기술해설

Amount of Information & Source of Information

情報量과 情報源

金 鎭 克

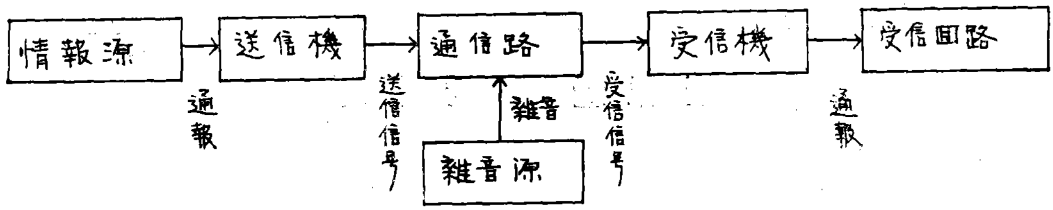
1. 序 言

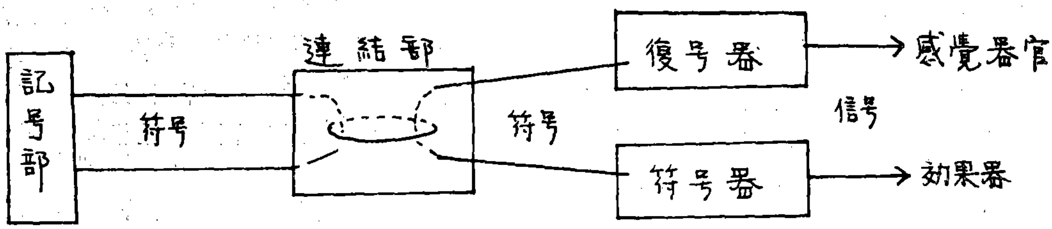
morse電信機 發明後 電氣通信技術은 二十世紀에 들어오면서 부터 長足의 發展을 거듭했다. 電話의 發明과 電話交換機의 發明 眞空管의 登場에 依한 長距離通信의 發達等 通信技術의 發展은 通信이라는 것을 情報傳送의 立場에서 考慮하겠끔 하였는 데 이 方面에 關心을 가진 사람들中 比較的 整理된 理論을 發表한 사람은 H. Nyguist와 K. Küpfmuller 이다. 이들의 1924年의 發表에 뒤이어 1928年 R.V.L.Hartley는 이들의 理論을 一般化하였는데 이 Hartley의 思考는 後日 情報理論의 根底를 이룬 Shannon의 理論의 母體가 된 것이다. 이 論文에서 Hartley는 情報의 量을 數量的으로 提示하였고 一定한 情報의 量을 傳送하기 爲해서는 周波數帶域幅과 傳送時間의 積이 一定해야 한다는 事實을 發表했다. 그는 確率的인 思考는 導入하지 않았으나 現代情報理論의 基礎를 이룩한 것이다.

世界第二次大戰後부터 戰後에 걸쳐 發達한 通信技術 및 電子工學의 發展으로 通信方式에 對한 研究에는 通信을 妨害하는 雜音에 對한 統計數學的 解析이 導入되어야 한다는 것을 認識하게 되었다. 그리하여 情報傳送形式의 研究에는 雜音을 考慮할 統計確率論的 立場에서의 檢討가 始作된 것이다. 이 時代에 發展된 通信方式으로서는 爲先 Edwin H. Armstrong이 1936年에 發表한 周波數變調方式을 들을 수 있는데 여기서 그는 信號對 雜音比가 入力信號電力을 增加시킴은 勿論 變調指數를 크게 하여도 改善될 수 있다는 事實을 發見했다. 戰時中의 Radar의 發達에 關聯된 pulse通信技術의 發達은 PAM PWM PFM PPM 等의 通信方式을 發展시켰고, 等히 1948年에 發表된 PCM은 雜音의 妨害를 效果的으로 排除할 수 있다는 面에서 情報理論에서 重要한 役割을 하고 있다.

雜音에 對한 研究로는 1945年과 1948年에 發表된 Pice의 確率統計論的 Random 雜音에 關한 研究가 있고 1946年의 Gabor의 時間의 周波數의 不確定性에 對한 研究가 有名하다.

이러한 發達에 併行해서 電子計算機와 自動制御技術의 發展도 情報理論發展에 拍車를 加하였다. 이보다 앞서 N. Wiener와 A. Resenblueth는 世界第二次大戰이 始作될 무렵부터 現在의 諸科學中에서 通信과 自動制御를 根底로하는 分野를 綜合하여 大學間 大系를 이룩하고저 各界의 學者들과 研究會를 거듭하여 이것을 1947年에 이르러 Cybernetics라고 命名했다. 이러한 雰圍氣속에서 情報量은 統計力學에서의 entropy形式으로 주어지게 된 것이다. 또한 Wiener는 anti-aircraft gun director control問題에서 Radar로 부터의 情報를 統計的으로 定常的인 時間의 函數로 取扱함으로서 容易하게 다룰 수 있다는 것을 指摘했다. 여기서 그는 모든 通信問題는 本質에 있어 根本的으로 統計的이고 成功的으로 取扱하기 爲해서는 統計的 方法으로 다루어야 한다는 것을 實證하였다. Wiener는 信號와 雜音을 確率過程이라 보고 雜音을 抑制하기 爲해 信號波形과 信號上에 雜音이 重疊된 波形과의 自乘平均誤差를 最小로 하는 filter의 設計와 入力信號부터 어느時間後의 信號値를 어느 自乘平均誤差





第2図

로서 豫測할 수 있는 豫測器의 回路가 可能하다는 것을 提示했는데 이것은 1950年 Y. W. Lee에 依해 實現되었다. 이 硏究는 또 Kolmogorov에 依해서도 行하여졌고 같은 結果를 얻었다.

이러한 過程에서 1948年 C. E. Shannon은 情報理論을 거이 完全한 形으로 統一된 立場에서 綜合했다. 그는 情報量을 entropy로 定義했고 雜音의 重要性도 考慮에 넣고 通信路의 通信容量이란 새로운 槪念을 導入했다. 그後 情報理論은 Fano等에 依해서 그 內容이 豊富해지고 여러 條件下에서 通信을 能率좋게 行하기 爲한 符號化의 問題等도 發展을 거듭하여졌고 이러한 硏究는 아직도 行하여지고 있다.

Shannon이 提示한 通信系는 大略 다음과 같이 5個의 要素로 構成되어 있는데 雜音이 通信系로 介入되고 있다. 生物과 같이 社會生活을 하고 있는 경우는 二者以上이 情報를 接受하는데 이와 같은 境遇는 다음과 같은 模型을 하고 있다고 Marko가 提示하고 있다.

2. 情報量과 Entropy

어떠한 現象이든 이를 數理的으로 取扱하기爲해서는 定量的이 表現이 必要하다. 情報理論에서 가장 基本이 되는 量은 情報의 量으로 이는 根本的으로 各情報의 發生確率에 依해 表示된다

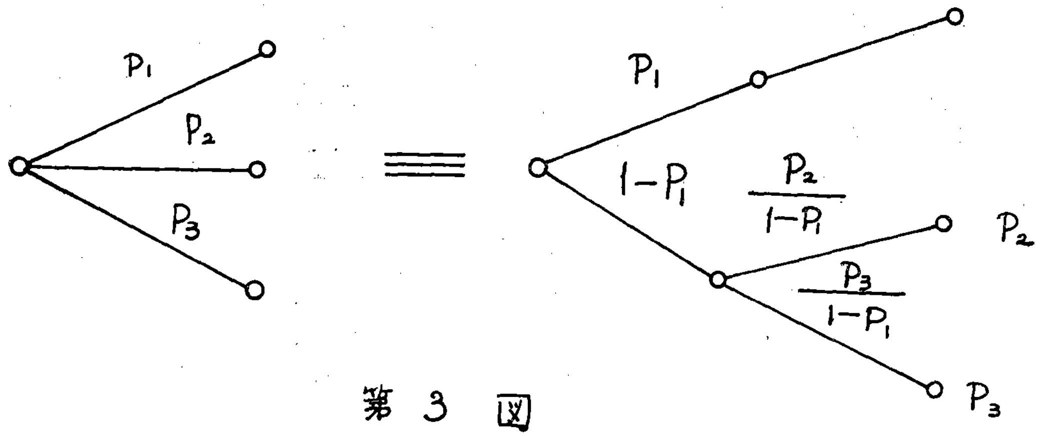
n個의 相異한 値 x_1, x_2, ……x_n만을 取하는 確率變數 X에서 $P(x=x_i)=P_i(i=1,2,\cdots\cdots,n)$라고 할때 다음과 같은 性質을 갖는 函數 $H(X)=H(P_1,P_2\cdots\cdots P_n)$를 定義해 본다.

(1) $H(X)$는 $P_1,P_2,\cdots,P_n$ 어느 것에 對해서도 連續函數이다.

(2) $H(X)$는 P_i가 모두 $1/n$일 때 n에 對한 單調增加函數이다.

(3) 하나의 事象의 發生을 繼續되는 두個의段階로 나누어 생각하면 나누기前의 $H(X)$는 나눈後의 $H(X)$의 荷重和로 表示된다.

x_1,x_2,x_3 3個의 境遇 (3)의 意味는 다음 그림



第 3 図

에서

$$H(P_1,P_2,P_3)=H(P_1,1-P_1)+(1-P_1)H\left(\frac{P_2}{1-P_1},\ \frac{P_3}{1-P_1}\right)$$

라는 것을 意味한다. 이러한 3號件을 滿足시키는 H(X)를 確率變數 X에 對한 情報量으로 定義하고 다음과 같은 形式을 갖는 것이 證明되어 있다.

$$H(X)=-\sum_{i=1}^{n} P_i \log P_i \quad \cdots\cdots(1)$$

이 式은 統計力學에서의 entropy와 같은 形式으로 事象 $x_i(i=1,2,\cdots,n)$가 平均해서 가지고 있는 情報量이며 單一確率變數를 갖는 情報源에 있어서의 平均情報量이다. $H(X)$의 單位로 對數의 底를 2로 했을때의 情報量은 bit(binaryunit)로 表示하고 底가 自然對數 e일때는 nat(natural unit), 10을 底로 했을 때는 Hartley 또는 decimalunit라고 한다.

(1)式에서 派生되는 情報量으로 自己情報量(Amount of self information)을 定義할 수 있다. H(X)는 $-\log P_i$의 平均値(期待値)로 表示되어 있는데 이 $-\log P_i$를 自己情報量이라 한다. 即

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

$I(x_i)$는 情報源에서 特殊한 記號 x_i가 出現할 때에 情報量이다.

確率變數가 X, Y 두개있을 때 即 X의 成分을 $x_i(i=1,2,\cdots,n)$ Y의 成分을 $y_j(j=1,2,\cdots m)$라면 X+Y의 事象($x_i\ y_j$)에 對한 確率分布를 P(ij)라 하면 이에 對한 entropy H(XY)는

$$H(XY)=-\sum_{i=1}^{n}\sum_{i=1}^{m} P(ij)\log P(ij) \quad \cdots\cdots(3)$$

으로 表示되고 이것을 結合 entropy라고 한다.

X의 成分 x_i의 確率 P(i) 및 Y의 成分 y_j의 確率 P(j)는 各各

$$P(i)=\sum_{i=1}^{n} P(ij), \quad P(j)=\sum_{i=1}^{m} P(ij)$$

이고 X, Y에 對한 entropy H(X) H(Y)는

$$H(X)=-\sum_{i=1}^{n} P(i)\log P(i) \quad \cdots\cdots(4)$$

$$H(Y)=-\sum_{j=1}^{m} P(j)\log P(j) \quad \cdots\cdots(5)$$

이다.

X, Y가 統計的으로 獨立이 아닐때 即 x_i가 發生後 y_j가 發生하는 確率 및 y_j가 發生後 x_i가 發生하는 確率을 各各 $P(j|i)$, $P(i|j)$라고 하면 이들 條件附確率은

$$P(j|i)=\frac{P(i,j)}{P(i)}, \quad P(i|j)=\frac{P(i,j)}{P(j)}$$

임으로 $P(j|i)$의 自己情報量은

$$I(y_i|x_i)=-\log P(j|i) \quad \cdots\cdots(6)$$

이고 이의 平均情報量 $H(Y|X_i)$는

$$H(Y|x_i)=-\sum_{j=1}^{m} P(j|i)\log P(j|i) \quad \cdots\cdots(7)$$

임으로 條件附 entropy H(Y|X)는

$$H(X|Y)=-\sum_{i=1}^{n}\sum_{j=1}^{m} p(ij)\log p(j|i) \quad \cdots\cdots(8)$$

로 表示된다. 이상에서 定義한 entropy 間에는 다음 關係가 있다.

(i) (1)式에서 $pi(i=1,2,\cdots n)$中 하나만이 1이고 나머지가 零일 때 H(X)=0이다.

(ii) (1)式에서 $pi=\frac{1}{n}(i=1,2,\cdots n)$일 때 H(X)는 最大가 되고 그 値는 log n이다.

(iii) $H(X)+H(Y)\geqq H(XY)$ ……………(9)

(iv) $H(XY)=H(X)+H(Y|X)$

$=H(Y)+H(X|Y)$ ………(10)

(v) $H(Y|X)\leqq H(Y)$ ……………………(11)

(9)式과 (11)式은 다음과 같이 一般化 된다.

$$H(X_1X_2\cdots X_n)\leqq H(X_1)+H(X_2)+\cdots+H(X_n) \quad \cdots(12)$$

$$H(X_1|X_2X_3\cdots X_{n+1})\leqq H(X_1|X_2X_3\cdots X_n) \quad \cdots\cdots(13)$$

(12)式은 n個의 情報源에서 個別的으로 얻어진 情報量의 和는 各情報源에서 同時에 얻어진 報量보다 크다는 것을 意味하고 있고 (13)式은 情報源 X_1 以外의 他情報源에서 情報가 많이 얻어지면 얻어질수록 X_1에 對한 情報量은 減少한다는 것을 表示하고 있다.

(9)式에서 또 相互 entropy I(X;Y)가 定義된다. 即

$$\begin{aligned} I(X;Y)&=H(X)+H(Y)-H(XY)\\ &=H(Y)-H(Y|X)\\ &=H(X)-H(X|Y)\\ &=-\sum_{i=1}^{n}\sum_{j=1}^{m}p(ij)\,log\frac{p(i|j)}{p(i)} \quad \cdots\cdots(14)\end{aligned}$$

이고 I(X;Y)는 通信容量을 定義하는데 重要한 要素가 된다.

以上의 離散的인 境遇에 entropy에 對한 定義는 連續變數 即 繼續되는 電氣的信號와 같은 境遇에도 擴張된다. 即 連續確率變數 X의 確率密度函數를 $p(x)$ Y의 確率密度函數를 $P(y)$라고 하고 同時確率密度函數를 $p(xy)$라고 하면

$$H(X)=-\int_{-\infty}^{\infty}p(x)\,logp(x)\,dx \quad \cdots\cdots(15)$$

$$H(Y)=-\int_{-\infty}^{\infty}p(y)\,logp(y)\,dy \quad \cdots\cdots(16)$$

$$H(X|Y)=-\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}p(xy)\,log\frac{p(xy)}{p(y)}dxdy \quad (17)$$

$$H(Y|X)=-\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}p(xy)\,log\frac{p(xy)}{p(x)}dxdy \quad (18)$$

$$H(XY)=-\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}p(xy)\,log\,p(xy)\,dxdy \quad (19)$$

等과 같이 定義되고 이들은 (9)式, (10)式, (11)式과 같은 性質을 가지고 있다. 連續的確率變數와 離散的確率度數에서의 entropy의 差異는 後者는 確率이 주어지면 一意的으로 entropy가 定해지나 前者의 境遇는 座標系에 따라 entropy가 다르다는 點이다. 確率變數 X가 Y로 變換되면 X에서의 確率密度函數 $P(x)$와 Y에서의 確率密度函數 $q(y)$間에는

$$q(y)=P(x)\,J\left(\frac{x}{y}\right)$$

의 關係가 成立되어

$$H(Y)=H(X)+\int_{-\infty}^{\infty}p(x)\,log\,J^{-1}\left(\frac{x}{y}\right)dx \quad \cdots\cdots(20)$$

이 되기 때문이다. 그러나 連續確率變數에서 定義된 相互 entropy에서는 이러한 問題點이 解消된다.

連續的 確率變數에 對한 entropy의 定義에는 몇가지 數學的인 問題點이 있는데 前述한 相互 entropy와 連續信號의 時間에 對한 連續性을 一定間隔을 두고 取한 標本値의 量子化에 依해서 解決되고 있다. 連續函數의 標本化問題는 標本化定理에 依한다. 即 連續函數 f(t)가 W(HZ)以上의 周波數를 包含하지 않을 때는

$$f(t)=\sum_{n=-\infty}^{\infty}f\left(\frac{n}{2W}\right)\frac{sin\pi(2Wt-n)}{\pi(2Wt-n)} \quad \cdots\cdots(21)$$

가 成立됨으로 T=1/2W(sec)의 時間間隔을 두고 標本値를 取함으로서 f(t)는 完全히 決定된다는 것이다. 即 f(t)를 $f\left(\frac{n}{2W}\right)$ $(n=-\infty,\cdots\infty)$의 離散的인 集合으로 取扱할 수 있게 된다. 이 連續信號의 標本化에서는 量子化가 問題가 되는데 量子化된 level의 數는 雜音이 많을 경우는 그 數가 적으나 雜音이 적을 경우는 그 數가 많아지며 情報源의 符號化에서 問題가 되고 있다.

3. 情報源

情報源으로서는 우리의 日常言語가 그 代表的인 것이라 할 수 있다. 이 情報源으로 부터의 情報는 우리의 日常文字인 한글, 數字, 英字等으로 表示되는데 이러한 文字(記號)의 系列은

通報를 構成한다. 이 文字의 數는 各言語에 따라 制限되어 있고 各言語에 따라 文字가 出現하는 頻度는 다를뿐더러 通報에서 서로 隣接하고 있는 文字들은 相互間 어떤 確率에 依해서 出現하는 것이다. 假量 한글의 24文字中(한글의 基本子母 24個는 文字라기보다는 文字를 構成하고 있는 要素라 하겠으나 한글을 情報源으로 考慮할 때는 이 基本子母 24個로 構成되는 모든 完成된 글자를 情報源의 基本文字로 取扱하는 것보다는 各各의 子母를 文字로 看做하여 取扱하는 것이 現行打字機나 테레타이프라이타의 文字盤을 考慮할 때 妥當하다고 생각된다) "ㄱ"이 出現하는 頻度와 "ㄴ"이 出現하는 頻度는 다를 것이다. 初聲의 "ㄱ"이 온다음에 "ㅓ"가 나타나는 確率과 "ㅛ"가 나타나는 確率은 다를 것이다 即 情報源에서 發生하는 記號(文字)는 先行하는 記號에 依해 影響을 받는다. 또 우리의 한글의 境遇는 文字 24個가 子音 14個 母音 10個로 區別되고 重子音, 重母音이 있으며 한글의 完全한 一文字는 初聲, 中聲 或은 初聲, 中聲, 終聲으로 區分되며 初聲, 終聲은 반드시 子音이고 中聲은 반드시 母音이 와야 한다. 이와 같이 情報源에는 確率的拘束條件과 確定的拘束條件이 있고 大體的으로 Markoff 過程으로 近似식힐 수 있다.

確率過程 X(t)에서 時刻 $t_n<t_{n-1}<t_{n-2}\cdots<t_1<t$에서 X(t)의 値가 $X(t_n)=x_n$ $X(t_{n-1})=x_{n-1}, \cdots, X(t_1)=x_1$ 임을 알았을 때 $X(t)=x$가 되는 條件附確率(推移確率)을 $p(x|x_n, x_{n-1}, \cdots, x_1)$로 表示할 때

$$p(x|x_{n+m}, x_{n+m-1}, \cdots, x_1) = p(x|x_n, x_{n-1}, \cdots x_1) \cdots\cdots (22)$$

가 成立하면 X(t)를 n重 Markoff過程이라 하며 言語의 情報源에서는 大體로 (22)式이 成立한다 即 記號 x의 出現確率은 有限個(n個) 比前에 出現한 記號에는 影響을 받지 않는다는 것이다.

情報源의 條件附確率間에

$$p(x_i|x_{j-1}, x_{i-2}, \cdots, x_{i-n}) = p(x_j|x_{j-1}, x_{j-2}, \cdots, x_{j-n}) \cdots\cdots (23)$$

가 成立하면 이 過程은 定常過程이 된다. 即 先行되는 n個의 事象이 주어지면 事象 x_i가 發生하는 推移確率은 i>n인 모든 i에 關係없이 一定하다는 것이다.

n重 Markoff過程 X(t)가 取할 수 있는 値가 $L_1, L_2, \cdots, L_q$인 q個의 記號에 限定되었다고 할 때 X(t)의 實現値의 系列인 n次空間의 Vector $(x_1, x_2, \cdots, x_n)$는 最高 q^n個의 系列이 可能하고 이들 Vector의 各各을 狀態(State)라고 부른다. 이들 狀態는 最高 q^n個가 全部 存在하는 것이아니고 狀態의 推移에는 强한 制限이 있다. 任意의 狀態$(L_0, L_1, L_2, \cdots, L_{n-1})$에서 推移할 수 있는 狀態는 $(L_1, L_2, \cdots L_{n-1}, L_k)$ $(R=1, 2, \cdots q)$의 q個 以上은 생각할 수 없다. 이와 같이 n重 Markoff 過程에는 系列의 發生에 依해 狀態라는 概念이 導入된다.

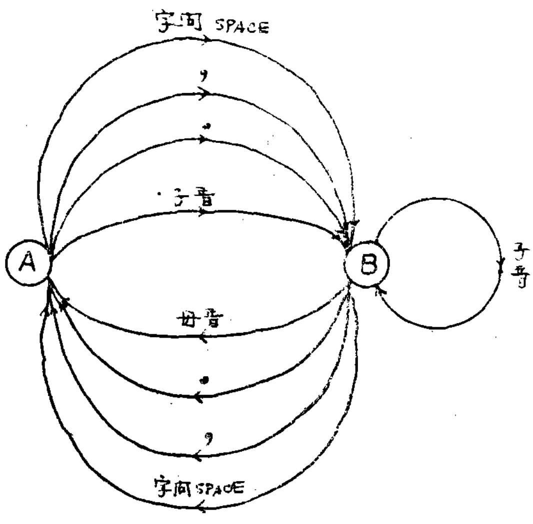
情報源으로서의 定常 Markoff過程에서 다음과 같은 試行을 考慮해 본다. n回의 推移가 發生하는 동안 狀態 E_j가 y回發生했다고 假定한다. 이 回數는 確率變數로 狀態 E_j와 回數 n의 函數임으로 $y_j(n)$로 表示하기로 한다. k回째의 狀態가 E_j일 때 1이고 其他에서는 0이 되는 確率變數 $U_j(k)$를 定義하면

$$y_i(n)=\sum_{k=1}^{m} U_j(k)$$

가 되고 다시

$$V_j(n)=\frac{1}{n}y_j(n) \cdots\cdots (24)$$

를 定義하면 $V_j(n)$는 n回의 推移에서 狀態가 E_j가 되는 回數의 比率을 表示하는 確率變數가



된다. (24)式에서 $n\to\infty$로 하면 $V_j(n)$의 期待値는

$$E(V_j(n))=q_j \quad\cdots\cdots\cdots(25)$$

가 되는 것이 證明되어 있다. 이 q_j는 正則 Markoff過程의 定常分布에서 狀態의 數를 r라 할 때 狀態 E_j가 發生하는 確率을 나타내고 있다. 이러한 條件을 滿足하는 情報源을 Ergodic 情報源이라 부르고 우리의 言語는 이러한 情報源인 것이다.

이러한 情報源의 推移確率을 圖式으로 表示한 것이 Shannon의 線圖이다. 이것은 各狀態를 點으로 表示하고 各狀態에서 一回의 推移가 發生하면 到達할 수 있는 다른 狀態까지 화살표의 線으로 表示한 것이다. 이 有向連結線에는 이미 定해진 發生될 記號가 定해져 있고 各記號의 推移確率도 아울러 考慮할 수 있다. 例로서 우리 한글을 子音과 母音만으로 區別하고(重母音이나 重子音도 하나의 母音과 子音으로 생각한다)"." 와 "," 및 字間 Space만을 考慮하면 다음과 같은 線圖를 그린다.

와 같은 狀態를 取하여 推移할 것이다. 우리의 한글文字의 現行綴字法은 比較的 그 構成에 있어 科學的으로 되어 있어 이 方面으로 부터의 研究를 爲해 資料蒐集과 統計集計에 民族的인 後援이 있어야 할 것이다.

다음 情報態이 $L=(L_1, L_2, \cdots, L_q)$인 q個의 記號로 構成되어 있고 이들의 生起確率을 $\begin{pmatrix} L_1, L_2, \cdots, L_q \\ p_1, p_2, \cdots, p_q \end{pmatrix}$라고 하면 自己情報量 $I(L_i)$는

$$I(L_i)=-\log p_i \quad\cdots\cdots\cdots(26)$$

이고 期待値는

$$E(I(L_i))=-\sum_{i=1}^{q} p_i \log p_i=H(L) \quad\cdots\cdots(27)$$

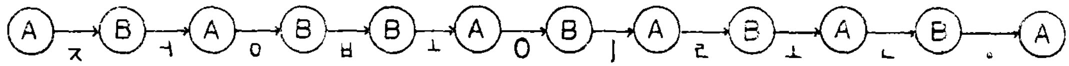
分散은

$$\sigma^2=\sum_{i=1}^{p} p_i[I(L^i)-H(L)]^2$$

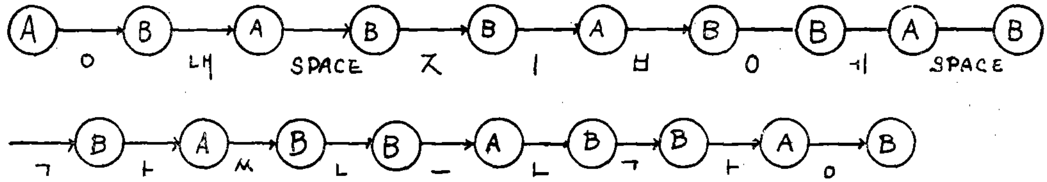
$$=\sum_{i=1}^{q}[I(L_i)]^2 p_i-[H(L)]^2 \quad\cdots\cdots(28)$$

이 情報態에서 繼續해서 情報가 發生할때 第k번째에 發生하는 記號를 確率變數 X_k로 表示하면 X_k의 確率分布는

"정보이론"이란 文字系列이 取하는 狀態의 推移는



와 같은 狀態에 따라 推移할 것이고 "왜 집에 갔는가"는



$X_k=\begin{pmatrix} L_1, L_2, \cdots, L^q \\ p_1, p_2, \cdots, p_q \end{pmatrix}$이다.

第k번째에서 L_i가 發生하는 自己情報量은 (26)式과 같고 따라서

$$p[I(L_i)=I(x_i)]=p_i$$

임으로 X_k의 期待値와 分散은 (27)式, (28)式과 一致한다. 即 確率變數 $Y(X_1, X_2, \cdots, X_k, \cdots)$로 表示되는 通報에서의 自己情報量$[I(X_1), I(X_2), \cdots, I(X_k), \cdots]$는 모두 同一한 平均値 $H(L)$와 分散 σ^2을 가짐으로 TchebyCheff의 不等式

$$p\left[\left|\frac{\sum_{k=1}^{m} I(X_k)}{M}-H(L)\right|>\delta\right]<\frac{\sigma^2}{M\delta^2} \quad\cdots\cdots(29)$$

이 成立된다. 이 不等式을 變形하고 $Y=X_1X_2\cdots X_M$의 自己情報量을 $I(Y)$라고 하면

$$I(Y)=\sum_{k=1}^{M} I(X_k) \quad\cdots\cdots\cdots(30)$$

임으로

$$|-\log p(Y=y)-MH(L)|\leqq M\delta \quad\cdots\cdots(31)$$

即

$$M(-H(L)-\delta)<\log p(y)$$
$$<M(-H(L)+\delta) \cdots\cdots (32)$$

M個의 記號로 構成되어 있는 通報의 集合에서 各通報의 生起確率 $p(y)$의 和가 $1-\varepsilon$보다 큰 部分集合 S에 屬하는 通報의 個數를 N라고 하면

$$1-\varepsilon \leqq \sum_{yes} p(y)=P(s)1$$

이고

$$N_0 2^{M(-H(L)-\delta)}<P(s) N, 2^{M(-H(L)+\delta)}$$

임으로

$$(1-\varepsilon)\cdot 2^{M(H(L)-\delta)}<N<2^{N(L)+\delta)} \cdots\cdots\cdots (33)$$

이 不等式으로 S에 屬하는 通報의 個數가 評價된다. entropy가 H인 情報源에서 길이 M인 通報는 大略 2^{-MH}이고 이 通報의 生起確率은 大略 2^{-MH}程度이고 나머지 通報들은 全體的으로 적은 確率로 發生하고 있다.

이것을 더욱 擴大하여 情報源의 記號 $L=(L_1, L_2, \cdots, L_g)$에서 重複을 許容하여 n個를 羅列한 系列의 集合$(\sigma_1, \sigma_2, \cdots, \sigma_{qn})$를 基本記號로 하는 情報源을 n次의 擴大情報源이라 하고 L^n로 表示하면 即

$$L^n=\begin{pmatrix}\sigma_1, \sigma_2, \cdots, \sigma g_n \\ p(\sigma_1), p(\sigma_2), \cdots, p(\sigma_n)\end{pmatrix}\text{에서는}$$

$$H(L^n)=nH(L) \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots (34)$$

가 成立된다. 即 擴大情報源에서의 1記號當의 entropy $H(L^n)/n$는 n如何를 不問하고 元來의 情報源의 entropy와 같다.

이 情報源이 ergodic이고 m重 markoff過程으로 構成되어 있으면 推移確率

$$p(x_i|x_{j_1}, x_{j_2}, \cdots, x_{j_m}) \quad i=1,2,\cdots q,$$
$$jp=1,2,\cdots q \cdots\cdots (35)$$

가 規定되고 또한

$$p(x_i, x_{j_1}, x_{j_2}, \cdots, x_{j_m})=p(x_i|x_{j_1}, x_{j_2}, \cdots, x_{j_m})$$
$$\times p(x_{j_1}, x_{j_2}, \cdots, x_{j_m}) \cdots (36)$$

가 成立하고 q^m個의 可能한 狀態가 있음으로 이 m重 markoff情報源의 平均情報量 H(L)는

$$H(L)=-\sum_{L^{m+1}} p(x_i, x_{j_1}, x_{j_2}, \cdots x_{j_m})$$
$$\times \log p(x_i|x_{j_1}, x_{j_2}, \cdots, x_{j_m}) \cdots\cdots\cdots (37)$$

號一 情報源이 記憶을 하지 않으면

$$p(x_i|x_{j_1}, x_{j_2}, \cdots, x_{j^m})=p(x_i) \cdots\cdots\cdots\cdots\cdots (38)$$

이 되어 (34)式과 一致되는 結果를 갖는다. 이것은 系列의 平均情報量과 狀態의 平均情報量이 갖다는 것을 意味하고 있어 이 情報源이 ergodic 라는 것을 表示하고 있다.

情報源의 記號 $L=(L_1, L_2, \cdots, L_q)$에서 發生할 수 있는 n個記號의 通報의 數는 q^n이고 各通報는 定해진 確率測度를 가지고 하나의 集合 S를 構成하고 있다. S의 測度를 μ(S)라고 하고 이 q^n個의 記號系列의 平均 entropy를

$$H_n=-\sum_s \mu(S)\log\mu(S) \cdots\cdots\cdots\cdots\cdots\cdots\cdots (39)$$

로 定義하면

$$\lim_{n\to\infty}\frac{H_n}{n}\to H(L) \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots (40)$$

인 H로 收斂한다. 또 $L=(L_1, L_2, \cdots, L_q)$에서 特定한 記號에 依한 n個의 通報의 集合을 C, 이의 確率測度를 μ(C)로 하고

$$Z_n=-\frac{1}{n}\log\mu(C) \cdots\cdots\cdots\cdots\cdots\cdots\cdots (41)$$

라 하면 $\varepsilon>0$, $\delta>0$에 對해

$$P(|Z_n-H|>\varepsilon)<\delta \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots (42)$$

가 成立하여 n가 充分히 클때는 通報 C는 다음의 2組로 分離된다.

(1) 第1組의 C는 모두

$$\left|\frac{\log\mu(C)}{n}+H\right|<\varepsilon \cdots\cdots\cdots\cdots\cdots\cdots\cdots (43)$$

을 滿足하고

(2) 第2組의 通報의 確率總和는 δ보다 적다는 것이 證明되어 있다.

Shannon線圖에서 狀態 E_i에서 有向線을 따라 推移確率 $P_i(j,S)$로 記號 $\overset{(s)}{L_{ij}}$를 發生하고 狀態 Ej로 變했다고 하면 各 n記號長의 通報가 發生하는 確率을 P라고 할때 n가 充分히 크면 $-\log p$의 確率分布는 $a<b$ 丶 對해

$$P_{rob}(a<-\log p<b)$$
$$\doteqdot\frac{1}{\sqrt{2\pi}\,\sigma_H}\int_a^b e\times t\left(-\frac{(x-nH)^2}{2n\sigma_H^2}\right)dx$$
$$\cdots\cdots\cdots\cdots (44)$$

但 $$H=-\sum_i P_i\sum_{is} P_i(js)(\log P_i(js)^2-H^2$$
$$+2\sum_{ijs} P_i P_i(js)\log P_i(j,s)\sum_{k,l,m}\Omega_{jk}$$
$$\times P_k(l,m)\log P_K(,lm)$$

$$\Omega_{jk}=\lim_{n\to\infty}\sum_{t=0}^{n-1}(P_j^{(t)}(k)-P_k)$$

여기서 P_i는 狀態 E_i의 定常確率, $P_j^{(t)}(k)$는 狀態에 E_j에서 t個의 有向線을 따라 狀態 E_k로

到達하는 確率이다.

連續的信號에 對한 情報源에서는 相互Entropy로서 情報量을 表示할 수 있음으로 이에 對한例를 들어본다. x를 送信信號의 瞬間値 y를 受信信號의 瞬間値라고 하면 $y-x$가 通信路에서의 雜音이 된다. 信號와 雜音의 推移分布가 Gauss分布를 이루고 있고 送信信號 및 雜音의 分散을 S, N라 하면

$$P(x)=\frac{1}{\sqrt{2\pi S}}\,exp\left(-\frac{x^2}{2S}\right)$$

$$P(y)=\frac{1}{\sqrt{2\pi(S+N)}}\,exp\left(-\frac{y^2}{2(S+N)}\right)$$

$$P(y|x)=\frac{1}{\sqrt{2\pi N}}\,exp\left(-\frac{(y-x)^2}{2N}\right)$$

임으로

$$I(X;Y)=\frac{1}{2}\,log\left(1+\frac{S}{N}\right)$$

$$(na^{+}/\text{自由度})\cdots\cdots\cdots(45)$$

가 된다. 한편 送信信號가 W(Hz)以下의 周波數만을 包含하고 있다면 標本化定理에 依해 每秒 2W個의 標本値만으로 이 信號를 表現시킬 수 있음으로 每秒當의 相互 Entropy는

$$I(X;Y)=W\ log\left(1+\frac{S}{N}\right)(na^{+}/sec)\quad(46)$$

가 된다. 이 (46)式은 通信容量으로서 重要한 意味를 가지고 있다.