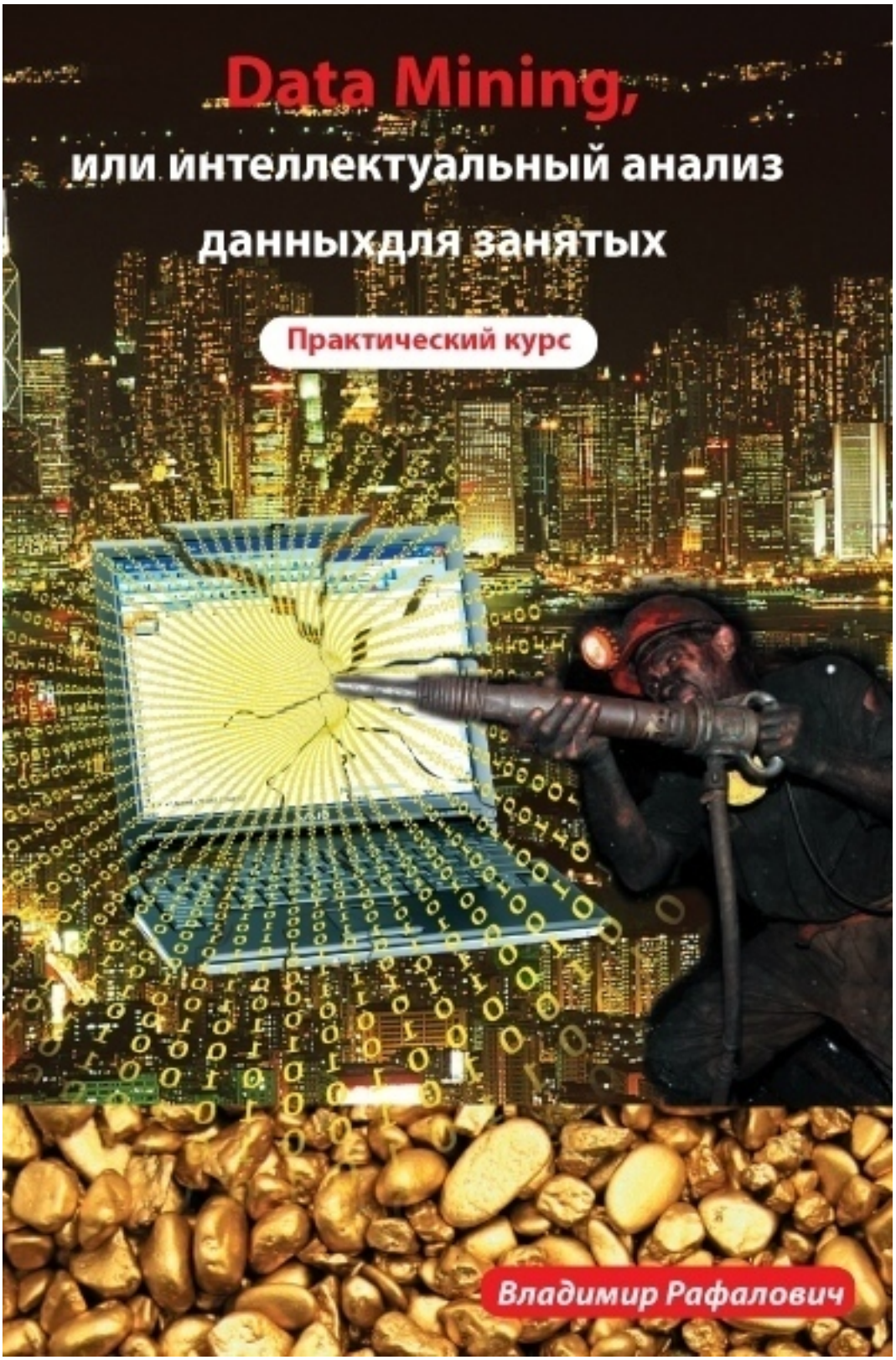
Data Mining,
или интеллектуальный анализ
данныхдля занятых
Практический курс
Владимир Рафалович

Владимир Рафалович

Data mining, или Интеллектуальный анализ данных для занятых. Практический курс

«И-трейд»
2014

УДК 316.77
ББК 88.53

Рафалович В.

Data mining, или Интеллектуальный анализ данных для занятых. Практический курс / В. Рафалович — «И-трейд», 2014

ISBN 978-5-9791-0311-2

Что такое информация? Как можно проанализировать данные, которые у вас есть? А если данных очень много и они требуют вычислительной мощи современных компьютеров? Какие выводы можно сделать из этого массива данных? Может – никаких, а может – это неиссякаемый источник, приносящий все новые возможности. Самое ценное, что есть у любого человека, это его знания, помноженные на опыт. Эта книга помогает занятому человеку быстро погрузиться в увлекательный мир интеллектуального анализа данных с целью извлечения полезной информации, которую можно использовать в дальнейшем, например, в бизнесе или в принятии решений. Эта деятельность по-английски называется Data mining и содержит методы, используемые самыми разными специалистами-аналитиками, исследующими медицинские, политические, экономические и другие всевозможные источники данных. Предполагается, что читатель более-менее знаком с Excel и пользуется им время от времени. Знания SQL-сервера не требуется, но полезно иметь.

УДК 316.77
ББК 88.53

ISBN 978-5-9791-0311-2

Содержание

Владимир Рафалович Data Mining, или интеллектуальный анализ данных для занятых. Практический курс

«Моему отцу Игорю Рафаловичу, который всегда понимал, что информация правит миром»

Предисловие

Мир, в котором мы живем, сконцентрирован вокруг информации, которая обрушивает на нас огромное количество битов ежесекундно. Наша вселенная колоссальный производитель информации, она же – его обработчик. Пришло понимание того, что законы физики не столько описывают объекты вселенной, сколько информацию о самих объектах вселенной. Долгое время полагали, например, что скорость света есть максимально допустимая скорость движения объектов (основной постулат специальной теории относительности). Но эффект Вавилова-Черенкова, когда элементарные частицы двигаются в среде быстрее скорости света в этой же среде, теория инфляции вселенной, которая предсказывает скорость расширения вселенной много превышающей скорость света, или скорость точки пересечения двух скрещенных лучей света легко может превышать скорость света – показывают, что это не так. Значит, речь шла не о скорости самих объектов. Хотите или нет, специальная теория относительности ставит ограничение на скорость распространения информации. Вот она-то не может превышать скорость света. Объект, движущийся быстрее света не может нести в себе информацию. Мы даже не касаемся термодинамики, когда законы физики не только по существу, но и по форме описывают информационные процессы. Вспомните хотя бы такое важнейшее понятие термодинамики, как энтропия.

Но достаточно. Чтобы разобраться в таком объеме информации, ее систематизация и изучение уже необходимость для нас. Огромные объемы информации, даже те, которые накапливаются (генерируются) бизнес-производством переходят те количественные пороги, которые предвосхищают качественные изменения и позволяют находить новые закономерности, доселе неуловимые в небольших накопленных объемах данных.

Эта книга для тех, кто интересуется темой, кто хочет быть в ладу с современностью и прикоснуться к поверхности огромной и быстроразвивающейся науки – интеллектуальный анализ данных. Книга написана максимально просто, с уклоном в практику и с большим количеством иллюстраций. Прочтя ее, вы, несомненно, сможете сами сразу же попытаться проанализировать имеющиеся данные.

Автор выражает благодарность Ивану Гриненко (г. Ростов-на-Дону), за помощь в снабжении данными для примеров в книге, редактору и издателю Ивану Закаряну (г. Москва) за поддержку и интерес, а также всем музам, вдохновляющим меня.

Предмет книги

Призрак бродит по России, призрак разработки данных. Фраза «разработка данных» происходит от английского Data Mining и в этой книге мы будем использовать оба термина. Кроме того имеется термин интеллектуальный анализ данных, который мы тоже будем часто использовать как эквивалентный. Разработка данных и обработка данных хотя звучат похоже, но вещи очень разные.

Таким образом сформулирован предмет книги: мы будем говорить о практических методах интеллектуального анализа данных. Эта книга не является учебным пособием, так как она не содержит систематического изложения использования таких приложений как Excel или SQL-сервер, книга предполагает, что читатель более-менее знаком с Excel и пользуется им время от времени. Знание SQL-сервера не требуется, но полезно иметь. В то же время, эта книга – не справочник, поскольку не содержит богатого фактического материала, хотя, как и справочник, она отличается краткостью изложения материала. Мы избегаем длинных пространных рассуждений и в каждой главе подводим читателя к самой сути проблемы и ее решению. Скорее всего, эта книга есть вводный курс к практическому интеллектуальному анализу данных. Если читателя захватит этот чарующий мир, он увидит насколько сильным инструментом он может овладеть, миссия книги будет считаться выполненной.

Для кого эта книга

Эта книга написана для тех, кто хочет быстро научиться анализировать данные подручными средствами, не приобретая дополнительных дорогих программ. Книга для людей, занятых и деловых, которые хотят войти сразу в суть проблемы и выяснить для себя как это делается, а потом решить, нужно ли им это или нет, и если нужно, то изучить другие, более детальные книги, с теоретическими основами. Эту книгу будет легко читать профессиональным программистам, SQL-разработчикам, администраторам баз данных, но не только. Самим выбором инструмента для разработки данных мы хотим довести методы интеллектуального анализа данных до самых широких слоев специалистов, включая аналитиков, исследующих медицинские, полицейские, политические, экономические и другие всевозможные источники данных. Мы намеренно опустили детальные математические обоснования конкретных алгоритмов, лежащих в основе изучаемых инструментов, поскольку не каждый аналитик, да и программист, имеет необходимую математическую подготовку. Мы концентрируемся в книге на практическом применении, понимании и анализе результатов. Книг на эту тему практически нет, в то время как хороших теоретических книг имеется большое количество. Предварительных знаний и умения навыков работы с Excel и SQL-сервером не требуется.

Почему Excel

Уже сегодня существует достаточно много приложений позволяющих разрабатывать данные. Microsoft (SQL Server), Oracle, SAP, TeraData, R и другие. Однако, все они предполагают серьезную программистскую подготовку и владение соответствующими языками, встроенными в эти приложения.

Заслуга компании Microsoft в том, что она революционизировала подход к этой проблеме, сделав ее доступной практически всем, не только программистам, но и аналитикам, интересующимся темой. Это стало возможным именно благодаря наличию Excel. Именно через него Microsoft двинула интеллектуальный анализ данных в массы. Теперь, пользователю Excel нет нужды знать математические тонкости алгоритмов и выбора моделей и нет нужды строить хранилища данных (что разумно в случае наличия огромного, исчисляемого сотнями тысяч и более записей, источника данных), что требует углубленного знания SQL-сервера. Наконец, тот самый факт что программа Excel de-facto уже используется многими миллионами специалистов, является очень популярной, самой распространенной и общедоступной не оставило нам сомнений, что вводную книгу, понятную не только программистам, на тему разработки данных, надо писать, основываясь на Excel.

Мы также убеждены, что лучший способ изучить новую область знаний – это начать самому анализировать свои данные. Трудно представить себе, что-нибудь более простое или более доступное, чем Excel. Главное – начать, войти в курс дела, разобраться с сутью, а затем можно выбирать другие инструменты по своему усмотрению. Например PolyAnalyst или R.

Естественно, владение SQL-ом очень поможет читателю для манипулирования данными, особенно на этапе их очистки, когда это легко сделать средствами SQL-сервера, но это необязательно. Можно обойтись самим Excel. В целом эта книга будет понятна аналитикам и всем тем, кто не имеет специального математического или программистского образования.

Данные и Информация

Почему разработка данных становится все более актуальной задачей с каждым днем? Да просто потому, что все окружающее нас, весь внешний мир это сплошной поток информации, которую наш мозг постоянно перерабатывает. В самом деле, даже такие казалось бы вещи, как касание другого человека, слушание его речи, купание в море – это все, не более чем, просто данные о температуре, твердости, цвете, вязкости и так далее, о среде или собеседнике. Весь внешний мир по сути это набор данных для нас, не более того. Вдумайтесь! Надо заметить, что, вообще говоря, понятия "данные" и "информация" не идентичны. Мы именно перерабатываем огромный набор зрительных, слуховых, осязательных и прочих данных. Когда в результате обработки мы находим похожие сегменты, мы выделяем их в одну сущность. Наш друг Петя, это определенный образ, характеризующийся более-менее неизменными характеристиками – зрительные данные (цвет волос, глаз, овал лица и т. д.), слуховые (тембр (частота) голоса) и прочее. Итак, благодаря значительной тавтологии в потоке данных, мы в состоянии выделять закономерности. Если бы не было повторяемости данных, то не было бы законов природы, так как невозможно было обобщить данные в лаконичную форму – закономерность. На самом деле все обстоит наоборот: наличие в природе закономерностей обуславливает повторяемость данных. Закон притяжения зарядов Кулона, например, обобщает огромный набор отдельных данных, связывающих между собой размер зарядов, расстояний между ними и силой, действующей на них. Вместо того, чтобы заполнять огромные таблицы в базах данных для разных сочетаний зарядов, расстояний и сил, значительно удобней и проще записать закон и рассчитывать из него силу, действующую между зарядами. В этом законе нет ничего лишнего, нет повторяемости. Он минимален и из него ничего нельзя убрать. Он содержит квинтэссенцию огромного набора данных. Он и есть информация. Информация в сущности это тот минимальный набор данных, который уменьшить нельзя, иначе данные невозможно будет узнать/восстановить. *Знчит, вжно умть выделть инфрмцю ради оббщния огрмнго обема днных.* Из предыдущей строки мы убрали лишние данные (лишние буквы), но информационная суть сохранилась. Почему? Благодаря высокому уровню тавтологии в русском (и любом другом) языке.

Так, разработка данных как раз и занимается тем, что обрабатывая объемные массивы данных, она пытается обнаружить более емкие закономерности. Выхолощить повторяемость и обнаружить действительно полезную информацию. А в наш век это очень необходимо, дабы не потеряться в дебрях огромного потока данных, проливающегося на нас.

Интеллектуальный анализ данных, что это

Разработка данных (Data Mining) иногда еще называемая обнаружением знаний из баз данных (KDD – knowledge discovery in databases), по сути, заключается в нахождении повторяющихся элементов (сегментов) в источнике данных. Когда данных собрано очень много, их количество позволяет обнаружить неизвестные до сих пор закономерности, которые не были заметны когда данных было мало. Огромное количество данных позволяет сделать качественный скачок и обнаружить новые закономерности. С другой стороны, что по сути означают физические законы? В результате наблюдений огромного количества повторяющихся явлений, люди были в состоянии резюмировать их в короткие по форме математические формулы, которые представляют собой информационную квинтэссенцию явлений. Поясним эту мысль. Данные в базах данных, даже в нормированных, еще не являются информацией как таковой, поскольку содержат большое количество явных и неявных повторений. Большое количество повторений, большая удаленность от чистой информации, как раз и позволяет находить в данных закономерности, то есть приводить систему данных к более близкому к информации состоянию, понижать энтропию данных, так сказать. Извлечение из совокупности данных повторяющихся закономерностей, сродни нахождению новых закономерностей (пусть и не выраженных в виде математической формулы), то есть извлечению новых знаний.

Исходные данные часто требуется подчистить перед разработкой, поскольку они могут содержать разного сорта мусор, шум. Например, всякого рода аномалии могут быть результатом случайной ошибки, хотя могут указывать и на специфику системы, описываемой данными. Данные могут содержать не имеющие отношения к делу параметры и поля. Или поля, которые мы не хотим по каким-либо причинам учитывать в анализе.

Эта книга отличается от большинства других по этой теме тем, что мы не углубляемся в суть математического обоснования или объяснения тех или иных моделей и алгоритмов. На эту тему написано огромное количество хороших книг. Но вот книг о практическом применения этих методов очень мало, если не сказать, что почти нет ни на русском, ни на английском языках. Для этого есть ряд объективных причин. Дело в том, что пользователи Excel редко имеют представление о том, что такое базы данных и как ими манипулировать. Специалисты работающие с SQL-сервером не нуждаются в Excel для разработки данных, поскольку в самом SQL-сервере имеются серьезные инструменты для интеллектуального анализа данных (SSAS – SQL Server Analysis Service, аналитические сервисы SQL-сервера), требующие значительных профессиональных знаний. Тема же нашей книги лежит как раз на стыке этих двух приложений. В результате, многочисленные книги об Excel, концентрируются в основном на использовании встроенных статистических функций, формулах, на вопросах о том, как создавать макросы и писать их на языке VBA и, как правило, обходят тему разработки данных стороной. Книги же по SQL-серверу вообще ориентированы обычно на специалистов и довольно глубоко входят в тему интеллектульного анализа данных в рамках самого SQL-сервера. Но при этом делается упор на построении хранилищ данных (Data Warehouse), так называемых кубов, выбора моделей и алгоритмов, на которых затем и базируется разработка данных.

Мы писали книгу для людей, которые работают с Excel, которые по природе своих занятий обрабатывают большие объемы данных и которым просто еще не пришлось обнаружить скрытые ресурсы находящиеся во взаимодействии Excel с SQL-сервером.

Введение

Обработка данных область далеко не новая, хотя наиболее интенсивно она стала развиваться в конце 20 века, когда персональный компьютер стал так же доступен как и телевизор. Статистической обработкой данных занимались люди тоже давно. Тем не менее, интеллектуальный анализ данных с помощью методов Data Mining (разработка данных) это нечто другое, чем просто статистическая обработка данных, хотя последняя лежит в ее основе. Прежде всего Разработка данных не сводится к статистической обработке данных, но содержит последнюю, скорее как внутренний инструмент. Когда у нас слишком много данных и очень много коррелирующих между собой параметров, то анализировать такие объемы вручную или традиционными методами становится проблематично. Традиционные методы не срабатывают в условиях сложных нелинейных и многочисленных комбинаций, либо требуют неадекватных затрат. Принципиальное отличие Разработки данных от статистической Обработки данных заключается в том, что первое позволят извлечь из груды данных новое знание (KDD – Knowledge Discovery from Database), новую закономерность, ранее неизвестную в принципе. Путем нахождения типичных повторений (pattern) или образцов. Разработка данных указывает на новые зависимости между входными параметрами и искомыми переменными. Довольно ярким примером подобного извлечения знаний является такой факт: обработка закупок в супермаркетах показала, что вместе с пивом люди часто покупают поленья для пикника и мясо. В результате в супермаркетах эти товары находятся в непосредственной близости, подсказывая и подталкивая покупателя на дополнительные покупки.

> *Отличие обработки данных (обычно статистической) от разработки данных (Data Mining) заключается в том, что первая, подготовив нужным образом данные, дает пользователю возможность делать свои заключения и выводы относительно полученных результатов обработки исходных данных. При разработке данных, сама машина предлагает пользователю свои выводы, сделанные относительно исходного набора данных на основе используемых алгоритмов и моделей.*

Существуют и другие многочисленные примеры практического применения результатов разработки данных. Конкурентная борьба между транснациональными американскими сетями магазинов заставляет их бороться за каждого покупателя и не давать ему переходить в другую торговую сеть. Американская торговая сеть Target, основной конкурент Walmart, понимала, что если в семье рождается ребенок, то главное затащить родителей в свой магазин и предложить им дайперсы, если не бесплатно, то по очень низкой цене. Дальше родители купят все остальное и вообще станут покупателями этой торговой сети. Но как узнать, когда в семье родится ребенок? Очевидно, что беременные женщины имеют тенденцию питаться несколько отличным образом от других. Они употребляют больше витаминов, молочных продуктов и т. д. Разработка данных и классификация покупателей методами интеллектуального анализа данных позволила определить группу беременных покупательниц. Им были разосланы приглашения посетить магазин с дисконтными купонами. Для этого использовался аналитический процесс **"Detect Categories" (Определить категории)**. Как это делается вы узнаете из главы 4–2.

Другой яркий пример работы ассоциативного алгоритма это компания Amazon.com. Она анализирует предметы покупок, книги, в частности, которые обычно покупаются вместе, а затем подсказывает покупателю те предметы (книги), которые обычно покупаются попутно. Подобная стратегия очевидно приводит к увеличению объема продаж. Для этого

используется аналитический процесс **"Shopping Basket Analysis" (Анализ покупательской корзины,** см. главу 4–7)

Наконец подозрительная активность с банковскими картами или слишком необычные для данного клиента покупки, нехарактерные для его привычек, позволяют банкам не пропускать транзакции, пока они не получат письменного или устного разрешения клиента. Скажем владелец карты использовал ее в течении нескольких лет для покупок питания на не более, чем 2000 рублей за раз и для эпизодической оплаты книжных покупок размером до 1000 руб. Однажды, банк получил требование на оплату счета в ресторане в размере 10.000 рублей. Банк заблокировал эту транзакцию и правильно, карта оказалось утерянной, но владелец еще не успел этого обнаружить. Для этого используется аналитический процесс **"Highlight Exceptions" (Выделение исключений,** см. главу 4–5)

Что нужно для работы

1. Вам необходимо разумеется иметь Excel 2007 или более позднее издание (2010 или 2013). В этой книге, однако, мы будем использовать Excel 2010 и все примеры будут иллюстрироваться из него. Excel обычно является частью Microsoft Office 2007, 2010 или 2013. Excel должен быть установлен на том компьютере, на котором вы работаете.

2. Нужно иметь стандартное или Enterprise издание SQL-сервера 2005, 2008 или более позднее. SQL-сервер не обязательно должен находиться на том компьютере на котором вы непосредственно работаете, но с ним должна быть хотя бы интернет-связь, поскольку вся обработка данных происходит именно на SQL-сервере. Кроме того, на самом SQL-сервере должна быть установлена компонента SQL Analysis Service (SSAS). Этот продукт хотя и является частью SQL-сервера, не устанавливается по умолчанию и должен быть установлен дополнительно. Это именно тот сервис SQL-сервера, где находятся все алгоритмы и где будет происходить расчет моделей и обработка данных.

3. Для Excel необходимо также иметь Data Mining Add-in. Это бесплатная подпрограмма-надстройка, которая естественным образом внедряется в Excel после установки и нужна для коммуникаций между Excel и SQL-сервером. К тому же она добавляет в Excel дополнительную линейку меню, необходимую для интеллектуальной разработки данных, выбора инструментов и манипулированния данными. Как устанавливается и откуда берется эта важная подпрограмма рассматривается в главе 1.

Глава 1
Установка подпрограммы-надстройки

Прежде, чем использовать Excel для разработки данных, необходимо провести установку подпрограммы Data Mining Add-in. Хотя она бесплатна, но она не устанавливается по умолчанию при первой установке Excel на компьютере. Сначала Add-In надо загрузить. Проще всего рекомендуем произвести в Google следующий поиск «sql server 2008 data mining add-ins». Результат поиска приведет вас на страницу компании Microsoft (Рис. 1–1). В самом низу страницы находится собственно линк на скачивание Add-in. Еще раз напомним, что если вы пользуетесь Excel версии 2013 года, то установка этой подпрограммы не нужна. Она уже является неотъемлемой частью Excel! Речь идет только об Excel версиях 2007 или 2010. Для загрузки можно воспользоваться также сайтом http://www.sqlserverdatamining.com

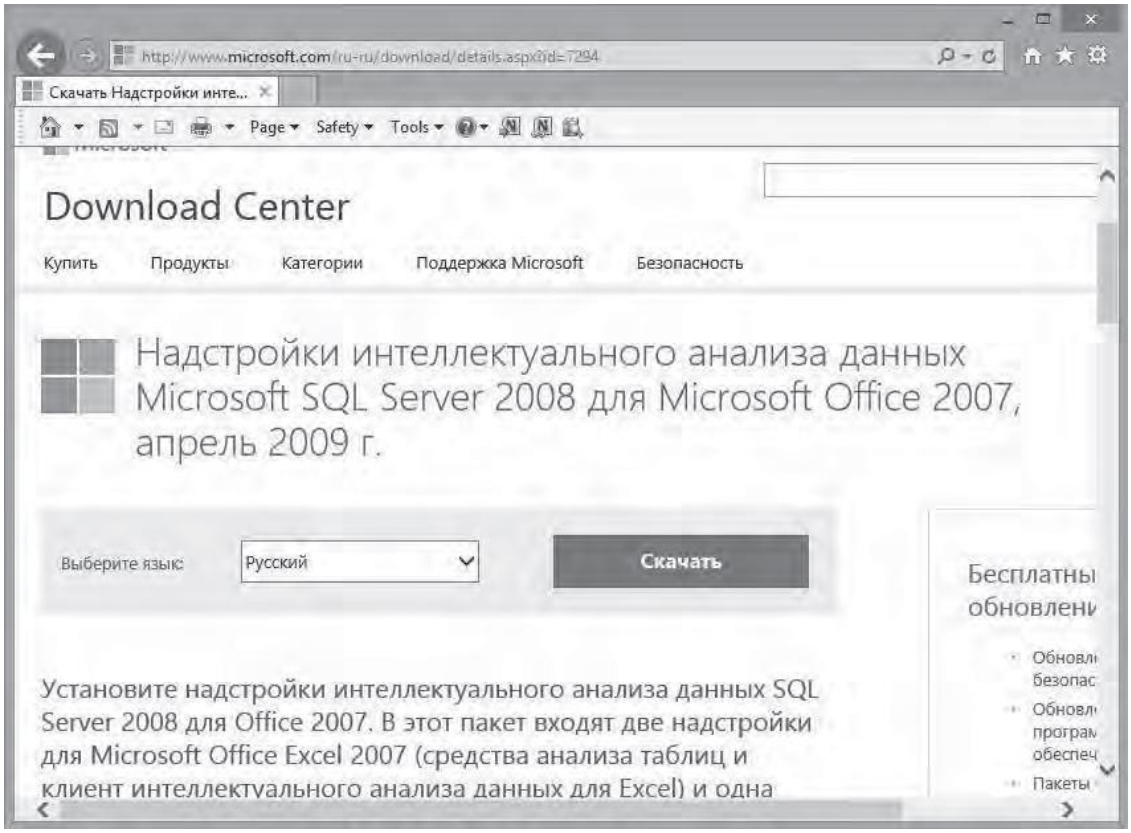

Рис. 1–1. Страница загрузки подпрограммы Data mining Add-In

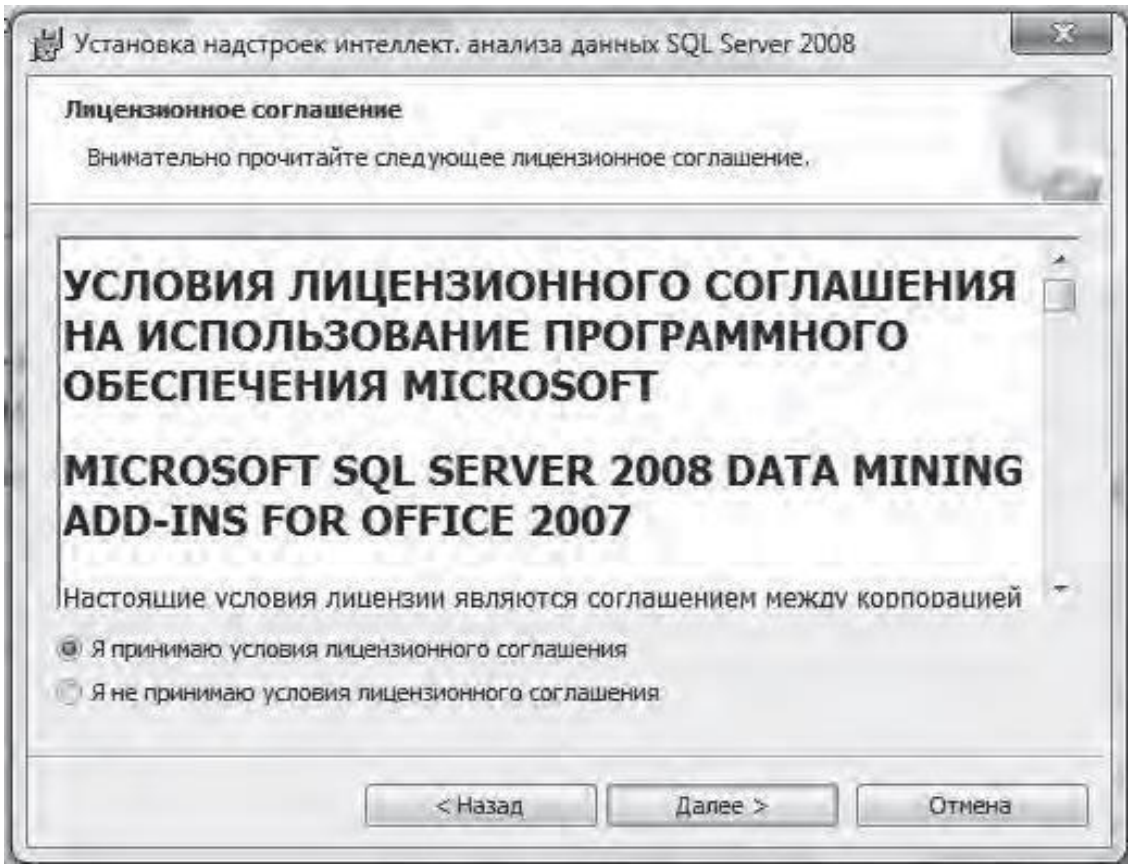

Рис. 1–2. Первый шаг мастера установки надстройки.

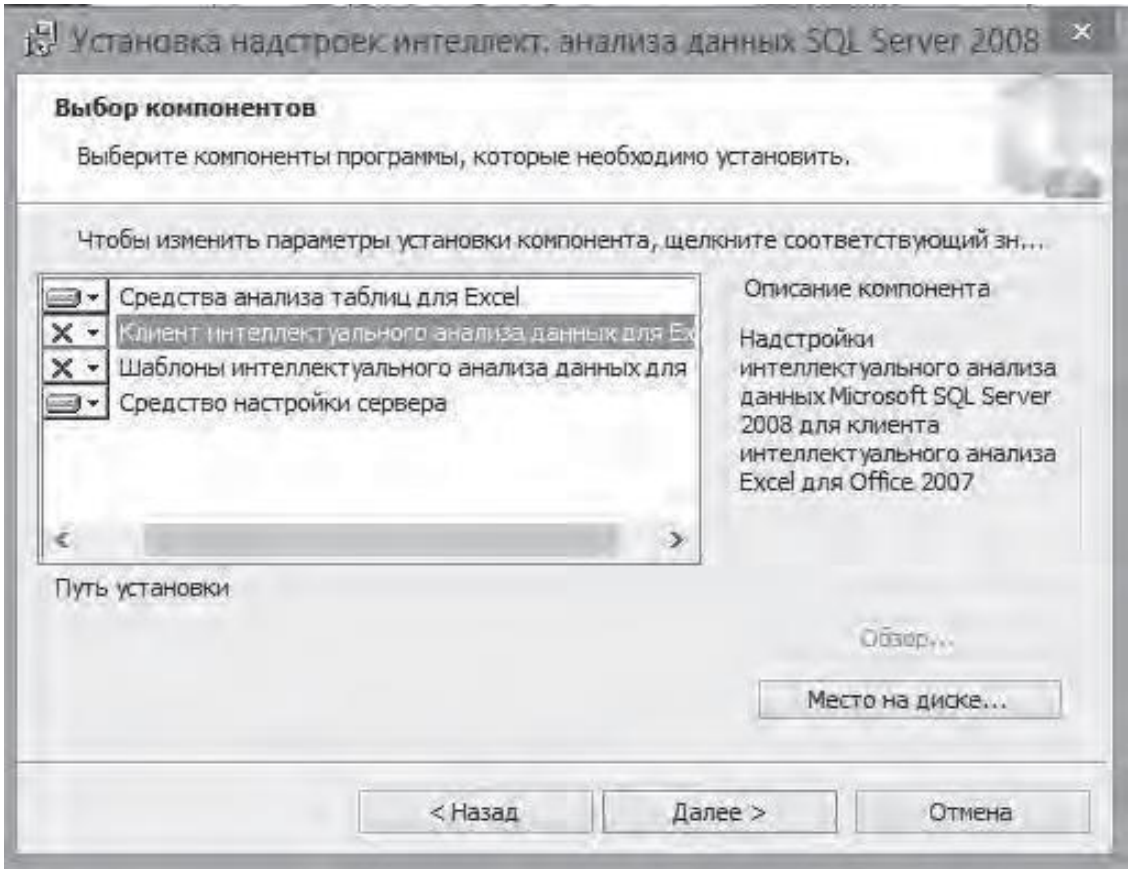

Рис. 1–3. Выбор необходимой опции: "Клиент интеллектуального анализа данных".

После загрузки подпрограммы-надстройки традиционно откроется мастер ее установки. После принятия условий использования программы (Рис. 1–2) будет предложено ввести ваше имя, а затем откроется окно опций установки (Рис. 1–3).

Далее можно нажать на кнопку «Далее». Однако, хотим обратить ваше внимание на опцию «Клиент интеллектуального анализа данных для Excel», которая по умолчанию не отмечена! Если ее выбрать и продолжить установку, то Excel будет иметь дополнительное меню: Data Mining. Опции под этим меню будут в принципе аналогичны тем, которые существуют в интересующем нас в этой книге анализе таблиц, но рассчитаны на более продвинутых пользователей, которые могут по своему усмотрению выбирать алгоритмы и модели для расчетов. Разница в меню представлена на Рис. 1–4 и 1–5.

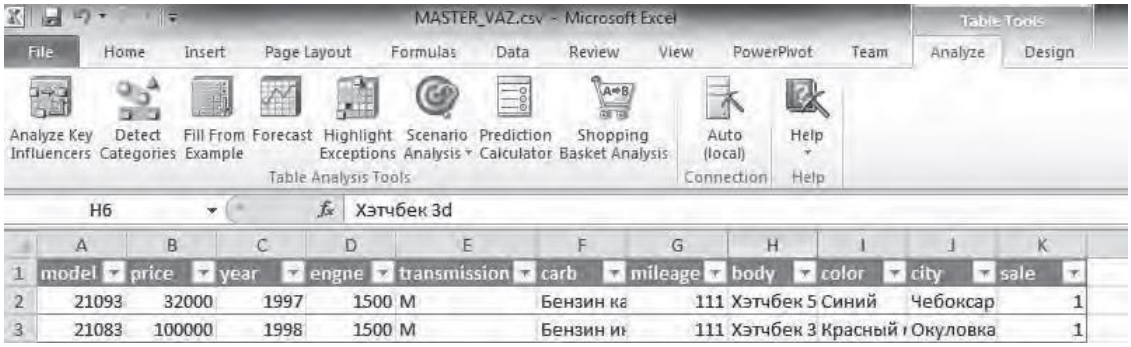

Рис. 1–4. Меню Excel, когда опция «Клиент интеллектуального анализа данных для Excel» не выбрана.

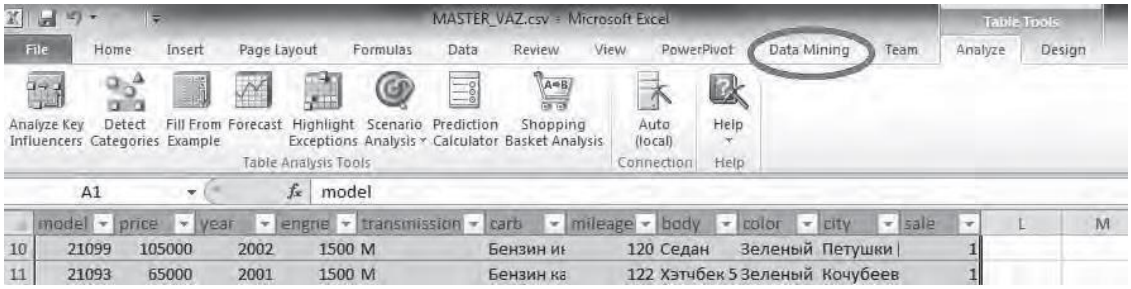

Рис. 1–5. Меню Excel, когда опция «Клиент интеллектуального анализа данных для Excel» выбрана. Видно дополнительное меню Data Mining для продвинутых пользователей.

Выбирайте эту опцию, или нет, по своему усмотрению и переходите к окончанию установки. Теперь сам Add-in установлен. Но это еще не все. Надо соединить его с SQL-сервером и провести конфигурацию последнего.

Подсоединение надстройки к SQL-серверу

Когда вы откроете Excel в первый раз после установки надстройки, скорее всего, возникнет новый мастер подсоединения Excel к SQL-серверу. Excel должен будет инициировать подпрограмму Data Mining Add-in и проведет вас через мастер его конфигурации. Дело в том, что для работы этого Add-in мало иметь Excel; поскольку сама обработка данных происходит на SQL-сервере, то необходимо иметь подключение к нему. Поэтому мастер предложит вам установить пробную версию SQL-сервера или, если SQL-сервер уже установлен, надо будет отметить опцию использования существующего Анализа Данных на SQL-сервере. При этом, в реальности, имеется две, а не одна опции, в зависимости от того, являетесь ли вы администратором на установленном SQL-сервере или нет.

При переходе на следующий экран мастера надо будет щелкнуть по линку внизу окна. Дело в том, что необходимо произвести конфигурацию системы Анализа Данных самого SQL-сервера, что бы он мог успешно взаимодействовать с Data mining Add-in в Excel. При этом стартует Мастер конфигурации SQL-сервер 2008 Data mining Add-in.

На первом шаге необходимо ввести имя SQL-сервера. Обычно, это (local), но если ваш SQL-сервера не локален, то вводится его сетевое имя. На втором шаге надо отметить опцию позволяющую создание временных моделей.

Третий шаг попросит ввести название новой базы данных, которая будет создана и, собственно, будет взаимодействовать с надстройкой и где будут производиться все вычисления. Название базы данных, естественно, произвольно.

Конец ознакомительного фрагмента.