

Access Data Analysis Cookbook



Access Cookbook™ 中文版

——数据分析详解

Ken Bluttman & Wayne S. Freeze 著

付宇光 杨舸 译

O'REILLY®



清华大学出版社

TP311.138/583

2008

Access Cookbook 中文版

——数据分析详解

Ken Bluttman & Wayne S. Freeze 著

付宇光 杨舸 译

O'REILLY®

Beijing • Cambridge • Farnham • Köln • Paris • Sebastopol • Taipei • Tokyo

O'Reilly Media, Inc. 授权清华大学出版社出版

清华大学出版社

Copyright ©2007 by O'Reilly Media, Inc.

Authorized Simplified Chinese translation edition, by O'Reilly Media, Inc., is published by Tsinghua University Press, 2007. Authorized translation of the original English edition, 2007 O'Reilly Media, Inc., the owner of all rights to publish and sell the same.

All rights reserved including the rights of reproduction in whole or in part in any form.

本书之英文原版由 O'Reilly Media, Inc. 于 2007 年出版。

本中文简体翻译版由 O'Reilly Media, Inc. 授权清华大学出版社于 2007 年出版。此翻译版的出版和销售得到出版权和销售权的所有者——O'Reilly Media, Inc. 的许可。

版权所有, 未经书面许可, 本书的任何部分和全部不得以任何形式复制。

北京市版权局著作权合同登记

图字: 01-2007-3667 号

本书封面贴有清华大学出版社防伪标签, 无标签者不得销售。

版权所有, 侵权必究。侵权举报电话: 010-62782989 13701121933

图书在版编目(CIP)数据

Access Cookbook 中文版——数据分析详解 / (美) 布鲁特曼 (Bluttman, K.) 等著; 付宇光等译. —北京: 清华大学出版社, 2008.6

书名原文: Access Data Analysis Cookbook

ISBN 978-7-302-16912-3

I. A… II. ①布… ②付… III. 关系数据库—数据库管理系统, Access IV. TP311.138

中国版本图书馆 CIP 数据核字 (2008) 第 008939 号

责任编辑: 龙啟铭

封面设计: Karen Montgomery, 张 健

责任校对: 张 剑

责任印制: 王秀菊

出版发行: 清华大学出版社

地 址: 北京清华大学学研大厦 A 座

<http://www.tup.com.cn>

邮 编: 100084

社 总 机: 010-62770175

邮 购: 010-62786544

投稿与读者服务: 010-62776969, c-service@tup.tsinghua.edu.cn

质 量 反 馈: 010-62772015, zhiliang@tup.tsinghua.edu.cn

印 刷 者: 清华大学印刷厂

装 订 者: 三河市李旗庄少明装订厂

经 销: 全国新华书店

开 本: 178 毫米×233 毫米 21.5 印张 字 数: 480 千字

版 次: 2008 年 6 月第 1 版 印 次: 2008 年 6 月第 1 次印刷

印 数: 1~3000 册

定 价: 49.00 元(册)

本书如存在文字不清、漏印、缺页、倒页、脱页等印装质量问题, 请与清华大学出版社出版部联系调换。联系电话: (010)62770177 转 3103 产品编号: 026835-01

O'Reilly Media, Inc. 介绍

为了满足读者对网络和软件技术知识的迫切需求，世界著名计算机图书出版机构 O'Reilly Media, Inc. 授权清华大学出版社，翻译出版一批该公司久负盛名的英文经典技术专著。

O'Reilly Media, Inc. 是世界上在 Unix、X、Internet 和其他开放系统图书领域具有领导地位的出版公司，同时也是联机出版的先锋。

从最畅销的 *The Whole Internet User's Guide & Catalog* (被纽约公共图书馆评为 20 世纪最重要的 50 本书之一) 到 GNN (最早的 Internet 门户和商业网站)，再到 WebSite (第一个桌面 PC 的 Web 服务器软件)，O'Reilly Media, Inc. 一直处于 Internet 发展的最前沿。

许多书店的反馈表明，O'Reilly Media, Inc. 是最稳定的计算机图书出版商——每一本书都一版再版。与大多数计算机图书出版商相比，O'Reilly Media, Inc. 具有深厚的计算机专业背景，这使得 O'Reilly Media, Inc. 形成了一个非常不同于其他出版商的出版方针。O'Reilly Media, Inc. 所有的编辑人员以前都是程序员，或者是顶尖级的技术专家。O'Reilly Media, Inc. 还有许多固定的作者群体——他们本身是相关领域的技术专家、咨询专家，而现在编写著作，O'Reilly Media, Inc. 依靠他们及时地推出图书。因为 O'Reilly Media, Inc. 紧密地与计算机业界联系着，所以 O'Reilly Media, Inc. 知道市场上真正需要什么图书。

译者序

数据库有很多种，每种数据库的性能、售价和适用范围都不相同，甚至有很大差异。不能单纯地说哪种数据库是最好的，只能说，哪种数据库是最适合当前的用户需求。

现在的主流数据库包括：Access、Mysql、MS SQL Server、Oracle 和 DB2 等。各自有各自的优势和适用环境。在现实的开发中，程序员将面对的客户绝大部分是中小型企业。而 Access 和 MS SQL Server 都是面对中小型企业的（Mysql 是免费数据库，性能有限、应用范围较小；Oracle 和 DB2 是大型数据库，应用范围也相对较小）。MS SQL Server 在安全性和事务处理上要优于 Access（用户不必担心 Access 可处理的数据量过小，而更要考虑 MS SQL Server 的复杂操作）。Access 的优势是移动性强，一个 mdb 文件就保存了所有数据。这对于中小型企业而言是非常重要的。无论对于网站还是应用程序而言，只有一个数据库，不需要进行复杂的配置就可以让系统运转起来，对于客户和开发人员而言，都是一件非常开心的事情。

读者可以通过学习本书，来解决实际工作中最常遇到的问题，比如导入导出数据等（数据库管理员的必备技能）。书中介绍了很多算法的实例，由于数据库都是相同的，所以只要读者理解了书中介绍的方法，就可以将该方法应用到其他数据库的开发上。通过阅读本书，读者一定可以对 Access 数据有一个相对全面的认识，可以在实际工作中灵活地使用 Access 数据库。

参加本书的翻译工作的人员包括：付宇光、杨舸、王萌、马金山、李健、胡嵬、王媛、陈洁容、杨喜华、李宁、李蕾、郑薇等人。感谢家人以及出版社有关人员对本次翻译工作的支持。没有他们的帮助，就不可能顺利完成本书的翻译工作。由于水平有限，翻译中难免有不妥之处，希望读者与我们共同探讨。

目录

前言	1
第 1 章 构建查询	7
1.1 查找不匹配的记录	7
1.2 使用 AND 和 OR	10
1.3 在 Criteria 中使用 IN 运算符	13
1.4 使用 NOT 运算符排除记录	16
1.5 为查询设置参数	17
1.6 使用 Top 和 Bottom 返回记录	22
1.7 删除重复的记录	24
1.8 返回随机记录	30
1.9 通过子查询对数据过滤进行微调	31
1.10 通过联合查询组合数据	36
1.11 在选择查询中插入任意字段	40
1.12 通过别名简化 SQL 语句	43
1.13 创建左联接	44
1.14 创建右联接	46
1.15 创建外联接	48

第 2 章 查询计算	51
2.1 在数据集中求和或者平均值	51
2.2 按组计数	55
2.3 在查询中使用表达式	57
2.4 在查询中使用自定义函数	59
2.5 在查询中使用正规表达式	62
2.6 使用 Cartesian 乘积返回数据的所有组合	66
 第 3 章 动作查询	 78
3.1 运行更新查询	78
3.2 追加数据	84
3.3 删除数据	88
3.4 通过生成表查询创建表	93
 第 4 章 管理表、字段、索引和查询	 96
4.1 通过程序创建表	96
4.2 改变表结构	103
4.3 创建和使用索引	105
4.4 通过程序删除一个表	107
4.5 通过程序创建一个查询	108
 第 5 章 字符串数据	 112
5.1 从字符串的左侧或者右侧返回字符	112
5.2 在已知起始位置和长度的情况下返回字符串中间的字符	114
5.3 在已知字符的情况下返回子字符串的起始位置	115
5.4 删除字符串末尾的空格	117
5.5 删除字符串中间的空格	119
5.6 用一个字符串替换另一个字符串	121
5.7 连接数据	123
5.8 排列作为文本格式存储的数字	126

5.9	通过 ASCII 码对字符进行分类	129
-----	-------------------------	-----

第 6 章 利用程序处理数据 133

6.1	在 Access 中调用 Excel 函数	133
6.2	处理内存数据	137
6.3	使用多维数组	141
6.4	数组排序	144
6.5	平面化数据	148
6.6	扩展数据 (Expanding data)	151
6.7	加密数据	153
6.8	应用近似匹配	157
6.9	应用事物处理 (Transaction Processing)	160
6.10	对 Windows 注册表进行读取和写入	162
6.11	创建图表	164
6.12	获取网络 HTML 资源	169
6.13	创建自定义报告格式	172
6.14	数值取整	176
6.15	运行 Word 邮件合并	179
6.16	建立一个复杂的查询选择窗体	181

第 7 章 导入和导出数据 186

7.1	创建一个导入 / 导出规格	186
7.2	自动导入和导出	192
7.3	通过 FileSystemObject 导出数据	195
7.4	通过 FileSystemObject 导入数据	197
7.5	使用 XML 导入和导出数据	201
7.6	生成 XML 架构	204
7.7	通过 XSLT 导入或者导出	206
7.8	通过 MSXML Parser 处理 XML	208
7.9	读写 XML 属性	213

7.10	创建一个 RSS 种子	214
7.11	传递参数到 SQL Server	217
7.12	处理从 SQL Server 存储过程中返回的数值	218
7.13	处理 SQL Server 数据类型	219
7.14	处理内嵌引号	221
7.15	从 Outlook Calendar 中导入约会	222
7.16	从 Outlook 中导入电子邮件	225
7.17	处理 Outlook 联系人名单	227
7.18	从 Excel 中导入数据	230
7.19	向 Excel 中导出数据	233
7.20	与 PowerPoint 协同工作	234
7.21	选择随机数据	237
第 8 章 日期和时间计算		240
8.1	计算经过时间	240
8.2	Counting Elapsed Time with Exceptions	244
8.3	时区	246
8.4	闰年	249
8.5	分离年、月和日	250
8.6	分离小时、分钟和秒	252
8.7	时间相加	253
第 9 章 商业及财务问题		257
9.1	计算加权平均值	257
9.2	计算移动平均线	259
9.3	计算回收期	260
9.4	计算投资收益	262
9.5	计算直线法折旧	263
9.6	创建贷款支付表	265
9.7	使用 PivotTable 和 PivotChart	267

9.8	创建 PivotTable	269
9.9	用数据制图	272
9.10	确定趋势	275
9.11	查找 Head and Shoulders 模式	279
9.12	布林线	289
9.13	计算邮编之间的距离	291

第 10 章 统计学 297

10.1	创建柱状图	297
10.2	查找和比较平均值、众数和中位数	299
10.3	计算一组数据的方差	303
10.4	计算两组数据的协方差	305
10.5	计算两组数据的相关性	307
10.6	返回一组数据中的所有排列	307
10.7	返回一组数据的所有组合	310
10.8	计算一组数据中的一个数值出现的频率	311
10.9	计算增长率	313
10.10	计算一组数据的概率群分布函数	315
10.11	计算峰值以了解一个概率群分布的峰度或平坦度	317
10.12	确定一组数据的倾斜度	320
10.13	通过百分比的形式返回数据	322
10.14	确定数据项目的排名	324
10.15	确定线性回归的斜率和截距	325
10.16	测量波动性	326

前言

本书的内容

商业用户面对 Access 的强大数据处理能力和灵活性的时候经常会大吃一惊。我的朋友经常对我说“Access 不能胜任，我们有 13000 条记录！”。每当听到类似这样的话时，我不得不笑出声来。正如我（当然也应该包括本书的读者）所知道的，Access 可以轻松处理这些数据，甚至更多的数据。

那么，Access 究竟有多么强大？Access 实际上可以帮我们完成很多工作，而大部分人并不了解。本书的目的就是让读者了解 Access 可以完成的工作。本书列举了 Access 的一些小型解决方案，为用户解决一些头疼的问题，帮助用户了解 Access 还提供了什么其他的数据类型。

通过阅读本书，读者可以解决实际工作中的问题。本书不涉及以下类似问题：如何设计窗体、学习主键或者讨论如何使用内置向导以进行简单查询并生成报表。本书的目的在于讲述如何将 Access 应用于实际工作中。

通过本书所讲述的大量精妙方法，读者可以学习到查询数据的新方法，如何通过不同的方法将数据导入或者导出 Access，如何计算出财务和投资中问题的答案，等等。

注意：在撰写本书的时候，微软已经发布了 Access 2007。本书中的解决方案适用于 Access 的所有版本，本书中的插图来自 Access 2007 和 Access 2003。本书的重点不在于描述 Access 的特殊性能，以及新版本的新界面，尽管 Access 有许多版本，但不变的是 SQL、VBA、DAO 和 ADO。

本书面向的读者

我不会将本书推荐给对 Access 一点都不了解的读者，但只要您拥有一些 Access 的使用经

验,那么我就会把本书推荐给您。如果您了解了 Access 的用户界面、基本的表结构和关系、如何构建简单的查询,那么您就应该阅读本书。对于经验丰富的开发者而言,新的窍门和技巧也是非常有用的。由于许多方法涉及的是商业问题而不是技术问题,所以我们应该通过本书学会如何分析和编写报告。

本书的内容

本书的重点在于数据。本书将提供查询实例、编程技巧,并涉及少量的数学方面的知识,一切以从数据中得到答案为目标。以下是各章的内容概述。

第1章“查询构建”

探讨进行基本查询和高级查询的具体方法。涉及大量不同的查询问题,包括使用 AND、OR、IN 和 NOT 运算符,创建联合查询和对连接类型的讨论。

第2章“计算查询”

进一步介绍通过查询解决实际问题的方法。介绍如何使用聚合函数、自定义函数、常规表达式和交叉表。

第3章“动作查询”

介绍如何执行主动查询,比如插入、更新和删除数据。

第4章“管理表、字段、索引和查询”

介绍如何通过程序创建并操作数据表和查询。

第5章“使用 String Data”

讲述一整套用于管理基于文本的数据的方法。还将介绍分离字符串的方法、删除字符串尾部和中间的空格的方法,以及讲述如何将数字存储为文本。

第6章,“利用程序处理数据”

讨论一些使用数组的方法,如何从 Windows 注册表中读取数据,或者将数据写入 Windows 注册表,如何进行数据加密,以及如何进行事务处理。还介绍查询方法、图表、数据关系操作等问题。

第7章,“导入和导出数据”

介绍从 Access 中导出或导入数据的方法:使用 FileSystemObject、带 XSLT 的 XML,以及和 SQL Server 之间的通信。还讲述与其他 Office 组件进行数据交换的方法。讨论如何创建 RSS 种子。

第8章,“日期和时间计算”

讨论关于基于时间的数据问题。该章将显示如何添加时间、计时,还要讨论闰年和时区问题。

第9章, “商业及财务问题”

该章涉及实际商务问题。计算贬值、投入产出比 (Return On Investment, ROI) 问题。还将涉及一些投资方面的问题, 比如移动平均值、头肩顶模式、布林线和趋势计算。还要介绍一个通过经纬度计算两点间距离的方法。许多相关的网站都可使用这种方法, 以帮助用户在指定范围内寻找医生、商店或者其他服务。

第10章, “统计学”

对于数学的狂热分子 (包括作者本人) 而言, 本章将是非常精彩的。将介绍许多统计技巧, 包括频率、差值、峰度、线性回归和排列。

本书约定

下面的一些字体格式是本书使用的一些惯例:

斜体 (*Italic*)

表示新的术语、URL、电子邮件、文件名、文件扩展名、路径名、目录、Unix 实用程序、命令和命令行参数。

等宽字体 (Constant Width)

指明其中的代码或片段。例如, 类名、方法名, 以及无论它们什么时候出现时呈现的常数宽度。

等宽粗体 (Constant Width Bold)

表明用户在混合输入/输出例子中的输入。

等宽斜体 (*Constant Width Italic*)

表明在语法例子中用户指定的项目。

使用代码实例

本书的目的就是帮助读者完成工作。一般情况下, 读者可以在自己的程序或者文档中使用本书中出现的代码。读者无须在获得我们的授权的前提下使用代码, 除非重新编写了代码的核心部分。读者可以从书中不同的位置摘取代码进行编程, 无须提前获得我们的授权。销售或者分发 O'Reilly 的带有代码的随书光盘当然需要得到授权。对本书进行评论, 以及对代码的引用也不需要得到我们的授权。将本书中讲述的大量实例代码用于实际产品的文档中, 一定要得到我们的授权。

我们非常高兴在其它书籍中被引用, 但并不要求您必须这样做。引用通常包括书名、作者、出版社和 ISBN。比如: *Access Data Analysis Cookbook* by Ken Bluttman and Wayne S. Freeze. Copyright 2007 O'Reilly Media, Inc., 978-0-596-10122-0。

如果您觉得您对代码的使用超出了我们约定的上述范围，那么请联系 *permissions@oreilly.com*。

意见和问题

有关这本书的意见和问题，请致信以下出版商。

美国：

O'Reilly Media, Inc.
1005 Gravenstein Highway North
Sebastopol, CA 95472

中国：

北京西城区西直门南大街2号成铭大厦C座807室
奥莱理软件（北京）有限公司

本书的 Web 页上列出了勘误表、示例和任何其他的信息。可登录以下网址查询：

<http://www.oreilly.com/catalog/9780596101220>

<http://www.oreilly.com.cn/book.php?bn=978-7-302-16912-3>

如果想就本书的技术问题发表评论或咨询，请发邮件至：

bookquestions@oreilly.com

info@mail.oreilly.com.cn

关于本书和其他书籍的更多信息，请访问 O'Reilly 的网站：

<http://www.oreilly.com>

<http://www.oreilly.com.cn>

致谢

来自 Ken Blattman

非常感谢 Simon St.Laurent 对我超出一般的耐心。Simon 为我付出了大量的工作，而他面对的是我疯狂的进度表和无休止的打断。当我撰写本书的时候，我的汽车撞上了一颗大树，我的肩膀受了伤，而我的妻子正在与癌症进行斗争，我的父亲过世了。生活真是变化莫测！

非常感谢 Wayne S. Freeze，当我不得不中断手头工作的时候他给予了我极大的帮助。Wayne 继续了我的工作，他撰写了本书第二部分的重要章节。出色的 Wayne!

特别感谢 O'Reilly 所有员工对我的支持。大量幕后的朋友为每本书都付出了辛勤的劳动。正是通过你们的双手才能使我的稿子印刷成文字，测试成功直至面对读者。

还要特别感谢 Doug Klippert 和 Michael Schmalz，他们在技术上对本书进行了检查。在多本书籍的写作中我都同这两个朋友进行了合作，他们的知识和洞察力令我钦佩。

我要将本书献给我的父亲——Herbert Bluttman，在很多事情上我们都持有截然相反的意见，当我已经出版了多本书籍之后，他仍然一无所获，但依旧努力工作。我曾经恐惧过他的电话，但是很久没有他的消息之后，我才意识到我非常想念他的声音。不愉快总是暂时的，爱是永恒的。

来自 Wayne S. Freeze

Ken（就是大名鼎鼎的 Access Analyst 先生）：我非常开心能与您共同参与本书的撰写工作，当我看到你的生活步入正轨的时候，我更加开心。Simon（就是大名鼎鼎的不知疲倦的编辑）：面对极大的挑战您总能体现出更大的耐心，非常感谢。总有一天，我们会更加轻松地与您共同完成另外一本书的撰写和出版工作。Sumita（就是大名鼎鼎的 Eagle Eye 女士）：非常感谢您把我晦涩难懂的语句修改为非常流畅的话语。为本书付出努力的所有 O'Reilly 员工（就是大名鼎鼎的质量保证小组）：你们完成了一项难以置信的工作，非常荣幸认识你们。

Christopher 和 Samantha（他们是我的孩子）：记得实现你们的梦想，如果没有梦想那么生活就会变得没有意义。Jill（是家中最擅长写作的，也是我最好的朋友）：就像所有的著名艺术家一样，你对工作从来没有真正满意过。我想在不久的将来你将找到新的梦想来替代已经实现的梦想。请牢记，我爱你。

构建查询

在任何数据库系统中选择查询都是最重要的部分。查询操作往往被动地收集一些我们关心的数据（不会改变原数据），依靠查询出的数据，我们就可以解决与数据相关的一些问题。在最常见的情况下，一个选择查询仅仅是从表中一字不差地返回每个记录。这不是我们感兴趣的内容，因为如果要浏览每条记录，可以直接查看表，而不必进行查询操作。但当用户希望添加搜索条件、进行连接以及使用实用的SQL函数时，选择查询才能变得有价值。

本章将讲述有关选择查询的一些技巧。本章讲述的方法将解释如何在运行时添加搜索条件、如何使用逻辑运算符添加搜索条件，以及如何处理重复记录。

为了能够更简单地阅读查询、使用查询进行工作，用户可以使用别名，该功能可以使用户轻松地将自己的表命名为另外一个名称。本章还将讲述如何使用联合查询，将不同表中的数据合并到一起（这些数据可以被看作一个数据源）的问题。

1.1 查找不匹配的记录

问题

假设存在一个员工的费用支出表。在费用支出表中的一些记录与员工表 (Employees) 中的记录不匹配。如何得到这些不匹配的费用支出记录，而不必去检查每个表中的数据？

解决方案

左联接（详见第1.13节）是一种特殊类型的连接模式，用于识别一个表与另外一个表中的不匹配记录。当然，匹配必须基于两个表中的公共字段——一般情况下是父表的唯一键字段。该查询将从子表中返回那些父表中不包含的记录。