정보흐름의 양적 표현

陳 庸 玉
〈경희대교수 · 통신공학 / 본지편집고문〉

1. 정보량의 정의

"이미 알고 있는 내용은 정보가 될 수 없다."이는 너무나 당연한 말이지만 이처럼 정보의 성질을 가장 극명하게 설명하는 말도 없을 것이다. 한 예를 들어보자 "오늘은 공휴일인데, 눈이온다"라고 말하면 이러한 상태는 얼마만한 정보가 포함되어 있을까? 일년중에 토요일을 포함해서 공휴일이 1/4이라고 가정하자(계산상 편의 때문임). 또한 맑고, 흐리고, 비오고, 눈오는 날이 일년중 각각 1/2, 1/4, 1/8, 1/8이 발생 한다고 하자 "오늘은 공휴일이다"라는 내용은 2비트의 정보량을 가진반면(계산방법 Box 1 참조) "눈이 온다"라는 내용은 3비트의 정보량을 가지고 있다. 따라서 "공휴일"이라는 사실과 "눈이 온다"라는 2개의 상황이 일어나는 경우에는 5비트의 정보량을 가지고 있다. 2개의 정보량을 합산한 것은 이들 상태가 각각 별개로 일어나기 때문이며, 이를 独立事象(Independent Event)이라 할 수 있다. 따라서 확률은 곱해지고 정보량은 합해진다. 만약 "비가 온다"고 했을 때도 똑같이 5비트 이지만 "흐린다"고 했을 때는 2비트가 합산되어 4비트, "맑다"고 했을 때는 1비트가 합산되어 3비트가 된다.

위와 같이 발생 확률이 높을수록 정보량은 줄어듬을 알수가 있다. 즉 발생확률이 높을수록 미리 내용을 알아버렸기 때문에 얻을 것이 없다는 뜻이 된다. 바꾸어 말하면 알고 있는 내용은 정보가 아니다 라는 명제가 성립한 것이다.

이들 설명내용은 〈그림1〉과 같다.

〔BOX 1〕
1. 정보량의 계산식
   1) 동전을 던졌을 때 앞면이 안나오면 뒷면이 나올 것이고 그확적 확률은 1/2이다. 만약 동전을 던져서 앞면이 나왔을 때를 생각하면 던지기 전의 확률이 1/2로써 애매하던 것이 그 확률이 1 로 증가되어 의심할 여지없이 확실해진다.
   이를 정보의 양적 관점에서 보면 초기사상(Event)의 확률이 적은 것으로부터 큰 것으로 변화시켰을 때 얻어지는 양으로써 판단할 수가 있다. 다시말하면 정보의 양적 표시는 확률의 변화비율로 표시할 수가 있다. 이러한 확률변화율을 대수함수적(log)인 관계로 표시하면,

$$I = \log \frac{P_o}{P_i} \cdots\cdots\cdots (1)$$

   가 된다. 여기에서 P_i는 초기상태의 확률이며, P_o는 최종상태의 확률이다. 즉, 최종상태의 확률은 결국은 그 확률이 1이 될 것이므로 정보량의 식은

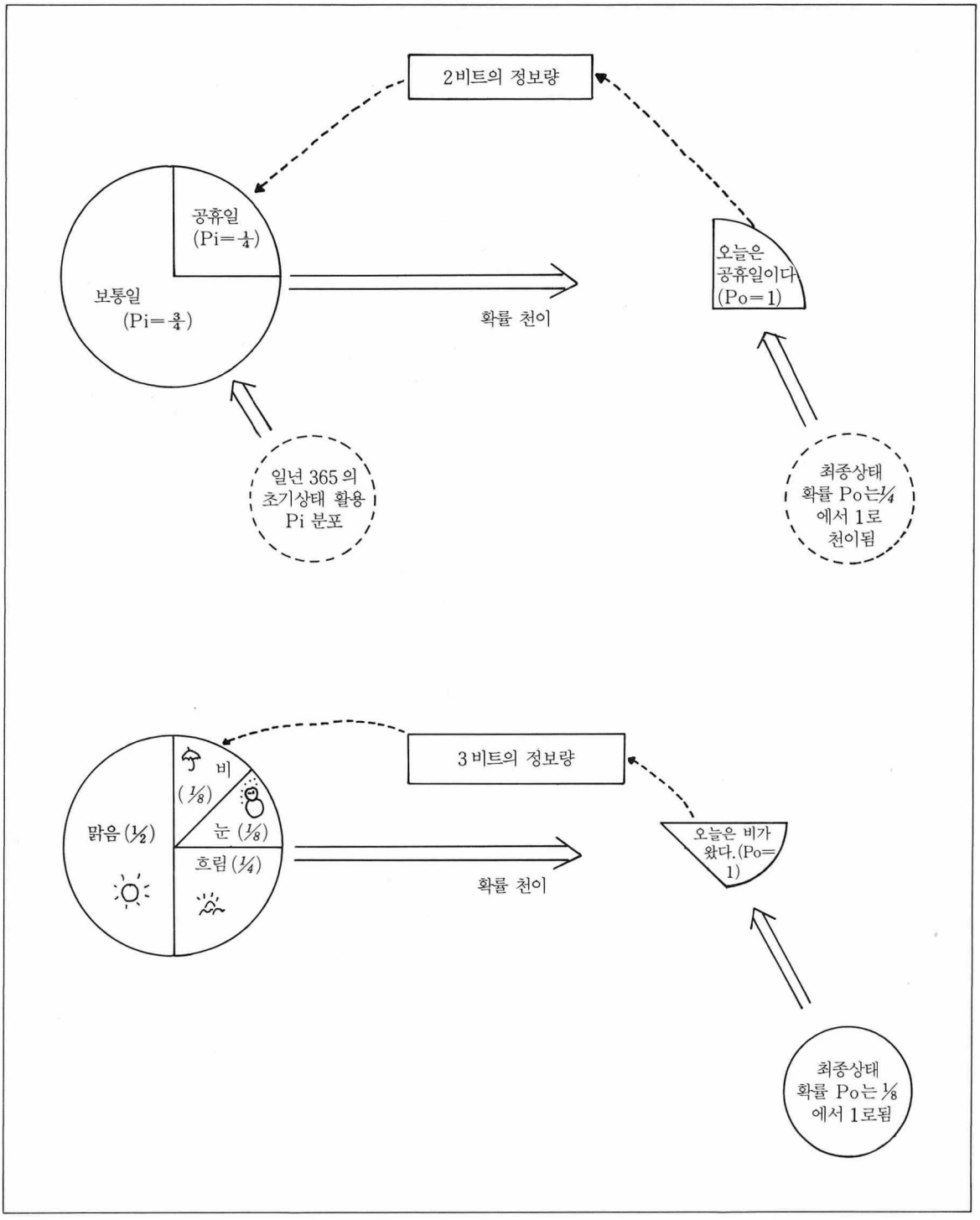
$$I = -\log P_i \cdots\cdots\cdots\cdots (2)$$

   가 된다.
   2) 이때 대수의 밑을 2로 할때 비트(Binary Unit)라는 단위로 규정하고 있다. 앞에서 예를든(상용대수일 때는 Decit이고 자연대수일 때는 Nat이다.) 동전의 경우는

$$I = -\log_2 \tfrac{1}{2} = -\log_2 1 + \log_2 2 = 0 + 1 = 1\text{비트}$$

   가 되며 2개의 동등한 출현 확률인 것으로 부터 양자택일로 얻어지는 정보량을 의미한다.

〈그림1〉 정보량의 성질과 양적표현

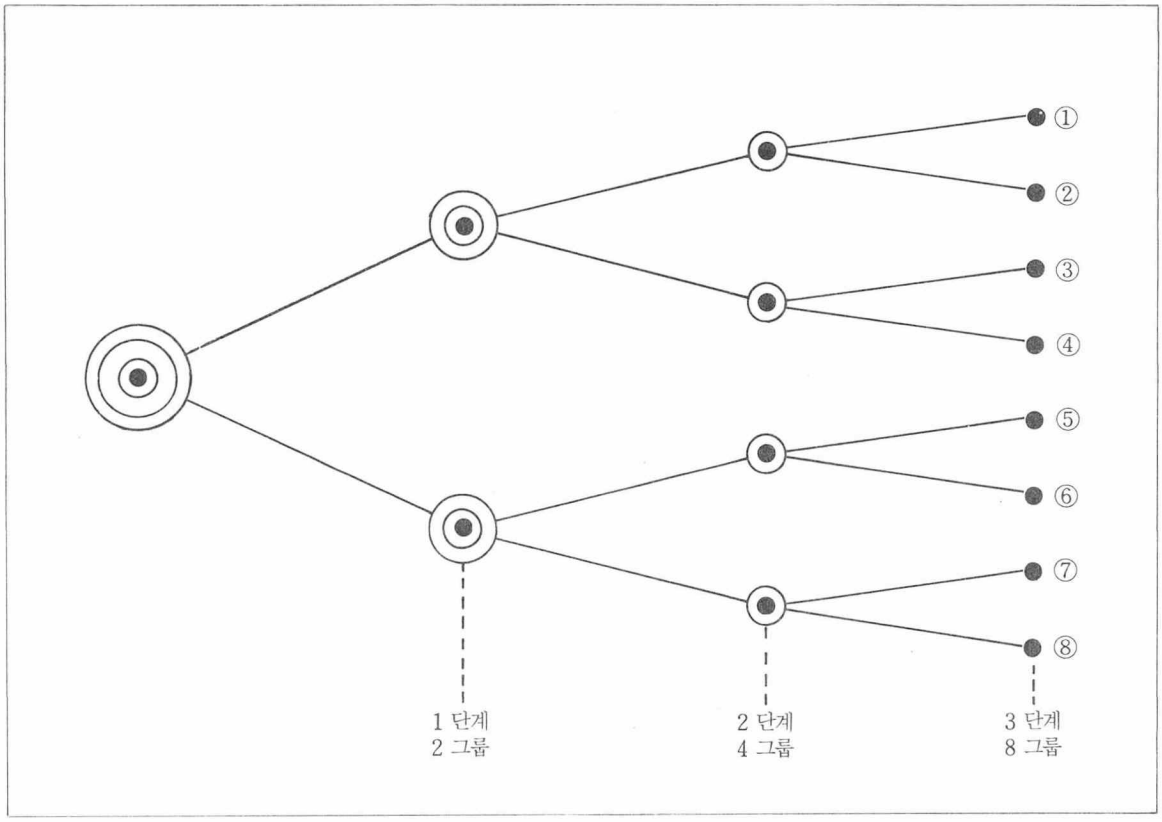


또 다른 예를 들어보자

가운데가 비어있는 구절판에 8가지의 각각 다른 다과가 들어 있다고 하자. 이 8가지의 다과의 종류를 구분해 보려면(즉, 정보를 알아내려면), 먼저 첫단계에서 4개로 2등분 시킨다음 이를 다시 2등분하여 4개의 그룹으로 나누고 3단계에서는 하나씩 분별해보면 된다. 결국 3번의 양자택일(2진판별) 과정을 실시하면 8개의 모든 다과의 종류를 확인할 수가 있다. 위의 경우 앞의 수식을 이용하여 정보량을 구하면 3비트가 된다. 이렇게 볼 때 정보량이란 「양자택일의 회수」와 동일한 의미가 된다고 하겠다.

이 과정을 〈그림2〉에 표시하였다.

〈그림2〉 3단계를 거친 3비트 정보량의 표현원리

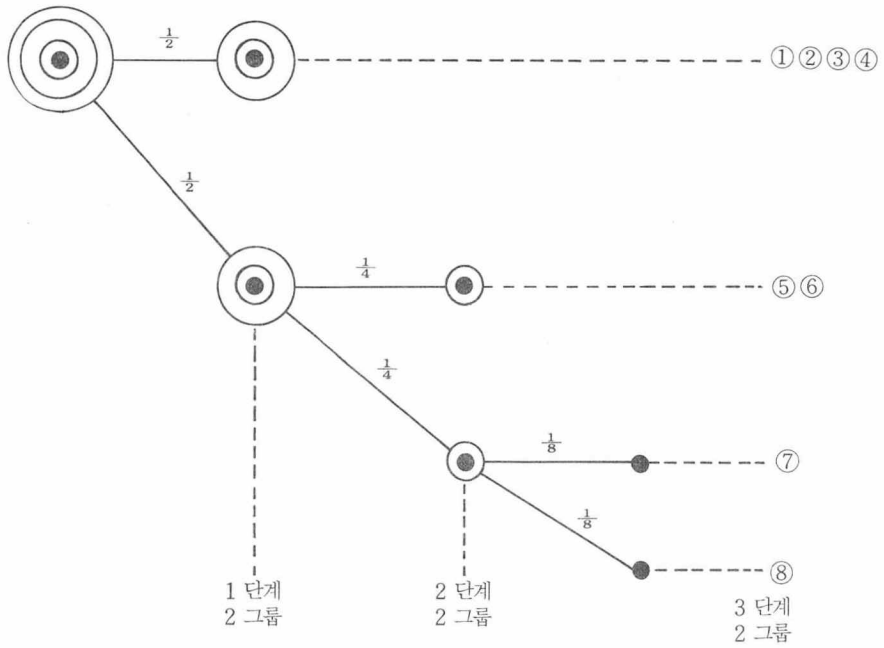


2. 평균정보량 : 엔트로피 (Entropy)

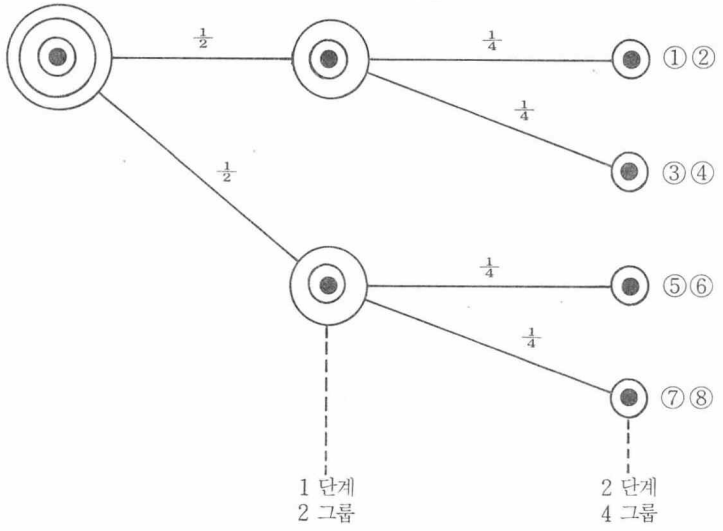
그러나 약과가 4개(확률$\frac{1}{2}$), 강정이 2개(확률$\frac{1}{4}$), 약식이 1개(확률이$\frac{1}{8}$), 유과가 1개(확률$\frac{1}{8}$) 등 모두 4종류 8개의 다과가 들어 있을 경우와 4종류가 각각 2개씩 있을 경우(평균확률$\frac{1}{4}$)에는 사정이 달라진다. 전자의 경우에는 종류가 같은 4개를 먼저 구분하고 그 다음 2개를, 그리고 최종으로 약식과 유과를 구분하면 4종류 8개의 다과가 있음을 알 수 있다. 즉, 약과를 고르는 것은 1단계에서 4개에 해당하므로 도합 4번, 강정은 2단계에서 구분 가능하므로 도합 4번, 약식과 유과는 각각 3단계에서 구분이 가능하므로 도합 6번의 양자택일을 한다. 평균하면 1개에 대해 1.75번의 양자택일의 과정이 필요하다. 그러나 후자의 경우에는 1단계에서는 그룹2로 구분하는데 8번, 2단계에서는 4그룹으로 구분하는데 8번씩 행하면 4종류 모두 구분된다. 따라서 모두 16번의 양자택일을 하고 이를 평균하면 2회 시행하는 꼴이 된다. 이와 같은 과정을 그림3에 표시하였다.

〈그림3〉 엔트로피 (평균정보량)의 변화과정

(a) 엔트로피 1.75비트인 경우: 분포가 편중되었을 때



(b) 엔트로피 2비트인 경우: 분포가 고르게 되었을 때



〈그림2〉와 〈그림3〉을 복잡한 정도로 나열한다면 그림2, 그림3-b, 그림3-a 순이다. 이 순서는 종류(상태)가 같을 때는 상태가 많을 때(그림2와 같이 발생 빈도가 낮다) 더 복잡하고, 종류가 일정할 때에는 분포가 편중되었을때 보다는(그림3-b와 같은 등확률분포) 골고루 분포(그림3-a와 같은 편증분포)되었을 때 더 복잡하게 된다. 즉 복잡성이 증가할 때마다 평균 정보량도 많아진다.

이와 같은 평균정보량을 엔트로피라고도 하는데, 여러 단계의 선택과정에서 각 단계의 평균한 값이 됨을 알 수 있다. 즉 그림2에서는 8개 모두가 3단계의 양자택일을 거치므로 평균하면 엔트로피는 3비트가 되지만 같은 8개가 있더라도 그림3-a에서와 같이 ①, ②, ③, ④는 1단계, ⑤와 ⑥은 2단계, ⑦과 ⑧은 각각 3단계를 거치므로 평균하면 $(4+4+6)\div 8=1.75$단계, 즉 1.75 비트가 된다. 그러나 그림3-b에서는 이같이 2단계를 거치므로 총 16단계를 거치고, 평균 2단계 즉 2비트의 엔트로피를 가진다는 뜻이다.

이렇게 볼 때 엔트로피란 「한개의 선택에 필요한 양자택일의 평균회수」를 나타내는 것이라 할 수 있으며, 회수가 많을수록 복잡한 단계를 거쳐야 한다. 이때 복잡한 단계란 쉽게 구별하기 어려운 무질서하게 나열되어 있는 상태를 뜻하는 것이다.

엔트로피란 원래 열역학에서 나온 개념으로 기본적으로 무질서의 정도(Degree of Randomness)라는 의미를 가지고 있다. 정보이론에서는 앞에서 살펴본 바와 같이 복잡한 단계 즉 무질서의 정도라는 점에서 동일한 의미로 사용할 수가 있다.

이러한 엔트로피(평균정보량)의 개념을 문자에 적용하여 보면, 한글의 경우 스페이스까지 포함할 때 25소자가 된다. 따라서 동일한 확률로 될때 4.644비트/자소가 된다. 실제 발생빈도로 평균할때는 4.0767비트이다(Box2 참조).

〔Box 2〕

엔트로피에 대한 정의는 $H(x)=\sum_{i=1}^{N} P\ I(x)$. 따라서 $H(x)=-\sum_{i=1}^{N} Pi\ \log_2 Pi$이다.

한글은 공백(스페이스)까지 포함할때 25자소라 할 수 있고 각 자소가 동확률로 발생한다면(확률$=\frac{1}{25}$) 엔트로피는 $\sum_{i=1}^{25} \frac{1}{2}\log_2 25=4.644$비트/자소가 된다. 영문은 log 27=4.75비트/자소가 되여, 26자의 서양문자는 모두 같다. 이때 한글의 공백소와 영문의 공백소는 단어를 구분한다는 면에서 그 역할은 같으나 자소만을 취급할 때와 자소의 조합으로 된 문자만을 상대로 할 때는 성질이 달라진다. 따라서 자소만을생각할 때 공백소를 제외하는 경우 엔트로피는 log24=4.58비트/자소기 된다.

그러나 실제로 한글자소가 발생하는 확률은 차이가 있어서 공백소가 0.113, 0(0.1056), 1(0.0953)…… 등이고 F가 제일 출현빈도가 희박해서 약 0.0011 정도이다. 이들의 실제 엔트로피 H(x)는 4.0767 이며 영어의 경우 공백소(0.2), E(0.105), T(0.072)…… Q와 Z이 0.001 정도이며 엔트로피는 4.053이다.

즉, 한글 한자소를 이해하려면 평균 약 4단계 또는 5단계의 양자 택일의 관계를 거쳐야 한다. 이는 5단위 부호체계를 구성해야 함을 뜻한다. 5단위 부호체계에서는 $2^5=32$ 상태가 표현 가능하다(이에 비해 일본 문자는 50음이므로 7단위 체계가 필요하다).

3. 군더더기율 (Redundancy)

등확률 엔트로피와(이를 최대 엔트로피라 한다) 실제 발생확률 엔트로피와의 비를 상대 엔트로피(Relativ Entropy : box4 참조)라 하는데 앞의 예에서 한글은 0.88, 영어인 경우 0.8522이며 항상 1보다 작은 값이다. 1에서 뺀 나머지 0.12와 0.15의 값을 군더더기율(Redundancy : box4 참조)이라 하는데 이는 12%와 15%의정보손실이 있어도 해독이 가능하다는 뜻이다. 이 수치는 자소를 기준으로 한 것이다. 영어의 경우 군더더기 율은 거의 50% 정도라고 한다. 이에 비해 한글의 경우에 는 음절(평균 2.64개의 자소가 모여 음절을 이룬다) 이는 영어보다 14% 정도 짧은 표현이 가능하다고 보겠다.

그러나 단어와 음절간에는 차이가 있기 때문에 이들을 상호 비교하는데 무리가 있으므로 좀더 고찰을 요한다. (이상은 스페이스를 포함하지 않은 이주근 교수의 논문과 스페이스를 고려한 안수길 교수의 논문 자료에서 인용한 것임.)

〔Box 3〕
정보량의 계량단위 : Bit, B/s, Baud의 관계 운송시의 정보량을 표현하는 단위로는 초당 비트수(bit/sec)를 사용한다. 가령 1초당 한글 다섯자를 보낸다면 1자당 8비트일때 5x 8=40 bit/sec의 정보량이 이동한 셈이다. 그러나 정보가 전송로에 공급되어 이동할 때에는 보오(Baud)라는 단위를 사용한다. 보오는 「최단 펄스 지속시간의 역수」로 규정되는데 1초간 운송될 수 있는 펄스의 갯수에 해당한다고 하겠다. 보통은 B/s와 Baud는 같이 사용되지만 1개의 펄수에 2개 이상의 비트를 보낼수도 있기 때문에 B/s는 보오에 2배 이상의 속도로 정보량이 이동한다. 이러한 이유로 보오는 변조신호 속도라 하며 B/s는 단순히 이송속도라고도 한다.

4. 정보흐름의 계량화 문제

앞에서 정보흐름에 대한 양적 표현을 위한 기초적인 개념과 단위량을 정의하였다. 다음 호에서는 실제적인 문제적용에 대한 계량화, 예를 들면 컴퓨터의 처리속도, 데이터의 전송속도와 변저속도, 입출력의 처리속도 등에대하여 기술하고자 한다. 그러기 위해서는 정보량 계량단위 (Box1 및 3 참조)에 대한 이해가 선행되어야 한다.

〔Box 4〕
군더더기율과 상대 엔트로피 공식

1. 군더더기율 : $r = 1 - \frac{H(x)}{H(x)\,max}$ ← 엔트로피 / ← 최대 엔트로피

2. 상대엔트로피 : $r = \frac{H(x)}{H(x)\,max}$

5. 결 어

지금까지 정보의 계량화를 위한 기본적 이론에 대하여 정보량과 엔트로피의 정의 단위 및 군더더기율 등에 대하여 논의하였다. 진정한 의미에서 정보의 계량화 문제는 정보를 이해함에 있어 가장 초기의 문제이다. 결국 정보의 양적 표시는 초기의 불확실성이 해소되는 정보를말하며 이러한 면에서 정보란 「불확실성이 해소되는 과정(Process)」을 지칭하기도 한다. 그러므로 초기 발생가능확률이 희박할수록 얻어지는 정보량은 많아진다. 다시 말하면 불확실성이 높을수록, 즉 원천발생확률이 낮을수록 얻어질 수 있는 정보량은 증가하며 정보가치는 증가할 수 있다는 것이다. 이런 측면에서 본다면 정보란 '불확실성을 제거했을 때 얻어지는 가치'로 정의할 수가 있을 것이다. 이와 같은 개념은 추상적인 정보개념을 보다 구체적으로 계량화 하였다는 점에서 높이 평가를 받을수 있을 것이다. 그러나 구체적으로 얻어지는 가치에 대해서는 계산할 수 없다는 것과 발생확률 자체를 명확하게 계량할 수 없다는 취약점을 동시에 갖고 있다. 정보의 계량적 단위가 비트로 정의 된 것은 km나 kg 및초 등이 거리, 무게, 시간의 기본단위로 정의된것과 같기때문에 정보이론상 획기적 개념상의 진보가 아닐 수 없다. ■