主编/卞毓麟

JINPINGGUO WENKU

金苹果文库

陈念贻

数据淘金

金苹果文库

主编　卞毓麟

数据淘金

陈念贻　著

江苏教育出版社是1994年受到中共中央宣传部和新闻出版署表彰的全国优秀出版社之一。

《金苹果文库》是江苏教育出版社出版的大型科学普及丛书，共出版5辑50种，印刷发行100万册，并多次获奖。

优秀科普作家和热心科普的科学家构成《金苹果文库》的作者阵容，他们奉献给广大读者的是"真正看得懂的科普，中国人自己的科普"！

《金苹果文库》第5辑书目

王直华著	《科学人文对话》	定价5.00元
樊洪业著	《欢迎"赛先生"》	定价6.00元
赵寿元著	《基因造福》	定价4.00元
陈念贻著	《数据淘金》	定价6.50元
甄朔南等	《走近恐龙》	定价6.50元
卞毓麟著	《群星灿烂》	定价6.50元
李葆明著	《大脑如何记忆》	定价5.50元
吕学诗著	《进化中的机器人》	定价6.50元
吴玉虎著	《昆仑探险》	定价6.00元
汪宗俊著	《警惕药害》	定价7.00元

金苹果文库
数据淘金
陈念贻 著
插图 白庚和
责任编辑 喻 纬 责任校对 周 萍

出版发行：江苏教育出版社
（南京市马家街31号，邮政编码：210009）
经 销：江苏省新华书店
照 排：南京展望照排印刷有限公司
印 刷：淮阴新华印刷厂
（淮安市淮海北路44号，邮政编码：223001）

开本 850×1168 毫米 1/32 印张 6.5 插页 5 字数 156 000
2003年12月第1版 2003年12月第1次印刷
印数 1－10 000册

ISBN 7－5343－5295－9

G·4990 定价：6.50元

主编的话

世纪之交，果园飘香，灿烂的阳光下，百万只“金苹果”挂满枝头。面对此情此景，你将有何感受？

这片果园，展现在中国的科普田野上；这每一只“金苹果”，就是我们这套《金苹果文库》的一册书。

《金苹果文库》列入国家重点图书出版规划后，编写出版工作进展顺利。全部5辑共50种图书，按每辑10种依次出版。前4辑40种出版后，至今已累计印行90万册，让全国数以百万计的读者品尝到了它们的芳香与甜美。现在，随着第5辑10种正式付印，“金苹果”的产量也真的上了百万。

我们在第1、2辑《主编的话》中说过，科学的发展是一代又一代富有献身精神的人不断努力、不断拼搏的结果。对此，科学巨匠牛顿有一句广泛流传的名言：“如果我比别人看得远些，那是因为我站在巨人们的肩上。”

从牛顿的时代至今的三个多世纪中，科学发展越来越迅速，也越来越复杂，所以科学家、科学教育家们就有义务向社会公众，特别是向青少年们尽可能通俗地宣传普及科学精神、科学思想、科学方法和科学知识，这就是我们主编这套《金苹果文库》的宗旨。

“金苹果”首先是为青少年朋友编写的，具有初中文化程度的读者基本上就可以看懂。当然，它们一定同样会受到渴

求加深了解科学技术的成年读者的青睐。“金苹果”的作者们有一个共同的心愿，那就是使读者充分体验到，阅读科学书籍实在是一种妙不可言的美的享受。

几年来的事实业已表明，“金苹果”很受读者欢迎，先期出版的第1、2、3辑已经多次获奖。例如，第3辑获第12届中国图书奖、江苏省第4届“五个一工程”图书奖，第1、2辑均被评为全国优秀畅销书、获华东地区优秀教育图书奖，第1辑获江苏省优秀图书一等奖。在许多地方，“金苹果”还被教育、科技部门推荐给广大中小学生，成为他们喜爱的课外读物。

“金苹果”为什么会取得成功？原因很多，其中有一条很值得一提，那就是我们组建了一支很优秀的作者队伍。这些作者大多获得过中国科普作家协会的表彰，而且有丰富的科研经验，这就为科普作品的科学性、新颖性和深刻性提供了有力的保证。同时，他们也了解中国读者对科普的需求，熟悉中国读者的阅读习惯和思维方式，他们乐意尽力用自己的智慧和笔墨，和读者一同赏析蕴藏在真实的科学精神、科学思想、科学方法和科学知识中的永恒魅力和无穷乐趣。

“金苹果”在选择作者和确定选题时，突破了严格按学科分类和强调覆盖主要学科门类的思维模式，而是先确保作者队伍的“整齐”，再由作者提出最“拿手”的选题，从而确保整套丛书的质量，突显丛书的特色。我想，这样培育出来的“金苹果”，大概是很难“克隆”的吧。

培育“金苹果”的历程，是一次“集结中国优秀科普作家队伍，展现中国优秀原创科普成果”的过程。如今，随着“金苹果”第5辑的问世，编辑出版这套文库的任务算是圆满完成了。然而，“金苹果”的生命力仍将与时俱增，为此，我们再次诚恳地请读者朋友将品尝“金苹果”的感受告诉我们，帮助我们不断地总结经验教训，不断地开拓进取，不断地为我国的科

普事业提供更加美好的新作品。

对我本人而言，和众多的作者、编者、读者一起，共同培育我们的“金苹果”，实在是一段非常值得回忆的美好经历。亲爱的朋友们，我衷心地期待着：有朝一日，在祖国的科普田野上，在一片新的果园中，我们大家再次来相聚。

卞毓麟

2003 年 8 月 31 日

目 录

社会应用篇

科研应用篇

展望篇

我与科学世界

我生于1931年。我出生才一个多月，日本军阀就在中国东北制造九一八事变，出兵占领了东北三省。1937年，就在我6岁那年，日本又制造七七事变，出兵向中国全面进攻。我的童年就是在民族灾难的年代里度过的。那时的情况，就像《黄河大合唱》中的《黄水谣》唱的那样："自从鬼子来，百姓遭了殃！奸淫烧杀，一片凄凉，扶老携幼，四处逃亡，丢掉了爹娘，回不了家乡！"当时我问我的父亲：为什么中国人这样倒霉，这样受人欺负？父亲回答说：因为中国人不掌握科学技术，所以日本鬼子能用飞机大炮打我们。我当时就暗暗下决心：我长大了以后一定要研究科学技术，帮助我们的国家强大起来。

我从那时起一直到现在，都在为实现这个目标而努力。

我小时候挨日本人的飞机轰炸，对那种身在防空洞，耳听敌人飞机俯冲时的怪叫声，不知下一秒钟炸弹是否就会在自己头上爆炸的恐怖，印象很深。所以我献身科学的第一志愿就是要研究造飞机的金属——铝的生产技术，帮助我们国家造飞机。我大学毕业后的第一个十年，就是从事炼铝的研究。我也确实配合我国铝工业的建设做出了一批成果，出版了我的第一部学术著作——《氧化铝生产的物理化学》。后来，我被聘为中国有色金属总公司铝冶炼技术中心的成员。1984

年在桂林开会，我听说我国铝产量即将突破年产 100 万吨的大关，成为位居世界第四的炼铝大国，非常欣慰，当即赋诗一首述怀：

铝业方兴战鼓频，漓江西畔会群英。
百万大关新技术，汇集众智建中心。
大河上下新槽列，长城内外炼轻金①。
车船建筑节能业，会看此物用途新。

自从 20 世纪 50 年代起，我认识到新材料比铝更加重要。因为原子能、计算机、航天等新技术，处处要用新材料。1956 年我参观过苏联的一个新材料研究所。给我印象最深的是他们安排了一百多名科研人员成天在配料，配出成千上万种成分的料去做试验，以便将来有一天能从中挑出性能好的新材料来。这简直就像“沙里淘金”，我觉得盲目性太大。恰好这时国际科学界已经提出了“材料设计”的想法，研究以计算机为工具，利用人类对物质结构的新知识来预报新材料，减少盲目性，少做试验，更快、更好地找到合用的各种新材料。我感到这种想法正合我意。从 1964 年开始，我就开始学习这方面的基础知识，打算开展“材料设计”研究。从那时起，我开始对计算机感兴趣了——这是我接近“数据挖掘”的第一步。

正当我为“材料设计”这一新方向做学术准备的时候，一场铺天盖地的政治运动——被称为“文化大革命”的十年动乱冲击了我的科学行程。当时上海是“四人帮”统治下的“重灾区”，科学家无论大小都横遭迫害。我也被安上了许多“罪名”，如“崇洋媚外”啦（我向中学生介绍过自己学外文的经验，

① “轻金”指“轻金属”，密度小于 4.5×10^3 kg/m^3 的金属为轻金属，包括铝、镁等，因为轻，故可造飞机。

鼓励他们学外文)、“个人奋斗”啦(我总想多出科研成果,不能不努力)之类。我被赶出实验室,辛辛苦苦建立起来的试验设备也被“砸烂”了。他们砸烂了我的试验设备,却砸烂不了我献身祖国科学的意志。我没有了在实验室中测量数据的条件,但我在过去15年中已积累了不少数据,也从国外文献中查到大量数据。于是我决心利用这些数据来总结其中的科学规律。当时我没有电脑,但我有纸、笔和计算尺,还有“人脑”。“塞翁失马,焉知非福”,我从此走上了从数据中“挖掘”宝贵的科学信息,也就是今天我们称为“数据挖掘”的道路。

当时我被送到上海冶炼厂“劳动改造”,和工人一起在车间里劳动。白天我在工厂和工人一起搞技术革新,晚上回家就“一张纸,一枝笔,一把计算尺”,做起计算来。记得1968年冬的一天,“造反派”们逼迫我承认自己是“反革命”,威胁说如果我不肯“交代”就叫我“坐牢”。当天我回到家,窗外大雪纷飞,室内滴水成冰。我照样算题到四更天,毫不懈怠。后来我的这些研究成果终于写成书,由科学出版社出版时,我感慨地在书的扉页上题了一首诗:

踏过岷山路更长,又迎急雨渡春江。
难忘滴水成冰夜,四更灯下演题忙。

1976年“四人帮”被粉碎后,我国走上了繁荣富强的道路。我也有了使用计算机的条件,并承担了国家高技术重点项目——材料设计和工业生产优化的研究工作,多年来做数据挖掘积累的知识和经验有了用武之地。

经过我和我的合作者多年的努力研究,我们已经发展了一套国外没有的新算法,为我国研制和生产多种新材料做了配合工作,也为我国宝钢(宝山钢铁公司)等一批大中型钢铁、炼油、化工企业节约能源、节约原材料、提高产品质量、突破生

图 1　我们的数据挖掘技术优化我国工业生产的成果分布图

产瓶颈和降低生产成本等做出了贡献(图 1)。如今我们的数据挖掘技术已不限于国内应用,连美国、欧洲和东南亚的一些单位也开始采用。我们已是一支国际知名的材料设计、计算化学研究的“正规军”。

科学技术包括许多学科,真是浩瀚的知识海洋。几十年来,我好比在知识的海洋里游泳,总想游向一块利国利民、可

以钻研出新水平的“宝地”。经过曲折的摸索，我找到了这样一块“宝地”——数据挖掘技术和它的应用。在这块“宝地”上工作，已经有 20 多年了。道路越走越宽，越往前看，越感到精彩纷呈。我虽已年过 70，还是“手提电脑走天下”——从上海到北京、从广州到新加坡、从硅谷到波士顿……到处交流和合作，越做越有劲。这就像我青年时写的一首诗那样：

安逸舒适非吾愿，尽情工作度今生。
科学大路多宽广，策马高歌万里征。

我说数据挖掘是块学科上的“宝地”，有以下一些理由：

(1) 当前世界上科技领域的信息革命正在走向高潮。数据挖掘已成为信息革命的一个新热点。在我国，各行各业在这方面基本上还没有动起来。但是，根据国际发展趋势预测，将来中国在这方面肯定会掀起高潮。如能早点着手，将会大有可为。我和我的同事通过长期的科研工作，已开发了一批国际上有特色的新算法和相应的软件。即使是从国际范围来看，我们也是占有某种有利地位的。

(2) 我国大部分工厂生产管理比较粗放，利用数据挖掘优化生产的潜力很大。特别是在我国加入 WTO 以后，面临国际市场的竞争，生产和经营都会需要数据挖掘。数据挖掘如果开展得好，不需要多少投资，就可以为我国工业每年增加数十亿元的利润。我国正从计划经济走向市场经济，企业管理既是薄弱环节也是改革热点。金融财务科学化管理问题也很多。这些都需要数据挖掘来“帮忙”。

(3) 目前科学上好几个热点领域，包括分子和材料设计、生物学中的基因研究、环境科学研究、宇宙科学研究以及金融科学研究，都已积累了大量复杂数据，都需要数据挖掘解决它们的关键问题。占领了数据挖掘这块宝地，随时可以向这些

热点领域进军，探讨自然界和社会现象的奥秘。

（4）就数据挖掘这门学科本身而言，它的理论和方法还都很不成熟，还没有完全脱离古典统计数学的框框。如果说古典统计数学家从事的是简单数据处理研究的话（历史上许多概率论权威从与赌徒交朋友、研究赌博中的统计规律入手，就是因为赌博中许多事件是简单的重复事件的缘故），那么我们也可以希望，面对现代社会大量的复杂数据，将来总可以总结出一套处理复杂数据的新理论、新方法。

我还要说明的是，在我的数据挖掘等领域的科研成果中，也凝集了我的同事和合作人员的心血。其中特别要提到的是我在上海大学的同事陆文聪、阎立诚，在上海冶金研究所和交通大学的同事陈瑞亮、钦佩、杨杰和姚莉秀，以及在工业优化方面有特殊贡献的张未名等。宝钢的康复、彭平和南京炼油厂的邵永锡、高秉宏，美国福特汽车公司的丹尼斯、朱卫平和宋沐民，硅谷的朱东屏，瑞士 MPDS 组织的维勒斯等，也在合作中给我们以莫大的帮助。钦佩高级工程师除了参加科研做出了贡献外，还参与了本书的写作和编排工作。

亲爱的读者，当你们看完这本小书后，一定能了解到数据挖掘技术是一门“上至天文，下至地理”，对探索物质结构、神秘的生命现象、制造各种新物质、发展工农业生产都有用的新兴科学技术。

本书是一本科普著作。我们力图使数据挖掘这门博大精深的科学技术得以通俗有趣地呈现在读者面前。但由于我们水平所限，更由于科普书不可能运用更多的公式和名词，这本书只能勾画出数据挖掘学科的概貌。如果读者有意进一步认识和掌握这门学科的准确知识，请你阅读我们的下列著作：

《模式识别方法在化学化工中的应用》，科学出版社 2000 年 9 月版；

《模式识别优化技术及其应用》，中国石油化工出版社 1997 年 10 月版；

《计算化学及其应用》，上海科学技术出版社 1987 年 10 月版。

未来属于你们青年人。希望你们把我们开创的学术领域进一步发扬光大，使我们的祖国更加繁荣昌盛！