомерный куб может быть «разрезан» по делениям одного из измерений и распределен между несколькими серверами. Например, можно разделить куб на временные периоды (по годам и месяцам), по территориальному признаку (каждый сервер отвечает за свой регион) и так далее. Критерием для разделения куба является следующий принцип: выполнение многомерного запроса должно ложиться не на один сервер, а на несколько, после чего полученные результаты собираются в единое целое. Например, если пользователь запрашивает статистику продаж по стране за указанный промежуток времени, а данные распределены по нескольким региональным OLAP-серверам, то каждый сервер возвращает свой собственный ответ, которые затем собираются воедино. Если же данные будут распределены по временному критерию, то при выполнении рассматриваемого примера запроса вся нагрузка ляжет на один сервер.

Проблема в том, что, во-первых, очень трудно заранее определить оптимальное распределение данных по серверам, а во-вторых, для части аналитических запросов может быть заранее неизвестно, какие данные и с каких серверов понадобятся.

Применительно к Большим Данным это означает, что существующие подходы для многомерного анализа могут хорошо масштабироваться и что они допускают распределенный сбор информации (рис. 1) — каждый сервер может самостоятельно собирать информацию, осуществлять ее очистку и загрузку в локальную базу.

Регрессия

Под регрессией понимают построение параметрической функции, описывающей изменение указанной числовой величины в указанный промежуток времени. Эта функция

строится на основе известных данных, а затем используется для предсказания дальнейших значений этой же величины. На вход метода поступает последовательность пар вида «время — значение», описывающая поведение этой величины при заданных условиях, например количество продаж конкретного вида товара в конкретном регионе. На выходе — параметры функции, описывающей поведение исследуемой величины.

Независимо от вида используемой параметрической функции подбор значений ее параметров осуществляется одним и тем же способом. Вычисляется суммарная разница между наблюдаемыми (то есть поданными на вход метода) значениями величины и значениями, которые дает функция при текущих значениях ее параметров. Затем определяется, как следует подкорректировать значения параметров для того, чтобы уменьшить текущую суммарную разницу. Эти операции повторяются до тех пор, пока суммарная разница не достигнет необходимого минимума или ее дальнейшее уменьшение станет невозможным.

С точки зрения обработки данных при регрессионном анализе ключевыми операциями являются вычисление текущей суммарной разницы и корректировка значений параметров. Если первая операция распараллеливается очевидным образом (сумма вычисляется по частям на отдельных серверах, а затем суммируется на центральном сервере), то со второй сложнее. В наиболее общем случае при корректировке весов используют общеизвестный математический факт: функция нескольких параметров возрастает в направлении градиента и убывает в направлении, обратном градиенту. В свою очередь, вычисление градиента состоит в вычислении частных производных функции по каждому из параметров, что сводится к дискретному дифференцированию, основанному на вычислении взвешенных сумм. В результате корректировка значений параметров также сводится к суммированию, которое может быть распараллелено.

Если регрессионный анализ сводится к вычислению взвешенных сумм, то он обладает примерно той же степенью применимости и при работе с Большими Данными, что и многомерный анализ. То есть системы регрессионного анализа волне могут масштабироваться и работать в условиях распределенного сбора информации.

Классификация

Задача классификации отчасти похожа на задачу регрессии и заключается в попытке построения и использования зависимости одной переменной от нескольких других. Например, имея базу данных о цене объектов недвижимости, можно построить систему правил, позволяющую на основе параметров нового объекта предсказать его примерную цену. Отличие классификации от регрессии состоит в том, что анализируется не временной ряд — подаваемые на вход значения никак не могут быть упорядочены.

На текущий момент разработано множество методов классификации (функции Байеса, нейронные сети, машины поддерживающих векторов, деревья решений и т. д.), каждый из которых имеет под собой хорошо проработанную научную теорию (самообучающиеся системы, метод обучения с учителем). Вместе с тем все методы классификации строятся по одной и той же схеме. Сначала производится обучение алгоритма на сравнительно небольшой выборке, а затем — применение полученных правил к остальной выборке. На первом этапе возможно копирование массива данных на один сервер для запуска «классического» алгоритма обучения без распараллеливания работы. Однако на втором этапе данные могут обрабатываться независимо — система правил, полученная по итогам самообучения, копируется на каждый сервер, и через нее прогоняется весь массив данных, хранящийся на этом сервере. Полученные результаты могут либо сохраняться там же на сервере, либо отправляться для дальнейшей обработки.

Таким образом, на этапе обучения классификаторов о работе с Большими Данными пока речи не идет — не существует выборок такого объема, подготовленных для обучения систем, а на этапе классификации отдельные порции данных обрабатываются независимо друг от друга. Это означает, что существующие методы классификации также применимы для работы с Большими Данными.

Кластеризация

Задача кластеризации состоит в разбиении множества информационных сущностей на группы, при этом члены одной группы более похожи друг на друга, чем члены из разных. В качестве критерия похожести используется функция-расстояние, на вход которой поступают две сущности, а на выходе — степень их похожести (ноль для полностью идентичных сущностей). Известно множество различных способов кластеризации (графовые, иерархические, итеративные, сети Кохонена и т. д.).

Проблема кластеризации Больших Данных состоит в том, что имеющиеся алгоритмы предполагают возможность непосредственного обращения к любой информационной сущности в исходных данных (заранее невозможно предугадать, какие именно сущности понадобятся алгоритму). В свою очередь, исходные данные могут быть распределены по разным серверам, и при этом не гарантируется, что каждый кластер хранится строго на одном сервере. Если распределение данных по серверам делать прозрачным для алгоритма кластеризации (он считает, что данные расположены в некоторой распределенной виртуальной памяти), то это неизбежно приведет к копированию больших объемов с одного сервера на другой.

Решение проблемы может быть следующим. На каждом сервере запускается свой алгоритм, который оперирует только данными этого сервера, а на выходе дает параметры найденных кластеров и их веса, оцениваемые исходя из количества элементов внутри кластера. Затем полученная информация собирается на центральном сервере и производится метакластеризация — выделение групп близко расположенных кластеров с учетом их весов. Этот метод универсален, хорошо распараллеливается и может использовать любые другие алгоритмы кластеризации, однако он требует проведения серьезных научных исследований, тестирования на реальных данных и сравнения полученных результатов с другими «локальными» методами.

Таким образом, для анализа Больших Данных подавляющая часть методов кластеризации неприменима в чистом виде и необходимы дополнительные исследования.

Поиск закономерностей

Суть метода заключается в нахождении правил, описывающих взаимозависимости между внутренними элементами данных. Классическим примером является анализ покупок в супермаркете и выявление правил вида «если человек покупает пельмени, то обычно он покупает еще и сметану». На вход задачи поиска закономерностей поступает неупорядоченное множество сущностей, для каждой из которых известен набор присутствующих информационных признаков: например, такими сущностями могут быть чеки на покупки, а признаками — купленные товары. Задача поиска закономерностей сводится к выявлению правил вида «если присутствуют признаки A1, A2, ..., AN, то присутствуют и признаки B1,B2,…, BM», при этом каждое правило характеризуется двумя параметрами: вероятностью срабатывания и поддержкой. Первый параметр показывает, как часто выполняется данное правило, а второй — как часто применимо данное правило, то есть как часто встречается сочетание признаков A1, A2, ..., AN. С практической точки зрения аналитика интересуют правила с высокой поддержкой и вероятностью срабатывания, но каков должен быть баланс между этими показателями — это вопрос конкретной практической задачи.

Задача поиска закономерностей решается с помощью алгоритма Apriori, и очевидно, что с точки зрения обработки Больших Данных ключевой операцией здесь является вычисление агрегирующих функций (значений F(D1,…, Dk)), что выполняется средствами многомерного анализа. Другим важным моментом алгоритма Apriori, препятствующим обработке именно Больших Данных, может показаться работа с наборами информационных признаков, но тут необходимо учесть следующее обстоятельство: количество рассматриваемых наборов зависит от количества информационных признаков, то есть от концептуальной модели данных, а не от их объема.

Визуализация результатов

Большинство публикаций, посвященных анализу Больших Данных, сфокусированы непосредственно на анализе и оставляют за кадром обработку полученных результатов, предполагая, что будут применяться уже существующие методы в виде генерации отчетов, а также построения различного рода диаграмм или графиков. Однако для просмотра результатов анализа существующие методы могут быть неприменимы сразу по нескольким причинам. Во-первых, большое количество данных на входе порождает большое число результатов анализа на выходе — если раньше многие закономерности оказывались за пределами статистической погрешности, то теперь они отчетливо преодолевают этот барьер. Может показаться, что этим можно пренебречь и для принятия решений ограничиться только ключевыми закономерностями, но это не так. В случае Больших Данных для достижения максимальной эффективности принимаемых решений требуется учитывать даже едва различимые закономерности и тренды, иначе вообще нет смысла в обработке больших потоков самых разнообразных сведений.

Во-вторых, значительно усложняется концептуальная модель выходной информации. Если в информационных хранилищах типичный отчет имел не более десятка входных параметров (например, временной срез, регион и т.д.), а более точная параметризация была не нужна, поскольку отчет банально становился вырожденным и состоящим из пустых строк и нулей, то для Больших Данных такая вырожденность исчезает.

Задачи анализа данных можно представить в виде равнобедренного треугольника, в основании которого объем исходных данных, а любая горизонтальная прямая, проведенная на некотором уровне, показывает, как много результатов будут давать соответствующие методы анализа. По мере роста объема исходных данных вершина треугольника сначала проходит уровень простого поиска и просмотра данных, затем многомерного и статистического анализа, и лишь затем выходит на уровень Data Mining. Однако по мере роста объемов исходных данных и на уровне Data Mining становится слишком много выходной информации. Получается, что если раньше для принятия решений необходимо было просмотреть всего несколько листов отчета, то в случае Больших Данных это не так. На человека, отвечающего за принятие решений, сваливается груда данных, из которой необходимо выделить наиболее важные. Эту проблему можно решить двумя способами.

Первый заключается в использовании автоматизированных средств, позволяющих выбрать наиболее важные отчеты, — например, в случае мониторинга динамики продаж отбирать отчеты, в которых наблюдается наиболее резкое изменение показателей по сравнению с предыдущим периодом. Такой метод применим не всегда, поскольку изменение показателей может просто объясняться внешними факторами, неизвестными системе: например, резкий рост спроса на газированную воду может объясняться жарой и совсем не говорит о том, что на следующий период необходимо планировать такие же объемы продаж.

Второй способ борьбы с большим количеством отчетов заключается в реорганизации работы — выделяются отдельные люди, в обязанности которых входит просмотр отчетов и формирование резюме, направляемых лицам, принимающим решения (рис. 3). Формирование таких резюме в значительной мере отличается от существующих средств генерации отчетов по следующим причинам. Во-первых, в резюме может входить самая разнородная информация. То есть один и тот же документ может включать в себя и показатели продаж, и динамику роста цен, и изменения в штатном расписании, и что угодно еще, вплоть до фотографий торговых залов. Во-вторых, все резюме уникальны и даже могут не иметь общего формата. В-третьих, резюме всегда готовятся для конкретного человека по конкретному случаю, а следовательно, учитывают и специфику случая, и особенности этого человека. Когда-то может быть удобен распечатанный текстовый документ, когда-то — презентация в формате для выступления перед

аудиторией, а когда-то — аудиозапись или видео. Одним словом, резюме должно позволять максимально быстро и наглядно получить всю информацию, необходимую для принятия решений.

Из перечисленных требований вырисовывается концепция автоматизированных метааналитических систем, позволяющих: визуализировать результаты анализа и на их основе создавать документы и презентации, предназначенные для просмотра вручную; монтировать аудиопотоки и видеоролики; создавать Flash-анимацию. Однако эти средства и методы не так просты, как кажется на первый взгляд, и сфокусированы на отбор и максимально наглядное представление разнородной информации. Решение этой задачи вручную создает высокую нагрузку на персонал, а хотя бы частичная автоматизация работы требует междисциплинарных научных исследований, касающихся нахождения наиболее эффективных способов представления сведений в резюме.

На данный момент подобные системы кажутся излишествами, но по мере роста числа практических задач, связанных с анализом Больших Данных, быстрая и удобная подготовка отчетов станет жизненно необходима. В противном случае принимающий решения персонал может попросту захлебнуться в океане отчетов и результатах анализа.

Для работы с Большими Данными в ряде случаев применимы методы, разработанные в информационных хранилищах и на деле доказавшие свою эффективность, — некоторые из существующих алгоритмов могут быть адаптированы для обработки больших распределенных информационных массивов. Вместе с тем серьезные затруднения могут возникнуть при наглядном представлении полученных результатов — из-за огромных объемов информации, поступающей на вход, резко возрастает количество разнородных отчетов на выходе. Для их удобного представления необходимы новые программные средства, принципиально отличающиеся от генераторов отчетов, используемых для традиционных хранилищ.

1.4. Лекция № 9, 10, 11 (6 часов)

Тема: «Современные программные средства анализа больших данных».

1.4.1. Вопросы лекции:

1. SPSS Statistics - компьютерная программа для статистической обработки данных.

2. Применение SPSS Statistics для решения прикладных задач прогнозирования: ввод и хранение данных; возможность использования переменных разных типов; частотность признаков.

3. Таблицы, графики, таблицы сопряжённости, диаграммы; первичная описательная статистика в SPSS Statistics.

4. Маркетинговые и медиа исследования; анализ данных маркетинговых и медиа исследований.

1.4.2. Краткое содержание вопросов:

Потребность в средствах статистического анализа данных в различных областях деятельности, особенно в науке, очень велика, что и послужило причиной развития рынка компьютерных программ для статистической обработки данных. За последние 20 лет активное развитие получили компьютерные программы, позволяющие проводить статистический анализ больших объемов данных с целью выявления закономерностей, сравнения вероятных альтернатив выбора, построения прогнозов развития событий, обнаружения связей между явлениями и процессами и пр. Существующие программы постоянно совершенствуются в части ускорения работы с данными, улучшения представления результатов анализа данных, повышения удобства интерфейса, совершенствования справочной системы, увеличения числа встроенных в программу статистических процедур, средств обработки данных и пр.

Отрасль развивается стремительными темпами. На сегодняшний день на рынке представлено около тысячи компьютерных программ для статистической обработки данных (далее – статистические пакеты). Разнообразие статистических пакетов обусловлено многоплановостью задач обработки данных с применением различных типов статистических процедур анализа для поиска ответов на вопросы из различных областей человеческой деятельности.

Рынок компьютерных программ для статистического анализа данных характеризуется высокой конкуренцией, нередки случаи консолидации и поглощений компаний-разработчиков. Например, один из самых активных игроков на рынке компания SPSS Inc. в 1994 г. поглотила компанию SYSTAT Software Inc., а в 1996 г. – BMDP Statistical Software Inc. Эти приобретения позволили компании усовершенствовать собственные программные продукты. В частности, поглощение BMDP Software позволило усилить графические инструменты представления данных в SPSS, а поглощение SYSTAT – технологии обработки и анализа данных, полученных при биологических и медицинских исследованиях1. В 2009 году компания IBM Inc. поглотила компанию SPSS Inc.

Перед пользователями различных категорий встает вопрос выбора оптимального статистического пакета для поиска верных ответов на существующие вопросы. Очевидно, что оптимальным является вариант, сочетающий в себе необходимые функциональные возможности, высокое качество работы и умеренную цену. При выборе пакета учитываются следующие параметры:

- соответствие характеру решаемых задач;
- объем обрабатываемых данных;
- требования, предъявляемые к квалификации пользователя (уровень знаний в области статистики);
- имеющееся в наличии компьютерное оборудование.

Статистические пакеты по признаку функциональности могут быть разделены на 3 основные группы.

1) Универсальные пакеты, или пакеты общего назначения (например, SPSS, STATA, STATISTICA, S-PLUS, Stadia, STATGRAPHICS, SYSTAT, Minitab).

Эти пакеты не ориентированы на специфическую предметную область и могут применяться для анализа данных из различных областей деятельности. Как правило, они предлагают широкий диапазон статистических методов и имеют относительно простой интерфейс. С такими пакетами рекомендуется работать начинающим пользователям, владеющим лишь базовыми знаниями в области статистики, а также опытным пользователям на начальных этапах работы с данными, когда еще четко не определены статистические методы, которые будут применяться для решения того или иного вопроса. Многопрофильность универсального пакета позволяет провести пробный анализ различных типов данных с использованием широкого диапазона статистических методов. Большинство существующих универсальных пакетов имеют много пересечений по составу встроенных статистических процедур.

- Для того чтобы статистический пакет считался универсальным, он должен удовлетворять ряду требований:
- содержать достаточно широкий набор стандартных статистических методов;
- быть достаточно простым для быстрого освоения и использования непрофессиональным пользователем;
- работать с достаточно большими базами данных и отвечать высоким требованиям к вводу, преобразованию и организации хранения данных;
- осуществлять обмен данными с широко распространенными пакетами и базами данных;

- иметь обширный набор средств графического представления данных и результатов их анализа;
- иметь подробное документационное сопровождение и справочную систему, позволяющую начинающему пользователю с легкостью находить ответы на вопросы, связанные с работой программы и возможностями применения средств анализа данных.

2) Профессиональные пакеты (например, SAS, BMDP).

Профессиональные пакеты отличаются от универсальных тем, что позволяют работать со сверхбольшими объемами данных, применять узкоспециализированные методы анализа, создавать собственную систему обработки данных. Как правило, подобные пакеты сложны в освоении для непрофессионалов. В то же время подготовленным пользователям работа с профессиональным пакетом предоставит больше возможностей для глубокого и детального анализа данных, построения сложных моделей и адаптации системы к собственным потребностям. Профессиональные пакеты более дорогостоящи, чем универсальные. Например, стоимость покупки SAS Analytics Pro на один год для индивидуального пользования составляет 5 360 EUR2. Эти факторы делают современные профессиональные статистические пакеты слишком тяжеловесными для массового применения в различных областях деятельности.

3) Специализированные пакеты (например, BioStat, MESOSAUR, DATASCOPE).

В некоторых областях деятельности анализируемые данные настолько специфичны, что к ним следует применять особые методы статистического анализа, как правило, не представленные в универсальных пакетах.

Специализированные пакеты позволяют проводить анализ с использованием ограниченного числа специализированных статистических методов или применимы к использованию для решения вопросов, относящихся к отдельно взятой предметной области. Как правило, с подобными статистическими пакетами работают специалисты, хорошо знакомые с методами анализа данных в той области, на которую ориентирован пакет. Так, статистический пакет BioStat создан для анализа данных в области биологии и медицины и будет подробнее рассмотрен ниже. Российский статистический пакет MESOSAUR специализируется на анализе одномерных и многомерных временных рядов и построении регрессионных моделей. Еще один российский статистический пакет DATASCOPE специализируется на проведении анализа многомерных данных.

Целесообразно пользоваться соответствующими специализированными пакетами, когда требуется систематически решать задачи из конкретной области или применять ограниченный круг сложных статистических процедур для анализа данных из нескольких областей человеческой деятельности.

Большинство представленных на рынке статистических пакетов обладают гибкой модульной структурой, которая может пополняться и расширяться за счет пользовательских модулей, дополнительно закупаемых или находящихся в свободном доступе в Интернете. Подобная гибкость позволяет адаптировать большинство пакетов к потребностям конкретного пользователя.

По мнению профессионалов, статистический пакет должен удовлетворять следующему минимальному набору требований:

- модульность;
- ассистирование при выборе способа обработки данных;
- использование простого проблемно-ориентированного языка для формулировки задания пользователя;
- автоматическая организация процесса обработки данных;
- ведение банка данных пользователя и составление отчета о результатах проделанного анализа;

- диалоговый режим работы пользователя с пакетом;
- совместимость с другим программным обеспечением.

Как правило, представленные на рынке статистические пакеты регулярно обновляются. При этом в новой версии сохраняются или совершенствуются возможности предыдущей, а также добавляются новые возможности работы с данными. В большинстве случаев обновленные версии пакета сохраняют исходное название, изменяется лишь порядковый номер, присваиваемый конкретной версии. Самые распространенные пакеты имеют русскоязычную версию.

Разработчики большинства статистических пакетов часто утверждают, что разработанная ими программа является наилучшей для обработки данных. Учитывая многообразие предложения, подчас бывает сложно сделать правильный выбор. По мнению М. Митчелла, имеющего 20-летний опыт работы со статистическими пакетами и 11-летний опыт работы в качестве консультанта по статистике в Калифорнийском университете в Лос-Анджелесе, статистический пакет – всего лишь инструмент в руках мастера. Если специалист не обладает достаточными знаниями и компетенциями, то даже самый совершенный программный продукт не позволит провести качественный анализ данных. В то же время неправильно подобранный пакет, не обладающий необходимыми для анализа техническими характеристиками, способен замедлить работу даже выдающегося ученого, затруднив выявление необходимых закономерностей и получение верных результатов анализа данных.

Необходимо отметить, что существует минимальный набор статистических методов анализа, который включен во все рассмотренные пакеты:

- описательная статистика (базовые статистические методы, проверка нормальности распределения данных);
- дисперсионный анализ;
- непараметрическая статистика (анализ таблиц сопряженности, непараметрические сравнения, дисперсионный анализ);
- контроль качества;
- анализ выживаемости;
- кластерный анализ;
- факторный анализ;
- дискриминантный анализ;
- регрессионный анализ;
- обработка данных (сортировка, отбор, трансформация данных).

Пакет SPSS

Пакет SPSS (Statistical Package for the Social Sciences) – универсальный статистический пакет компании SPSS Inc5. Первая версия пакета была выпущена в 1968 г. В 2009 г. компания IBM поглотила SPSS Inc., поэтому новая версия пакета включает в свое название аббревиатуру IBM (IBM SPSS Statistics 19).

По мнению разработчиков пакета, SPSS является одним из лидирующих программных продуктов в области статистического анализа данных для решения вопросов в правительственной, академической и бизнессфере.

SPSS является модульной программой. Ее основу составляет базовый модуль (SPSS Base), позволяющий осуществлять управление данными и содержащий наиболее распространенные методы статистического анализа данных: проведение описательной статистики; построение линейных и нелинейных моделей; осуществление преобразования данных; проведение факторного, кластерного, дисперсионного анализов; вычисление корреляций; построение графиков; подготовка отчетов и пр.

Для проведения расширенного и углубленного анализа данных могут быть установлены дополнительные модули пакета. Для пакета IBM SPSS Statistics 19

разработаны 16 различных модулей. Например, модуль IBM SPSS Advanced Statistics предназначен для проведения анализа сложных взаимосвязей при помощи процедур, учитывающих свойства исследуемых данных, что позволяет продвинуться за рамки базового анализа данных. В модуль встроены мощные инструменты построения моделей. Модуль IBM SPSS Bootstrapping ("Самогенерация") позволяет аналитикам проверять устойчивость построенных моделей, а модуль IBM SPSS Direct Marketing ("Прямой маркетинг") предоставляет возможность маркетологам самостоятельно выполнять основные виды анализа. Модуль IBM SPSS Data Entry автоматизирует процесс разработки анкеты и ввода результатов опросов.

Достоинства SPSS:

- развитый аппарат статистического анализа;
- универсальность (может быть использован для решения широкого круга вопросов из различных предметных областей, требующих проведения статистического анализа данных);
- широкий набор статистических и графических процедур (более 50 типов диаграмм) анализа данных, а также процедур создания отчетов;
- высокая скорость вычислений, простой и удобный интерфейс;
- детальная контекстно-ориентированная справочная система, позволяющая неопытному пользователю с большей легкостью ориентироваться в программе;
- возможность свободного скачивания демонстрационной версии продукта на официальном сайте компании, наличие версий продукта на различных языках;
- совместимость с операционными системами Windows, Mac, Linux;
- наличие значительного количества литературы по работе с пакетом.

Недостатки SPSS:

- высокие требования к системе компьютера (требуется 1GB оперативной памяти, 800MB памяти на жестком диске и процессор с частотой 1GHz и выше);
- высокая цена по сравнению со статистическими пакетами аналогичного уровня (стоимость покупки для индивидуального пользования сроком на год составляет около 1000 долл.6).

Последняя версия SPSS включает в себя следующие новые возможности:

- импорт данных из Excel и SAS;
- экспорт результатов в MS Office, PDF; сохранение результатов в формате HTML;
- одновременная работа с несколькими наборами данных;
- построение диаграммы для переменных с множественными ответами;
- построение диаграммы с двумя осями Y;
- улучшенный редактор синтаксиса с поддержкой автозавершения и цветового кодирования команд;
- быстрая подготовка данных к анализу посредством Автоматизированной подготовки данных (IBM SPSS Data Preparation), позволяющей облегчить процесс интеллектуального анализа данных, выявляя и исправляя ошибки в данных и объясняя пропущенные значения. Также посредством этой функции можно подготовить отчет с рекомендациями о возможности использования данных для анализа.

2. МЕТОДИЧЕСКИЕ МАТЕРИАЛЫ ПО ПРОВЕДЕНИЮ ПРАКТИЧЕСКИХ ЗАНЯТИЙ

2.1. Практическое занятие № 1, 2, 3, 4 (8 часов).

Тема: «Определение больших данных. Технологии хранения больших данных».

2.1.1. Вопросы к занятию:

1. Большие данные (big data) в информационных технологиях.
2. Совокупность подходов, инструментов и методов обработки структурированных и неструктурированных данных огромных объёмов и значительного многообразия.
3. Средства массово-параллельной обработки неопределённо структурированных данных решениями категории NoSQL, алгоритмами MapReduce, программными каркасами и библиотеками проекта Hadoop.
4. Определяющие характеристики для больших данных: объём, скорость, многообразие.

2.1.2. Краткое описание проводимого занятия:

2.1.2.1. Проведение текущего контроля успеваемости

Задания для проведения текущего контроля успеваемости

1. Большинство данных в мире в 2011 году содержалось:

+ a) В цифровом виде

b) В аналоговом виде

2. В каком веке произошёл перевес объёмов накопленных человечеством данных в сторону цифровых?

Ответ: в двадцатом

3. Объём накопленных человечеством цифровых данных на 2012 год измеряется:

a) Петабайтами

+ b) Зеттабайтами

c) Экзабайтами

d) Йоттабайтами

4. Сколько Петабайт в Зеттабайте?

Ответ: 1024

2.2. Практическое занятие № 5, 6, 7, 8 (8 часов).

Тема: «Технологии обработки больших данных. Технологии анализа больших данных. Научные проблемы в области больших данных».

2.2.1. Вопросы к занятию:

1. Методы и техники анализа, применимые к большим данным.
2. Методы класса Data Mining: обучение ассоциативным правилам.
3. Классификация, кластерный анализ, регрессионный анализ.
4. Краудсорсинг, смешение и интеграция данных.
5. Машинное обучение, сетевой анализ, оптимизация.

2.2.2. Краткое описание проводимого занятия:

2.2.2.1. Проведение текущего контроля успеваемости

Задания для проведения текущего контроля успеваемости

1. Укажите фактор, способствовавший появлению тренда больших данных

+ a) Маркетинговые кампании крупных корпораций

+ b) Снижение издержек на хранение данных

c) Появление новых технологий обработки потоковых данных

d) Выпуск баз данных с обработкой данных в памяти

2. Какие вероятные разочарования тренда больших данных?

Ответ: Из-за угрозы безопасности личной жизни граждан будут усложнены процедуры сбора данных, что приведёт к падению ценности больших данных

3. Отметьте значимые события, повлиявшие на формирование тренда больших данных:

+ a) Разработка Hadoop

+ b) Изобретение принципа MapReduce

c) Разработка языка Pyhton

d) Победа Deepblue в матче с Г.Каспаровым

4. Выберите верный ответ:

a) Большие данные - это обработка или хранение более 1 Тб информации

+ b) Проблема больших данных - это такая проблема, когда при существующих технологиях хранения и обработки сущностная обработка данных затруднена или невозможна

c) Большие данные - это огромная PR- акция крупных вендоров и не более того

d) Большие данные - это явление, когда цифровые данные наиболее полно представляют изучаемый объект

2.3. Практическое занятие № 9, 10, 11, 12 (8 часов).

Тема: «Статистические методы анализа больших данных».

2.3.1. Вопросы к занятию:

1. Понятие прогноза и предвидения. Отличие прогнозирования от предвидения.

2. Закон распределения случайной величины. Статистические оценки параметров. Доверительные области. Теория моментов.

3. Корреляционный анализ. Модель множественной линейной регрессии. Доверительные интервалы для зависимой переменной. Сглаживание временных рядов.

4. Многомерный анализ. Регрессия. Классификация. Кластеризация.

5. Адаптивные и мультипликативные методы прогнозирования. Экспоненциальное сглаживание. Авторегрессионные модели. Модели скользящего среднего.

2.3.2. Краткое описание проводимого занятия:

2.3.2.1. Проведение текущего контроля успеваемости

Задания для проведения текущего контроля успеваемости

1. Выберите неверный ответ:

+ a) Большие данные - это данные объёма свыше 1 Тб

b) Проблема больших данных - это проблема, когда при существующих технологиях хранения и обработки сущностная обработка данных затруднена или невозможна

c) Большие данные - это тренд в области ИТ, подогреваемый маркетинговыми кампаниями крупных вендоров

d) Большие данные как правило не структурированы

2. Отметьте те из вариантов, в которых данные структурированы:

a) Данные о продажах компании, представленные в виде помесячных отчётов в формате MS Word

+ b) Таблица с ежедневными показаниями температуры помещения за год в файле формата csv

c) Текст педагогической поэмы А.С. Макаренко, представленный в формате PDF

d) Библиотека фильмов, представленных в формате mpeg на одном жестком диске

3. Перечислите четыре основных характеристики больших данных:

a) Virtualization, Volume, Variability, Vehicle

+ b) Variety, Velocity, Volume, Value

c) Verification, Volume, Velocity, Visualization

d) Video, Value, Variety, Volume

4. Выберите неверное высказывание:

+ a) Большие объёмы данных приводят к слабой их структуризации, поэтому появляется такое разнообразие данных

b) Увеличившаяся производительность телекоммуникационных каналов привела к росту объёмов передаваемой информации

c) Удешевление систем хранения на единицу информации привело к росту рынка больших данных

d) Большое разнообразие источников данных

2.4. Практическое занятие № 13, 14, 15, 16 (8 часов).

Тема: «Современные программные средства анализа больших данных».

2.4.1. Вопросы к занятию:

1. SPSS Statistics - компьютерная программа для статистической обработки данных.

2. Применение SPSS Statistics для решения прикладных задач прогнозирования: ввод и хранение данных; возможность использования переменных разных типов; частотность признаков.

3. Таблицы, графики, таблицы сопряжённости, диаграммы; первичная описательная статистика в SPSS Statistics.

4. Маркетинговые и медиа исследования; анализ данных маркетинговых и медиа исследований.

2.4.2. Краткое описание проводимого занятия:

2.4.2.1. Проведение текущего контроля успеваемости

Задания для проведения текущего контроля успеваемости

1. Отметьте неверное понимание Variety в контексте характеристик больших данных:

+ a) Высокая скорость генерирования данных

+ b) Разные типы данных в колонках таблиц реляционных СУБД

+ c) Разнообразие отраслей, являющихся источниками данных

d) Разнообразие типов данных, включающих в себя структурированные, полуструктурированные и неструктурированные

2. Принцип MapReduce состоит в том, чтобы

+ a) Производить вычисления на узлах, где информация изначально была сохранена

+ b)Использовать вычислительные мощности систем хранения

c) Использовать функциональное программирование для решения задач массивно-параллельной обработки

3. Выберите одно неверное высказывание про MapReduce:

+ a) Интерфейс для массово-параллельной обработки данных, где вычисления производятся на узлах, где информация изначально была сохранена

b) MapReduce - это две операции: распределения и сборки данных

+ c) MapReduce был придуман разработчиками Hadoop

d) MapReduce был анонсирован разработчиками Go