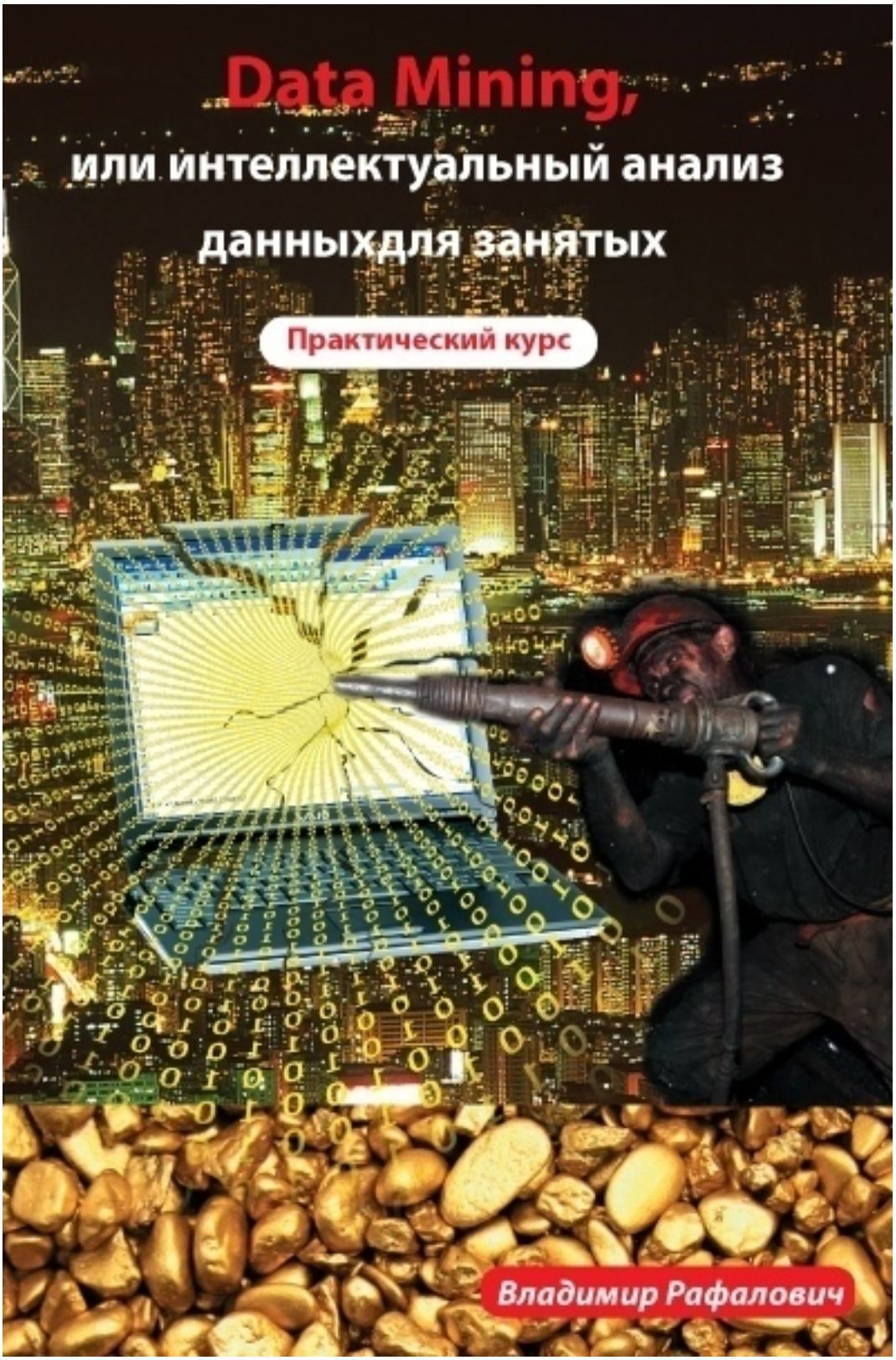
Data Mining,
или интеллектуальный анализ
данныхдля занятых
Практический курс
Владимир Рафалович

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 днн*х. 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.

Конец ознакомительного фрагмента.