信息计量学导论

XINXIJILIANGXUEDAOLUN

主编 郭强 刘俊友

合肥工业大学出版社

信息计量学导论

主　编　郭　强　刘俊友

副主编　赵　瑾　郑　斌

编　者　（按姓氏画划排序）

刘亚迅　刘俊友　刘思源　李伟芬

陈龙强　张　芳　郑　斌　郭　强

赵　瑾　程　浩

合肥工业大学出版社

内容提要

信息计量学是目前国内新兴的交叉边缘学科，处于起步阶段，目前主要有两个研究方向：即广义信息计量学和狭义信息计量学。本书侧重于如何将广义信息计量问题转化为一般计量技术问题的分析过程，形成新的内容结构体系，更注重可操作性，强调为信息决策服务。

全书共由三部分组成，第一部分介绍了信息计量的基本概念、相关问题和技术准备；第二部分介绍了信息计量的图书情报学技术、系统分析技术、决策分析技术和模拟分析技术；第三部分为附录，主要介绍信息的基本知识和信息分析软件的使用。

本书可作为信息管理与信息系统、信息决策、情报学、图书馆学、档案学、信息资源管理、电子商务等相关专业的教材使用，也可作为从事信息管理与运用工作的实用性的参考书籍。

图书在版编目(CIP)数据

信息计量学导论/郭强，刘俊友主编．—合肥：合肥工业大学出版社，2007.8

ISBN 978-7-81093-656-9

Ⅰ.信… Ⅱ.①郭…②刘… Ⅲ.文献计量学 Ⅳ.G257

中国版本图书馆 CIP 数据核字(2007)第 131007 号

信息计量学导论

郭 强 刘俊友 主编　　责任编辑 权 怡 刘亚宁

出 版	合肥工业大学出版社	版 次	2007 年 8 月第 1 版
地 址	合肥市屯溪路 193 号	印 次	2007 年 8 月第 1 次印刷
邮 编	230009	开 本	710×1000 1/16
电 话	总编室：0551-2903038	印 张	12.75
	发行部：0551-2903198	字 数	240 千字
网 址	www.hfutpress.com.cn	印 刷	合肥创新印务有限公司
E-mail	press@hfutpress.com.cn	发 行	全国新华书店

ISBN 978-7-81093-656-9　　定价：28.00 元

前　言

一般来说，我们对“信息”的理解有广义和狭义之分，前者认为“信息是能够用来消除不确定性的东西”（申农，1948 年），这是一个具有共性的抽象的概念；后者则着眼于文献、情报、图片、声音等内容，他们基本上都可以“看得见”、“摸得着”，或者二者兼而有之。既然如此，“信息计量学”也就相应的有了“广义信息计量学”和“狭义信息计量学”之分。前者主要探讨“以广义信息论为基础的广义信息的计量问题”（邱均平，2007 年）；后者则主要研究情报信息或文献情报的计量问题。

目前，国内出版的关于信息计量学的书籍并不算多。它们大致可以分为侧重点迥异的两类：一类侧重于“狭义信息计量学”的研究范畴，涉及文献计量学的相关内容、文献信息统计分析、引文分析和计算机辅助分析等部分，不涉及信息计量的本质问题和技术措施；另一类内容正好相反，只强调基于信息论、控制论和系统论的信息计量问题，以严密的数学分析和大量实例计算分析为显著特点。当然，这只是大致的说法，互有交叉的情形也是存在的。

编者一直从事信息（情报）分析和信息系统模型研究等相关领域的教学和研究工作，深感应对这些内容有所侧重，即如何将广义的信息计量问题转化为一般计量技术问题的分析过程。但又担心这部分内容与数学分析在论述风格上反差过大，所以在设计纲目时予以折衷，让两方面都靠近一些。作为一种尝试与探索，编者编写了这本教材，起名为《信息计量学导论》。

本书内容框架是否妥当，尚祈读者指正。

编　者

2007 年 6 月

目　录

第一章 绪 言

第一节 信息计量学的定义

我们在几千年前就拥有了关于文字、数字、图画、声音等知识，但对如何统一表述这些知识、如何统一地计量它们的数量的问题迟至19世纪末还没有正确地提出来，更谈不上如何去解决了。直至20世纪中期，随着电报、照片、无线电、电话、雷达、电视等等的出现和发展，如何计量各种信号中信息多少的问题才开始被提上日程。

对信息的计量就是要从数量关系上把握信息。对信息的定量把握，是进一步探讨信息的科学规律的基础，也是信息处理和应用的基础。信息的定量描述来自于对信息本质的认识，对信息的本质认识到什么程度，就会出现与之相适应的信息计量方法。遗憾的是，现在人们对信息的本质虽然有了一定程度的认识和把握，但仍然是不完整的。在各个不同的层次上，信息的理论研究发展是不平衡的，也造成对关于信息的计量问题的阐释，会经常出现一些非严格、非完整、似是而非的缺憾，这些又反过来影响人们对信息概念的准确认识。

目前，国内外关于信息计量学较为成熟的观点如下。

一、信息计量学的概念

所谓信息计量学，就是一门采用定量方法来描述和研究情报（信息）的现象、过程和规律的学科，是情报学关于定量分析的分支学科，它是由数学、统计学、运筹学等与情报学紧密结合而形成的，具有交叉学科的性质。

邱均平教授认为：信息计量学应分为“广义信息计量学”与“狭义信息计量学”。

“广义信息计量学”主要探讨以广义信息论为基础的广义信息的计量问题，其范围非常广泛。大家知道，信息与物质、能量构成客观世界的三个基本要素。信息是物质的普遍属性，是一个系统，通过感觉器官与外界交换的一切内容，它能够减少或消除系统的不确定性。正如1948年申农在《通信的数学理论》中所指出的，一个系统所接收的“信息是能够用来消除不确定性的东西”。所以，信

息是可以度量(计量)的。如同物质、能量的度量一样,信息的计量关键在于得出了几乎一致的结论。维纳在1948年发表的《控制论》和1950年的《人当作人来使用》的著作中进一步拓宽了信息的概念,提出了信息量的定义和计算公式。弗希尔则从古典统计理论的角度研究了信息量度问题。申农在1948发表的著名论文《通信的数学理论》和1949年的论文《噪声中的通信》中,集中阐述了他的研究成果。他把各类信息系统中的不同信号的共同特征抽象出来,略去其具体内容,当作抽象的随机事件处理,单纯从量的角度来描述信息,在语法和传输上进行定量研究,建立了统一的通信理论,解决了同一信息可用不同的信道传输或不同的信息可用同一信道传输的问题,并且提出了信息量的数学公式,定义信息量等于被消除的不确定性的数量,从而创立了作为一门独立学科的信息论。

"狭义信息计量学"就是我们通常讲的"信息计量学(或情报计量学)",主要研究情报信息(或文献情报)的计量问题,是运用数学、统计学等定量方法来分析和处理信息过程中的种种矛盾;从定量的角度分析和研究信息的动态特性,并找出其中的内在规律。

信息计量学的研究对象目前主要是在各种事物信息的数量方面。其内容包括消息(Message)、数据、事件、实物、文本和文献等,其中既有正式交流的事物信息,也有非正式交流的事物信息。当然,从广义的信息计量学来看,它还包括过程信息和知识信息的数量方面。

二、信息计量学的研究目的与意义

信息计量学研究的基本目的就是:要引进量的概念和定量分析方法,进一步揭示信息单元(包括文献、数据、实物、消息、事件等)的体系结构和数量变化规律,从理论上提高情报学及信息管理学科的科学性和精确性,促使这些学科向定量阶段发展。同时,为改善情报信息系统提供定量依据,达到高效能的科学管理,使信息交流系统处于最佳运行状态,提供最优化的信息服务,以便更好地解决信息服务工作中的基本矛盾,克服"信息危机",使信息管理工作更有效地为科学技术、经济和社会发展服务。

信息计量学研究的最大意义在于:从理论上继续总结各种经验定律,使经验层次上的信息(情报)"工作"上升到理论层次上的信息(情报)"科学",从而充实其理论的广度和深度,同时将各种经验定律在新的信息单元条件下进行检验和修正,探讨它新的适用性,从而大大提高情报学的科学性,同时又能为实际工作提供理论指导。

三、信息计量的哈特莱对数计算方法

哈特莱早在20世纪20年代就提出用对数作为信息量的测度。哈特莱认

为：消息和信息不同，多种多样、千姿百态的消息是信息的载体，消息究竟包含了多少信息，应该用消息出现的概率的对数来计算，从而使他为信息度量找到了对数这一数学理论。

哈特莱对数计算方法的基本机理，即对某个被传递的消息的要素数量和这些要素出现的可能性利用0、1二进制算法对其排列。比如，某个被传递的消息有8个要素，即由8个可能性（或不定性）组成，采用0、1二进制算法对其排列，那么每个不定性需要两个数字排列，则这8个可能性便可排列为：111，000，110，001，101，010，100，011。由此排列说明，该被传递信息用二进制的三位数字就可传递完毕。由于$8=2^3$，取对数，则$\log_2 8=3$。因为二进制的信息单位称比特（Bit），这就是说，一个有8个不确定因素的信息（假设其为等概率）含有3比特的信息量。比如，有一则信息，有9个可能性要素，那么每个不确定性需要3个数字排列，则可排列为00，01，02，10，11，12，20，21，22，由此排列说明，被传递的信息是用三进（两位数）就可传递完毕。由于$9=3^2$，那么，$\log_3 9=2$，说明该信息含2铁特（Tet）的信息量。

再如，基于某天天气情况的分析。根据天气预报情况，通常有四种可能性，即：阴、晴、雨、雪。如果其概率相同，那么$4=2^2$，取对数，$\log_2 4=2$，即这一则信息有4个不确定因素，含有2比特的信息量。同理，如果气象台已将雨、雪排除，那么只剩下两种可能性，即阴、晴，因为$2=2^1$，则$\log_2 2=1$，其信息量为1比特。据此，我们可以依据类推原理，即有m个要素的不确定消息，如果采用二进制，每个要素的信息量为：$H=\log_2 m$，其中H代表信息量。可见H单位将随m的增加而增加，具有累加性。由此就又可推出，两个消息加在一起的总信息量等于每个信息各自信息量之和，即：

$$\log_2(m_1 \cdot m_2)=\log_2 m_1+\log_2 m_2$$

采用哈特莱的对数计算方法需要注意：一是要注意知识被传递的消息的要素；二是要根据要素数量确定采用进制单位；三是要注意各要素（不定性）应是等概率。

此外，采用对数方法时，计算信息量的单位由于其底数不同，其单位也不同，如以2为底数，信息单位称比特（Bit），二进制单位；以3为底数，则称铁特（Tet），三进制单位；以e为底数，称奈特（Net），为自然单位；而以10为底数，称笛特（Det），即十进制单位。

四、信息计量的申农概率论统计方法

众所周知，概率论是研究随机现象的数学分支学科。数理统计又是以概率论为理论基础，以统计数据（实验数据）为对象来进行统计分析推断、研究统计规律性的。可见概率统计是研究事件的必然与偶然这一矛盾的数学分支。申

农在其《通信的数学理论》一书中对哈特莱的对数计算方法给予肯定的同时，提出“实际的信息量是接收前具有的不定度减去实际收到信号的不定度的差”的观点，因此，申农在哈特莱对数计算方法的基础上，为了把握被传播消息要素的不同概率，以及如何正确计算并排除误差，提出了用概率统计法计算信息量的方法。而这种方法的理论基础就是概率论。

1. 概率统计信息度量法的基本机理

由于信息的要素和各要素产生的概率都不同，信息在从信源经信道到达信宿的传播过程中，受干扰因素的影响，信宿接收到的信息往往又小于信源信息。因此，申农根据信息的随机特性，提出信息的概率统计模式，即：

$$H = -\sum p_i \log p_i$$

他认为，这一模式在信息论中起着主要作用，“它们为信息选择和不确定性的度量”。

2. 应用实例

为了说明问题，我们不妨以掷骰子为例：大量实验证明，一颗结构均匀的骰子，出现 1 点的概率为 1/6，不出现 1 点的概率为 $1-1/6=5/6$，其他 2 点，3 点，…，6 点出现与不出现的概率都与 1 点的情况相同。因此，就各点出现的概率而言，1 点 = 2 点 = 3 点 = 4 点 = 5 点 = 6 点。由于概率论一般以 p 表示概率，p 通常有两个条件：既不能是负值，又不能超过 1($0 \leqslant p \leqslant 1$)。正如上述掷骰子的 6 个点，尽管其各点的概率相同，即均为 1/6，但各点的概率总量之和等于 1。由此出发，骰子 6 个面，其不定性总量为 6，用对数表示，为 log6。由于各点是等概率，故骰子每面朝上的概率为 1/6，其量为 1/6 乘以 log6，这样每面朝上提供的信息量为 $-1/6 \times \log \frac{1}{6}$，同时，整个骰子的不定量为 6 面信息之和，即它可以提供的信息量，由此，我们可以把掷骰子的整个信息统计模式转换为：

$$H = -\sum_{i=1}^{6} p_i \log p_i = -\sum_{i=1}^{6} \frac{1}{6} \log \frac{1}{6} = 2.585(\text{比特 / 点})$$

如果我们把上式符号化，可为下式：

$$H = k\sum_{i=1}^{N} p_i \log p_i$$

式中 H——代表某个事件的平均信息量

K——系数(取决于变量单位)

N——代表某事件各种信号之和

$\sum$——总和符号

p_i——代表该事件中某一信号可能出现的概率

第二节 信息计量学的发展简史

最早的信息计量研究始于20世纪初，1917年，由文献学家科尔(F. T. Cole)和伊尔斯(N. B. Eales)进行的文献统计研究为起点。

1922年，英国图书馆学家休姆(E. W. Hulme)在其编著的《统计目录学与现代文明增长的关系》中首次使用了“统计目录学”(Statistical Bibliography)的名称。

1969年，英国著名情报学家阿伦·普里查德(Alan Pritchard)首次提出用术语“Bibliometrics”取代“统计目录学”的名称。他的建议很快得到图书馆学、情报学界的普遍承认。这一术语的出现标志着文献计量学的正式诞生。

1978年，匈牙利学者蒂博尔·布劳温(Tibor Braun)创办的《科学计量学》杂志，为国际上从事科学计量学的学者提供了一个学术交流的平台，促进了科学计量学的发展。

1979年，德国学者昂托·纳克教授(Otto Nacke)最早提出德文表述的“信息计量学”(Informetrie)，随后文献中很快就出现了与之对应的英文术语 Informetrics。

1980年，信息科学家说服国际文献联合会(FID)设立了“信息计量学委员会”(Committee On Informetrics)，并拟定了一个长期信息计量学教学与研究工作计划。

1980年9月，在德国法兰克福召开了第一次情报计量学(含科学计量学)研讨会，纳克教授在会上宣传了他提出的“情报计量学”术语。

1981年，在我国的期刊上也出现了上述德文和英文术语，并将其译为情报计量学。

1985年7月，在印度出版了非正式的《信息计量学研究通讯》(Informetrics Newsleuer)杂志。

1987年，第一届文献计量学与情报检索理论国际研讨会在比利时举行，布鲁克斯在会上提议，应将 Infolmetrics 术语补充到拟于1989年在加拿大召开的第二届国际学术会议的名称中去，这一提议得到了与会学者的普遍赞同和支持。

1991年在印度、1993年在德国召开的第三届、第四届国际会议上，布鲁克斯的意见都在一定程度上被接受了，至1995年6月，在美国芝加哥召开的学术会议就改名为“第五届科学计量学和情报计量学国际会议”，文献计量学虽被包括在内，但在会议名称中被取消了。现名为“国际科学计量学和信息计量学学会”(International Society for Scientometrics and Informetrics，缩写为 ISSI)主办的两年一次的国际研讨会名称的变化也说明“情报计量学”得到了国际学术

界的认可，并且其学科地位越来越突出了。参见表1-1。

表1-1 部分科学计量学与信息计量学国际学术研讨会概况

届数	会议名称	时间	地点	主题
一	文献计量学和信息检索的理论问题国际研讨会	1987年8月25日～28日	比利时迪彭贝克	1. 几大规律的深入探讨 2. 引文分析的应用
二	文献计量学、科学计量学与信息计量学国际研讨会	1989年7月5日～7日	加拿大伦敦	1. 三计学范围的界定 2. 几大规律的推广
三	信息计量学国际研讨会（印度统计协会）	1991年8月9日～12日	印度班加罗尔	1. 统计方法在信息计量学中的应用 2. 数学方法的应用
四	文献计量学、科学计量学与信息计量学国际研讨会	1993年9月11日～15日	德国柏林	1. 三计学研究内容的关系 2. 引文分析的应用
五	科学计量学与信息计量学国际学术研讨会	1995年6月7日～10日	美国伊利诺斯州	1. 期刊评价的探讨 2. 几个规律的新发展
六	同上	1997年6月16日～19日	以色列耶路撒冷	1. 引文分析的应用 2. 文献老化和离散规律研究 3. 数据压缩 4. R&D管理研究等
七	同上	1999年6月5日～8日	墨西哥科利马	1. 学术期刊评价 2. 内容分析 3. 引文分析和数学模型 4. 定律分布和论证等
八	同上	2001年6月16日～20日	澳大利亚悉尼	1. 科学领域的规律及分布 2. 信息计量规律的数学模型 3. 引证动机和科研评价 4. 知识地图及可视化 5. 科技政策的分析和预测 6. 图书馆管理等

（续表）

届数	会议名称	时间	地点	主题
九	同上	2003年8月25日～29日	中国北京	1. 信息计量规律的数学建模 2. 科研评价和大学排序方法论 3. 引文分析及数据库 4. 科技创新（专利）的定量分析 5. 网络信息检索研究等
十	同上	2005年7月24日～28日	瑞典斯德哥尔摩	1. 科学计量学的历史 2. 引用动机研究 3. 知识地图 4. 网络信息计量学 5. 科学政策分析和预测

时至今日，我们对信息计量学的研究还是较多地侧重在“狭义信息计量学”上，还带有较为浓厚的基于文献情报的文献计量学色彩。如前所述，信息计量学具有一定的交叉学科性质。可以预计，未来的信息计量学的发展将更加依靠现代计算技术和更深层次的数学分析。进一步探索信息的本质、信息计量的基本模式、信息计量的方法与技术、信息计量的模拟与验证、网络信息计量学将成为信息计量学下一步的发展方向。

这或许表明，作为一个年轻的学科，信息计量学依然处于方兴未艾的阶段。因此，愿意对这门学科的发展进行努力的人，是会大有可为的。

习　题

1. 什么是信息计量学？
2. 信息计量学的研究对象、研究目的分别是什么？
3. 信息计量的方法有哪些，它们的基本机理是什么？

第二章　信息计量中的熵描述

第一节　熵与信息、自信息

一、基本概念

在引入一些定义之前，首先澄清容易发生误解的一些概念是有好处的。

1. 主观与客观

例如，字母 a、n 和 t 可能组合成 tan、ant、nat，这些字有一定的含义，对每一个人的主观作用也是不一样的，是随着读者的主观反应而变的。一般来讲，这样所携带的主观信息是不相同的，而且不能加以定量的度量。信息理论不包含这种主观信息，最少目前还不能讨论这些问题。申农(Shannon)和韦弗(Weaver)曾将信息理论划分为三个层次：一是实用性层次，可将信息论应用于各个领域；二是实效性层次，研究信息产生与传输的实际效果与效率问题；三是意义性层次，研究信息的意义和对信息的理解。现在仅讨论第一层次的问题，其他问题留待以后发展。例如，心理学家研究并发现了在一种刺激中的信息量与对该刺激的反应之间存在很有趣的关系。如可以设置一个实验，包含四盏灯和四个相应的开关，这些灯以随机的顺序打开，要求实验者在灯亮之后尽快地按下相应的按钮。由此可以发现，反应所需的时间是随着灯所传达的信息量的增加而成线性地增加。实验结果在各种情况变动下仍然是相同的，如改变灯数和概率甚至相邻灯之间的相关量。这些结果揭示，在一定条件下，人们在操纵信息时，可随主观的愿望而使用信息理论中类似于代码和其他方法。这些问题在本书中不再涉及，本书只讨论客观的信息量度。

2. 信息与消息

这是两个截然不同的概念，不可混淆。消息是由符号、文字、数字或语音组成的序列。一份电报、一句话、一段文字和报纸上登载的新闻都是消息。消息中不确定的内容，构成信息。消息是信息的载体，信息是消息的内涵。例如，在小孩子未出生之前，是男是女尚不知道，如果这时医生告诉你一定生男孩，这条消息就很有信息。但在小孩子出生之后，已经知道是男孩，如果护士再来告诉

你生的是男孩，这条消息就一点信息也没有了。因为事情已经确定了，就不再有信息了。所以说，同样的消息所含的信息可能很不一样。这是不容混淆的，否则要出问题。如果把没有信息的消息当作是很大的信息，这是要上当的。昨天的报纸一点信息都没有了，因为报上的消息已经为大家所知道，人们往往把没有任何价值的情报、消息，视为宝贵难得的信息，上了当还不知道，就是因为不懂什么是信息、什么是消息的缘故。

3. **自然信息、社会信息和知识信息**

信息的来源是多种多样的，在第一章里已经讲过了。人们往往认为自然信息才是信息，而从社会中、知识中得来的信息不算是信息，这是不对的。当然，从不同的信息源得来的信息虽各有其特点，但都是信息，这一点却是一样的。电信号、光信号中含有信息，商品信息、经济信息、市场信息、考古信息也都含有一定的信息，在一定条件下，也都可以加以度量。

4. **信息与通讯**

无疑，信息与通讯有着密切的关系，通讯的目的就在于传送信息，但二者是有区别的。只能说通讯是信息理论的一种重要的应用，但不能说通讯就是信息，也不能说信息理论就是通讯理论。

二、自信息与熵

令 S 代表一组事件 $E_1, E_2, \cdots, E_n$，它们出现的概率分别为 $P(E_k) = p_k$，$0 \leqslant p_k \leqslant 1$并有：

$$p_1 + p_2 + \cdots + p_n = 1$$

现引入如下的定义：

定义 2-1 事件 E_k 的自信息记作 $I(E_k)$，并定义为：

$$I(K_k) = -\log p_k$$

在这定义中，对数的底没有加以指明，选择什么样的底对我们来说是无关紧要的，没有什么关系，因为底的改变仅仅变动了计量的尺度——单位。最常遇到的底是 2 和 e，当底选为 2 时，I 是以比特(bits)度量(bits 是 binary digits 的缩写，即二进制数字的缩写)；当底选为 e 时，I 的单位为纳特(nats，含有自然对数 natural logarithm 的意思)。nats 数是 bits 数的 0.693 倍。通常，对于底不做特殊的选择(除非要求大于 1 的情况)，当选用 2 为底时，写成 $\log_2$，或在适当的地方附加上字 bits。自然对数则总是写成 ln。

$$p_k = \frac{1}{2}, -\log_2 p_k = 1, I(E_k) = 1 \text{ bits}$$

因此，1 比特的信息是从两个相等而又相似的可能发生的事件中任选一个时所含的信息。如生男生女问题的信息，就是 1 比特。对数底为 2 特别适合于讨论二进制数字，因而在计算和编码的各种应用中是重要的，多采用这种度量。

底为 2 的对数是有表可查的，如果手边无表可用，可按如下关系进行计算：

$$\log_2 x = \frac{\log_{10} x}{\log_{10} 2} = \frac{\ln x}{\ln 2} = \ln x \log_2 e \tag{2-1}$$

一般情况：

$$\log_2 x = \frac{\ln x}{\ln 2} \tag{2-2}$$

因为上面已附加限制条件 $a > 1$，$\ln a > 0$，因而所遇到的对数总是自然对数乘以一个正数，这情况将在以后的分析中经常用到。p_k 愈小则 $I(E_k)$ 愈大，这是和我们的感觉相符合的。因为，一事件愈罕见，则其出现所带来的信息就愈多。

[例 2-1] 从英文字母中任意选取一个字母时给出的信息是多少呢？因有 26 个可能情况，取出一个字母的概率为 $\frac{1}{26}$，所以，

$$I = -\log_2 \frac{1}{26} = 4.7 \text{ bits}$$

是选择的这个字母所给出的信息。

[例 2-2] 设随机选择 m 个数字的二进制数。该数的每一位可从两个不同的数字中取一个，因此有 2^m 个等概率的可能组合。因此得：

$$I = -\log_2 \frac{1}{2^m} = m \text{ bits}$$

这样，需要 m 比特的信息来指明一个 m 二进制数字的序列。

[例 2-3] 64 个点被均匀排列于一个正方形格子里，令 E_j、F_k 是这样一个事件，是随意拾取落于 j 列 k 行的一个点。于是有：

$$P(E_j) = P(F_k) = \frac{1}{8}$$

$$I(E_j) = 3 \text{ bits}$$

$$I(F_k) = 3 \text{ bits}$$

这就告诉我们，落于 j 列的点给出 3 比特的信息，同样多的信息来自 k 行的点。同时在 j 列 k 行的点所给的知识意味着这个点是 64 个等同相似的可能情况中的一个，即：

$$I(E_j \cap F_k) = -\log_2 \frac{1}{64} = 6 \text{ bits}$$

因而得到：

$$I(E_j \cap F_k) = I(E_j) + I(E_k) \tag{2-3}$$

反映了这一事实，该点落于各行和落于各列是独立的事件。公式(2-3)对于[例2-3]不是特殊情况，对于统计独立情况都适用。假设 S 包括事件 $E_j \cap F_k$，$P(E_j) = p_j$，$(p_1 + p_2 + \cdots + p_n = 1)$ 和 $P(F_k) = q_k$，$(q_1 + q_2 + \cdots + q_n = 1)$。如果对全部 j 和 k 而言，E_j 和 F_k 是统计独立的，则根据统计独立的定义，如果事件 E_1 和 E_2 是统计独立的，则 $P(E_1 \cap E_2) = P(E_1)P(E_2)$，因此 $P(E_j \cap F_k) = p_j q_k$。

从定义2-1，得

$$I(E_j \cap F_k) = -\log p_j q_k = I(E_j) + I(F_k) \tag{2-4}$$

$$(j = 1, \cdots, n, k = 1, \cdots, m)$$

函数 f 具有与系统 S 中的事件 E_k 相对的数值 f_k，则 f 的期望值或平均值定义为：

$$E(f) = \sum_{k=1}^{n} p_k f_k$$

这一概念使得我们可以引出如下的定义。

定义 2-2 S 的熵，称为 $H(S)$，是自信息的统计平均值，即

$$H(S) = E(I) = -\sum_{k=1}^{n} p_k \log p_k$$

因为 p_k 可以为零，在此定义中 $p_k \log p_k$ 成为不定式，所以当 $p_k = 0$ 时，规定 $p_k \log p_k$ 等于零。

字母 H 是用来纪念波尔兹曼(Boltzmann)的，他是第一个给出这种类型(气体统计力学)的定义，并指定用 H 来表示。

业已指出，一个事件的自信息是随其不定程度(uncertainty)的增长而加大的。所以熵可被认为是一系统不定程度的度量，从而可导出并能证实这一观点的熵的性质。

首先，应指出 $\log p_k \leqslant 0$，当 $0 \leqslant p_k \leqslant 1$，因而 $H(S) \geqslant 0$，所以熵不可能是负值，但可以等于零。令 $p_1 = 1, p_2 = \cdots = p_n = 0$，按规定，当 $p_k = 0$ 时，$p_k \log p_k = 0$。于是在定义2-2中，从总和里移去所有 $p_k = 0$ 的各项，只余 $p_1 \log p_1$ 一项；又因 $\log 1 = 0$，所以该项也为零，因而在此情况下 $H(S) = 0$。这是必然的，这时