数据思维

SHUJUSIWEI

樊 敏 / 著

电子科技大学出版社

数据思维

樊　敏／著

电子科技大学出版社

图书在版编目(CIP)数据

数据思维 / 樊敏著. -- 成都 : 电子科技大学出版社, 2016.7

ISBN 978-7-5647-3653-8

Ⅰ. ①数… Ⅱ. ①樊… Ⅲ. ①数据处理 Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字(2016)第 103456 号

数据思维

樊 敏 著

出　　版:电子科技大学出版社(成都市一环路东一段 159 号电子信息产业大厦　邮编:610051)

策划编辑:李波翔

责任编辑:马　瑶

主　　页:www.uestcp.com.cn

电子邮箱:uestcp@ uestcp.com.cn

发　　行:新华书店经销

印　　刷:成都市火炬印务有限公司

成品尺寸:148mm×210mm　　印张 7　　字数 170 千字

版　　次:2016 年 7 月第一版

印　　次:2016 年 7 月第一次印刷

书　　号:ISBN 978-7-5647-3653-8

定　　价:28.00 元

前　言

伴随着互联网产业的崛起，大数据浪潮汹涌来袭，人们开始重新审视数据的巨大价值，力图充分利用数据中有价值的信息。人们对于数据并不陌生，但认识不太全面。

现实世界中的多种信息，如数字、字符、文字、图形图像、声音、视频等作为计算机中的数据在计算机内部是如何表示和存储的，在生动的外表之下，计算机如何能够认识这些数据并进行相应的处理。数据思维的形成首先要让大家认识计算机中的数据。不同的信息对应着不同类型的数据，不同数据类型所占的存储空间也不尽相同，数据的值也是不相同的。

事物之间联系是普遍存在的，同样数据之间也存在一定的联系，为了表达数据之间的联系，节省存储空间，有利于算法的实现，提高数据处理的效率，所以数据要以一定的组织结构存储于计算机中，这是数据结构的问题。在各个领域中需要对现有的数据管理，对数据进行事务处理，如插入、删除、修改、查询信息等。这些关于数据的管理需要首先将现实世界中信息通过数据建模在计算机中以一定的数据结构来表示其数学模型及操作的，并将具有统一

数据结构的数据的集合，存放在统一的存储介质中为多个应用程序所共享就构成数据库。随着市场经济竞争的加剧和企业数据量的积累，大数据时代汹涌来袭，传统数据库只保留了当前的业务处理信息，缺乏决策分析所需要的大量的历史信息。人们不满足于利用数据库对数据的处理和管理，更希望能够将多方面、多渠道的数据综合处理，从而在数据库的基础上集成能够进行决策分析的数据环境——数据仓库。发现有价值的信息来指导决策需要功能强大和通用的工具进行数据分析，把这些数据转化成有价值的知识，这是数据挖掘研究的内容。

对数据的认识、收集、整理和分析离不开数据思维。本书从计算机中数据的表示、数据的组织、数据的管理、数据集成以及数据的分析和挖掘等多方面详细系统地介绍了计算机领域信息数字化方法、数据结构理论、数据库、数据仓库、数据挖掘等相关技术，循序渐进地全面阐述了数据思维的内涵，着力于培养系统的数据思维，有助于理解大数据技术，进而为大数据管理提供新的解决思路，进行大数据分析，发现有价值的信息，指导人们生活实践。

随着互联网数据的膨胀式发展，大规模数据的聚集和交换形成了“大数据”。新时代的来临将影响着我们的思维，将为人们开启一扇窗，抓住了它也就抓住了机遇，当代人应该对数据理论和技术有所了解。

由于作者本身的水平有限，书中难免有不妥之处，恳请读者批评指正。

编　　者

目　录

第一章
认识数据

1.1 数的起源与发展

人类从原始的“数”到抽象的“数”概念的形成，是一个缓慢、渐进的过程。人们在生产活动中认识到了具体的数，通过在实践中的表达逐渐形成了记数法。

在远古时代，人类的祖先为了生存，抵御野兽的袭击，过着群居的生活。他们狩猎而归，猎物或有或无，于是有了“有”与“无”两个概念。后来，群居发展为部落。部落由一些成员很少的家庭组成。所谓“有”，就分为“一”、“二”、“三”、“多”四种。任何大于“三”的数量，他们都理解为“多”或者“一堆”、“一群”。然而，不管怎样，他们已经可以用双手说清这样的话（用一个指头指鹿，三个指头指箭）：“要换我一头鹿，你得给我三枝箭。”这是他们当时仅有的数的概念。

大约在1万年以前，冰河退却，一些从事游牧的石器时代的狩猎者在中东的山区内，开始了一种新的生活方式——农耕生活。他们碰到了记录日期、季节，计算收藏谷物数、种子数等问题，原有

“一”、“二”、“三”、“多”，已远远不够用了。

底格里斯河与幼发拉底河之间及两河周围，叫作美索不达米亚，美索不达米亚人和埃及人虽然相距很远，但却以同样的方式建立了最早的书写自然数的系统——在树木或者石头上刻痕划印来记录流逝的日子。尽管数的形状不同，但又有共同之处，他们都是用单划表示“一”。他们重复地使用这些单划和符号，以表示所需要的数字。

公元前1500年，南美洲秘鲁印加族（印第安人的一部分）习惯于“结绳记数”——每收进一捆庄稼，就在绳子上打个结，用结的多少来记录收成。“结”与痕有一样的作用，也是用来表示自然数的。根据我国古书《易经》的记载，上古时期的中国人也是“结绳而治”，就是用在绳上打结的办法来记事表数。后来又改为“书契”，即用刀在竹片或木头上刻痕记数，用一画代表“一”。直到今天，我们中国人还常用“正”字来记数，每一画代表“一”。当然，这个“正”字还包含着“逢五进一”的意思。

1.1.1 自然数

数的概念最初不论在哪个地区都是1、2、3、4……这样的自然数开始的，但是记数的符号却大不相同。

中国古代数字符号有甲骨文上的数字，在殷商之前，我国人民把文字写在乌龟甲和牛骨上，如图1-1所示。甲骨文上的数字，分别表示1～10和100，1000，10 000。我国古代还用算筹来表示数，分为纵式和横式两种方法，如图1-2算筹所示。用算筹计数时，个位、百位、万位都用纵式；十位、千位都用横式；高位在左，低位在右；遇到数字0时，就用一个空位表示。后来，编写上书时，就约定俗成以符号〇代表数字0，这恰好与今天阿拉伯数字0的形态

相近。中国古代十进位制的算筹记数法，在世界数学史上是一个伟大的创造，把它与世界其他古老民族的记数法做一比较，其优越性是显而易见的。

图 1-1　甲骨文上的数字

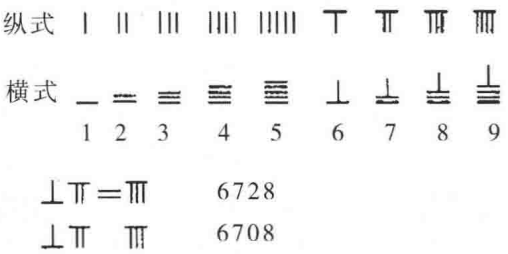

图 1-2　算筹

古罗马的数字现在许多老式挂钟上还常常使用。实际上，罗马数字的符号一共只有 7 个：I、V、X、L、C、D、M 分别代表数字 1、5、10、50、100、500、1000。利用这 7 个符号按照一定规则组合来表示数，如重复次数表示几倍、右加左减等。如果要记稍大一点的数目就相当繁难，而且罗马数字中没有“0”。其实在公元 5 世纪时，“0”已经传入罗马。但罗马教皇凶残而且守旧。他不允许任何人使用“0”。

由于人类在计数时自然而然地使用十个手指，所以绝大部分国家和地区的人民普遍认同十位进制的记数符号，现在世界通用的数符号 0、1、2、3、4、5、6、7、8、9，人们称之为阿拉伯数字。实际上最早使用阿拉伯数字的是古印度人，后来阿拉伯人把古希腊的数学融进了自己的数学中去，又把这一简便易写的十进制记数法传遍了欧洲，逐渐演变成今天的阿拉伯数字。

1.1.2 分数

后来人们发现，仅有自然数是远远不够的。如果分配猎获物时，5 个人分 4 件东西，每个人该得多少呢？于是分数就产生了。公元前 2500 年，毕达哥拉斯的学生在研究 1 与 2 的比例中项时，发现没有一个能用整数比例写成的数可以表示它，这个新数的出现使毕达哥拉斯感到震惊，紧接着人们又发现了很多不能用两整数之比写出来的数，如圆周率就是最重要的一个，人们就把这些数称作无理数。早在公元 3 世纪，我国古代数学家刘徽在解决数学难题时就提出了把整数个位以下无法标出名称的部分称为徽数。最初，人们表示小数只是用文字。小数的名称是公元 13 世纪我国元代数字家朱世杰提出的，在 13 世纪中叶我国出现了低一格表示小数的记法。古代，还有人记小数是将小数部分的各个数字用圆圈圈起来，例如：1.5 记作 1⑤，这么一圈，就把整数部分和小数部分分开了。这种记法后来传到了中亚和欧洲。直到 16 世纪，德国数学家克拉维斯首先使用了小数点作为整数部分与小数部分分界的记号，于是，小数的写法就成了我们现在的表示方法。但是，用小数点表示，在不同的国家也有不同的方法。现在，小数点的写法有两种：一种是用“,”；一种是用小黑点“.”。在德国、法国等国家常用“,”，写出的小数如3,42、7,51……而英国和北欧一些国家则和我国一样，用“.”表示小数点，如 1.3、4.5……

1.1.3 负数

我国是世界上首先发明和使用负数的国家。战国时法家李悝（约公元前 455—前 395 年）曾任魏文侯相，主持变法，我国第一部比较完整的法典《法经》（现已失传）中已应用了负数，“衣五人终岁用千五百不足四百五十”，意思是说，5 个人一年开支 1500 钱，

差 450 钱。在甘肃居延出土的汉简中，出现了大量的“负算”，如“相除以负百二十四算”、“负二千二百四十五算”、“负四算，得七算，相除得三算”。

我国古代数学家刘徽在建立负数的概念上也有重大贡献。公元 3 世纪刘徽在注解《九章算术》时率先给出了负数的定义：“两算得矢相反，要以正负以名之”，并辩证地阐明：“言负者未必少，言正者未必正于多”。刘徽第一次给出了正负区分正负数的方法。他说：“正算赤，负算黑；否则以邪正为异。”意思是说，用红色的算筹摆出的数表示正数，用黑色的算筹摆出的数表示负数；也可以用斜摆的算筹表示负数，用正摆的算筹表示正数。用不同颜色的数表示正负数的习惯，一直保留到现在。现在一般用红色表示负数，财政赤字表明支出大于收入，财政上亏了钱。

1.1.4 其他数

有理数和无理数一起统称为实数。但在解方程的时候常常需要开平方，如果被开方数为负数，这道题还有解吗？于是数学家们就规定用符号“i”表示“-1”的平方根，虚数就这样诞生了。

数的概念发展到虚数以后，在很长一段时间内，连某些数学家也认为数的概念已经十分完善了，数学家族的成员已经都到齐了。可是 1843 年 10 月 16 日，英国数学家哈密尔顿又提出了“四元数”的概念。所谓四元数，就是由一个标量（实数）和一个向量（其中 x、y、z 为实数）组成的数。四元数在数论、群论、量子理论以及相对论等方面有广泛的应用。与此同时，人们还开展了对“多元数”理论的研究。到目前为止，数的家庭已发展得十分庞大。

1.2 计数制

计数制也称数制，指用一组固定的符号和统一的规则来表示数值的方法。在日常生活中会使用十进制来记数，计算机中采用二进制数进行信息编码，利用八进制数和十六进制数来表达二进制数等等。

1.2.1 常用的计数制

十进制数（Decimal Number）由十个不同的数字符号0、1、2、…、9中的某些数字符号构成的，按照“逢十进一、退一当十”的规则进行计数。这十个数字符号称为数码，用K表示。数码的个数称为基数，用R表示。数码在排列中所处的位置，即数位，用i表示。整数部分数位从低位向高位依次为0、1、2…，小数部分的数位从高位向低位依次为－1、－2…。一个数码在不同数位上，具有不同的位置值，即位权，值为R^i。

所以，R进制数有如下的特点：

- 数码的个数等于基数R；
- 最大的数码比基数小1；
- 第i位上的权为R^i；
- 运算规则是“逢R进一”，“退一当R”。

二进制、八进制和十六进制是在计算机领域广泛使用的数制。为区分不同的计数制，可以在数的右下角注明数制，也可在数字后面加一个字母，如字母B（binary）表示二进制数；字母Q（octal）表示八进制数，这是为了不与数字O相混淆，把字母O改成Q；字母D（decimal）或不加字母表示十进制数；字母H（hexadecimal）表示十六进制数。例如，$1001B$、$47.3Q$、$386.5D$、