Принципы анализа данных [M.002]

Процесс анализа

В информационном подходе к анализу данных помимо модели присутствуют еще три важные составляющие: эксперт, гипотеза и аналитик.

Определение

> *Эксперт — специалист в предметной области, профессионал, который за годы обучения и практической деятельности научился эффективно решать задачи, относящиеся к конкретной предметной области.*

Эксперт — ключевая фигура в процессе анализа. По-настоящему эффективные аналитические решения можно получить не на основе одних лишь компьютерных программ, а в результате сочетания лучшего из того, что может человек и компьютер. Эксперт выдвигает гипотезы (предположения) и для проверки их достоверности либо просматривает некие выборки различными способами, либо строит те или иные модели.

Пример

> *Гипотезой в анализе данных часто выступает предположение о влиянии какого-либо фактора или группы факторов на результат. К примеру, при построении прогноза продаж допускается предположение, что на величину будущих продаж существенно влияют продажи за предыдущие периоды и остатки на складе. При оценке кредитоспособности потенциального заемщика выдвигается гипотеза, что на кредитоспособность влияют социально-экономические характеристики клиента: возраст, образование, семейное положение и т. п.*

В крупных проектах по созданию прикладных аналитических решений участвуют, как правило, несколько экспертов, а кроме того, а также аналитик.

Определение

> *Аналитик — специалист в области анализа и моделирования. Аналитик на достаточном уровне владеет какими-либо инструментальными и программными средствами анализа данных, например методами Data Mining. Кроме того, в обязанности аналитика входят функции систематизации данных, опроса мнений экспертов, координации действий всех участников проекта по анализу данных.*

Аналитик играет роль «мостика» между экспертами, то есть является связующим звеном между специалистами разных уровней и областей. Он собирает у экспертов различные гипотезы, выдвигает требования к данным, проверяет гипотезы и вместе с экспертами анализирует полученные результаты. Аналитик должен обладать системными знаниями, так как помимо задач анализа на его плечи часто ложатся технические вопросы, связанные с базами данных, интеграцией с источниками данных и производительностью.

Поэтому в дальнейшем главным лицом в анализе данных мы будем считать аналитика, предполагая, что он тесно сотрудничает с экспертами предметных областей.

Пример

В организации создается законченное аналитическое решение в области отчетности и прогнозирования продаж. Оно включает в себя консолидацию данных, настройку отчетов, построение моделей прогнозирования и др. В реализации проекта участвуют специалисты из нескольких подразделений предприятия: высшее руководство, экономисты, логистики, программисты, администраторы баз данных. Аналитик обеспечивает связь между всеми участниками проекта и координирует проект в целом.

Несмотря на то что существует множество аналитических задач, методы их решения можно разделить на две основные группы методов их решения (рисунок 1):

- извлечение и визуализация данных;
- построение и использование моделей.

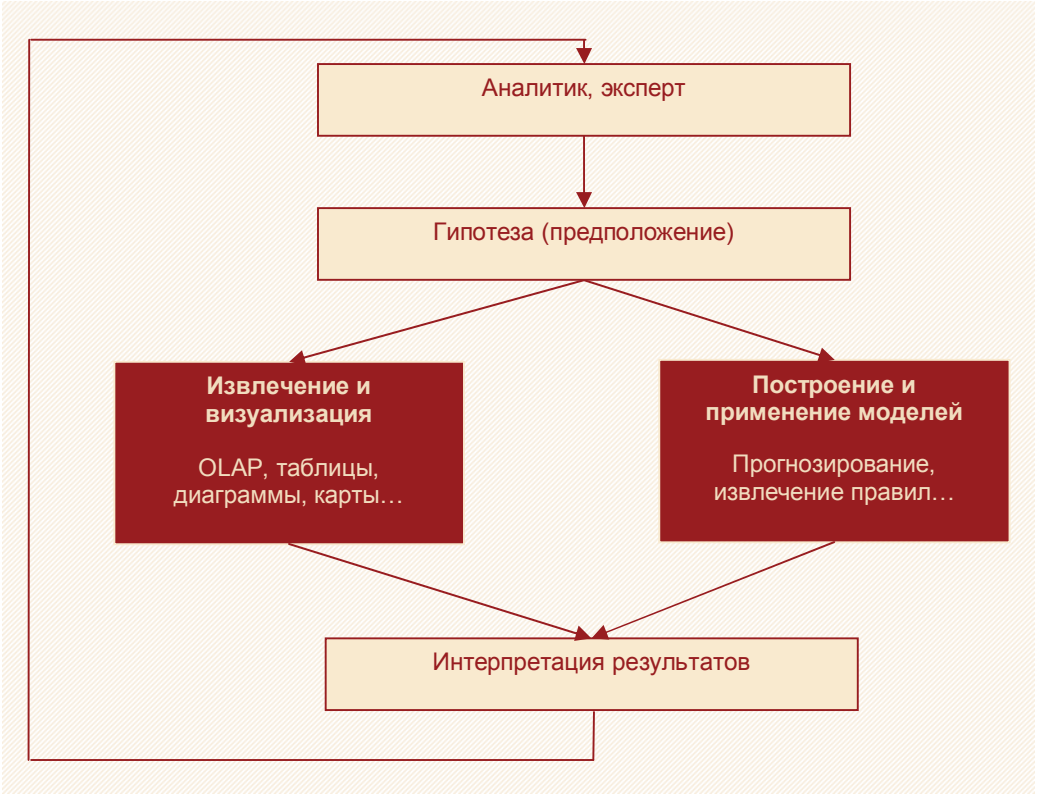

Рисунок 1 – Общая схема анализа

Извлечение и визуализация данных

Чтобы получить новые знания об исследуемом объекте или явлении, не обязательно строить сложные модели. Часто достаточно «посмотреть» на данные в нужном виде, чтобы сделать определенные выводы или выдвинуть предположение о характере зависимостей в системе, получить ответ на интересующий вопрос. Это помогает сделать визуализация.

В случае визуализации аналитик некоторым образом формулирует запрос к информационной системе, извлекает нужную информацию из различных источников и просматривает полученные результаты. На их основе он делает выводы, которые и являются результатом анализа. Существует множество способов визуализации данных:

- OLAP (кросс-таблицы и кросс-диаграммы);
- таблицы;
- диаграммы, гистограммы;
- карты, проекции, срезы и т. п.

Проиллюстрируем вышесказанное. На рисунке 2 приведены два способа визуализации одних и тех же данных по продажам в аптечной сети: в виде таблицы и в виде графика.

Дата	Отдел.Наименование	Группа.Наименование	Количество	Сумма
18.06.2004	Аптека 1	Антисептики и дезинфицирующие средства	3	86.37
18.06.2004	Аптека 1	Витамины и витаминоподобные средства	3	512.23
18.06.2004	Аптека 1	Иммуномодуляторы	2	56.26
18.06.2004	Аптека 1	Местные анестетики	1	4.93
18.06.2004	Аптека 2	Антисептики и дезинфицирующие средства	1	122.45
18.06.2004	Аптека 2	Витамины и витаминоподобные средства	1	68.5
19.06.2004	Аптека 1	Антисептики и дезинфицирующие средства	4	52.3
19.06.2004	Аптека 1	Витамины и витаминоподобные средства	3	151.41
19.06.2004	Аптека 1	Желчегонные средства и препараты желчи	1	31.3
19.06.2004	Аптека 1	Местные анестетики	1	8.14
19.06.2004	Аптека 1	Микро- и макроэлементы	1	2.24
19.06.2004	Аптека 1	Общетонизирующие средства и адаптогены	1	353.46

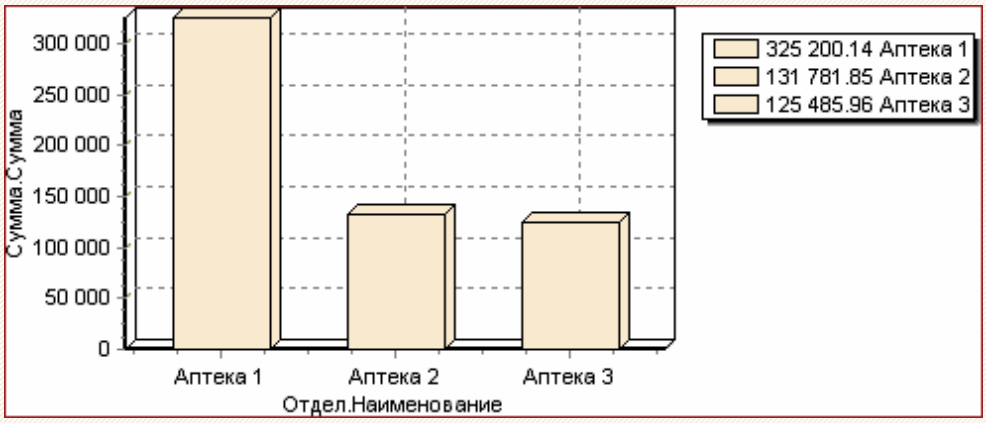

Рисунок 2 – Два способа представления данных: табличный и графический

В первом случае, глядя на таблицу, нам трудно сделать какие-либо выводы, относительно динамики продаж.

Во втором варианте, представив те же данные в виде сумм продаж в разрезе аптек и построив столбчатую диаграмму, мы видим, что самые большие продажи приходятся на Аптека 1.

Несомненными достоинствами визуализации являются относительная простота создания и введения в эксплуатацию подобных систем и возможность их применения практически в любой сфере деятельности. Кроме того, в этом случае по максимуму используются знания эксперта в предметной области и его способность принимать во внимание многие трудно формализуемые факторы, влияющие на бизнес.

Недостатками визуализации являются неспособность людей обнаружить достаточно сложные и нетривиальные зависимости, а также невозможность отделить знания от эксперта и тиражировать знания.

Этапы моделирования

Построение моделей — универсальный способ изучения окружающего мира, позволяющий обнаруживать зависимости, прогнозировать, разбивать на группы и решать множество других важных задач. Но самое главное: полученные таким образом знания можно тиражировать.

Определение

> *Тиражирование знаний — совокупность методологических и инструментальных средств создания моделей, которые обеспечивают конечным пользователям возможность использовать результаты моделирования, для принятия решений без необходимости понимания методик, при помощи которых эти результаты получены.*

Процесс построения моделей состоит из нескольких шагов (рисунок 3).

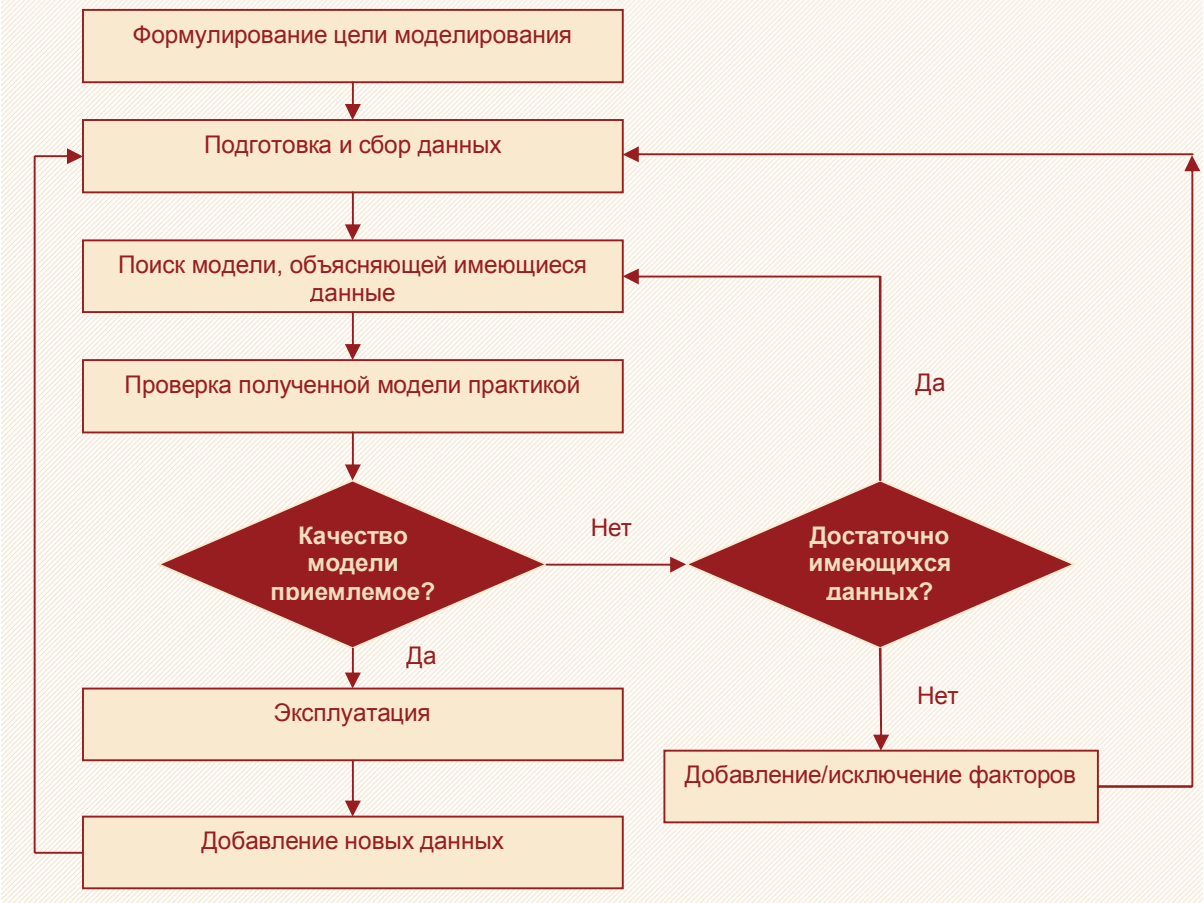

Рисунок 3 – Процесс построения модели

Формулирование цели моделирования. При построении модели следует отталкиваться от задачи, которую можно рассматривать как получение ответа на интересующий заказчика вопрос.

Например, в розничной торговле к таким вопросам относятся следующие.

- Какова структура продаж за определенный период?
- Какие клиенты приносят наибольшую прибыль?
- Какие товары продаются или заказываются вместе?
- Как оптимизировать товарные остатки на складах?

В этом случае можно говорить о создании модели прогнозирования продаж, модели выявления ассоциаций и т. д. Данный этап также называют *анализом проблемной ситуации*.

Подготовка и сбор данных. Информационный подход к моделированию основан на использовании данных, подготовить и систематизировать которые — отдельная задача. Принципам подготовки данных, а также их очистке и обогащению посвящены отдельные главы.

Поиск модели. После сбора и систематизации данных переходят к поиску модели, которая объясняла бы имеющиеся данные, позволила бы добиться эмпирически обоснованных ответов на интересующие вопросы. В промышленном анализе данных предпочтение отдается самообучающимся алгоритмам, машинному обучению, методам Data Mining.

Если построенная модель показывает приемлемые результаты на практике (например, в тестовой эксплуатации), ее запускают в промышленную эксплуатацию. Так, при тестовой эксплуатации скоринговой модели, рассчитывающей кредитный рейтинг клиента и принимающей решение о выдаче кредита, каждое решение может подтверждаться человеком — кредитным экспертом. При запуске кредитного скоринга в промышленную эксплуатацию человеческий фактор удаляется — теперь решение принимает только компьютер.

Если качество модели неудовлетворительное, то процесс построения модели повторяется, как это показано на рисунке 3.

Моделирование позволяет получать новые знания, которые невозможно извлечь каким-либо другим способом. Кроме того, полученные результаты представляют собой формализованное описание некоего процесса, вследствие чего поддаются автоматической обработке. Однако результаты, полученные при использовании моделей, более чувствительны к качеству данных, к знаниям аналитика и экспертов и к формализации самого изучаемого процесса. К тому же почти всегда имеются случаи, не укладывающиеся ни в какие модели.

На практике подходы комбинируются. Например, визуализация данных наводит аналитика на некоторые идеи, которые он пробует проверить при помощи различных моделей, а к полученным результатам применяются методы визуализации.

Полнофункциональная система анализа не должна замыкаться на применении только одного подхода или одной методики. Механизмы визуализации и построения моделей должны дополнять друг друга. Максимальную отдачу можно получить, комбинируя методы и подходы к анализу данных.