第3章

数据思维原理：信息学视角

数据时代，“数据思维”作为“数据密集型科学发现第四范式”的思维方式，需要一些普遍意义的、基础性的思想和原则来主导其在各领域的应用和发展，从而成为其应用的基本原理或逻辑起点。数据思维的基本原理由数据产生及应用过程的特征和规律决定，是其他思维不具备的性质，它又反过来影响数据的产生、处理和利用过程，从而对数据思维产生重大影响。对这些基本原理进行探讨与学习，可以帮助我们更好地理解与运用数据思维来解决问题。

本章将从信息学视角出发，来探讨数据思维的基本原理——最大熵原理、最小努力原理、信息生命周期理论、对数透视现象和小世界现象，帮助读者更加深入地理解数据思维。

3.1 最大熵原理

3.1.1 熵及信息熵的概念

熵在不同的学科中有着不同的表达方式。热力学中，对于每一个热力学平衡状态，都有状态函数——熵(S)：从一个状态 O 到另一个状态 A，S 的变化定义为：

$$S - S_o = \int_O^A \frac{\mathrm{d}Q}{T}$$

其中，$\mathrm{d}Q$ 为流入系统的热量，T 为热力学温度，积分可以沿着连接 O 与 A 的任意可逆变化过程进行。S_o 为一常数，对应于状态 O 的熵值。通常取 O 为绝对零度状态，此状态 $S_o = 0$。对于一个给定的孤立系统，任何变化不可能导致熵的总值减少，即 $\mathrm{d}S \geqslant 0$——这就是利用熵概念表述的热力学第二定律。通过玻尔兹曼关系式 $S = k \cdot \ln W$（k 为玻尔兹曼常数，W 为与某一宏观状态所对应的微观状态数），宏观量熵 S 与微观状态数 W 联系起来，明确表达出了熵函数的微观意义（统计解释），解释了熵的本质：熵代表了一个系统的混乱程度。

1948年，美国电气工程师香农在其《通信的数学原理》一书中首次提出了**“信息熵”**的概念，把熵作为一个随机事件的**“不确定性”**或**信息量**的量度，从而奠定了现代信息论的科学理论基础，大大促进了信息论的发展。

信息量是信息论的中心概念。信息论量度信息的基本出发点是把获得的信息看作用以消除不确定性的东西。因此信息数量的多少，可以用被消除的不确定性的大小来表示，而随机事件不确定性的大小可以用其概率分布函数来描述。

考虑一个随机实验 A，设它有 n 个可能的(独立的)结果：$a_1, a_2, \cdots, a_n$；每一个结果出现的概率分别是：$p_1, p_2, \cdots, p_n$；它们满足以下条件：

$$0 \leqslant p_i \leqslant 1 \quad (i=1,2,\cdots,n) \quad 及 \quad \sum_{i=1}^{n} p_i = 1$$

对于随机事件，其主要性质是：对它们的出现与否没有完全把握，当进行和这些事件有关的多次实验时，它们的出现与否具有一定的不确定性。随机实验先验地含有的这一不确定性，本质上是和该实验可能结果的分布概率有关的。为了量度概率实验先验地含有的不确定性，引入了函数：

$$S = S(p_1, p_2, \cdots, p_n) = -K \sum_{i=1}^{n} p_i \ln p_i$$

作为随机实验 A 的结果不确定性的量度。式中，K 是一个与度量单位有关的正常数，因此 $S \geqslant 0$。上式中的 S 称为信息熵或 Shannon 熵。

它具有这样的意义：**在实验开始之前，它是实验结果不确定性的度量；在实验完成之后，它是从该实验中所得到的信息量**。S 越大，表示实验结果的不确定性越大，实验结束后，从中得到的信息量也越大。

例如，设在实验 A 中，某个 $p_i = 1$，而其余的都等于 0，则 $S = 0$，这时可以对实验结果做出确定性的预言，而不存在任何的不确定性；反之，如果事先对实验结果一无所知，则所有的 p_i 都相等($p_i = 1/n, i = 1,2,\cdots,n$)，这时 S 达到极大值 $S_{\max} = K \ln n$。很明显，在这一情况下，实验结果具有最大的不确定性，实验结束后，从中得到的信息量也最大。

基于此，**可以类推对数据所包含的“信息”或“价值”的度量方法**。

3.1.2 最大熵原理的内涵

Shannon 很好地解决了关于不确定性的度量问题，但没有解决如何进行概率分配的问题。后一个问题是由 Jaynes(杰恩斯)解决的。

设想有一个可观测的概率过程，其中的随机变量 X 取离散值 $x_1, x_2, \cdots, x_n$。如果从观测的结果知道了这个随机变量的均值与方差等值，怎样才能确定它取各离散值的概率 $p_1, p_2, \cdots, p_n$ 呢？一般地，满足可观测值的概率分配，可以有无限多组。那么究竟应该选哪一组呢？即在什么意义下，所选出的一组概率才是最可能接近实际的呢？

Jaynes 在《信息论与统计力学》一文中提出一个准则：“在根据部分信息进行推理时，我们应使用的概率分布，必须是在服从所有已知观测数据的前提下，使熵函数取得最大值的那个概率分布。这是我们能够做出的仅有的无偏分配。使用其他任何分布，则相当于对我们未知的信息做了任意性的假设。”这一理论称为**最大熵原理**，它为我们如何从满足约束条件

的诸多相容分布中，挑选“最佳”“最合理”的分布提供了一个选择标准。尽管这个准则在性质上也有主观的一面，但却是一个最“客观”的主观准则。因此，由此得出的估计，人为偏差最小。

在数学上，把最大熵原理表示为如下的优化问题。

$$(E_0)\begin{cases}\max\limits_{p} \quad S=-k\sum\limits_{i=1}^{0}p_i\ln p_i & (1)\\ \text{s.t.} \quad \sum\limits_{i=1}^{0}p_i g_i(x_i)=E[g_i(x)] \quad j=1,2,\cdots,m & (2)\\ \sum\limits_{i=1}^{0}p_i=1 & (3)\\ p_i\geqslant 0 \qquad\qquad i=1,2,\cdots,n & (4)\end{cases}$$

其中，$g_i(x)$代表可观测的函数值，$E[g_i(x)]$代表相应的均值。

Tribus 曾经证明，正态分布、伽玛分布及指数分布等都是最大熵原理的特殊情况。例如，在知道均值与方差的情况下，求解问题 E_0 得到正态分布。这就是说，正态分布包含与观测量一致的最大的不确定性，即含有最大的熵。如果对一个随机过程，任何可观测量也得不到时，则约束(2)不再存在，由问题 E_0 得到的解是一个均匀分布，这与人们的直观认识是相等的。

3.1.3　最大熵原理的应用

最大熵原理和方法的应用范围非常广泛，目前它已经渗透到信息论、工程优化、气象学、热力学、统计力、天文学、生物学、社会学、管理学、经济学等各个领域，在学科交叉和结合中起到了桥梁和纽带的作用。下面以统计力学为例，来讲解最大熵原理的应用。

在统计力学中，研究由大量分子组成的系统，设系统可以处于编号为 $1,2,\cdots,i,\cdots$的微观状态，并设 $P_i(i=1,2,\cdots)$是发现系统处于状态 i 的概率，每一种概率分布都对应着一个熵值，这个熵值反映了其内部分子热运动的不确定性。

对于处在给定宏观条件下的系统，它按微观状态的分布概率是与某些约束条件有关的，这些约束条件就是某些给定的宏观物理量，而宏观量是相应微观量(随机变量)的统计平均值。当然，当一个或几个随机变量的平均值给定时，还可以有许多的概率分布与这些平均值相容，问题是如何从这些相容的分布中挑选出“最佳”的分布作为系统处于平衡时的最常见分布。根据最大熵原理，最佳分布应是具有最大熵的分布，由此可以求出处于各种宏观条件下的系统按其微观状态的分布概率。下面以正则系统为例。

正则系统的研究对象是与大热源相接触而达到平衡态的系统。由于与热源相互作用，系统的能量是可变的(随机变量)，但其温度受热源控制，所以系统的平均能量是给定的，设系统的微观状态由能量 E_i 确定，E_i 还可以是某一参数 y 的函数：$E_i=E_i(y)$。分布概率满足的约束条件是：

$$\sum_i p_i=1 \qquad \text{(归一化条件)} \tag{5}$$

$$\sum_i E_i p_i = U \quad (\text{给定的恒量}) \tag{6}$$

系统的信息熵为

$$S = -K \sum_i p_i \ln p_i$$

在约束条件(5)和(6)下,引入 Lagrange 乘子 α 和 β,由熵的极值条件可得分布概率为:

$$p_i = \frac{1}{Z} e^{-\beta E_i}$$

上式就是吉布斯正则分布,式中,Z 为正则配分函数 $Z = Z(\beta, y) = \sum_i e^{-\beta E_i}$。

最大熵原理可以用来解决随机性或不确定性问题。**应用其解决问题的思路是**:先将所研究的问题转换为一个概率模型。这样,问题的随机性就表现为概率分布(每种概率分布对应一个熵值,熵值的大小就表示了不确定性的大小或状态的丰富程度),问题的解决就归结为求一种最佳的概率分布,然后采用最大熵原理求出最佳分布。**由此得到启发**:凡是带有随机性的问题(不论是哪一领域的),都可以尝试用最大熵的方法加以解决。这就为一些优化、决策、预测问题的解决提供了新的途径和方法。

大数据时代,事物状态的量化数据更为丰富,应用最大熵原理的场景也更加广泛。

3.2 最小努力原理

3.2.1 最小努力原理的内涵

最小努力原理是人类生态的基本规律之一,它体现在人类社会的各个方面。心理学家们对人类行为进行了大量深入的研究,其研究结果对于人类行为的控制提供了方法论的指导。

学术界通常认为,美国哈佛大学教授、著名语言学家和心理学家乔治·金斯利·齐夫(George K. Zipf,1902—1950)在研究自然语言词汇使用时提出了“最小努力原理”(Principle of Least Effort)。1948 年 4 月,46 岁的齐夫完成了他的专著《人类行为与最小努力原则——人类生态学引论》(*Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*,见图 3-1)。这部著作影响很大,有些学者用了一句古老的拉丁语称它为 Magnumopus(巨著、杰作)。齐夫在书中引用了大量的统计数据,对“最小努力原理”做了精辟的阐述。

齐夫认为,每一个人在日常生活中都必定要在他所处的环境里进行一定程度的运动,他把这样的运动视为走某种道路。然而,人们在自己的环境里所走的道路并非就是他的全部活动。对于一个处于相对静止状态的人来说,他要完成新陈代谢,就要有连续不断的物质和能的运动,进入、通过和输出他的系统,这个物质和能的运动也是在一定的道路上进行的。的确,人的全部机体都可以视为物质的聚合,正以不同的速度在不同的道路上穿过人的系统。而人的系统继而又作为一个整体在他的外部环境的道路上运动。齐夫强调道路和运动的概念,是要证明每一个人的运动,不管属于哪种类型,都将是在一定的道路上进行的,而且都将受到一个简单的基本原则的制约。齐夫把这样一个简单的基本原则称为“最小努力原

理”。在这一原则的制约下，人们力图把他们可能做出的平均工作消耗最小化，即人类行为建立在最小努力原理的基础上。

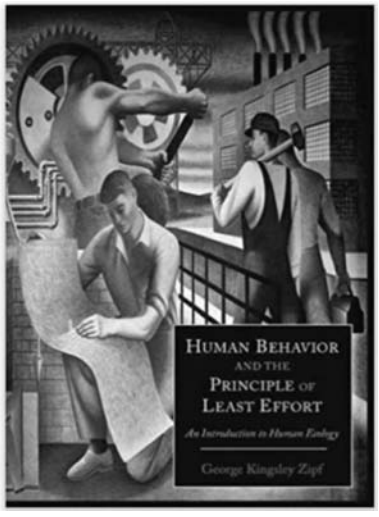

图 3-1　*Human Behavior and the Principle of Least Effort*

怎样理解齐夫的“最小努力原理”？下面举个简单的例子。一个人在解决他面前的问题时，要把这个问题放到他所考虑到的将来还会出现的问题的整体背景中去考虑。这样，当他着手解决这个问题时，就会想方设法寻求一种途径，以至于把解决面前的问题和将来可能出现的问题所要付出的全部工作最少化。也就是说，他要把他**可能会付出的平均工作消耗最小化**。做到了这一点，就可以把他的努力最小化了。这就是“最小努力原理”，是指一个人努力把他可能的平均劳动支出额降低到最低限度。该理论认为，人们的各种社会活动均受此原则支配，总想以最小的代价获得最大的效益。注意，这里的努力的概念，是齐夫专门定义的努力，所谓最小努力是最少工作的变种。

3.2.2　最小努力原理的应用

1. 齐夫定律及应用

人类交流、获取和利用信息（Information）、知识（Knowledge）、情报（Intelligence）总是趋向简捷、方便、易用、省力。研究和揭示人类情报行为追求易用与省力的特征、规律可以使情报获取和情报服务的成本最小，效益最大。这里的情报与信息具有不同的含义，是为实现主体某种特定目的，有意识地对有关的事实、数据、信息、知识等要素进行劳动加工的产物。目的性、意识性、附属性和劳动加工性是情报最基本的属性，它们相互联系、缺一不可，情报的其他特性则都是这些基本属性的衍生物。

齐夫在“最小努力原理”思想的指导下，首先对语言学进行了研究，因为这是人类活动很重要的一个方面。按齐夫的说法，当我们用语言表达思想时，就像受到两个方向相反的力的作用——“单一化的力”和“多样化的力”的作用。这两种力的作用表现在人的谈话或写文章时，一方希望尽量简短，另一方面又希望尽量详尽。如从这一观点出发，说话者以只用一个词表达概念最为省力，而听话者以每个概念都能用一个词表达理解起来最为省力。“单一化的力”和“多样化的力”相互作用，取得平衡，使自然语言的词汇出现频次呈双曲线。在现实生活中，人们在读写时越来越多地使用缩略词便是信息交流追求省力的一个很好的例证。

齐夫通过对较长文章中的词进行统计，发现词在一篇文章中的出现次数（频率）按照递减的顺序排列起来（高频词在先，低频词在后），并用自然数从小到大给词频的倒顺序给予等级（高频词等级值小，低频词等级值大），就会发现，等级值和频率值相乘是一个常数，即：

$$fr = c$$

（f 表示词在文章中出现的频数，r 表示词的等级序号，c 为常数）

当然，这里的常数并不是绝对不变的恒等值，而是围绕一个中心数值上下波动的，有时相差很大。

齐夫揭示的这种词频规律，后来被人们称为“齐夫定律”。不过，当时齐夫本人是按“最小努力原理”去解释的。其理由是，在任何语言中，凡是高频使用的词，功能总是不会太大，即词意本身在这个场合中价值小，因而传递它们所需要的努力就不大。

长期以来，**齐夫定律被视为文献计量学的基本规律之一被广泛应用**。研究证实，齐夫定律不仅适用于自然语言，而且适用于人工语言，因而又被应用于情报的组织、存储和检索领域。例如，怎样进行词汇控制、编制什么规模的词表、选用多少词、根据什么选词都必然涉及齐夫定律。学者们按照齐夫的词频分布方法，通过标引实验，找出被标引文献与叙词使用频率的分布特征，最后确定符合使用频率的词，编入词表，再不断根据标引实践反馈修改，使词表既满足实用，又不致规模过大。在自动分类（Automatic Classification）和标引（Indexing）中，频率太高的词和频率太低的词因其在检索中的价值不大都不能用于标引或词表编制，也需要通过对词频进行统计分析，筛选出适于标引的词，或者与一个特定的分类系统比较，进行分类处理。在情报组织中，不同属性的字段（著者字段、篇名字段、主题字段等）都是由词组成的，为了控制一个倒排档的大小，就要考虑倒排档中每一个词在不同记录中出现的次数，加以统计排序，选出最适合的词，将倒排档控制在对信息组织和用户来说都是“最省力”的规模。

齐夫定律所描述的省力法则虽然发源于语言应用领域，但他实际上注意到了这一法则更为一般的意义，各个不同领域中最短路径（Shortest Path）的选择和确定问题都与此有关。例如，企业供应商和库存地点的选择、社区供货点的位置、交通路线安排、通信路线架设等，都涉及最短路径寻求的解决方法。就是在图书情报领域，齐夫定律也不仅应用于语言文字有关的问题，也涉及最短路径的选择问题。已有学者将其应用于设计图书馆、文献中心资料库的排架，以使得资料出纳员在存取资料时所走路径最短。这方面最富想象力和创意的就是提出图书馆或文献中心不按传统的分类排架，而按资料使用次数多少以出纳员为中心按辐射状排架，使用频率最高的资料离出纳员最近，使用频率较少的资料则放到离出纳员较远的位置。齐夫定律还可以帮助我们合理地选择公共图书馆和情报中心的地点位置，使得各类用户能方便到达。现代运筹学已经介入这些问题的研究，并取得了很好的成果，使得齐夫描述的省力法则，可能从经验观察统计上升到严密的科学抽象。

2. 穆尔斯定律及应用

“最小努力原理”一出现，国内外的学者们，尤其是西方的学者纷纷对此进行研究。有的提出了异议，有的则运用于实践，涉及包括信息科学在内的很多领域。在信息研究领域，可以将“最小努力原理”理解为：**人类总是通过信息进行交流，并千方百计采取简单、方便、快捷、易用的手段来获取和利用信息、知识和情报**。

在研究信息用户与用户信息需求过程中，信息的易用性是一个重要因素。为此，美国著名情报专家穆尔斯（Calvin N. Mooers，1919—1994 年）在研究用户利用情报检索系统时概括出一条定律：**“一个情报检索系统，如果用户从它取得情报比不取得情报更伤脑筋和麻烦的话，这个系统就不会得到利用。”**这就是著名的穆尔斯定律。该定律揭示了用户利用情报检索系统的规律。导致用户这种选择规律的原因在于：

（1）使用情报资料首先必须获得情报资料，这是情报资料的有效性的前提和基础。

（2）情报资料的有效性对用户来说是一个模糊的概念，用户不可能确定一个关于有效性的明确的检索目标，检索结束以后也无法断定是否检索到最有用的资料。

（3）用户存在取得的资料总有用的观念。虽然经常发生误检，用户的这种观念却不易改变。

用户对信息的选择几乎都是建立在易于存取、易于利用基础之上的，最便于存取的信息源（或渠道）首先被选用，对质量的要求则是第二位的。可以说，穆尔斯定律完全是依据最小努力原理演化出来的。

许多典型的信息用户调查表明，用户对信息源的选择几乎都是建立在易用性的前提下，便于接近的信息源，将被优先利用，而该信息的质量乃至可靠性却成了次要问题。1996 年，艾伦·福斯特（Allen Foster）建立了寻找情报行为的模型。该模型主要有以下几点结论。

（1）易用性是用户利用情报的一个决定性因素。

（2）易用性和明显的技术质量影响着用户对情报源的选择。

（3）用户对易用性的认识受经验影响，用户熟悉的情报渠道，易用性比较强。

（4）对一份情报是利用还是谢绝，在易用性后考虑的是质量标准。如果质量上不符合要求，易用性就要贬值。由此可见，情报质量可能而且势必反作用于易用性的评价和看法。

显然，最小努力原则在艾伦模型中得到了集中体现。

1972 年，索普（M. E. Soper）的研究与艾伦的结论也并行不悖。索普利用引文分析法逐一检查、分析了自然科学与社会科学中的每条引文，弄清受引论文位于何处。根据调查结果，受引论文的位置不外乎是：个人藏书、服务单位馆藏、所在单位馆藏、所在城市藏书、外地藏书等。他们采用信函方法，得到 178 件回函，包括受引论文 5175 篇。其中，用户使用的情报资料中有 57％来源于个人信息库，大约 26％来自其单位的图书情报中心，大约 10％来自地理上较难存取的图书情报中心。用户搜寻情报的过程常常是，首先从自己已有的资料中查找，然后转向非正式渠道，取得同行帮助，在用尽了这些办法还不能解决问题之后，才考虑求助于图书馆或情报中心。

1975 年，美国建筑师沃尔曼（Richard S. Wurman，1935—）提出“信息构建”（Information Architecture）这一概念，强调清晰、美观和易用。随后很快将此概念用到互联网上，人们更强调信息构建的可用性，例如，可记忆、可学习、可靠、有效、满意，并将信息构建定义为“组织信息和设计信息环境、信息空间或信息体系结构，以满足需求者的信息需求的一门艺术和科学”。信息构建的核心要素是信息组织系统、信息标识系统、信息导航系统、信息搜索系统。现在，人们又在知识构建上努力，使人们以最小的努力获取、利用信息、知识和情报。

3.3　对数透视现象

3.3.1　对数透视现象的内涵

人类获取和接收信息、知识和情报的认知过程遵循对数转换机制。研究这一转换机制可以揭示物理空间的信息与进入认识空间中的信息、知识和情报之间，信息载体和信息内容

之间在数量和特征上的差异，为情报、情报学的定量化提供理论、方法和途径。这一原理看起来似乎很神秘，实际上是普遍存在的人类感官系统对外界物理刺激的反应机制，它描述物理空间的对象特征在人的感觉系统中的影像之间的差异符合对数转换律。

对数透视原理实质上源于实验心理学中的韦伯-费希纳定律。19 世纪中叶，德国著名心理物理学家韦伯(E. H. Webor)和费希纳(G. T. Fechner)通过实验验证后提出心理量是刺激量的对数函数，具体公式可表示为：

$$S = k \lg R$$

其中，S 是由外部物理刺激引起的人的感觉量质，R 是物理刺激量，如声音高低、光的强弱、颜色深浅等，k 是常数。

这个定律说明在人类运用感官系统或者神经系统进行认识的过程里，**人的一切感觉，包括视觉、听觉、味觉等，都与对应物理量的强度的常用对数成正比**，而非与对应物理量的强度成正比。

受韦伯-费希纳定律的启发，20 世纪 70 年代末，贝特拉姆·C·布鲁克斯(Bertram C. Brookes，1910—1991)在研究人的信息获取和吸收过程时，引入韦伯-费希纳定律，并进行了大胆的拓展。他写道："如果我们的感官系统按照对数规则工作，那么我们所有的神经系统，包括脑神经系统，都可能按某种对数方式工作。"布鲁克斯称之为"**对数透视原理**"(**Logrithmic Perspective**)，**即对象的观察长度 Z 与从观察者到被观察对象之间的物理距离 X 成反比**，并提出了 $Z = \log X$ 的对数假说，在一定程度上较好地说明了知识、信息传递中随时间、空间、学科(领域)的不同而呈现的对数变换。

同时，布鲁克斯假设了一个抽象的信息空间，其潜在的情报均匀分布在其物理线的线性尺度上。令潜在的情报密度为 ρ，观察者处于该空间，并感知过程受到对数透视原理的影响。观察者在物理线上由 a 到 $a+n$，长度为 n 的区间里，该区间的感知信息量为：

$$I_1 = \rho[\log(a+n) - \log a]$$

上式为一维物理线的情报容量计算公式，同理，二维平面、三维空间的感知信息量计算公式分别为 $I_2 = 2\pi\rho[\log(a+n) - \log a]$；$I_3 = 4\pi\rho[\log(a+n) - \log a]$。

事实上，布鲁克斯提出的对数透视原理是建立在一系列假设前提之上的，归纳起来有：

(1) 宏观上信息空间密度均匀，微观上每篇文献包含的知识、信息量相等。

(2) 知识、信息获取者的接受能力相同。

(3) 知识、信息的获取没有其他辅助工具或技术的支持。

(4) 知识的继承性好。

很明显，这些假设都是在当时的信息环境下提出的，属于对数透视原理的经典理论，它很好地解释了传统信息环境下人们信息行为的现象和规律。但是，当前信息网络的出现与信息环境的变化会导致这些假设前提的改变，但传统对数透视原理在解释网络环境下的信息现象和规律方面仍然具有一定的客观性与普适性，而且会有新的表现形式和发展趋势。

简言之，**对数透视原理解释了人们遵循最小努力原理进行信息、知识、情报的获取和吸收这一现象，即时间上寻求最新、空间上寻求最近、学科(领域)上寻求从自己最擅长和最熟悉的领域来查询并获取知识和信息**。产生对数透视的根本原因在于信息的功利性。一般来说，人们最关心、最重视的是与自己的切身利益有关的信息。因此，如果时间上、空间上、学

科知识方面或是经济利益方面，距 R 越近、关系越密切，被重视的可能性就越大，信息的表现感觉也就越高。

3.3.2　网络环境下的对数透视现象

1. 空间对数透视现象

空间对数透视描述信息流经由一定空间到达信息接收者 R 后，R 对信息产生的主观感觉。R 在接收信息时一般只关心来自最近的信息源的信息。布鲁克斯的对数假说同样适用于社会信息空间。在社会信息空间中，社会信息流流动的距离不仅是物理空间中的物理距离，而且还包括随信息流从 S 流至 R 所经历的信息栈的多少。根据对数透视原理，在获取信息时，人们平衡物理距离的远近和获取信息的难易程度，来选择一个最佳路径。

传统的空间对数透视原理揭示了在信息空间无论是物理空间还是社会信息空间，人们都对较近的信息给予了较大的关注。一般情况下，人们不会接收较远的信息，其根本原因是功利性决定了人们在获取信息时会遵循最低消耗原则。获取较近信息的比例较高，获取较远信息的比例较低。这个原理是建立在上述提到的 4 个假设条件之上的。但在现代网络环境下，信息和信息环境都发生了很大的变化。**归纳起来，主要有以下一些变化：**

(1) 信息的不同。在传统的信息环境下，信息的更新换代较快，信息呈指数增长，而传统媒体环境则相对成熟，使人们对信息质量有了一定的要求和规范；而在网络环境下，网络信息的数量是急剧增长的，网络的开放性和不规范性，又使信息的质量很难得到保证。

(2) 信息交流条件的不同。在传统信息环境下，人们之间的交流方式受到空间上的限制，随着现代通信技术和网络技术的发展，人与人之间的交流变得越来越便利，越来越高效。越来越多的人通过网络联系，如电子邮箱、MSN 等。交流便利的同时，信息获取和信息共享也变得更加方便，越来越多的信息可以直接通过网络传播。

(3) 信息获取方式的不同。在传统信息环境下，人们获取信息的渠道是有限的，而且效率较低。在网络环境下，获取信息的途径越来越多，如搜索引擎、网络数据库等。而且随着技术的不断进步，搜索工具的检索效率会越来越高。

(4) 信息接收者的能力不同。社会的不断发展也促使信息接收者的能力不断地发生变化，现代社会人在接收信息时具有与较早时期的人不同的特点。

在传统条件下，人们获取信息主要是通过报纸、杂志等纸质载体，想要了解某个知识需要去书店购买或到图书馆借阅某方面的书籍。因此，物理距离的远近是人们选择获取信息渠道的条件之一。例如，对于同样的书籍，人们会选择路程最近的本单位图书馆去借阅，而不会浪费时间去其他远距离的图书馆。而在今天，由于计算机和互联网日新月异的发展，使人与人之间的沟通变得更为便捷，各种各样的信息充斥在网络中，人们足不出户就可以遨游信息的海洋。人与信息源的物理距离已经不再是人们选择信息源的先决条件。在网络条件下，空间的距离可以表现为获得某信息的超级链接的多少以及在搜索结果中的排名等。

根据最小努力原理，通过越多的链接才能获取到的信息，被需求者关注到的概率就越小。也就是说，获取到某信息所经过的超级链接的多少与它被获取到的次数成反比。人们利用搜索工具检索相关信息时，都倾向于先查看排在检索结果中前几页的信息，如果这些信息能够满足检索者的需要，他们就往往就不会再去浏览之后的信息。这就是说，信息在检索结果中排名的先后与它被获取到的次数成反比，排名越靠前的信息，被获取的几率就越大。

因此，空间对数原理在网络条件下同样适用，只是空间的表现形式与传统条件下的空间表现形式有所不同。网络环境对传统的空间对数透视提出了挑战，我们现在的研究已经不能再局限于物理空间和社会信息空间，应该把注意力转移到网络信息空间上来，即研究人们在网络上获取信息的具体行为。

2. 时间对数透视现象

时间变换主要是描述信息沿时间轴的变换。**假定时间轴上的信息都是均匀分布的，根据收信者 *R* 对信息的接收情况，*R* 对时间较远的信息接收较少，而对时间轴上较近的信息接收较多**。虽然客观上各时间点上的信息是均匀分布的，但是通过对数变换，比较“年轻”的信息仅仅由于时间较近，在 *R* 接收信息时就获得了更大优先级。

由于时间的对数变换，会出现两种情况，一是虽然两种或几种信息客观上具有相同的重要性，但在 *R* 接收时，因为时间相关性的差异，会产生不同重要性的主观感觉。如在人类航天史上，滑翔机的研究成功与喷气式飞机的研制成功具有相当的重要性，但是对于现代科技人员而言，滑翔机的技术信息相对于喷气式飞机的技术信息在重要性方面就相差很多。二是由于对数变换，现代或近代的一些并不重要的信息会被视为与古代的重要信息同等重要。例如，在计算机发展史上，人们常将公元前 1100 年发明的算盘、1642 年 Pascal 发明的加法机以及 IBM8100 相提并论，而事实上，算盘之于人类较之 IBM704 FORTRAN[①] 更加重要。

随着网络的发展，网络信息的更新速度是传统条件下所不能及的。互联网上每天都有海量的信息产生，同样也有大量的信息老化。在网络条件下，人们对信息的新颖性和时效性相比以前要求更高。

下面做一个简单的抽样统计，从情报学核心期刊《情报杂志》2008 年刊载的所有论文中随机选取 20 篇论文作为样本，统计分析其引用文献的时间，统计结果如表 3-1 所示。通过统计数字可以看出，作者引用距离时间较近，也就是较新的论文篇数较多，发表时间在 5 年之内的引文数目占了总数的大部分比重，而发表时间越长的论文被引用的次数越少。其中，引用 2008 年发表的文章少是由该刊物发文的时滞性引起的，2008 年度刊载的许多论文是 2007 年或更早的时间截稿。

这些数字可以从一方面反映现在的网络环境中，人们对信息的选择依然是符合时间的对数透视原理，在选择信息时，人们依然首先关注到较新颖的信息。因此，无论是过去还是现在，时间对数透视原理都是适用的。

---

① 1954 年，IBM 704 计算机的诞生，是计算机历史上最伟大的进步，并催生了 FORTRAN 语言。FORTRAN 源自于“公式翻译”(FormulaTranslation)的英文缩写，是一种编程语言。它是世界上最早出现的计算机高级程序设计语言，广泛应用于科学和工程计算领域。

表 3-1　《情报杂志》2008 年载文的引文统计

时间(年)	引文篇数	时间(年)	引文篇数
2008	1	1998	6
2007	23	1997	1
2006	45	1996	1
2005	35	1995	1
2004	31	1994	0
2003	14	1993	1
2002	8	1988	1
2001	12	1980	1
2000	4	1977	1
1999	1	1972	1
论文总数	20	引文总数	188

3. 学科对数透视现象

学科对数透视的直接解释是：**各行各业及各种类型的信息接收者 *R* 在选择信息时，最常先选择的是本行业、本学科、本领域的信息，其次是关系最紧密的邻近学科、行业或领域，再次是有一定关系的学科、行业或领域，最后是那些关系很疏远的领域**。如果我们将这种学科(行业或领域)根据相关性做成一幅学科行业地图，同样会发现，在信息交流中，对接收者 *R* 而言，它们也都符合对数透视原则。研究学科对数透视，主要分析的是它与学科知识相关性的关系。

表 3-2 给出了情报学文献引用其他学科文献统计结果。从表 3-2 中可以看出，在情报学中，情报学研究人员在获取信息时，66.76%的信息来自于学科自身，18.51%来自于关系最密切的图书馆学、计算机科学、经济学、哲学、数学及系统科学，余下的 14.73%来自于近 48 个学科，这种分布正是对数透视原理的具体表现。学科的对数变换和科技情报中的核心期刊效应是不谋而合的。它描述了信息接收者在获取信息时的不均匀性，而这为我们根据读者或用户群选择文献或馆藏提供了理论依据。

表 3-2　情报学文献引用其他学科文献统计表

被引学科名称	引文数/条	被引学科名称	引文数/条
情报学	1553	软科学	5
图书馆学	147	生理学	5
计算机科学	76	时事	5
经济学	70	统计学	5
哲学	54	机器翻译	4
数学	44	光学技术	4
系统科学	41	运筹学	3
信息科学	39	思维科学	3
科学学	39	化学	3
语言学	37	工业技术	3
管理学	29	文化	3
中文信息处理	27	传播学	2

续表

被引学科名称	引文数/条	被引学科名称	引文数/条
知识学	18	潜科学	2
社会学	16	历史	2
心理学	15	政治学	2
马列毛著作	13	法律	2
综合科学	11	逻辑学	2
物理学	10	预测学	2
社会科学	9	医学	1
专利	9	档案学	1
声像技术	8	军事科学	1
控制论	7	教育学	1
未来学	7	地球科学	1
新三论	7	文学	1
行为科学	7	摄影技术	1
文献学	6	人才学	1
通信技术	6	其他	1
咨询学	5	—	—

在网络环境下，很多科研工作者已经习惯了从网上直接获取学术资源。由于网络上的信息关联性更强，学科之间的交叉融合就更为明显。同时由于获取工具的影响，与所要检索的主题相近的文献也更容易被检索到，这也增加了利用的可能性。通过互联网可以很方便地查阅各个学科领域的知识，学习和利用其他学科的知识也更为便捷。互联网加速了学科之间的交叉和相互渗透，也使各学科之间的联系更密切。而在传统的信息环境下，实现这些功能是不太可能的。因为人们在查询信息时，会直接查找相关度高的文献源，例如本领域的专业期刊等，而不会费力去找其他专业领域的信息源，所以一些或许有价值但相关度较低的文献不会被利用，例如所谓的“睡美人文献”。

因此可以认为在网络环境下，某学科获取信息时引用其他学科的信息比例会相对较高。为此，可参考一个简单的实证研究。从情报学学科的权威期刊《情报学报》的 2006 年下半年发表的全部文献中随机抽取 20 篇，对其后的参考文献进行初步分析，见表 3-3。

表 3-3　随机抽取《情报学报》20 篇文章参考文献分析结果

统计项	引文数/条	占引文数总数的比例/%
引文数总数	135	100
属于本学科的引文数	60	44.4
不属于本学科的引文数	75	55.6
直接从网络获取的引文数	51	37.8
不从网络直接获取的引文数	84	62.2

由表 3-3 可以得出，情报学领域的科研人员在获取和利用信息时，引用本学科领域的文献数占总的引文数的 44.4%，而引用其他学科的引文数占到 55.6%，并且，直接从网上获取的文献引文数比例就达 37.8%。这说明了网络环境对传统学科对数透视产生了重要影响。

学科之间引文数量越来越多，学科交叉趋势越加明显。

3.4　信息生命周期理论

3.4.1　信息生命周期的内涵

生命周期是生命科学的术语，其本义是指生物体从出生、成长、成熟、衰退到死亡的全部过程。20世纪50年代中期，美国的波兹（Booz）和艾伦（Allen）在《新产品管理》（*New Products Management*）一书中，首次将“生命周期”引入企业管理理论中，提出“产品生命周期”，并将其分为投入期、成长期、成熟期和衰退期等不同销售时期。而后，英国的戈波兹（Kuznets）等人，提出了戈波兹曲线数学模型，开启了产品生命周期理论定量研究阶段。1966年5月，美国哈佛大学的雷蒙德·弗农（Raymond Vernon，1913—1999）教授在《产品周期中的国际投资与国际贸易》（*International Investment and International Trade in the Product Cycle*）一文中提出国际产品生命周期理论，并将新产品的生命周期划分为产品创新、产品成熟和标准化三个阶段。

信息生命周期研究始于20世纪80年代。1981年，美国学者列维坦（K. B. Levitan）首次将“生命周期”引入信息管理理论中，认为**信息或信息资源是特殊的商品，也具有生命周期特征，其包括信息的生产、组织、维护、增长和分配**。1982年，美国学者泰勒（R. S. Taylor）认为信息生命周期应该是包含数据、信息、告知的知识、生产性知识和实际行动的过程。

1985年，美国学者霍顿（F. W. Horton）在《信息资源管理》（*Information Resources Management*）一书中指出，信息是一种具有生命周期的资源，其生命周期由一系列逻辑上相关联的阶段或步骤组成，认为信息资源生命周期体现了信息运动的自然规律，并据此定义了两种不同形态的信息生命周期：**一是由需求、收集、传递、处理、存储、传播、利用7个阶段组成的信息利用和管理需求信息生命周期；二是由创造、交流、利用、维护、恢复、再利用、再包装、再交流、降低使用等级、处置10个阶段组成的信息载体与信息交流信息生命周期**。

然而，信息生命周期理论真正进入主流视野还是源于ISO/TC171文件成像应用技术委员会于2000年10月12日召开的伦敦年会，会议通过的405号决议将ISO/TC46“信息与文献技术委员会”的一个分委员会改为“信息生命周期管理”技术委员会。该决议进一步明确了信息生命周期的概念。指出：**“信息无论是以物理形式还是数字形式管理，其信息生命周期均包括信息的生成、获取、标引、存储、检索、分发、呈现、迁移、交换、保护与最后处置或废弃。”**

此后，EMC（易安信，一家美国信息存储资讯科技公司）、StorageTek、DC等存储服务商也纷纷基于组织管理需求与数据服务层级变化，提出面向企业级数据/信息存储的信息生命周期管理理念，并推出了基于该理念的数据存储与管理解决方案。例如，2004年，世界知名的IT设备生产商EMC公司开始将信息生命周期管理（Information Lifecycle Management，ILM）引入数字存储领域，推出了一系列具有ILM特征的IT产品（存储设备和存储系统）。EMC认为，**数据价值与管理成本随时间发生变化，信息生命周期包括数据的创建、保护、访问、迁移、归档以及回收（销毁）6个阶段**。

信息生命周期理论在信息管理领域得到了蓬勃的发展，各种生命周期模型及基于生命周期的信息与数据资源管理策略与手段纷纷出现。

3.4.2 信息生命周期运动的认识

信息是物质内部结构与外部联系运动的状态和方式。实际上，信息运动既是客观存在的，又是极其复杂的，它由信息内在价值和外部环境等多重因素决定，并具有抽象性、多样性、周期性和阶段性的特征。

1. 信息运动的抽象性

信息运动的抽象性是指信息运动更多的是一种抽象运动而非具体的载体形式变化或物理空间改变，因此无法通过观察直观地看到信息运动，而只能通过信息运动过程中一些外部特征的变化间接对其进行研究。例如，承载信息的图书、期刊从甲地被移动到乙地，此时信息只是随载体介质发生了物理位移，其内容及所包含的价值并未发生变化，信息并未发生运动；而如果图书、期刊所包含的内容被阅读、参考与引用，则认为信息发生了运动。

2. 信息运动的多样性

虽然目前还无法完整地解释信息在其生命周期中的运动轨迹，但从本质上讲，信息生命周期中的信息运动是一种客观状态。在实际信息活动中，这一抽象运动过程又表现为载体变化、空间移动、价值衰减等多种具体形式。因此，通过考察信息在其生命周期中的多样化运动方式，可以分析出信息生命周期运动的阶段性特征与内在运动规律。

3. 信息运动的周期性

信息生命周期之所以被称为“周期”，在于它并非单向单次运动，而是一个周期性循环往复的运动过程。信息自创建到传播再到利用直至处置的完整生命周期中，其价值始终随着生命周期阶段的演进不断发生变化。总体而言，该趋势应当是一个价值逐渐衰减的过程，但此种规律性变化并非一成不变的，而是存在许多不确定性。如处于生命周期晚期价值已严重衰减的信息随时有可能随着某一学科领域甚至某一知识点的突破创新而在极短时间内重新跃迁至活跃期，从而开始一轮新的生命周期循环。随着信息资源数字化与网络化趋势的日益深入，这一现象在数字资源中可能体现得更为明显。因此，要完整地理解信息生命周期，就必须动态地看待信息运动，将其视为一个不断变化的周期性循环过程。

4. 信息运动的阶段性

以往研究者大多是将信息作为一种具有生命的资源，并从管理角度将信息生命周期划分为若干相互关联的阶段。事实上，信息生命周期与信息管理周期就内部阶段组成而言，显然是有差异的；但就其周期的时间跨度来说，通常却是一致的。因此，许多人往往将信息管理阶段混同为信息运动阶段。从本质上讲，两者既存在显著差异，又有着一定关联。

一方面，信息管理活动的阶段性是由信息运动的阶段性决定的，信息的阶段性管理必须依据信息运动的阶段性规律来进行；另一方面，尽管信息运动是客观的，是由内因决定的，但对信息实施不同管理方式与手段却可以影响信息的运动过程。例如，印刷型文献经过数字化加工并通过互联网发布后，其传播方式、范围、速度均会有所改变。

总之，信息生命周期考察的是信息在生命周期不同运动阶段的内在规律，信息生命周期管理探究的则是信息在生命周期中不同阶段的管理方法与策略。与此同时，由于信息生命周期与信息管理周期以及信息生命周期管理均存在着相互关联，因此信息运动阶段与信息

管理阶段也存在着密切联系。

3.4.3　信息生命周期理论

1. 信息生命周期的研究对象

信息生命周期的研究对象是信息，其核心是对信息从产生到消亡整个生命周期过程中的运动与变化规律进行研究。

万里鹏的《信息生命周期研究范式及理论缺失》一文认为，**信息生命周期的研究对象是信息运动**。然而信息运动作为信息存在的表现形式，即信息如何随时间、空间而运动变化，其主体归根结底仍然是信息，因此信息生命周期本质上观察的也仍然是信息。所以可以认为，信息生命周期的研究对象既不是信息运动，也不是信息管理，而就是信息本身。

2. 信息生命周期的研究内容

目前来看，**信息生命周期的相关理论问题主要包括**：①不同类型信息的生命周期有何异同；②信息生命周期存在哪些阶段；③信息在生命周期中的阶段递进与跃迁机理如何。对于上述问题，目前还缺乏深入研究，仅停留在初级探索阶段。

信息生命周期的核心在于如何科学地揭示信息运动的内在规律，具体内容包括信息自产生到消亡生命周期中的内在运动规律、运动轨迹及描述方式等。具体来讲，信息生命周期至少应包括以下几方面研究内容。

（1）**信息生命周期运动阶段性理论研究**。关于信息生命周期阶段的理论研究始终是信息生命周期研究的热点问题；Taylor、Horton 等相关学者以及国际科技信息委员会、ISO/TC171 文件成像应用技术委员会等国际机构乃至 EMC 等信息存储服务商均对此有过论述。目前，该领域研究的核心问题包括：信息生命周期的阶段划分及依据、阶段之间的内部关联、阶段递进与跃迁的机理、信息生命周期与信息管理周期各阶段的异同等。

（2）**信息生命周期运动影响因素分析**。由于信息运动的抽象性、多样性特征，目前还无法直接地改变信息的运动轨迹，但却能够利用信息运动的影响因素分析，结合信息利用结果以及信息之间的显性与隐性关联等，分析信息的生命周期运动并通过改变其影响因素间接对其施加积极影响，从而最终实现科学有效的信息生命周期管理。

（3）**信息生命周期测度研究**。目前国内外对于信息生命周期测度问题的研究还相对较少，然而该领域既是信息生命周期理论研究的重要内容，同时也是研究的关键性难点。如果不能在信息生命周期测度领域有所突破，那么相关研究就难以深入开展。既然信息从最初产生到最终消亡构成了一个完整的生命期，那么该生命期的长度应该是多少，如何对其进行测度，这些问题都值得探讨。

（4）**不同类型信息生命周期差异性研究**。信息依据其内容、载体形式、传播渠道等可分为多种类型，不同类型信息的生命周期存在显著差异。具体来说，不同载体类型信息的运动轨迹有何异同，不同学科领域信息的生命周期存在何种差异，不同类型信息生命周期的影响因素存在哪些区别，都是需要研究的问题。

（5）**信息生命周期理论的应用领域研究**。信息生命周期理论科学地揭示了信息的内在运动规律，为基于生命周期的信息管理与信息利用奠定了坚实的理论基础，因此具有广阔的

应用领域。例如，基于信息生命周期内的运动规律对信息采取科学的管理方式和手段，可以改变信息的传播与利用状况，从而促进信息的价值实现，提高信息利用效率。

3. 信息生命周期的研究方法

信息生命周期的研究对象涵盖多种印刷型与数字资源类型，同时研究内容涉及信息老化测度、影响因素分析、信息内在关联的揭示描述等多个领域，因此研究对象的多样性与研究内容的广泛性要求我们必须综合运用各种定性与定量研究方法推进信息生命周期研究，具体方法如下。

1）因素分析法

信息运动及其生命周期均受多种因素影响，如学科领域范围与发展状况、信息增长与老化规律、信息介质类型、信息管理方式方法、用户利用习惯等。因此，必须借助因素分析方法针对各种纷繁复杂的影响因素及其作用方式与影响效果展开综合分析，从而揭示信息运动与生命周期的内在机理，具体分析方法包括：层次分析法、模糊分析法、因子分析法、灰度关联分析法等。

2）信息计量学方法

信息生命周期与信息老化、半衰期之间存在密切关系，同时信息老化曲线也是信息生命周期变化的最直观表征。因此，可以借助引文分析法、电子资源在线使用统计等信息计量学方法，观察和测度信息的老化速度、价值变化趋势等问题，从而为信息生命周期测度进行前期准备与探索。

3）社会网络分析法

社会网络研究发端于二十世纪二三十年代的英国人类学研究，其基本事实是每个行动者都与其他行动者有或多或少的关系，社会网络分析就是要建立这些关系的模型，力图描述群体关系的结构，研究这种结构对群体功能或者群体内部个体的影响。发展至今，社会网络分析已被广泛应用于网络社会关系发掘、支配类型发现（关键因素）以及信息流跟踪，通过社会网络信息来判断和解释信息行为和信息态度。作为一种跨学科研究方法，在社会学、心理学、经济学、信息科学、系统科学与计算机科学的共同努力下，社会网络分析已从一种隐喻成为一种现实的研究范式。利用社会网络分析，能够针对信息生命周期中信息单元之间的内在关联展开分析，从而构建基于信息生命周期的知识网络。此外，还能够研究知识生产者在信息生命周期运动中的角色关联，考察信息价值在生命周期中的转移与流动情况等。

3.4.4 大数据与信息生命周期理论

信息时代下，大数据与信息生命周期理论的联系主要在于以下几点。

（1）大数据技术是一系列收集、存储、管理、处理、分析、共享和可视化技术的集合。而纵观信息生命周期理论的发展及其定义，信息生命周期总会经历信息采集、处理、存储、传播、利用和处置等阶段。大数据的各项技术是信息生命周期阶段推进和周期更替的动力，大数据时代下离开大数据技术，信息生命周期将无法运行，可以说：大数据时代下，大数据技术是信息生命周期的动力和技术支撑。

（2）信息生命周期是以信息采集开始，信息采集最关键是选取合适的信息源，从中获取满足个人需求或企业决策的信息。而在庞大的数据中，对每个信息采集者来说，大部分信息

是没有价值的，有用的信息只是其中的很小部分，采集到需要的信息越来越难。并且庞大的数据量仅仅是大数据的重要特征之一，大数据的集成价值、处理效率和持续存取才是关键。大数据技术则会实现对动态、异构、庞大数据的存储和管理，并从中提取出简约的数据集，从而节约信息采集时间，提高信息采集的效率和所得信息的质量，为信息采集人员提供了有别于传统信息源的大数据时代信息源。

（3）理论指导实践，实践又会反作用于理论。信息生命周期理论揭示了信息价值或利用率在时间上变化的客观规律。而大数据进行信息处理是致力于采用数据实时处理技术，尽早尽快地处理最新鲜的数据，对其进行数据分析，最终输出处理结果。例如，流处理，是大数据信息处理技术之一，其理论支撑便是随着时间的推移数据的价值会不断减少，这些数据所蕴含的知识价值往往也在衰减。而随着大数据时代的到来，离线数据分析向在线实时数据处理分析转变。很多实例则证实，数据的价值会随着不断地被利用而增长，有违信息生命周期理论中有关信息价值的阐述。由此可见，大数据是信息生命周期理论的实践，信息生命周期理论是指导大数据理念产生及其发展的基础理论之一。

3.5　小世界现象

3.5.1　小世界现象的由来

1967年，美国哈佛大学斯坦利·米尔格拉姆（Stanley Milgram，1933—1984，见图3-2）在《今日心理学》杂志上提出了“六度分隔”（Six Degrees of Separation）理论，大意为任何两个欲取得联系的陌生人之间最多只隔着6个人，便可完成两人之间的联系（见图3-3）。今天，该理论也被称为“六度空间理论”“小世界理论”等。

图3-2　米尔格拉姆（Milgram）

图3-3　小世界现象

米尔格拉姆的这一假说是在其完成了一项实验的基础上提出的。当年，米尔格拉姆给内布拉斯加州奥马哈市随意选择的三十多人发信，要求他们把他的这封信寄给波士顿市一个独一无二的“目标”人，分别由每个人独自联系。米尔格拉姆告诉每个发信人通过人传人

的送信方式来统计人与人之间的联系程度。首先把信交给实验者 A，告诉他信件最终要送到目标人 S 那里，如果他不认识 S，那么便把信送到某个他认识的人 B 那里，理由是 A 认为在他的交际圈里 B 是最可能认识 S 的。但是如果 B 也不认识 S，那么 B 同样把信送到他的一个朋友 C 那里，……以此类推，信件一步步到达 S 那里。于是信件从 A 到 B 到 C 到……最后到 S 连成一条链，链条上的每个成员都力图把信件寄给他们的朋友、家庭成员、商业同事或偶然的熟人，以便使信件尽快到达目标人。米尔格拉姆发现 60 个链条最终到达目标人，链条中平均步骤大约为 6，由此得出结论：任意两个人都可通过平均 6 个熟人联系起来。这就是六度分隔理论的产生经过。

从严格的科学角度讲，六度分隔理论猜想的成分居多。米尔格拉姆的实验并不十分理想，实验的结果是有 4/5 的信件没有到达目标人。多年来，六度分隔理论虽然一再被提及，但一直没有获得颇具说服力的佐证，六度分隔与其说是一种理论，不如说是一种假说。

2002 年 1 月，也就是在米尔格拉姆进行实验的 30 年后，为了验证“六度分隔”理论，纽约州康乃尔大学的社会学家邓肯·瓦茨（Duncan Watts）和哥伦比亚大学社会学系（Department of Sociology，Columbia University）合作开展了一个“小世界研究计划”（Small World Research Project，SWRP），准备再次重复米尔格拉姆当年的实验。不同的是，这次实验借助的是现代高科技互联网，而且实验规模也扩展到了全球范围。

来自全球 16 个国家超过 6 万人参与其中。所有这些参与者的任务就是，发送成千上万封 E-mail，并让这些 E-mail 最终能够到达指定的 18 名“目标接收者”。但前提是，每封邮件你只能发给认识的人，而且每次只能发一封，而这 18 名“目标接收者”全都是随机选出来的，他们的职业、性别、地理位置各不相同，其中包括一名美国教授、一名澳洲警察，以及一名挪威兽医。

研究人员发现，在绝大多数情况下，人们只需通过 5～7 封 E-mail 就可以联系到“目标接收者”，显然“六度分隔”理论所言不虚。不仅如此，实验还表明 E-mail 和互联网的出现并未令人类传统的社交关系发生根本的改变。在这个实验里，互联网只是我们用于传递信息的一个简单工具而已。和那些网络之外的人际关系——例如工作、学校、家庭和社区等相比，E-mail 作为一种社交媒介并无任何特别之处。研究人员还发现，到目前为止，人们用得最多的人际关系是工作关系，而且如果 E-mail 是被转送到某个同性那里，它到达“目标接收者”的可能性会更大。

当“六度分隔”假说被提出的时候，互联网的前身才刚刚诞生，米尔格拉姆可能没有想过有朝一日互联网能够被用于验证他的理论。不论“六度分隔”的确切数字到底是多少，这个世界看上去确实很小，“小世界”由此得名。**小世界现象提示了客观生物运动中某种最为快捷的信息传达传递方式和传导路径**，可用来描述在一定时期内发生的、引人关注的诸多生活事件。小世界现象实质上揭示的是人类信息联系和信息对象之间的相关性，亦即无论世界多么大，人口怎样多，分布如何广，网络结构多么复杂、节点数量如何巨大，都可以通过相关的信息达到最短的路径联系。

3.5.2 小世界现象的研究类型

“六度分隔”现象在学术上称为小世界效应（Small World Effect）。小世界效应的定义是：若网络中任意两点间的平均距离 L 随网络节点数 N 的增加呈对数增长，即 $L \sim \mathrm{In}N$，

且网络的局部结构上仍具有较明显的集团化特征，则称该网络具有小世界效应。这里的平均距离具有广泛的含义，例如，在信件传递实验中，平均距离就是平均传递次数 6，具有“六度分隔”现象的网络的 L 值都是 6。

对于小世界效应的研究大致可分为两类：一是随机网络；二是著名的 W-S 小世界网络模型[①]及转化类型。

1．随机网络

对小世界效应最简单的解释是用到已在数学领域得到深入研究的随机网络原理，即网络中节点间的平均距离 L 随网络大小 N 呈对数增长，则随机网络具有小世界效应。因此也有人将小世界效应定义为：具有小世界效应的网络其节点间的平均距离与随机网络中的平均距离之比。

然而，要将随机网络作为现实网络的抽象模型尚存在很大的问题，假设随机网络中某人 A 有 2 个熟人，每个熟人又分别有 Z 个熟人，那么 A 有 Z^2 个熟人，以此类推。但在现实生活中，人们的朋友圈子在很大程度上有重叠性，也就是说，你的两个朋友有可能互相又是朋友。这个重叠特征被称为集团化，而随机网络没有显示集团化或者集团化程度很低，不符合现实网络特征，因此无法用随机网络研究现实网络。

2．W-S 小世界网络模型

随机网络不是小世界网络，我们将既具有小世界效应，又具有集团化特征的网络称为小世界网络（Small-World Network，SWN），它是具有高度的局部集团化和较短的全局平均路径长度的网络，介于高度有序和高度随机之间的一种抽象图。研究者通过理论分析和数值模拟证明：只要使大世界中各连接以很小的平均概率“断链重连”，就可以实现大世界向小世界的渡越，从而基本保持大世界的结构而实现小世界的功能。

W-S 模型实质上是具有一定随机性的规则点阵。构建方法是：在环状规则点阵中用“断链重连”的方法，即顺序浏览每条边，以较小的概率 p（$p \approx 0.1$）将边的一端移到另一个随机选取的位置上，即形成了所谓的小世界网络（SWN），如图 3-4 所示。

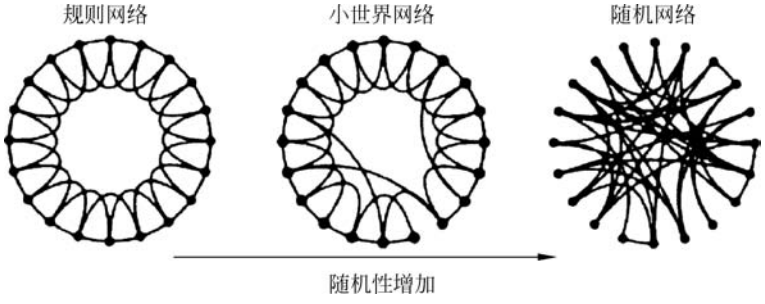

图 3-4　规则网络、小世界网络与随机网络示意图

虽然少数边伸展到较远的地方（这些较长的边称为捷径），但由于 p 很小，模型仍大致维持规则结构，具有较高的聚类系数。另一方面，加入捷径使特征路径长度下降很快，这使得小世界网络的特征路径长度与随机网络的特征路径长度相当。研究显示，现实网络（例

① W-S 是两位发现者的名字的首字母。

如，神经网络、社会关系、输电线路）当中，出现远距离连接的捷径的现象很普遍。通过对特征路径长度和聚类系数的测量，瓦茨和史蒂文·斯托加茨（Steven Strogatz）发现，许多领域的合作网络都存在小世界现象，于是断定小世界现象是大型现实网络的内在属性。后来许多学者对 W-S 模型加以改进，提出以较小的概率 p 在网络中将少量边“断链重连”或直接加入少量捷径，保持网络基本不变，而节点之间的特征路径长度则下降很快。这种网络就同时具有短特征路径长度和高聚类系数，实现了由大世界向小世界的转换。

研究表明，小世界网络可以较好地反映现实网络特征，有助于探讨网络结构对网络功能的影响。小世界网络之所以引起各学术界的关注，是因为研究小世界模型有助于理解大型系统的动态属性和其结构特征之间的关系，如小世界网络理论等。

3.5.3 小世界网络现象的应用

小世界网络提出后引起各学术界的关注，在物理、数学、生物等自然科学领域都有较广泛的应用。SWN 以全新的理论思路和有效的技术工具，展现出很强的适用性和广阔的发展前景。

1. SWN 在物理学中的应用

小世界问题研究在物理学上取得了丰硕的成果，例如，传播介质在一个要素间平均分离为 6 的网络中扩散要比在平均分离度为 100 或一百万的网络中快得多，这对于研究信息、疾病等传播具有指导意义。除了理论研究方面，科学家们发现在小世界网络中研究物理问题能够解释许多实际现象。

莫纳森（Monasson）用转移矩阵的方法研究了小世界网络结构上的拉普拉斯算子（Laplace）特征谱。这个特征谱告诉人们，建立在小世界网络结构上的动力体系的普通形式以及动力扩散在小世界网络中的产生方式，而扩散运动可能提供某种社会网络信息传播的简单模型。国内研究者朱陈平和熊诗杰提出了无序量子小世界网络模型，发现存在局域化——退局域化相变，并以此解释了掺杂高聚物中电导变现象。

2. SWN 在生物学中的应用

研究人员通常运用 Bak-Sneppen 物种进化模型（模拟大数量物种间相互作用对进化产生的影响）来描述生态系统，库尔卡尼（Kulkarni）建立了小世界网络结构模型对相同的问题进行了研究，结果表明，基于网络的网络结构，其小世界网络结构模型比 Bak-Sneppen 低维规则模型更接近真实的生态网络。

2000 年，费尔南德克（Lago Fernandek）等研究了 Hodgkin-Huxley 神经元系统的各种基本图形，发现由于网络结构的高度集团化引起系统相干振荡，网络中各点间较短的平均间隔距离使得网络对外部刺激快速做出反应。同时，具有这两个特征的小世界网络（高度的局部集团化和较短的全局平均路径长度）是他们发现的唯一的、但同时具有相干性和快速反应的网络结构形式。

3. SWN 在医学中的应用

目前，运用小世界理论最多、最具成效的研究是疾病传播问题，研究表明：病毒在小世界网络中传播很快，这与实际情况很接近。

库珀曼(Kuperman)和阿布拉姆森(Abramson)建立了 SIRS 动态模型，研究社会结构对疾病动态传播的影响。他们发现对应于一定的人群结构，网络中的连接依概率 p 断开并重新与其他点相连时，被传染的人数从不规则的、小幅度的增加(p 很小)发展到自发的、大范围的振荡状态(p 较大)。其中，当 p 值在 0.1 附近时，传染人数明显增加，显示出小世界效应。由于小世界网络结构是目前描述社会网络结构较好的工具，因此，在小世界网络中研究疾病传播问题极具现实意义。

4. SWN 在经济学中的应用

长期以来，许多经济与管理学家致力于从纷繁多变的经济、管理现象中寻找可能存在的定量规律来指导实践。研究表明，小世界现象同样广泛存在于经济与管理领域中，因此，SWN 模型也是研究经济与管理问题的有效工具。

人们可以将经济与管理等抽象问题转换为 SWN 模型，运用 SWN 分析方法研究模型中网络结构参数对网络功能的影响，以寻求网络功能优化的途径。例如，①分析动态联盟企业的内外部合作关系，其结果表明动态联盟企业具有小世界效应，这是在该领域运用 SWN 进行深入研究的基础；②通过对 SWN 结构及数字特征的分析，人们能定性定量地解释现实博弈问题，解释了双方合作是最佳联合策略的前提下，人们仍会选择背叛的原因，并对不同博弈策略进行了比较；③提出了动态有向 SWN 新产品市场扩散随机响应模型，并对其进行了定量分析。

5. SWN 在交通管理中的应用

我国学者将 SWN 模型引入交通管理领域，以研究小世界网络的数量特征。用成熟的数学与物理理论方法，结合我国社会交通网络的特征，提出符合国情的降低网络平均路径长度的策略方案，开发并实现了基于网络效率理论的网络评估方法和辅助规则技术。这表明：小世界网络理论可以对现实生活中的交通网络规划和测评起到极好的辅助效果。

小结

数据思维需要一些普遍意义的、基础性的思想和原则来指导其在各领域的应用和发展，从而成为其应用的基本原理或逻辑起点。从信息学视角来看，其基本原理主要有最大熵原理、最小努力原理、信息生命周期理论、对数透视现象和小世界现象。

最大熵原理的应用范围非常广泛，可以用来解决随机性或不确定性问题。应用其解决问题的思路是：先将所研究的问题转换为一个概率模型。这样，问题的随机性就表现为概率分布，问题的解决就归结为求一种最佳的概率分布，然后采用最大熵原理求出最佳分布。由此得到启发：凡是带有随机性的问题(不论是哪一领域的)，都可以尝试用最大熵的方法加以解决。这就为一些优化、决策、预测问题的解决提供了新的途径和方法。

最小努力原理是人类生态的基本规律之一，它体现在人类社会的各方面。齐夫所描述的省力法则虽然发源于语言应用领域，但各个不同领域中最短路线的选择和确定问题都与这一法则有关。例如，企业供应商和库存地点的选择、社区供货点的位置、交通路线安排、通信路线架设等，都涉及最短路径寻求的解决方法。

对数透视现象则解释了人们遵循最小努力原理进行信息、知识、情报的获取和吸收这一

现象，即时间上寻求最新、空间上寻求最近、学科（领域）上寻求从自己最擅长和最熟悉的领域来查询并获取知识和信息。

信息生命周期理论的研究对象是信息，其核心是对信息从产生到消亡整个生命周期过程中的运动与变化规律进行研究。信息时代下，大数据与信息周期理论联系紧密。大数据时代下，大数据技术是信息生命周期的动力和技术支撑，大数据是信息生命周期理论的实践，信息生命周期理论则是指导大数据理念产生及其发展的基础理论之一。

小世界现象广泛存在于信息生产、信息系统、信息获取、信息传递和信息利用过程及信息对象的分布特征中，具有普遍意义和广泛应用性。互联网上的各类网站、网页、网络目录和上网用户之间的有效链接更加展现了任何一种信息载体和信息传递方式都构成小世界网络的强大功能。

讨论与实践

1. 结合自己的思考与理解，谈谈对“熵”“信息熵”及“最大熵原理”的认识。
2. 结合自己的思考与理解，谈谈大数据时代“齐夫定律”的应用场景。
3. 你认为“信息生命周期”应该划分为哪几个阶段？数据也有生命周期吗？
4. 举 2～3 例说明大数据时代“对数透视原理”的应用场景。
5. 结合医学和生物学的应用，分析“随机网络”及“小世界网络”对实践的指导作用。

参考文献

[1] 程鹏，李勇. 情报概念及相关问题之辨析[J]. 情报学报，2009，28(6)：809-814.
[2] 梁战平. 开创情报学的未来——争论的焦点问题研究[J]. 情报学报，2007，(1)：14-19.
[3] 彭知辉. 数据：大数据环境下情报学的研究对象[J]. 情报学报，2017，36(2)：123-131.
[4] 李建东，王永茂，胡林敏. 最大熵原理及其应用[J]. 硅谷，2009，(4)：42-43.
[5] 曲英杰，孙光亮，李志敏. 最大熵原理及应用[J]. 青岛建筑工程学院学报，1996，(2)：94-100.
[6] 王姿砚. 论最小努力原则及其应用[J]. 中国图书馆学报，1991，(4)：6-8，87.
[7] 王洵. 最小努力原则与齐夫定律[J]. 情报科学，1981，(2)：32-36.
[8] 马费成. 论情报学的基本原理及理论体系构建[J]. 情报学报，2007，26(1)：3-13.
[9] 杜彦峰，相丽玲，李文龙. 大数据背景下信息生命周期理论的再思考[J]. 情报理论与实践，2015，38(5)：25-29.
[10] 索传军. 试论信息生命周期的概念及研究内容[J]. 图书情报工作，2010，54(13)：5-9.
[11] 张俊娜. 浅谈网络环境下的空间和学科对数透视原理[J]. 科技情报开发与经济，2007，(25)：116-117，124.
[12] 肖楠，任全娥，胡凤. 网络环境下的对数透视原理[J]. 图书情报知识，2007，(3)：60-64.
[13] 翟文姣. 网络条件下的对数透视原理[J]. 中国商界（下半月），2010，(11)：383，385.
[14] 朱亚丽. “六度分离”假说的信息学意义[J]. 图书情报工作，2005，(6)：59-61，32.