作为 Shannon 理论和 Popper 理论桥梁的广义信息公式

鲁晨光 (独立研究者)
Email:：survival99@hotmail.com
个人主页：http://survivor99.com/lcg

摘要：哲学家 K. R. Popper 提出评价科学命题的信息准则。按照这一准则，命题的先验逻辑概率越小，后验逻辑概率越大（经得起检验），信息量和命题的价值就越大；越是把原以为偶然的事件预测为必然，信息量和命题的价值就越大；不可证伪的永真命题不提供信息，因而没有价值。然而，Shannon 信息公式并不能反映 Popper 所说的信息－－语义信息或广义信息。本文介绍了从 Hartley 信息公式到 Shannon 信息公式的演化，分析了 Shannon 信息公式的缺陷。文中兼用逻辑概率和统计概率而不是仅仅使用统计概率，借助集合 Bayes 公式，从经典信息公式推导出广义信息公式－－新的信息公式正巧反映了 Popper 的基本思想，成为从 Shannon 信息论到 Popper 知识进化论的桥梁。同时文中也讨论了概率命题和模糊命题评价问题以及对 Popper 理论的改进。

关键词: Shannon 信息论，Popper 理论，广义信息，语义信息，知识进化，逻辑概率，信息准则

A Generalized Information Formula Bridging Shannon's Theory and Popper's Theory

Lu， Chenguang （independent researcher)
Email:：survival99@hotmail.com
Personal Website: http://survivor99.com/lcg

Abstract: Philosopher K. R. Popper proposed the information criterion for accessing scientific propositions. According to this criterion, the less the prior logical probability of a proposition is, the larger the information or value of the proposition is; the more prior occasional a event that is correctly forecasted definitely is, the larger the information or value of the proposition is; a always true proposition that cannot be falsified does not convey any information and hence is meaningless. However, Shannon's information theory cannot reflect this kind of information, semantic information or generalized information, Popper meant. This paper introduces the evolution from Hartley's information formula to Shannon's mutual information formula, and analyzes the defects of classical information formulas. The paper uses both logical probabilities and statistical probabilities instead of using statistical probabilities only to derives a generalized information formula from the classical information formula with the help of set-Bayesian formula. The new information formula just reflects Popper's primary thought, and becomes a bridge between Shannon's information theory and Popper's theory of knowledge evolution.

Key words: Shannon's information theory, Popper's theory, generalized information, semantic information, knowledge evolution, logical probability, information criterion

1 序言：寻求 Popper 理论和 Shannon 信息论之间的桥梁

关于科学理论的进步标准，Popper 写道：“凡是包含更大量的经验信息或内容的理论，也即在逻辑上更有力的理论，凡是具有更大的解释力和预测力的理论，从而可以通过把所预测的事实同观察加以比较而经得起更严格检验的理论，就更为可取。总之，我们宁取一种有趣、大胆、信息丰富的理论，而不取一种平庸的理论。”【1】可见，Popper 是用信息作为准则评价科学命题的价值的。根据 Popper 的证伪理论，命题在逻辑上越容易被证伪（先验逻辑概率越小），而事实上经得起检验（后验逻辑概率越大），命题提供的信息就越多，就越有意义。反之，在逻辑上不能被证伪的永真命题不含有信息，没有科学意义。

虽然 Popper 也很早就谈论概率和信息，但是，Popper 并没有提供自己的信息公式。而 Shannon 把概率和信息联系在一起，联系电子通信，成功地创立了 Shannon 信息论【2】——它是经典信息论的核心。然而， Shannon 信息公式并不能用来度量 Popper 所说的信息， 比如天气预报的信息。原因何在？ 原来在 Shannon 信息论中，命题的含义不被考虑。比方说，两个气象预报员， 一个总是把有雨预测为“无雨”（总是报错了）， 另一个相反（总是报对的），但是按 Shannon 互信息公式， 两个人提供的信息一样多。这是荒唐的。

正是由于这个原因， Shannon 反对把自己的理论应用于需要考虑语义的场合。这说明 Shannon 有自知之明。然而，Shannon 的教条主义继承者们因此反对任何度量语义信息的研究，以至于 IEEE 从来不发表研究语义信息的文章。

我于 1988 年开始研究广义信息，1993 年开始陆续发表专著《广义信息论》【3】和相关文章【4,5】。 我的研究表明，经典信息公式的一点小小改动就可以使其威力大增——不仅可以度量统计信息， 也可以度量语义信息或广义信息。新的信息公式完全可以反映 Popper 的信息准则，成为从 Shannon 理论到 Popper 理论的桥梁。

2 从 Hartley 信息公式到 Shannon 互信息公式

Hartley 信息公式【6】是

$$I = \log N \tag{1}$$

其中 I 表示确定 N 个等概率事件中的一个出现时提供的信息。假如一个学校有 N_1＝512 名学生。已知你要找的肇事学生在这个学校， 但不知道是谁。要确定他是谁， 需要 I=log(512)=9 比特的信息。

如果事件 y 把不确定范围从 N_1 个缩小为 N_2 个，那么信息就等于

$$I_r = \log N_1 - \log N_2 = \log \frac{N_1}{N_2} \tag{2}$$

我们且称这个公式是 Hartely 相对信息公式。假设肇事学生所在班学生数目是 N_2＝32。 如果有人告诉那个肇事学生在某个班，没告诉具体是谁， 那么信息就是 I_r=log(512/32)=4 比特。

下面我们用一种易于理解的方式推导出 Shannon 互信息公式。

用 Hartley 公式计算信息，要求 N 个事件是等概率的，即 P=1/N，但是通常的情况并不如此。这时候我们用实际的概率代替假设的相等的概率，即用 P_1 代替 1/ N_1，用 P_2 代替 1/ N_2。于是，Hartley 信息公式就变为：

$$I=\log(1/P)$$

(3)

上面的相对信息公式就变为：

$$I_r = \log N_1 - \log N_2 = \log \frac{P_2}{P_1}$$

(4)

我们把客观事件集合 $A=\{x_1, x_2, ...\}$ 中的一个表示为 x_i；把消息(message)集合 $B=\{y_1, y_2, ...\}$中的一个表示为 y_j，那么 y_j提供的关于 x_i的信息就是：

$$I(x_i; y_j) = \log \frac{P(x_i \mid y_j)}{P(x_i)}$$

(5)

其中 $P(x_i)$是 x_i发生的先验概率，$P(x_i|y_j)$是 y_j发生后 x_i的条件概率。因为有 Bayes 公式

$$P(y_j \mid x_i) = \frac{P(x_i \mid y_j)P(y_j)}{P(x_i)}$$

(6)

所以有

$$I(x_i; y_j) = \log \frac{P(x_i \mid y_j)}{P(x_i)} = \log \frac{P(y_j \mid x_i)}{P(y_i)}$$

（7）

我们用 X表示事件变量， 用 Y表示消息变量， 那么 y_j提供关于不同 X的平均信息就是

$$I(X; y_j) = \sum_i P(x_i \mid y_j) \log \frac{P(x_i \mid y_j)}{P(x_i)}$$

（8）

这个公式和 Kullback 公式形式相同，我们且称之为 Kullback 信息公式。 进一步求平均，我们得到 Y提供关于 X的平均信息

$$\begin{aligned} I(X;Y) &= \sum_j \sum_i P(y_i)P(x_i \mid y_j) \log \frac{P(x_i \mid y_j)}{P(x_i)} \\ &= H(X) - H(X \mid Y) \end{aligned}$$

（9）

其中 $H(X)$和 $H(X|Y)$分别是 Shannon 熵和 Shannon 条件熵：：

$$H(X) = -\sum_i P(x_i) \log P(x_i)$$

$$H(X \mid Y) = -\sum_j \sum_i P(x_i, y_j) \log P(x_i \mid y_j)$$

（10）

Shannon 互信息公式对于优化电子通信威力巨大。$H(X)$反映了无失真编码的平均码长；给定信道 $P(Y|X)$,改变信源使 $I(X;Y)$达最大， 这个最大值就是信道容量；给定信源 $P(X)$和平均误差上限 D, 改变信道 $P(Y|X)$,求最小互信息 $R=R(D)$（R是 D的函数）， 这个 R就反映信号传输的最短比特或必要的最小速率。

Shannon 不是通过方面步骤推导出互信息公式的， 而是先推导出熵 $H(X)$－－它反映信源无失真编码的最短平均码长－－和条件熵 $H(X|Y)$,再推导出互信息公式的。因为 Shannon 从来不谈单个事件的信息量，公式(5)也是后人从他的互信息公式中提取出来的

【7】。公式(5)可能算出负的信息－－比如：谎言使你要找的学生范围更大，信息就是负的。而负的信息是 Shannon 及其教条主义继承者不能接受的。

正是公式(5),给我们提供了推广 Shannon 信息论的突破口！

3 命题的逻辑概率及集合 Bayes 公式

经典信息论用到的概率可谓统计概率，而反映命题真假的概率是逻辑概率。我们用股市指数预测为例来说明两种概率的区别和联系(参看图 1)。

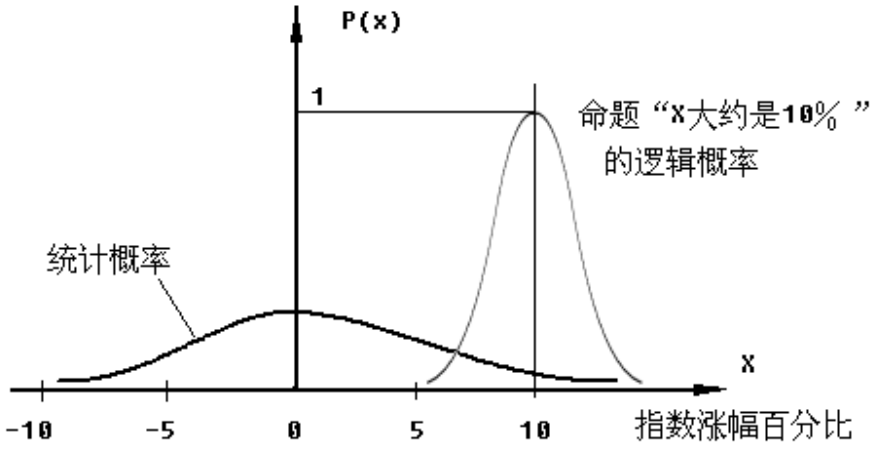

图 1 统计概率和逻辑概率比较

我们可以统计某个股市多年来日升跌幅的频率， 当天数大到一定程度后，相对频率（在 0 和 1 之间）就趋向某个确定值，这个值就是概率或统计概率。命题的逻辑概率是事实 x_i 一定时，命题 y_j 或 $y_j(x_i)$被不同的人判断为真的概率。比如命题 y_j=“指数升幅接近 10％”在指数实际升幅 X=10%时被判定为真的逻辑概率是 1；随着误差增大，逻辑概率渐渐变小；误差超过 5%时逻辑概率接近 0。

从数学的角度看，两种概率的区别是：对于 $P(x_i)$， 一定有

$$\sum_i P(x_i) = 1 \tag{11}$$

而逻辑概率 $Q(y_j$为真$|x_i)$的最大值是 1，求和之后一般大于 1。

假设使 y_j 为真的所有 X 构成模糊集合 A_j，逻辑概率 $Q(y_j$为真$|x_i)$就是 x_i 在 A_j 上的隶属度，那么我们也可以用 $Q(A_j|x_i)$表示 x_i 在 A_j 上的隶属度和命题 $y_j(x_i)$的逻辑概率。

命题 y_j 对于不同 X 的平均逻辑概率（也就是谓词 y_j(X)的逻辑概率）是：

$$Q(A_j) = \sum_i P(x_i) Q(A_j \mid x_i) \tag{12}$$

已知命题 y_j 为真，x_i 发生的概率是：

$$P(x_i \mid A_j) = \frac{Q(A_j \mid x_i) P(x_i)}{Q(A_j)} \tag{13}$$

这就是集合 Bayes 公式－－以事件 $x_i \in A_j$ 为条件的 Bayes 公式（证明见【3】，3.1 节）。如果条件 $x_i \in A_j$ 即 y_j 为真变为 y_j, 它就变成经典的 Bayes 公式。

4　广义信息公式－－经典信息公式的微妙改进

我们把相对信息公式（5）中的条件 y_j 改为 y_j 为真，即用 $x_i \in A_j$ 取代 y_j，那么公式(5)就变为：

$$I(x_i; y_j) = \log \frac{P(x_i \mid A_j)}{P(x_i)} \tag{14}$$

根据上式和集合 Bayes 公式(13), 我们得到

$$I(x_i; y_j) = \log \frac{Q(A_j \mid x_i)}{Q(A_j)} \tag{15}$$

其含义是：

$$\text{命题} y_j \text{为真在} x_i \text{发生时提供的信息} = \log \frac{\text{命题} y_j(x_i) \text{的逻辑概率}}{y_j(X) \text{的平均逻辑概率}} \tag{16}$$

其几何意义可由图 2 看出。由图可见，事实和预测完全一致时，即 x_i=x_j,，信息量达最大；随着误差增大， 信息量渐渐变小； 误差大到一定程度信息就是负的。这是符合常理的。

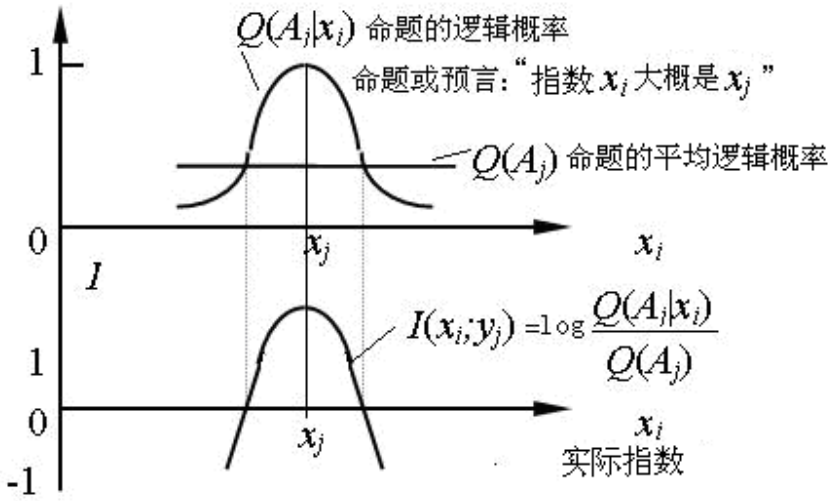

图 2 广义信息公式图解

这一公式同样可以度量测量信息，一般预言信息， 感觉信息【3】。

5　广义信息公式如何反映 Popper 思想

Popper 认为，所有科学命题都是猜测而不能证实，经得起检验的命题就是好命题，检

验越严格越好。在我的广义信息论中，最一般信息是预测信息，正确描述事实的命题信息是预测和事实相一致时的特例; 预测越是出人意外并且正确, 信息就越多。两种理论是一致的。

Popper 认为，永真命题不含有信息，因为无法检验或在逻辑上不可证伪，因而没有科学价值。根据上面广义信息公式有非常一致的结论：如果命题的逻辑概率总是 1，那么其平均逻辑概率也是 1， 信息 I 就总是 0。

Popper 一再强调，命题的逻辑概率越小，如果经得起检验， 其价值越大。

广义信息公式(15)的更多性质如图 3 所示。我们也可以把图中预测理解为对物价指数、降水量、温度……或产量的预测。道理同样。三个山型曲线表示三个不同预测－－用以比较精确预测（虚线表示）和模糊预测， 对更偶然事件的预测（右边曲线）和对不很偶然事件的预测。

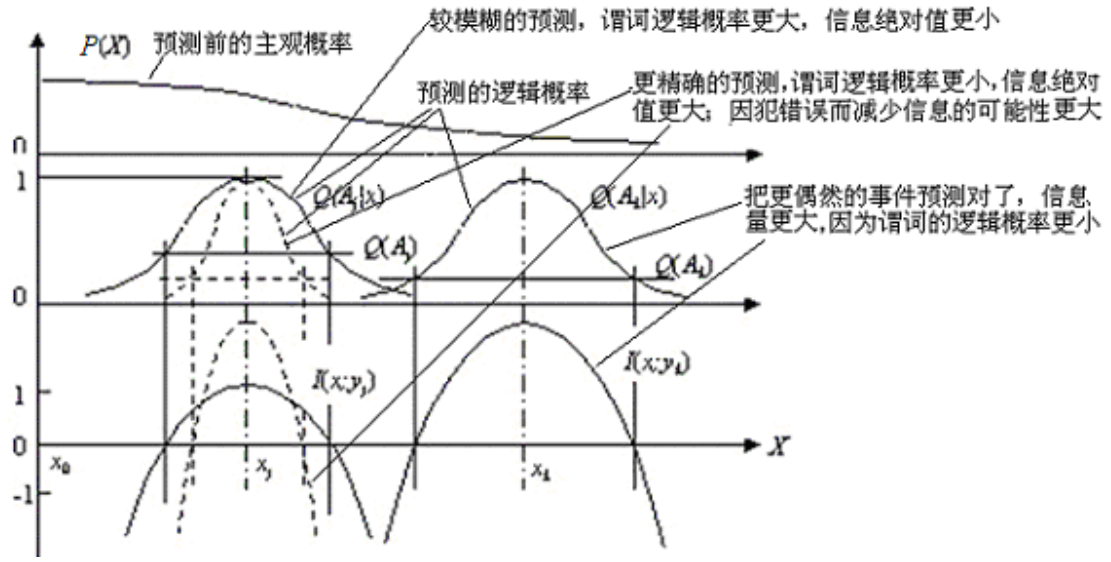

图 3 以股市指数预测为例说明广义信息公式的性质

图中显示：精确预测(逻辑概率曲线分布较窄)的信息绝对值更大，但精确预测因为容易出错而更容易导致负信息。精确预测也就是 Popper 说的逻辑概率更小的预测。Popper 认为，逻辑概率小的事件在逻辑上更容易证伪，如果它事实上没有被证伪，或者说经得起检验，它的意义就更大，就更具有科学价值。上面信息公式和 Popper 观点是一致的。

图中还显示，把更偶然事件（比如股市涨幅较大的事件）预测准了信息量更大。更偶然事件的概率小， 由图 1 和公式(12)可以看出相应的逻辑概率也小。

Popper 肯定，全称命题只要有一个反例就足以被证伪。广义信息公式也有同样结论。因为当命题的逻辑概率 $Q(A_j|x_i)=0$ 而 $Q(A_j)>0$ 时，信息为负无穷大。只要有一个 x_i 使得 $I(x_i; y_j)$=负无穷大，那么对 $I(x_i;y_j)$求平均得到的平均信息 $I(X; y_j)$(参看式(17))也是负无穷大。因而命题不可取。

6 关于概率命题和模糊命题的评价－－Popper 理论改进

如何评价概率命题（比如：“如果气压过低，天十有八九要下雨”）和模糊命题（比如：“明天有中到大雨”）？Popper 没有很好办法。而现实中和社会科学中，大多数预测

和命题都是概率的或者是模糊的。在这种情况下，个别反例并不能证伪或完全证伪一个命题。而广义互 Kullback 公式和广义互信息公式可以作为这类命题的合适评价准则。

对公式(14)求平均可以得到广义 Kullback 公式：

$$I(X;y_j)=\sum_i P(x_i\mid y_j)I(x_i;y_j)=\sum_i P(x_i\mid y_j)\log\frac{P(x_i\mid A_j)}{P(x_i)} \tag{17}$$

其中对数后面的 $P(x_i)$也可以不是来自统计，而是来自主观估计，这时候我们记它为 $Q(x_i)$，这时 $P(X|A_j)$变为 $Q(X|A_j)$。广义 Kullback 信息公式的性质可以通过图 4 说明。

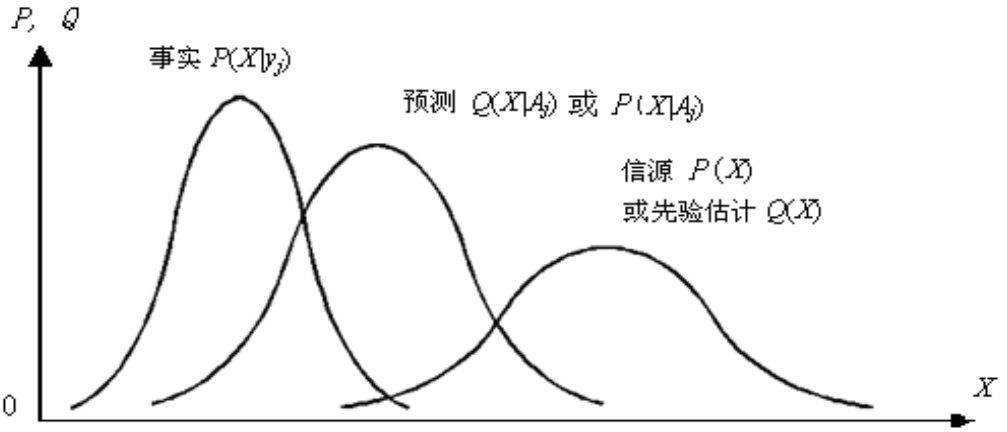

图 4 广义 Kullback 公式性质图解

可以证明，预测 $Q(X|A_j)$和事实 $P(X|y_j)$越是重合，信息量越大， 信源 $P(X)$或先验估计 $Q(X)$越是与事实 $P(X|y_j)$不同(表示预测的事情越是出乎意外)，信息量越大。这和 Popper 的科学进步评价准则是一致的。只是全称命题变为概率命题，清晰命题变为模糊命题， 其适用范围更广。

当 $Q(X|A_j)=P(X|y_j)$时，表示预测正确，上面公式就变为 Kullback 公式。这时信息 $I(X;y_j)$达最大。

对 $I(X;y_j)$求平均，我们得到广义互信息公式

$$I(X;Y)=\sum_j\sum_i P(x_i,y_j)\log\frac{Q(A_j\mid x_i)}{Q(A_j)} \tag{18}$$

它反映一组可选命题在不同事实发生时的平均信息，它与广义 Kullback 公式类似，和 Popper 基本思想兼容，但是更适合概率命题和模糊命题评价。

7 广义信息公式和广义互信息公式的应用

经济数值预测通常使用误差准则和均方误差准则， 广义信息准则与之相比，对小概率事件的正确预测评价更高。 比如对于均误差准则，总是预测股市涨幅为 0， 它的均方误差较低， 但是信息极少。而用广义信息准则，把涨幅 10％预测准了得分更高－－和把涨幅为 0%或为 1％预测准了相比。

广义信息准则也可以评价模糊预测－－比如：“今年股市指数将再涨 30%到 40％”，“明天有小到中雨”。 对于这些预测，用误差准则评价要么不方便－－因为没有中心点， 要么不合理－－如果假设有许多中心点的话。后者鼓励模糊的信息极少的预测。

广义互信息公式对优化广义通信也有重要意义。经典信息率失真理论研究的是电子通信优化问题：给定信源和失真限制，Shannon 信息（传输的比特）最少需要多少？或者是，在给定 Shannon 信息时，失真范围最小可达多少。可惜那里失真是人为定义的，没有客观标准。现在有了广义信息公式，我们就可以用广义信息量替代失真量作为评价通信质量的准则，把信息率失真论改进为保精度信息率论，由此可以得到许多重要结论，比如：听信算命先生信口开会减少我们已有的信息；要用谎言迷惑敌人，也需要一定的客观信息；人眼感官分辨率和图像精度之间存在最优匹配，视觉分辨率有限，图像精度太高反而不好【3-5】。

8 结束语

现在我们看到，Shannon 信息公式一点小小的改动就大大增强公式的解释力，并且使 Popper 的知识进步准则有了严格的数学形式。这无论是对 Shannon 理论还是对 Popper 理论，都使最好的继承。作为两种理论的桥梁，广义信息公式的更多应用有待进一步探讨。

以广义信息公式为基础的广义信息论不仅支持 Popper 的哲学理论， 还促使我们从新的角度反省哲学基本问题(参看【9-10】以及我的主页 http://survivor99.com/lcg) 。

参考文献

1 ［英］波普尔，付季重等译，猜想和反驳——科学知识的增长，上海译文出版社，1986

2 Shannon, C. E. A mathematical theory of communication，*Bell System Technical Journal*,*27* (1948)， 379—429，623—656

3 鲁晨光，广义信息论，中国科学技术大学出版社，1993

4 鲁晨光, 广义熵和广义互信息的编码意义, 通信学报, 5, 6(1994), 37-44.

5 Lu, Chenguang (鲁晨光) “A generalization of Shannon's information theory" , Int. J. of General Systems, 28: (6) 1999，453-490

6 Hartley, R. V. L. Transmission of information, *Bell System Technical Journal*，7 (1928)，535

7 ［英］罗斯，钟义信等译． 信息与通信理论，人民邮电出版社，1978

8 Kullback S. *Information and Statistics*，John Wiley & Sons Inc.，New York，1959

9 鲁晨光， 色觉奥妙和哲学基本问题，中国科学技术大学出版社，2003

10 鲁晨光，摒弃信息哲学中的朴素反映论思想，http://survivor99.com/lcg/books/GIT/bq.mht