[**快速排序里的学问：信息熵**](http://www.nowamagic.net/librarys/veda/detail/2389)

信息是个很抽象的概念。人们常常说信息很多，或者信息较少，但却很难说清楚信息到底有多少。比如一本五十万字的中文书到底有多少信息量。直到1948年，香农提出了“[信息熵](http://www.nowamagic.net/librarys/veda/tag/%E4%BF%A1%E6%81%AF%E7%86%B5)”的概念，才解决了对信息的量化度量问题。

一条信息的信息量大小和它的不确定性有直接的关系。比如说，我们要搞清楚一件非常非常不确定的事，或是我们一无所知的事情，就需要了解大量的信息。相反，如果我们对某件事已经有了较多的了解，我们不需要太多的信息就能把它搞清楚。所以，从这个角度，我们可以认为，信息量的度量就等于不确定性的多少。

香农指出的信息熵的计算公式如下（本文的对数一律以2为准）：

H(x) = -∑p(xi)log(p(xi)) (i=1,2,..n)    （其中p(x)是x事件出现的概率）单位为bit

在数学之美里是用赛后怎么知道32个球队里谁是冠军来讲解了这个信息熵的概念。

当概率相等时，每次询问用折半查找的原理（如“冠军队伍在1-16吗？”）可以减少一半的队伍，这样就需要5次就能知道结果了。这里就是log32 = 5。

而使用信息熵计算信息量，的确也是5。但是为什么信息熵这个公式会代表信息量呢

按我的理解，在等概率事件里，1/p(x) 代表那一次所有可能出现的量、在球队问题里，就是32种可能性。

而等概率事件里，因∑p（xi） = 1，所以信息熵可以看成：

信息熵H(x)= -∑p(xi)log(p(xi)) (i=1,2,..n) = -log(p(i)) = -（- log（1/p(x)））= log（1/p(x)）

也就是说等概率事件里的信息量可以看成：

H（x）= log（所有可能性）

为了加深对信息量的定义的理解，再回到上述32个球队的问题，我们已经知道他的信息量是5Bit。

问过一次之后，我们可以知道冠军在哪16个队伍中，也就是说我们获得了1bit的信息后不确定性减少，等于信息熵变成了log16 = 4bit =5bit -1bit 。

而最大熵模型呢？它的原理就是保留全部的不确定性，将风险降到最少。

最大熵原理指出，当我们需要对一个随机事件的概率分布进行预测时，我们的预测应当满足全部已知的条件，而对未知的情况不要做任何主观假设。（不做主观假设这点很重要。）在这种情况下，概率分布最均匀，预测的风险最小。因为这时概率分布的信息熵最大，所以人们称这种模型叫“最大熵模型”。

我们常说，“不要把所有的鸡蛋放在一个篮子里”，其实就是最大熵原理的一个朴素的说法，因为当我们遇到不确定性时，就要保留各种可能性。

也就是说发现不确定信息的时候，不要对不确定的产物任何主观假设使他们的概率分布均匀，则能获得最客观的结果。而这时风险会最小，我们就可以用这个结果来进行最客观的决策。数学上来说就是最优下界。

这种策略的本质可以概括成“让未知世界无机可乘”。它是没有“弱点的”，答案的任何一个分支都是等概率的。反之，一旦某个分支蕴含的可能性更多，当情况落到那个分支上的时候你就郁闷了。二分搜索为什么好，就是因为它每次都将可能性排除一半并且无论如何都能排除一半（它是最糟情况下表现最好的）。

我再用算法的时间[复杂度](http://www.nowamagic.net/librarys/veda/tag/%E5%A4%8D%E6%9D%82%E5%BA%A6)说明一下最大熵原理吧，用几个主流的算法对n个数据进行排序时间复杂度基本上都是从O(nlogn)到O(n2)。而一般情况下为什么O(nlogn)最优呢（透露下，快速排序的平均时间复杂度就是O(nlogn)），因为n个数据的先后顺序是随机的，我们可以看做不确定性相等，则可以用最大熵原理获得最优(最稳定)结果。则信息熵则为：

H(x)= log(所有可能性)= log(n!) 而n->00 则log(n!) 近似于lognn= nlogn

假设我们每次能获得1bit数据，就至少需要获得(nlogn)bit数据才能取消信息的不确定性，也就是要比较nlogn次。但因为各种排序算法策略不同，我们不可能每次都能获得1bit数据，所以按照信息熵的定义这是理论上最优的结果。而最佳的排序算法就是要每次获得1bit数据，越接近于1则越有效。

虽然快排和堆排序两个是都是时间复杂度O(nlogn)的算法，但是快速排序一般都会比堆排序快，就是因为堆排序每次获取的平均信息量比快排来的低。

而上面，我们根本没提到具体算法，就算到了最优的时间复杂度。在实际生活中很多时候我们虽然不会想到具体的策略，但我们至少可以知道极限在哪里，可以知道还有没有提高余地。任何排序和猜数字的算法可以理解为通过获得信息量去消减原来的熵。