

“信息量悖论”及其有关问题

屠德雍

(浙江工业大学, 浙江 杭州 310012)

摘 要: 该文首先提出了“信息量悖论”, 并以此展开讨论引出与此相关的几个问题, 提出一些新概念。这些问题和新的概念对进一步理解信息论及在方法论上用延伸类比等方法分析类似问题或许有启发作用。

关键词: 信息量悖论; 信息量原子论; 虚拟信息; 信息结构

中图分类号: B014 **文献标识码:** A **文章编号:** 1003 - 5680(2003) 02 - 0023 - 03

一 引 言

在现今信息社会里, 关于信息的哲学问题的讨论, 近年来在有关的文献中也常可见到^[1]。本文首先提出了一个新问题——“信息量悖论”, 并以此展开讨论了与此相关的几个问题, 提出了一些新的概念。如所周知, 自申农(C. E. Shannon)发表信息论半个多世纪以来, 众多文献资料及教科书中都引用着信息量的定义极其解释。它们都已成为经典。不仅如此, 正是由于信息量的定义由此引出“平均信息量”, “互信息”, “信息熵”……等等一系列重要概念, 从而形成信息论的最重要基础并再进而推出整个信息论的主要内容。这些理论在指导实践中应用获得了极为巨大的成功。无疑, 信息量的定义本身并无什么问题, 但在如何理解或解释这一定义的含义上却有值得推敲之处。信息量的定义或许在开始时是从信息内容出现的概率性提出的, 但在以后对它的推理极其应用却是在信息构成的概率性有关的问题上进行。这也许就是本文提出的“信息量悖论”的真正原因。由此看来, 从某种角度讲或许“信息量悖论”本身可能就是一个伪问题。不过把它提出来讨论或许对进一步理解信息论及类似问题有启发作用。

二 信息量 $I(P_K)$ 悖论

我们这里只讨论信源信息量。信息量通常是以概率空间的可能状态数目及其概率来描述的。设给定集合 $\{X, P(x)\}$, 其中 $X = \{x_K, k = 1, 2, \dots, n\}$, 为使信源的状态空间对每一事件 $x_K \in X$, 相应的概率为 $P(x_K)$ 且有 $\sum P(x_K) = 1$, 那

么事件 $x_K \in X$ 的信息量定义为

$$I(x_K) = -\log p(x_K)$$

通常取底数为 2, 信息量单位比特(bit)。可见 $I(x_K)$ 是集上的一个概率函数。对信息量意义的解释几乎所有有关信息论的书或资料中都有类似如下的说法: “事件发生的可能性愈小, 则一旦发生愈使人感到惊奇, 其所带来的信息量就愈多”。由上式定义也可见此时信息量就愈大。例如设有一条消息: “A 先生中了大奖”。因为中大奖的概率是很小的, 故这条消息的信息量应是很大的。又有一条消息: “B 先生中了小奖”。因中小奖的几率比大奖要大的多, 故这条消息的信息量并不大, 这看来没什么争议。接着 A 与 B 先生都到电信局拍电报告诉他们的夫人。最后在收费单上他们需交的费用竟然是一样的。他们都觉不太公平而去咨询电信局经理。经理告诉他们电信局当然也是以信息论作为根据来运行的。在收费时当然也要考虑信息量的问题, 例如你发两页纸的电报(设 200 个字)比发一页纸的电报信息量大约要多一倍, 因此收费也多一倍, 但一般为了方便收费标准, 只按字的个数来计算而不考虑信息的内容极其出现的概率等。他还特别强调指出: 申农的信息论中是不关心信息的具体内容的。上述解释更引起两位先生的纳闷, 因为他们也知道信息论确实不考虑(也很难考虑)信息的具体内容, 但却又以信息事件出现的概率作为信息量度的唯一函数变量。而该变量的定义又是整个信息论的基础。这种既“无关”又必须“有关”的矛盾我们称之为“信息量悖论”。在下面的继续分析中我们就可以看到其问题所在。

【收稿日期】 2002 - 02 - 04

【作者简介】 屠德雍(1934 -), 男, 浙江工业大学教授。

三 信息的可分性及信息原子论

大家知道一则报道、一篇文章都是由许许多多最简单的、但是有完整意义的句子组成的。这每一句子就是一条信息,如果要再分的话,那只有分成没有完整意义的字或字母等,这就不是信息了,所以这种最简单的信息单元我们可以叫它信息分子。一篇文章是信息分子的集合。当然,这集合中的每一元素不一定是相互独立的。它们之间可能具有“记忆”关系,上述这一事实可称之为“信息的可分性”。我们再继续讨论构成信息分子的字符(字汇可以看成是有记忆关系的字符的集合)。如果说信息分子的数量应是无限的,那么构成它们字符的数量应是有限的。正像世界的物质的分子种类是无限的,而构成它的原子种类却是有限的。这有限种构成信息分子字符可叫它为“信息原子”。例如英文字母只有 26 个,中文的文字虽然多得多,但常用的也就 6000 左右,当然,原子还有它的组成物,这就是各种基本粒子和量子。与此类似,作为信息原子的字符它又可以由更简单的代码组成,例如二进制,我们也可称它为一种“信息量子”,我们可再继续进行类比讨论,我们知道各个字符在一条信息中的出现是随机的,如果用上面的信息量定义来定义字符出现状况的一种度量,那么它可称之为“信息原子量”,而平均信息量则是“信息分子量”,它们之间是有紧密联系的,但其意义是不相同的,原子量与分子量是物质的重要参量,信息原子量与信息分子量也是信息的重要参量。上述的讨论,我们可称之为信息原子论的基本观点。

由上讨论我们既可说明了上述所谓“信息悖论”的产生原因。原来,所谓“信息量”的定义实际上应是信息原子量。个别元素的大小当然不一定决定分子量的大小。电信局是按照信息分子量及考虑其他一些操作方便及政策等因素来收费的。两者当然不会一致。而且也都与信息内容无关。其实,信息内容是很难量化的。但信息内容的构成(不管它用什么方式表述)是可以量化的,信息量的定义实际上是对信息内容的构成进行量化的基础。

申农的信息论首先是研究信源描述(信息编码等)问题,其它是信源的量子化问题,再是研究信息在信道中传输及到信宿前的信息译码等问题。也就是研究信息原子及量子在运动过程中的特性及相互作用等问题。这里我们可把嘈声也可看成一种信息量子。故申农的信息论实际上是一种信息原子论和量子论。至于更广义的信息论,例如社会信息论,生物信息论等等可认可是各种信息化学理论。

四 信息的可组合性及虚拟信息

信息既然具有可分性,那么它也具有可组合性,那是自然的事情。信息量子组成信息原子及信息原子组成信息单元,即信息分子。这是解码或译码问题,也是信息论的内容组成部分。而信息单元组成信息集合,这一般不包括在狭义信息论范畴之内了。它涉及范围非常之广,语文学当然是与其直接有关的。其他如文章,演讲及音像资料以及文学艺术都可看成是信息集合的各种形式。下面只讨论有关信息组

合性的几条重要特性:

1、信息组合的相关性。设各个信息分子之间和信息子集之间一般有记忆关系。信息集合与由其子集 $U_1, U_2, \dots, U_N$ 组成。可以证明:

$$H(S) \leq H(U_1) + H(U_2) + \dots + H(U_N)$$

其中 H 代表共熵。上式表示相关性使总的信息集合的共熵 $H(S)$ 要小于各个子集熵之和。只有当各子集相互独立时它们才相等。这是可以预料的。

2、信息组合的系统性。如果仅是一条消息,那么收信者只了解这一条消息的内容而已。若是两条或两条以上相关的信息,那就会产生两方面情况,一是除了知道这几条消息内容之外或许还能推出新的信息。例如刑警从所获的少量犯罪线索中可以进一步探测新的线索。不过这是收信人认识过程中大脑内部的信息加工问题,这里不去讨论。^[2]二是相关信息组合就构成信息系统。其本身就是一条全新的信息,它已不同于各个构成它的信息单元之和。即信息组合服从总体大于各组成部分之和的系统论原则。我们读一本书所获得的印象决不是书中的只言片语的总和。欣赏一件艺术品我们首先得到的是总体的概念。

3、信息组合的可虚拟性。这是信息组合的又一个具有重要意义性质。我们可以从来自不相关的原始信源的信息加以人工剪切和组合加工并使其相关而形成新的信息或信息集合。显然它已与实际信源信息大不相同。它是“虚拟”的。所有的科幻片、神话片及特技摄影都是虚拟信息作品。至于近代才发展起来的用计算机技术来实现的虚拟环境等不仅在娱乐更重要在训练和科研等方面正在发挥出愈来愈大的作用。

还可指出,广义的信息组合可应用到生物学、心理学及社会学等等许多领域。例如,我们可把遗传密码信息加到另一生物的 DNA 中使产生变异而得到新的物种。又如人们脑中储存的各种不同信息的合理组合可产生创新性的思维,而组合失控则会产生各种幻觉,甚至表现出精神病态^[2]等等。这些都已超过狭义信息论的范围,本文也不去进一步讨论了。

五 信息的结构特性

由上讨论可见,正像物质具有结构一样,信息也是具有自己的结构的。但信息结构特性与物质的结构特性既有共同之处,也有不同特点。现讨论如下:

1、结构的层次性。上述已指出,信息由“信息分子”组成,即它有分子层次。“信息分子”又由“信息原子”组成,而它们又可有“信息量子”即代码等来描述。所以信息也有其原子层次和量子层次。不同的层次的性质与功能是不同的。例如信息分子层次是组成信息内容的最基本层次。它是对主观或客观状态的一种最简单的表述。例如“我很累”,“天亮了”等等,信息原子是组成信息分子的基本元素,它已不是直接与信息内容有关,而是按一定的顺序构成信息分子的基本砖块。例如上例中的“我”“很”“累”等它们按一定的顺序排列才构成一个信息分子。至于信息量子(例如字母和代码

等)它又是代表信息原子的各种符号,它们也有不同的级别,而且也须按一定的顺序才能代表某一信息原子。应当说它们与信息内容更无直接关系了。

2、信息结构的非唯一性或多样性。如果说上述的信息结构的层次性至少与物质结构有类似之处,那么信息结构的非唯一性则不同了。我们知道一种物质的分子只能由某些确定的原子按一定结构组成。但同一内容的一条信息可用多种表述方法。甚至可以说有无限多种的表述方法(当然实际应用中只是有限几种)。例如可用不同的语言、文字及不同制式的代码等。即同一信息内容可以用完全不同的信息原子和信息量子来代表。这就是信息结构的多样化,也是与物质结构特性有本质区别的特点之一。之所以形成信息结构多样化的根本原因在于信息结构的描述一般带有主观性,而物质结构却是客观的,虽然你可以用不同的名称去称呼它,但水分子总是由两个氢原子和一个氧原子并按一定的结构所组成,否则就不是水分子了。

3、信息结构的可压缩(或简约)性。这有两方面情况。一是由于信息内容的可压缩性因而带来其结构的变化。例如听了一场报告或看了一本小说,既可详尽地把内容细节传达给别人,也可用三言两语或长话短说概要地叙述大概轮廓。通信的基本目的是接受端能给定的失真度条件下重现信源信息。但所谓“给定的失真度”都是(在客观条件允许下)人们主观意志所决定,而同时也决定着信息的简约程度。另一方面是信息分子中有强记忆关系的信息原子,例如有多个单字组成的字汇,例中“中华人民共和国”七个字组成字汇,它是一个整体,在传输过程中可仅用相当于一个单字的码长来代表。还有对于有较多重复结构的信息,例如一幅画

中的许多具有相同的色彩和亮度的采样点。在传输过程中也可用“压缩”的办法使总的码长大为减少,从而提高传输效率。以上这些都是狭义信息论的重要研究内容。在应用中已被大量的实践所证实。

4、信息结构的非重复性。一块物质,一般是由许多具有同一分子结构的物质组成。混合物也只是由一些不同物质分子简单集合组成的总体。对于这些物质,我们只要取一块足够小的(在分子数量级上)的物质,其他都是在结构上相同的小块物质的重复而已。这些都是众所周知的。但对信息结构却完全不同。一个信息集合,决不可能是许多同样信息或信息子集的重复所组成。因为重复的信息就不成其为信息。因此形成在信息分子结构上的非重复性。这也是信息结构特性与物质结构特性的又一重要差别。

以上讨论仅仅是从哲学和方法论上对理解“信息量”这一重要概念提供一些作者的看法。对于这一人们似乎天天在应用它,并确实早已成为“经典”的概念。忽然意外发现它对我们通常的“常识”理解之间还有一条缝隙。尽管这一缝隙很容易弥补,即使不去弥补它也不会影响我们使用。但这终究是一个人们认知上的一个小的缺憾。我们是否也可推想在其他问题上也存在类似的缺憾问题呢?

【参 考 文 献】

- [1]冯国瑞.系统论,信息论,控制论与马克思主义认识论[M].北京大学出版社,1991.
- [2]屠德雍,鲍健强.一种新的认知模型和信息加工规律[J].自然辩证法研究,1998(3).
- [3]常迦.信息理论基础[M].北京:清华大学出版社,1993.

(责任编辑 魏屹东)

书 讯 ·

随着科学技术的加速发展和广泛应用,科学技术的“双刃剑”效应也日益明显和尖锐地展现在我们面前。如何正确地认识现代科技的负面作用及其成因,又如何有效地规范和化解科学的负面作用?这正是《科技时代的反思》(作者:詹颂生)一书所着重探讨和解决的主要问题。该书是国家社会科学基金项目“现代科技发展和应用中的负面作用研究”的最终研究成果,由中山大学出版社于2002年8月出版。

该书的核心内容和主要见解体现在以下三个方面:首先,作者以现代科技对人类自身、人类社会和生态环境的负面影响为主线,在收集并分析了大量客观事实和数据材料的基础上,全面而深入地揭示了现代科技负面作用的主要表现形式及其严重后果。其次,作者分别从科技系统的内部机制和外部环境以及理论和现实的不同视角和层面上,详尽分析了现代科技的负面作用形成的主要根源,并在此基础上概括了现代科技负面作用的本质特征。最后,以对现代科技负面作用的成因以及科技发展和应用的现实状况的研究分析为基本依据,作者从指导原则、评审制度、科技法律、科技伦理、技术手段和公众教育等方面,系统地提出并论证了一套旨在有效防范和化解现代科技负面作用的对策性思路和建议。

该书具有选题新颖、视角独特、论据充分以及对策研究针对性等特点,对从事科学技术研究、应用、开发和管理的领导者和实际工作者有较突出的参考价值,对科学技术与社会(STS)、科学学和科学技术哲学等学科领域的研究工作有一定的借鉴意义,它将有助于我们进一步探索科技发展的规律和机制,不断追求并最终达致科学技术与人、社会和生态环境良性互补、协调发展这一理想目标。