Введение в визуализацию [M.050]

Одной из важнейших составляющих аналитических технологий является визуализация — представление данных в виде, который обеспечивает наиболее эффективную работу пользователя. Способ визуализации должен максимально полно отражать поведение данных, содержащуюся в них информацию, тенденции, закономерности и т. д. При этом выбор способа визуализации зависит от характера исследуемых данных и от задачи анализа, а также от предпочтений пользователя.

Многие связывают визуализацию только с интерпретацией, оценкой качества и достоверности результатов анализа. Однако это в корне неверно. Визуализацию необходимо применять на всех этапах аналитического процесса без исключения. На практике в процессе анализа данных пользователь непрерывно работает с различными визуализаторами.

Цели и задачи визуализации на разных этапах аналитического процесса иллюстрируются на рисунке 1.

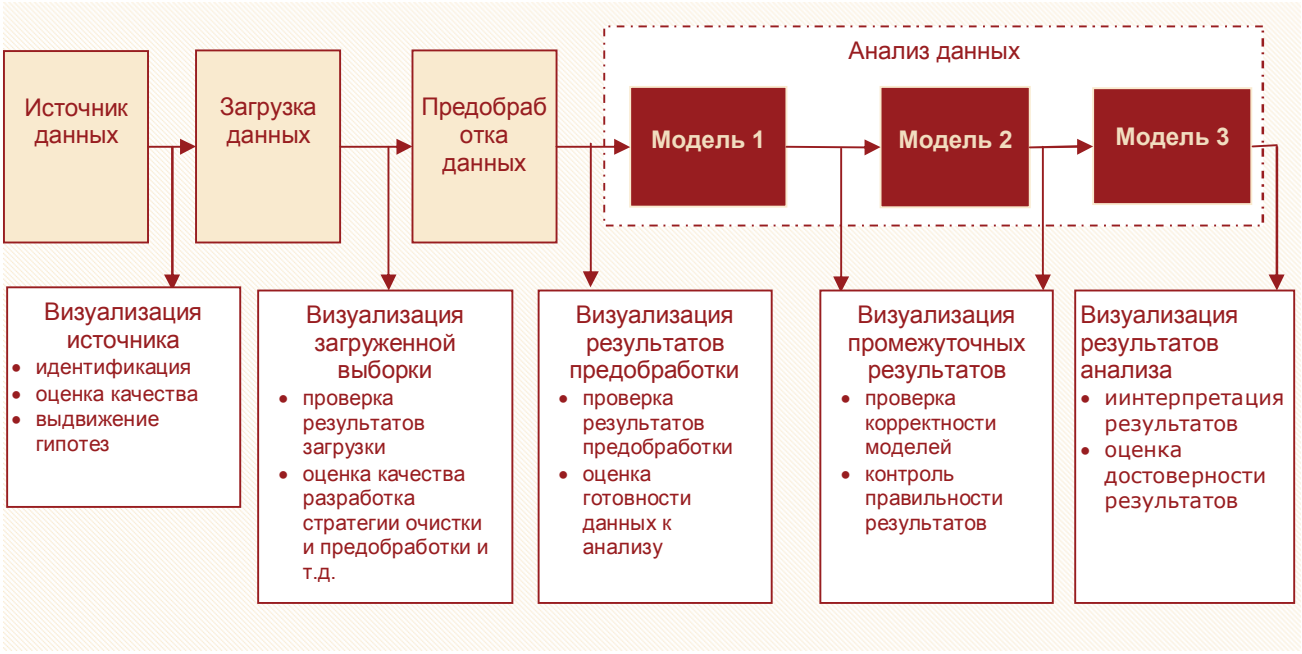

Рисунок 1 – Цели и задачи визуализации данных

Цели и задачи визуализации на разных этапах аналитического процесса

Визуализация используется на разных этапах аналитического процесса для достижения следующих целей и решения следующих задач.

Визуализация источников данных. В источнике данных перед их загрузкой в аналитическую систему аналитику требуется визуально оценить:

- характер, тип и поведение данных;
- динамический диапазон значений;
- степень гладкости;
- наличие факторов, снижающих качество данных, таких как шумы, аномальные и пропущенные значения.

Визуальный анализ источника данных позволяет:

- увидеть, соответствуют ли данные ожидаемым;
- оценить степень пригодности данных к анализу;
- выдвинуть гипотезы о закономерностях процессов, описываемых данными;
- определить, какие виды очистки и предобработки необходимо применить к данным.

Кроме того, визуализация источников данных позволяет определить метод загрузки данных в аналитическое приложение и параметры, которые при этом должны быть использованы. Например, для корректной загрузки данных из текстового файла с разделителями необходимо правильно определить символ-разделитель, используемый формат даты и времени, расположение заголовков столбцов и т. д. Неправильный выбор любого из этих параметров приведет к некорректной загрузке, что не позволит выполнять обработку данных в аналитическом приложении.

Замечание

> *Особенно важно правильно настроить параметры загрузки при импорте больших массивов данных, тем более при загрузке по сети. Если источник имеет большой объем, то процесс загрузки данных из него в аналитическое приложение может оказаться очень длительным. А после долгого ожидания может выясниться, что данные были загружены некорректно, поскольку в параметрах загрузки текстового файла был неправильно указан символ-разделитель.*

Для визуализации источников данных можно использовать приложения, в которых они были созданы (текстовые редакторы, СУБД, электронные таблицы и т. д.). Кроме того, большинство аналитических приложений содержат собственные средства предварительного просмотра источников данных.

Визуализация данных, загруженных в аналитическое приложение. После загрузки данных из источника в аналитическое приложение работа с выборкой также начинается с визуального анализа. Однако теперь цели, задачи и методы визуального анализа будут несколько другими.

Нужно убедиться, что данные загрузились правильно: не появились пропуски, сохранилась структура строк и столбцов и т. д. Искажение данных при загрузке может произойти из-за несоответствия их типов, неправильной настройки параметров загрузки и т. д.

Если данные загружены корректно, то, как правило, стараются оценить степень их гладкости, наличие шумов и аномальных выбросов. Интерес представляет поиск фрагментов данных с некоторыми особенностями. Кроме того, большинство аналитических систем предлагают пользователю возможность получить статистические характеристики — минимальное и максимальное значения, дисперсию и среднеквадратическое отклонение и др.

По результатам визуального анализа исходной выборки делаются выводы о целесообразности применения тех или иных видов очистки и трансформации данных, вырабатывается методика и стратегия их анализа.

Визуализация данных в процессе их аналитической обработки. Сложные аналитические процедуры являются многошаговыми. Это означает, что в процессе анализа к данным последовательно применяется несколько алгоритмов или моделей. Например, сначала данные подвергаются предобработке с целью сглаживания и нормализации, затем к результирующей выборке применяется та или иная модель. При этом выборка, формируемая на выходе каждого алгоритма или модели, может подаваться на вход следующего этапа обработки. Очевидно, что если данные, поступившие с предыдущего этапа, окажутся некорректными, то дальнейшая обработка теряет смысл. Поэтому очень важно предусмотреть визуализацию промежуточных результатов анализа с целью проверки корректности используемых моделей и алгоритмов.

Визуализация результатов анализа. После получения конечных результатов аналитической обработки на первый план выходит задача их интерпретации и оценки достоверности. И здесь не обойтись без визуализации. Следует заметить, что, даже если в процессе анализа были получены достоверные и ценные результаты, неудачный выбор визуализации не позволит их интерпретировать, увидеть в них зависимости и закономерности.

Группы методов визуализации

В настоящее время в корпоративных аналитических системах используется несколько десятков методов визуализации. Выбор метода определяется особенностями и характером данных, спецификой решаемой задачи и, наконец, предпочтениями пользователя. Рассмотрим основные методы визуализации.

Табличные и графические. Как правило, таблицы применяются в том случае, когда пользователю необходимо работать с отдельными значениями данных, вносить изменения, контролировать форматы данных, пропуски, противоречия и т. д. Графические методы позволяют лучше увидеть общий характер данных — закономерности, тенденции, периодические изменения. Кроме того, графические методы более эффективно сопоставляют данные: достаточно построить графики двух исследуемых процессов на одной системе координат, чтобы оценить степень их сходства и различия.

Одномерные и многомерные. Одномерные визуализаторы представляют информацию только об одном измерении данных, в то время как многомерные — о двух или более. Если график показывает зависимость суммы продаж от даты, то он будет одномерным, поскольку на нем будет отображаться только одно измерение — Дата, значениям которого будет соответствовать факт Цена. Если же информация о продажах приводится по датам и наименованиям товаров, то появляется еще одно измерение — Товар, и тогда для корректного представления данных используется многомерный визуализатор. Популярные многомерные визуализаторы: OLAP-куб, многомерная диаграмма, карта Кохонена и др.

Общего назначения и специализированные. Методы визуализации общего назначения не связаны с каким-либо определенным видом задач анализа или типом данных и могут использоваться на любом этапе аналитического процесса. Это своего рода типовые визуализаторы: графики и диаграммы, графы, гистограммы и их разновидности, статистические характеристики и др. В то же время существует ряд задач, специфика которых требует применения специализированных визуализаторов. Например, карты Кохонена специально разработаны для визуализации результатов кластеризации, матрицы классификации используются в основном для проверки состоятельности классификационных моделей, а с помощью диаграмм рассеяния оценивается корректность работы регрессионных моделей.

При изучении различных видов визуализации удобнее рассматривать их не по отдельности, а в контексте задач, для которых они наиболее часто применяются. Можно выделить следующие группы методов визуализации:

- общего назначения — применяются для решения типовых задач анализа данных: визуальной оценки качества и характера данных, распределения значений признаков, статистических характеристик и т. д.;
- OLAP-анализ — комплекс методов для визуализации многомерных данных;
- оценка качества моделей — позволяет оценивать различные характеристики моделей, такие как точность, эффективность, достоверность результатов, интерпретируемость, устойчивость и т. д.;
- интерпретация результатов анализа — служат для представления конечных результатов анализа в виде, наиболее удобном с точки зрения их интерпретации пользователем.

Подсистемы визуализации данных содержатся не только в специализированных аналитических платформах, но и практически во всех программных средствах, которые связаны с обработкой данных, — от офисных приложений до систем компьютерной математики. Однако в аналитических платформах визуализации данных уделяется особое внимание, поскольку она

является одной из составляющих аналитического процесса, без которой невозможно эффективно решить поставленные задачи.

Наилучших результатов можно добиться, если считать визуализацию не отдельной подсистемой, а такой же частью аналитического процесса, как, например, моделирование, очистка и трансформация. Это позволит получить максимум полезной информации в случаях, когда применение других методов неэффективно.

Даже если для построения качественной модели данных недостаточно, визуализация позволяет выдвигать гипотезы, делать выводы на основе экспертных оценок, разрабатывать способы повышения информативности данных.