한글 데이타 압축코드를 위한 엔트로피 감소

o
이재영*, 성굉모**, 이종각**
한림대 전산과*, 서울대 전자과**

Entropy Reduction for the Code of Hangul Data Compression

o
Jae-Young Lee*, Koeng-Mo Sung**, Chong-Kak Lee**
Dept. of Computer Science, Hallym Univ.*
Dept. of Electronics Eng., Seoul National Univ.**

요 약

본 논문에서는 정보원의 집합이 여러개의 부집합으로 나누어지며 정보 발생 방법은 이들 부집합을 순서적으로 선택한후 그 부집합에서 확률에 따라 원소를 발생시키는 성질을 갖는 정보원을 인코딩할때 데이타를 압축시킬 수 있는 엔트로피 감소 모델을 제시하였다. 그 방법은 정보를 전송하는 과정에서 정보원이 갖고있는 성질에 관한 정보를 제외시킴으로서 평균 정보량 즉 엔트로피를 줄이는 것이다. 이와 같은 원리로 인코딩을 하면 데이타를 압축시킬 수 있다.

I. 서론

정보란 여러개의 대상물로부터 한 대상물을 구별할 수 있는 상태에 관한 지식으로서 상태에 관한 불명확성을 제거하는 것이며, 정보의 엔트로피는 무질서한 정도를 나타내는 열역학적 엔트로피와 다른 입장으로 부터 출발하였으나 수학적 표현이 일치하며 서로 밀접한 관계를 갖고있다. Shannon[1948,12]은 그의 정보이론에 관한 논문에서 정보를 수량화 하였으며, 여기서 정보량은 한기호를 다른 기호로 부터 구별할 수 있는 선택의 회수를 나타낸다. 따라서 평균 정보량 즉 엔트로피는 평균 선택 회수를 나타내며, 정보원을 코드화했을 경우 문자당 평균 bit수를 나타낸다. 이와 같이 정보원을 코드화할 때 하나의 기호당 평균 bit 수 혹은 전체 데이타의 총 bit수를 줄이는 것이 텍스트 압축이다.

외국의 경우 텍스트 압축 방법에 대한 많은 연구가 있었으며 주요 연구 결과를 살펴보면 다음과 같다. Dewey [1923,18]는 영어 단어 약 100,000개 즉 문자 약 308,800개에 해당되는 데이타의 빈도수를 분석하였으며, Shannon [1951,13]은 영문의 엔트로피와 리던던시 (redundancy)를 평가하는 방법을 제시하였으며 그 엔트로피에 대한 결과는 1 문자의 엔트로피는 4.14 bit/letter, 2 문자의 엔트로피는 3.56 bit/letter, 3 문자의 엔트로피는 3.3 bit/letter로 나타났다. Huffmann [1952,15]은 문자의 빈도수에 따라 가변코드를 할당하여 텍스트를 압축하는 Huffmann 코드을 제안하였으며 이 코드는 빈도수가 높은 문자에 길이가 짧은 코드를 할당하고 빈도수가 낮은 문자에 길이가 긴 코드를 할당함으로써 약 50 %의 압축 효과를 얻었고, White [1967,19]와 Pike [1981,20]는 단어 인코딩방법과 Lea [1978,21]는 n-gram 코딩 방법 등 텍스트를 압축하는 방법을 제안하였다.

한글의 경우 문교부 주관으로 한글의 빈도수를 조사한 것[1956,2]이 있으며, 한글의 음절에 관한 엔트로피[1974,9], 한글 단어에 관한 엔트로피[1979,1], 공백소를 포함한 한글 자소에 관한 엔트로피[1980,4], 및 한글의 초, 중, 종성의 조건부 확률과 엔트로피[1987,6] 등의 연구가 있었으나, 그동안 한글 코드를 표준화하는 문제로 인하여 한글 데이타 압축에 관한 연구가 별로 없었다.

본 논문에서는 어떠한 규칙성을 갖는 정보원이 있을 때 이 정보원의 불확실성을 줄이는데 적합한 엔트로피 감소 모델을 제안하고, 이 모델에 규칙성을 갖는 정보원을 적용함으로써 일반적인 확률 모델에 적용하는 것보다 엔트로피를 줄일 수 있음을 보인다. 또한 한글 정보원을 이 엔트로피 감소 모델에 적용하고 한글 자소를 Huffmann 코드로 코드화하여 데이타를 압축함으로써 일반적 확률 모델에 적용하여 얻은 결과보다 더 좋은 결과를 얻었다.

II. 엔트로피 감소 모델

정보원의 문자의 발생은 하나의 확률 현상이라고 생각할 수 있으며, 한 순간의 문자 발생 확률은 일반적으로 그 직전에 발생한 여러개 문자의 영향을 받는다. 이러한 확률적인 정보원은 마코프 소스 (Markov source)로서 기술할 수 있으며, 임의의 정보원에서 발생한 문자의 시간 계열은 다음과 같이 표시한다.

$$Xi\text{-}m\text{-}n\ ,\ \ldots\ ,\ Xi\text{-}m,\ Xi\text{-}m+1,\ \ldots\ ,\ Xi\text{-}2,\ Xi\text{-}1$$

이와같은 시간계열 다음 순간 Xi 라는 문자가 발생할 조건부 확률이 다음과 같은 관계를 가질 때, 일반적으로 M 차 마코프 정보원이라고 한다.

$$P(Xi \mid Xi\text{-}1,\ Xi\text{-}2,\ \ldots\ ,\ Xi\text{-}m,\ \ldots\ ,\ Xi\text{-}m\text{-}n) = P(Xi \mid Xi\text{-}1,\ Xi\text{-}2,\ \ldots\ ,\ Xi\text{-}m)$$

m = 0 일 때, 문자가 발생할 확률은 P(Xi)가 되므로 각 문자(자모)의 출현 확률이 상호 독립인 영메모리 정보원 (zero memory source)이 된다. 여기서 N 개의 문자로 구성된 M 차 Markov 정보원의 경우 엔트로피는 감소할수 있으나 이 정보원을 코딩하는데 중복 조합 nCm 개의 문자 코드표가 필요하며 메모리 공간과 탐색 시간에 큰 오버헤드가 되기 때문에 본 논문에서는 0차 마코프 정보원 혹은 영 메모리 정보원같은 독립 확률 모델의 경우 를 고려하며 이 정보원보다 엔트로피를 줄일수 있는 정보원 모델을 제안하며 엔트로피 감소 모델이라고 명명한다. 이 모델에 적용할 수 있는 정보원은 다음과 같은 조건이 만족되어야 한다.

1. 정보원은 문자 집합 S가 있을 때 이 집합은 n개의 부집합 Si (i=1,2,...,n)으로 구성되며 각 부집합 Si는 mi (mi=1,2,3,...) 개로 문자로 이루어져 있다.
2. 정보원의 문자 발생 방법은 부집합 Si에서 문자 발생 확률에 따라 1문자를 선택하여 발생하며 부집합의 순서는 S1, S2, S3, ... , Sn 순으로 반복된다.
3. 해석을 간단하게 하기 위하여 서로 다른 부집합에 같은 문자가 있더라도 문자가 속해있는 부집합이 다르면 서로 다른 문자로 취급한다.

위와 같은 조건을 만족하는 정보원을 일반적 확률 모델에 적용할때 엔트로피를 구하여 본다. 우선 정보이론에 있어서 정보량이라는 개념은 정보원의 각 요소가 선택될 수 있는 가능한 방법수의 척도로 이진 시스템(binary system)일 경우 단위는 비트(bit)가 된다.

$$\text{Information} = \log_2 1/p = \log n \quad \text{(bit)}$$

여기서 p는 요소가 선택될 확률을 나타낸다. 엔트로피는 일반적으로 정보량의 척도로서 통계열력학에서의 엔트로피와 대응되는 것이다. 언어 정보원의 경우, 엔트로피는 정보원에서의 각 요소가 갖는 정보량의 평균값을 나타내는 통계학적인 파라메터로서 각 요소의 평균 디지트(digit) 수를 뜻하므로써 능률적인 코드화의 기초가 되며 다음과 같은 식으로 나타낸다.

$$H = -\sum_{i=1}^{N} Pi \log_2 Pi \quad \text{(bits/symbol)} \quad (1)$$

여기서 Pi는 각 요소의 발생 확률이며, 요소의 총수는 N 개이다.

1. 영 메모리 정보원 (확률 모델)

영 메모리 정보원은 k개의 문자로 구성된 집합 S에서 확률에 따라 문자를 선택하여 발생하며 다음과 같이 나타낼 수 있다.

$$S = \{ s1, s2, s3, \ldots, sk \}$$

$$P(si) = pi \quad (i=1,2,3,\ldots\ k)$$

$$\gamma n \equiv \{(p1,p2,p3,\ldots pk) : \sum_{q=1}^{k} pq=1,\ pq \geq 0 ;$$

$$q=1,2,3,\ldots k\}$$

γn : the set of all complete finite (k-ary) probability distribution

$$P = (p1, p2, p3, \ldots, pk) \in \gamma n$$

$$Hk (p1,p2,..,pk) = -\sum_{q=1}^{k} pq \log pq \quad (2)$$

p1=p2=...=pk=p = 1/n이 만족하는 경우는 다음과 같이 최대값을 갖는다.

$$Hk (p,p,\ldots p) = -\log p = \log n$$

성질

1) non negative

$$H(p) \geq 0 \quad \text{for all} \quad 0 \leq p \leq 1$$

2) additive

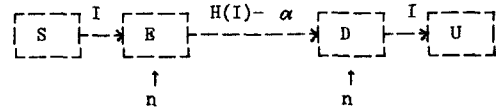
$$H(pq) = H(p) + H(q)$$

3) normalized

$$H(1/2) = 1$$

2. 엔트로피 감소 모델

문자 발생 정보원에서 나오는 문자열을 사용자에게 전달하는 과정을 살펴보면 송신측에서 송신할 문자를 인코딩(encoding)한 비트 스트림(bit stream)을 전송로로 보내면 수신측에서는 이 비트 스트림을 디코딩(decoding)함으로써 원하는 문자를 전달할 수있다. 이때 비트 스트림의 감소는 전송로의 비용의 절감을 나타내는 직접적인 관계가 있다. 이 비트수를 줄이는 기본 원리는 정보에 관한 특성을 최대한으로 이용하여 인코딩과정에 반영하는 것이다. 즉 송수신측이 서로 같이 알고있는 정보원의 특성이 있다면 이와 관련된 정보를 제외한 정보만 보냄으로써 평균 정보량 즉 엔트로피를 줄일수 있다. 본 논문에서 제안한 엔트로피 감소 정보원 모델의 특성은 위의 조건에서 기술한 바와 같이 한개의 집합은 n개의 부집합으로 구성되며 순서에 따라 선택된 부집합에서만 확률에 따라 문자 원소를 선택하여 문자를 발생시키는 것이다. 엔트로피 감소 모델의 인코딩 방법은 다른 방법에서 문자 원소의 발생 확률 p를 이용한 인코딩 방법 이외에 순서에 따라 같은 확률로 선택되는 부집합의 수 n을 이용하여 인코딩하는 것 이다. 즉 선택된 부집합에서 발생한 문자의 확률은 나머지 n-1 개의 부집합과는 독립된 확률을 가지기 때문에 전체집합의 관점에서 이문자의 확률보다 커짐으로 정보량은 상대적으로 감소하며 이와 관련하여 감소되는 정보량을 α라고 한다. 그러면 정보 전송 시스템에서 인코더와 디코더는 부집합의 수 n을 알고 있다면 인코더는 이와 관련된 정보 α는 코딩에서 제외한나머지 정보만 코딩해서 보내고 디코더는 수신한 정보에 α를 포함시켜 원래의 정보로 변환시키며 이러한 과정을 나타내는 블럭 다이어그램을 그림에 나타내였다.



S:정보원
U:사용자
E:인코더
D: 디코더
α: 감소되는 정보량

그림 . 엔트로피 감소 모델

전체 집합 S는 n개의 집합으로 구성되며 모든 i,j에 대하여 $Si \cap Sj = \phi$ $(i \neq j)$이 만족한다고 할때 전체 집합 S와 j번째 집합을 다음과 같이 표시한다.

$$S = S1 \cup S2 \cup \ldots \cup Sn = \bigcup_{i=1}^{n} Si$$

$$Si=\{ sil,si2,\ldots ,simi \}$$

여기서 i번째 집합에 있는 j번째 원소의 확률 p(sij)를 간단하게 pij로 표시하며 한다. 이 집합을 원소를 나열한 집합으로 표현하면 다음과 같이 나타낼 수있다.

$$S = \{s1,s2,s3,\ldots,sk\},\ k = \sum_{i=1}^{n} i*mi$$

$$p(sq) = pq$$

여기서 i,j,q는 다음과 같다.

$i=1,2,\ldots,n$

$j=1,2,\ldots,mi$ mi : i 번째 부집합의 원소수

$q=1,2,\ldots,k$

$$Pi = (p(sil),p(si2),\ldots,p(simi)) = (pil,pi2,\ldots,pimi)$$

$$\gamma imi \equiv \{ (pil,pi2,\ldots,pimi) : \sum_{j=1}^{mi} pij=1,\ pij \geq 0 ;$$

$$j=1,2,\ldots,mi \}$$

Pi ∈ $\mathcal{T}$imi

Hmi(pil,pi2,...pimi) = Hmi(Pi)

i 번째 부집합에 있는 j번째 문자에 대한 정보량은 log (1/pij)이며 이 문자의 평균 정보량은 전체 집합에 대한 이 문자의 확률 δi(j)에 log (1/pij)를 곱한 형태로 되며 ↓H로 표시하고, δi(j)만으로 표현된 식과 pij만으로 표현된 식은 다음과 같이 표현한다.

$$\downarrow H(pij) \equiv \delta i(j) \log \frac{1}{pij}$$

$$= \delta i(j) \log \frac{psi}{\delta i(j)}$$

$$= -\delta i(j) [\log psi + \log (1/\delta i(j))]$$

$$= H(\delta i(j)) - pij\ H(psi) \qquad (3)$$

$$\downarrow H(pij) \equiv \delta i(j) \log \frac{1}{pij}$$

$$= pij\ psi \log \frac{1}{pij}$$

$$= psi\ H(pij) \qquad (4)$$

여기서 δi(j)와 pij와의 관계는 다음과 같다.

$\delta i(j) = pij\ psi$
$pij = \delta i(j) / psi$
$psi = mi/k$

$$\sum_{q=1}^{k} pq = \sum_{i=1}^{n} \sum_{j=1}^{mi} \delta i(j) = 1$$

$$(q = \sum_{r=1}^{i-1} r*mr + (j-1))$$

$$\sum_{i=1}^{n} \sum_{j=1}^{mi} pij = n$$

전체 집합에 대한 i 번째 부집합의 엔트로피 Φi라고 하면 (3)식으로부터 구하면 다음과 같다.

$$\Phi i = \sum_{j=1}^{mi} \downarrow Hmi(pij) = \downarrow Hmi(Pi)$$

$$= \sum_{j=1}^{mi} \delta i(j) \log \frac{1}{pij}$$

$$= \sum_{j=1}^{mi} \delta i(j) \log \frac{psi}{\delta i(j)}$$

$$= \sum_{j=1}^{mi} H(\delta i(j)) - H(psi) \qquad (5)$$

(4)식으로 부터 Φi를 구하면 다음과 같다.

$$\Phi i = \sum_{j=1}^{mi} psi\ H(pij)$$

$$= psi \sum_{j=1}^{mi} H(pij) \qquad (6)$$

전체 집합에 대한 엔트로피 Φ는 다음과 같다.

$$\Phi = \sum_{i=1}^{n} \Phi i$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{mi} [\ \delta i(j) \log psi + \delta i(j) \log (1/\delta i(j))\ ]$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{mi} \delta i(j) \log psi + \sum_{i=1}^{n} \sum_{j=1}^{mi} H(\delta i(j))$$

$$= Hk(p1,p2,\ldots,pk) + \sum_{i=1}^{n} \sum_{j=1}^{mi} \delta i(j) \log psi$$

$$= Hk(p1,p2,\ldots,pk) - \sum_{i=1}^{n} \sum_{j=1}^{mi} pij\ H(psi)$$

$$= Hk(p1,p2,\ldots,pk) - \sum_{i=1}^{n} H(psi) \sum_{j=1}^{mi} pij$$

$$= Hk(p1,p2,\ldots,pk) - \alpha \qquad (7)$$

$$(\ \alpha = \sum_{i=1}^{n} H(psi))$$

성질

1) $\downarrow H(pij) \geq 0$
2) $\downarrow H(pij) = 0$ for n=k
3)psl=ps2=...=psn= 1/n 일때
Φ는 최소가 된다.

(7)식에서 엔트로피가 log n만큼 감소되는 것을 알 수 있고 n=k 즉 모든 원소가 일정한 순서대로 발생한다면 엔트로피가 0이 된다.

III. 한글의 각종 엔트로피 와 코드화한 결과

1. 샘플 데이타

본 논문에서 사용한 데이타의 샘플은 1956년 문교부 주관으로 조사한 한글의 빈도수 데이타이다. 여기서 조사된 데이타는 총 192464개의 문자를 분석하여 단음절 문자 1259의 빈도수, 초성 21, 19, 14개로 구성된 자소 빈도수, 중성 21, 14 ,10개 로 구성된 자소 빈도수, 종성 29, 17, 15개로 구성된 자소 빈도수, 초성과 종성을 합쳐서 20, 15개 로 구성된 자소 빈도수, 그리고 초성, 중성, 및 종성을 합한 34, 29, 25개로 구성된 자소의 빈도수등이 있다. 물론 이 데이타는 오래된 데이타로서 현재 사용하지 않는 문자 ㅇ피 (6), ㅇ버(4), ㅇ베 (4), ㅇ브 (1)이 포함되어 있으나, 이 문자를 제외한 데이타 를 요즘 조사된 데이타와 비교해볼때 문자의 빈도수순의 서열은 다를지라도 빈도수의 분포가 균일하지 않고 스큐드 되었다는 점은 같다. 엔트로피 감소 모델은 문자의 빈도수 서열보다는 스큐드(skewed)된 빈도수 분포와 여러개의 부집합이 교대로 선택되고 그 선택된 부집합에서 확률에 따라 자소를 발생하는 특성을 갖는 데이타에 적용하기 때문에 위와 같은 데이타를 사용하여도 실험 데이타에 큰 영향을 미치지 않는다. 여기에 사용된 데이타는 초, 중, 종성을 합한 54, 24개의 자소 빈도수,초성 19개의 빈도수, 중성 21개의 빈도수 및 종성 28개의 빈도수 (받침이 없는 경우 X로 표시하고 하나의 자소로 고려함)로서 현재 사용되지 않는 15개 의 문자에 포함된 자소를 제외한 것이다.

이러한 데이타를 사용하여 51개 자모, 24개 자모, 19개 초성, 21개 중성, 28개 종성에 대한 엔트로피와 코딩한 결과를 나타내었다.

2. 각종 엔트로피

1) 최대 엔트로피 (p1 = p2 = ... = pn =p)

$$H24(p1,p2,\ldots,p24) \mid_{p=1/24} = \log_2 n = 4.585 \text{ (bits/letter)}$$

$$H51(p1,p2,\ldots,p51) \mid_{p=1/51} = \log_2 n = 5.672 \text{ (bits/요소)}$$

2) 영 메모리 정보원 엔트로피

$$H24(p1,p2,\ldots,p24) = -\sum_{i=1}^{24} Pi \log_2 Pi = 4.038 \text{ (bits/letter)}$$

$$H51(p1,p2,\ldots,p51) = -\sum_{i=1}^{51} Pi \log_2 Pi = 4.408 \text{ (bits/요소)}$$

3) 엔트로피 감소 모델

$$\text{Ó} = -\sum_{j=1}^{19} \text{Ó}1 - \sum_{j=1}^{21} \text{Ó}2 - \sum_{j=1}^{28} \text{Ó}3$$
$$= 1.429 + 1.362 + 0.542$$
$$= 3.333 \text{ (bits/요소)}$$

3. 코드화한 결과

1) 24 개 자모의 평균 bit 수
(14 개 자음 + 10 개 모음)
자소당 평균 bit 수 : 4.062 bits/letter

2) 51 개 자모의 평균 bit 수
(30 개 자음 + 21 개 모음)
자소당 평균 bit 수 : 4.092 bits/letter

3) 51 개 요소당 평균 bit 수
(30 개 자음 요소 + 21 개 모음 요소)
요소당 평균 bit 수 : 4.435 bits/요소

4)엔트로피 감소 모델의 경우 요소당 평균 bit 수
(19 개 초성 + 21 개 중성 + 28 개 종성)
요소당 평균 bit수 : 3.769 bits/요소

5)엔트로피 감소 모델의 경우 평균 bit 수
(19 개 초성 + 21 개 중성 + 28 개 종성)
(받침 없음경우도 코드를 할당함)
자소당 평균 bit수 : 3.477 bits/letter

이 결과를 분석해보면 먼저 실험 결과에서 나온 감소된 정보량 α'=4.408-3.333=1.075 이다. 이론식 (7)에서 감소된 정보량 α를 구해면 다음과 같다.

ps1 = ps2 = 0.411
ps3 = 0.187
H(ps1) = H(ps2) = 0.527
H(ps3) = 0.443

α = H(ps1)+H(ps2)+H(ps3)=1.497

여기서 $\alpha-\alpha'$ = 0.422 이며 이것은 실험적으로 나타난 감소 정보량 α'이 이론적인 감소 정보량 α보다 적게 나타났으며 그 이유는 $s1 \cap s3 \neq \phi$이기 때문이다. 즉 한글을 51개 자모로 분류하였을때 초성 자음 집합과 종성 자음 집합의 교집합 원소가 존재하기 때문이다.

IV. 결 론

본 논문에서는 여러개의 부집합에서 일정한 규칙을 갖고 정보를 발생하는 정보원의 엔트로피를 감소시킬 수있는 모델을 제안하고, 이 모델에 규칙성을 갖는 정보원을 적용함으로써 일반적인 확률 모델에 적용하는 것보다 엔트로피를 줄일 수 있음을 보였다. 문교부 주관으로 한글의 빈도수를 조사한 데이타 중에서 51개 자모를 일반적인 확률 모델에 적용한 결과 엔트로피가 4.408 bits/요소로 나타났고, 이 데이타를 초성 자음 19개, 중성 모음 21개, 및 종성 자음 27개로 분류하여 엔트로피 감소 모델에 적용한 결과 3.333 bits/요소로 나타나 1.075 bits/요소 만큼 엔트로피가 감소되었음을 알 수 있었다. 또한 51개 자모를 Huffman 코드로 인코딩하였을 때 한글 데이타 압축 결과는 4.092 bits/letter이고, 초성, 중성, 종성으로 분리하여 각각을 Huffman코드로 인코딩했을 때 3.477 bits/letter로 나타나 자소당 평균 0.615 bit를 줄일 수 있었다.

V. 참고 문헌

[1]남궁건, "한글 낱말의 발생 빈도 분포와 Etropy에 관한 연구", 서울대학교 대학원 석사학위논문, 1979.
[2]문교부, "우리말 말수 사용의 잦기조사",문교부, 1956.
[3]송익호,안수길,"2진 1차 Markov 정보원의 엔트로피에 관한 연구",전자공학회지,vol.20,no.2,pp16-22,1983.
[4]안수길,안지환,"공백소를 포함한 한글 자소 발생 확률과 엔트로피",전자공학회지,1980-17-2-4,1980.
[5]이재영,성경모,이종각,"단어.어원 Dictionary에 의한 Text 압축 ", 전기.전자공학 학술대회논문집, pp.607-611,July,1988.
[6]이재홍,오상현, "한글의 초성, 중성, 종성단위의 조건적 발생확률과 엔트로피"한국통신학회,추계 학술 발표회 논문집,1987.
[7]이주근, "한글 문자의 인식에 관한 연구(IV)", 전자공학회지, vol.9, no.4,pp.197-204,Aug.1972.
[8]이주근,박종욱,김창선,"한국어 정보원의 구조 분석과 코오드 개선", 전자공학회지, vol.15, no.2, May,1978.
[9]이주근.최홍문, "한국어 음절의 엔트로피에 관한 연구", 전자공학회지, vol.11, no.3, pp.119-125,June,1974.
[10]A. Feinstein, foundations of Information Theory, McGraw-Hill,1968.
[11]E.S. Schwartz, "An Optimum Encoding with Minimum Longest Code and Total Number of Digits", Information and Control 7, pp.37-44,1964.
[12]C.E. Shannon, "A Mathematical Theory of Communication", Bell System Tech.J., vol.27, pp.379-423,623-656, July 1948.
[13]C.E. Shannon,"Prediction and Entropy of Printed English", Bell System Tech.J., vol.29, pp.147-160, Jan., 1951.
[14]C.E. Shannon, W. Weaver, The Mathematical Theory of Communication, Urbana, ill. Illinois Univ. Press, 1964.
[15]D.A. Huffman, "A Method for the Construction of Minimum Redundancy Codes",Proc.,IRE, vol.40, pp,1098-1101,Sept., 1952.
[16]E.N. Gilbert & E.F. Moore, "Variable-Length Binary Encodings",Bell System Technical Journal, pp.933-967,July, 1959.
[17]F. Jelinek, Probabilistic Information Theory, McGraw-Hill,1968.
[18]G. Dewey, Relative Frequency of English Speech Sounds,Harvard Univ. Press, Cambridge, Mass, 1923.
[19]H.E. White, "Printed English Compression by Dictionary Encoding",Proceeding of the IEEE vol.55, no.3,pp.390-396,1967.
[20]J. Pike, "Text Compression Using a 4-Bit Coding Scheme",The Computer Journal,vol.24, no.4, pp.324-330, Nov., 1981.
[21]R.M. Lea, "Text Compression with an Associative Parallel Processor",The Computer Journal, vol.21, no.1,pp.45-56,Feb., 1978.