

51.731 1
GSL

信息论入门

H. 郭斯勒 著
张宗炳 译



科学普及出版社

总号: 032

信息論入門

Henry Quastler A Primer

On Information Theory

原著者:	H. 郭	斯	勒
譯者:	张	宗	炳
校者:	蔡	长	年
出版者:	科	学	普 及 出 版 社

(北京市西直門外靜家灣)

北京市书刊出版业营业許可証出字第 112 号

发行者:	新	华	书	店
印刷者:	北	京	市	印 刷 一 厂

开 本: 787 × 1092 1/32 印张: 2 3/8

1963年12月第 1 版 字数: 48,000

1963年12月第 1 次印刷 印数: 5,200

統一书号: 13051 • 018

定 价: 0.27 元



信 息 論 入 門

H. 郭 斯 勒 著
张 宗 炳 譯
蔡 长 年 校

科 学 普 及 出 版 社

一九六三年·北京

目 次

提綱.....	3
一 緒論.....	5
二 信息的表示.....	9
三 信息或不定度的測度.....	28
四 关于两个相关变量的信息度量.....	39
五 組織(系統、結構与构型)	55
附录一 信息量的估值.....	60
附录二 习题答案.....	69
参考文献.....	74

提 綱

一 緒論 信息論的发展史；信息論目前如此广泛流行的原因。通用系統理論；信息論的具体作用。

二 信息的表示 Paul Revere 的密碼；情报表示的主要特点。表示信息的可能性；各种不同方法；数据和运算；表示的有意識的和无意識的作用；广义的‘信息’。‘真实’和‘符号’事件；表示之前的抽象化。符号、符号(字母)表、‘字’；表示的单位。二元表示或符号化的标准方法；(1)最简单的情況：‘真实’类别的数目是二的整数幂，等长的‘字’；(2)任何数目的类别：不等长的‘字’，Fano的不引起混淆的密碼，符号数最少的密碼；(3)单‘字’所表示的事件組；(4)不等的概率：获得最少符号数的密碼的一般規則；(5)任意概率：在标准表示法中最少符号数的一般公式；(6)表示定理。习题。

三 信息或不定度的测度 获得的信息和消除的不定度。不定度的量是事件的概率的函数，不是它們的属性、起因和后果；信息量和表示性； H 函数；二进单位。Shannon-Wiener 信息函数的某些特性：独立性，連續性，相加性，天然的标度；概率为 0 和 1 时的值；平均及汇总的效应。习题。

四 关于两个相关变量的信息度量 ‘通信’的一般意义。两个相关变量的举例：父亲和女儿的身高；組合不定度，內部制約； T 函数；标度对于 H 及 T 的影响。一般的二部系統：两个变量的 6 个信息函数。通信系統：名称。噪声；身高的相关性作为有噪声通路的一例；通路容量；信息的处理不增加信息

量。錯誤的鉴定和校正：剩余信息；噪声通路的定理；校核錯誤的經濟性。实际的通信系統：信号和通路作为实际存在的实物。习题。

五 組織（系統、結構与构型） 組織，通信，剩余。系統的分析：信息分析，信息要求和性能；信息处理的一般限制，多部系統。結合。結論。

附录一 信息量的估值 典型困难。信息测度的相对性；确定实际值的选择的任意性。近似法。举例：会話中的信息传递速率；每一个印刷字母的信息量。

附录二 习题答案

一 緒 論

在信息論的文獻中似乎有一個缺陷。我們可以找到許多說明什麼是信息論的文章，一個沒有多少數學知識的讀者，也能看得懂其中某些文章；其實，有些文章的水平相當於一般通俗雜誌的水平。我們也能找到許多說明如何應用信息論的文章和書籍，但是所有這些都是十分專門的。我還不知道有任何既不太專門和嚴密而又充份明確和實用的文章，使讀者懂得應用信息論。這本書^①就是為了填補這一個缺陷，以滿足當前的需要。我準備將它寫成一本具有初級教科書的特點的書，並包括一些習題。所有的舉例，大部分取自通信工程和工程心理學，因為它們是解釋信息論最方便的方式，雖然不涉及有意識的通信也可以闡明全部理論。

信息論建立在這樣一個概念的基礎上，即：信息是可測的。這個概念並不新。在物理學中，關於信息与不定度（熵）間的可測關係的概念遠在1872年 Boltzmann 的著作中已經提出，而在1929年為 Szilard(1) 所發展。在1918年，統計學家 R. A. Fisher (2) 因為需要一個標準來估計實驗數據內的信息被某一給定的統計方法所利用的程度，作出了信息的測度；這個

① 原載“生物學中的信息論論文集”(“Symposium on Information Theory in Biology” Gatlinburg, Tennessee, October 29—31, 1956)——譯者註

測度直到現在還在統計學中應用。後來，由於電報電話的巨大發展，又產生了度量載荷信息的潛在能力的需要；R. V. L. Hartley (3) 在 1928 年發表了這樣的一個測度。1948 年，Wiener (4) 注意到信息量的測度是研究通信的一個基本要素，而通信的研究又是最廣義的、控制學研究的一個基本要素。在同一年，通信工程師 G. E. Shannon (5) 發表了一篇關於通信的數學理論的文章，這篇文章在幾個方面都超出了以往的研究範圍。這篇文章十分專門，很難讀懂。它發表在一個專門性的雜誌 (*The Bell Systems Technical Journal*) 上，並且除了電信以外不牽涉到其他方面。它肯定地不像一篇註定要在心理學家、語言學家、數學家、生物學家、經濟學家、美學家、歷史學家、物理學家，……中廣泛流行的文章；事實恰恰正是如此。在 1949 年，美國依利諾大學出版社出了一本書 (6)，其中重印了 Shannon 原來的著作，並有 Warren Weaver 的一篇文章。在後一篇文章內，十分強調信息量這個概念的一般性。從此以後，關於“信息”的文獻幾乎是以爆炸式的速度在增加着。

什麼理由使這樣一篇高度專門性的文章產生如此驚人的後果呢？當然，這篇文章很好是一個理由。另一個理由是，信息測度的概念滿足了當代的一個普遍而深切的要求。現在可供利用的巨大信息量正在高速度地增長。因此，信息的表示方法成為一個越來越重要的問題，而信息論提出了表示的一般原則。同時，我們也在發展那些越來越複雜的組織，它們的機能取決於成功而有效的通信聯繫；而信息論提供通信的普遍原則。

信息論是與一類專門技術有關的，這些技術或者是全新的，或者是如同信息論本身一樣，最近大大地增加了普及性。舉幾個例來說，如：操作分析、經驗設計理論、決策論、綫性規劃理論、控制論、博弈論、管理理論、群動力學及其他。

只要略看一下上述名单，就可以提出以下几点通則：

(1) 所有这些学科都是数学的学科，它們經常要涉及到概率論及統計学領域；

(2) 每一学科都应用了一套对某种东西計值的系統^①；

(3) 每一学科都涉及复杂的情况。在每种复杂情况內，总是有多重的可能选择，形成某种排列；在多数情况中，有很多相互作用的因素影响选择。因此，它們可以被称为系統理論；

(4) 在这些学科中，沒有一个学科是主要論及所研究的系統的物理性质或其各部分相互联系的机理的。

这是它們的共同特点，但是，上面所列举的每一項目都是一門專門的学科，因为每一个項目都涉及到系統的一个不同方面。究竟这些不同的学科，由于它們的相似性，能不能合并成一个共同的学科，称为通用系統学說，还是一个公开討論的問題(1954年还因此而成立了一个学会)。

因此，信息論只是几个系統学說中的一个。与其他相关的专业对比，表征它的特点的是对于一个事物（或一个条件或一个事件）的确定程度的度量，也就是它与其他可能事物（条件，事件）的差別在确定程度上的度量。通信与組織都是用相互規定的关系來說明的。一个說明信息論的本质的方法，就是把它与統計学来比較。統計学与信息論都是研究一个集的諸元素的差异性的；但是方法不同。統計学认为差异性是一种妨碍，因此企图撇开它来确定可以說明些什么或者做些什么。信息論把差异性作为一个有价值的东西，沒有了它，象选择、通信、表示及規定等这样的操作都成为不可能；信息論试图藉一定程度的

① 在某些情况下，这些值是能明显地确定的；在另一些情况下，它們只能由观察到的行为来推断。这并不一定意味着这些数值一定都是有意識地确定的，有目的的行为也可以在沒有任何有意識地确定数值的动作下发生。

差异性来确定可以获得些什么。因此，在信息論中所計算的‘信息’不同于日常所說的信息。例如，消息中的‘信息’（作为事物的一种形式）是在理想情况下，这种消息以符号为媒介所可能传送的知識（情报）在量上的測度。

許多人开始接触信息論时，觉得它大有前途；有些人在認真研究之后更认为它大有希望。信息論在工程界已經被确认。在心理学中，也已佔有了一定的地位。在生物学、經济学、政治学、美学、語言学界，許多人对信息論感兴趣，但是实际应用它的还只是少数。

二 信息的表示

“1 代表从陆路，2 代表从海路”。Paul Revere 和他的同胞們并不知道信息論，但是他們知道并且利用了表示情报的原则，后者就是信息論的基础，Paul Revere 的密碼虽然达不到現代的标准，但是它具备了密碼的主要性质：

(1) 有关敌人侵犯的来路的新聞被譯成另一种情报，即尖塔上的灯光；这个翻譯是有用的，因为它把一件很难广播的情报轉变为一件容易广播的情报。

(2) 所有可能发生的事件的范围，可以区分为有意义的几类。这个密碼当然还可以簡化为只說明敌人的来与不来。它也可以扩展为一个表示更准确的进犯方向、或表明敌人的力量及其他有意义的詳情的密碼。在这种情况下，就需要一个較复杂的密碼，而这样作会增加誤解的可能性。适当經濟的分类是信息表示中的一个重要特点。

(3) 这个表示法应用了一个預先約好的密碼。沒有灯表示沒有敌人进犯，一个灯是敌人从陆上来，两个灯是从海上来。似乎沒有約定敌人从陆上和海上同时进犯的情况，但是以‘3 个灯’表示的消息一定会被有关的人正确地了解。

信息表示的可能性

一种信息用另一种信息来表現的可能性是无限多的。(可以看出‘信息’一詞比‘情报’含有更广泛的意义；一般說来，‘情报’这个名詞只局限于有意識的信息。) 信息表示的唯一条件是

一套預先約好的翻譯規則——密碼。●這套密碼只是在區別被示的信息的能力、準確地產生一個所需的表示的能力和密碼的範圍方面受到限制。

表示并不只限于离散类的情報。信息或事件的状态的連續性可以用一個物理的連續性來表示，如一個範圍內的電壓或一根軸的旋轉。任何一種信息，無論是离散的或連續的，都可以用一個電子管內的電荷、金屬表面上一點的磁化程度、陰極射綫管的射綫的偏轉、落在感光乳膠上的光等等來表示(7)。

不但可以表示出數據，並且也可以表示出對數據的運算。在計算尺上，數目(一種情報)是用綫條上的標度來表示的，這是一種編碼方法。在作乘法時，把滑尺和游標定位，就是編碼；而在定尺上讀出答數，就是解碼。因此，計算尺是一個早期的、簡單的处理信息設備的例子。目前已經有了能接受和儲存大量有關數據和運算的符號信息的機器，並且它們能夠用這些信息進行大規模的運算。有足夠理由使我們相信，我們可以真正製造出能計算任何可計算的數字的機器；並且，推而廣之，它們可以達到任意一種思維過程的結果，而這種思維過程是可以明確而一定的運算來描述的。

按照上述情況，原始的和譯為密碼後的情報之間的關係是以相互對照為依據的。但是情報的翻譯并不局限於這種情況；它可以根据單方對密碼的選擇，即對於情報的接受者不預先說明。條件反射的情況就是如此：每次喂狗時响鈴；過了一些時候，‘鈴响’就成為表示‘食物已备好’這樣一個情報，因而導使狗準備進食。這個密碼是由實驗者建立的，並沒有和狗商量。事實上，信息的表示在任何情況下不一定要求有意識的理解。例如，調節人的呼吸及氧的吸入量的系統，其調節并不取決於血液本身的含氧量，而是取決於其中的二氧化碳的濃度。一般

說來，血液中的 CO_2 量就是 O_2 量的一个可靠的表示；但是在某些情況下，當這種表示不再是正確時，就可能產生困難。在所有這些情況中，正如在任意規定了密碼的情況中一樣，信息論所研究的是約束從一種信息譯為另一種信息的可能性的一般定律。現在，讀者已可明白，‘信息’這個名詞的專門意義比日常語言中的含義要廣得多。

‘真實’和‘符號’事件

現在讓我們來對於表示信息的原理進行更正式的更普遍的討論。這裡我們將只涉及表現為离散元的信息；但是，過渡到處理連續函數形式的信息並不困難。

在我們的討論中，‘真實的’和‘符號的’這兩個詞將代替‘被表示的東西’和‘表示的東西’。應該記住：‘真實的’及‘符號的’並不是指事物的性質而言，而是指它們在給定情況下的功能；‘符號的’這個詞可以、但不一定意味著真正發生了一個自覺的符號化動作。以 CO_2 的含量水平作為需要氧氣的符號，就足以說明這裡所用的‘符號的’一詞比慣常的意義更廣。

‘信息’不是一個脫離具體的事物；它總是和某些具體的荷帶者（一樣東西或一個事件）有關的。我們可以用任何一個字，只要適合於給定的情況就行；但是‘事件’是最常用的一般名稱。應該永遠記住，它可能由別的名詞代替；信息論對於所有的信息載荷者都一樣適用。用形式語言來說，信息的荷載者是指話語的單元，或樣空間中的點或性質的構形 (configuration of properties)。

一個具體事件，由於它的細節的豐富性，是不可能完全表示出來的。個別的人在一瞬間的唯一完全的代表，就是在那個瞬間的本人。用符號能表示的只是具體事件的某些方面；例如這

个人在这时间属于工学院三年級、平均等級为 4.32 的男学生的一类。这样的信息本身就是可以表示的，例如，以 Hollerith 卡片上某种孔的位置来表示。一般說来，当我们談到了表示某一事件时，我們并不是說表示了具体事件的全部个性，而只是把它們的抽象作为一类事件的范例表示出来。用形式語言來說，一个事件的有信息意义的方面是它的‘所属类組’或是它的‘所属点集’的名称。

因此，表示信息的头几步是：（1）决定把什么作为信息的基本荷載者或基本事件；（2）确定这些事件的哪些特征与我們的考虑是有关系的；（3）把相应于各个特征或它們的組合的各类事件全部列出。理想的情况是，对实际情况的分析在用符号表示开始之前已經完成；实际上，通常比較方便的是根据一个不完全的分析定一个暫时的表示系統，然后在必要时再把它精細化或作些附加。图书馆內按题目的分类书目，就是一个有伸缩性的和允許发展的表示系統的例子。

符号、符号(字母)表、‘字’^①

符号化的基本单位称为符号。全部可用的符号組成符号表。在最簡單的表示方法中，每一个別事件譯成单一符号，这个符号就代表該事件的某一类型。这是可以做到的，只要字母表大到足以在将真实事件分类时，可以給予每一类事件一个不同的符号。Paul Revere 的密碼就是一个例子。

这种一对一的表示法缺少灵活性，并且十分煩瑣。經常我們只用一个小的符号表表示可能性很多的情况。在这情况下，单一事件須用几个符号的組合来表示；并称这个組合为碼組或

^① 这里和以后加引号的‘字’不是指普通意义的字，而是指几个符号組成的組合(或碼組)——譯者註。

是‘字’。例如，26个字母組成的符号(字母)表可以用来表示几十万个英文字，并且有足够的灵活性来安排任何数目的新字；这只要平均每一个字用4.5个字母就可以做到(这里字母还并没有最经济地加以利用)。这样，一个只有两个符号(Morse 电碼的点和划)的符号表就足以表示全部26个字母，以及10个数碼、标点符号及几个标准概念，而在每一碼組或‘字’(按上述定义，不是普通意义)中的符号从来不多于6个。根据 Gamow 和 Yčas 的著作，蛋白质中20个氨基酸的每一个，可以在 RNA 分子中用3个核苷酸組成的‘字’来表示；在这种情况下，‘字’内符号的排列次序被认为是沒有关系的。

二元表示

最简单的情形 为了发展表示信息的一般理論，把所有的表示都化为某种标准形式较为方便。任何一种标准形式都是可行的，但是习惯上采用最简单的符号表，即二元符号表。常用的两个符号是1和0。必須着重指出，0并不一定意味着沒有任何实际动作。例如，1和0可能代表左和右，正和負，点和划等等。任何事件在标准符号化之后，将是一个二进位数，如100, 101……，这里每一位数碼和数碼的組合的符号意义都由一些相互关系的定律所决定。

必須指出，如果考虑的是整个消息，Morse 碼便不是严格的二元表示法。Morse 碼实际上是一个四元碼。这是因为，除了点和划两个符号(‘黑’的)之外，还有两个不同长度的空白符号(‘白’的)，即字母間的空号和字間的空号。这两个符号都是这一編碼制度中的主要組成部分，否則，我們就不知道

.....—

究竟是指‘hen’，‘sue’，‘sin’，还是指‘site’。

用一定数目的二元符号所組成的‘字’可以表示多少事件呢？一个符号只能組成两个不同的‘字’（1 和 0）；它們可以代表把一組真实事件分为两类的情况。用两个符号可以組成 4 个不同的‘字’（11, 10, 01, 00）。一般說来， n 个二元符号可以組成 2^n 个碼組。因此：

由 n 个二元实验所組成的序列可以区分出 2^n 种可能性；

由 n 个二元选择所組成的序列可以在 2^n 个不同的选择中选出任何一个；

由 n 个二元說明所組成的序列可以在 2^n 个項目中任意指出一項；等等。

相反地，如果具有 r 个不同表示的密碼本是用标准二元碼来編制的，那么每一个‘字’将大約是具有 $\log_2 r$ 位的二进制数^①。例如，事件分为八类时，可以用含有 3 个二元符号的碼組来表示（ $3 = \log_2 8$ ）：

类别 A——111

类别 B——110

类别 C——101

类别 D——100

类别 E——011

类别 F——010

类别 G——001

类别 H——000

可以看到，每个二元符号的意义取决于它的位置和字中其他符号的性质，例如，第二位的 1 可以是 A、B、E 或 F；如果它的前面是 0，那么它只能是 E 或 F；如果它后面是一个 1，那么它

① 以 2 为底的对数可从已出版的对数表中查出；或者可以从一个具有对数-对数尺的計算尺上讀出，或者把以 10 为底的对数乘上 3.322 得出。

就毫无疑问地表示 E 了。这就暗示了符号‘字’中的每位数碼，就其本身来考虑，并不一定表示一个操作。例如，一套識別操作（如生物学家的检索表）可以排为一个二元檢驗序列。在这样的情况下，人們并不总是应用同一檢驗序列；在一般的情况下，第二个檢驗决定于第一个檢驗的結果，而第三个檢驗又决定于前面两个的結果等等；因此，密碼本必須規定一个特定符号所指的是那一种操作及那一种結果。

碼‘字’的序列将代表一串事件。設消息表示事件序列‘GCA’。用上述密碼，即得

001101111.

注意在这消息中的‘字’并没有用空号来分开。假如加一个具有一定符号意义的空号（如 Morse 碼中的空白符号），就使符号表成为三元的，不再是二元的了。受信者不会遺漏空号；預計他知道这一密碼，因而，他会把这个消息自左端第一符号起讀成三位一組。

任何类别数 通常要編碼的类别数目，不会正好都是 2 的整数幂（如 2、4、8、16、32……）。但是我們总是可以用 2 的較高次（最接近的）幂来作为編碼計劃的基础。例如，有 5 个类别要表示，我們可以简单地用 8 个类别的三位碼的一部分，

类别 A——111

类别 B——110

类别 C——101

类别 D——100

类别 E——011

在这一例子中，沒有充分利用三位符号的所有可能組合。这是不經濟的；人們将怀疑在位数上是否可以更經濟些。这只能意