디지털 방송용 한글 데이터의 엔트로피 부호화

진경식, 김충일, 황재정
군산대학교 전자정보공학부

Entropy Coding of Hangul Data for Digital Broadcasting

{acejin, pkceio, hwang}@ace.kunsan.ac.kr
Kyoung-Sig Jin, Chung-Il Kim, Jae-Jeong Hwang
Dept. of Radio communication Eng. Kunsan National Univ.

요 약

본 논문은 표준완성형코드를 표준으로 허프만 부호를 생성하기 위해 부호화 효율이 가장 높은 곳에서 예외 부호화를 통해 최적의 허프만 부호를 얻는다. 현재 우리나라의 DTV는 한글문자를 압축하지 않고 전송하는 형태이며, 본격적인 데이터 방송이 시작되면 한글 데이터가 차지하는 전송량으로 인한 심각한 문제가 야기된다. 본 논문에서는 데이터 방송에서 문제가 되는 전송량을 줄이기 위해 한글 전용 최적의 허프만 부호를 생성하여 일련의 해결책을 찾고자 하며 영문 위주인 데이터 압축기술을 한글에 맞게 적용하여 DTV 방송용 한글 전용 압축부호를 만드는데 있다.

1.서론

DTV(Digital TV)에서는 귀선기간에 전송하는 자막방송에 대한 9600bps 정보외에 프로그램 스트림으로 전송되는 데이터방송용 정보를 포함한다. 이것은 최대 19.4Mbps이내에서 데이터량의 크기에 따라 자유롭게 변화시켜 전송하는 것을 의미한다. 한편 미국의 ATSC 에서는 A/65를 제정하였다.[4] 이 표준에서는 영문자에 대한 압축으로 허프만 코드를 사용하고 있으며 또한 타이틀과 내용으로 테이블을 분리, 사용하고 있다.

현재 디지털 방송 위한 한글코드 압축 테이블이 마련되어 있지 않아 현재까지 한글문자를 압축하지 않고 전송하는 형태를 취하고 있어 2001년부터 디지털 방송이 시작되고 데이터방송이 본격화되면 한글데이터가 차지하는 전송량으로 인해 심각한 문제가 야기될 것으로 예상된다.

따라서 본 논문에서는 영문에 대한 허프만 압축기술을 응용하여 한글에 맞게 적용하도록 한다. 먼저 한글코드의 특징을 고찰하고 표준조합형과 표준완성형 및 유니코드를 비교·분석하며, 1차 허프만부호화와 2차 허프만 부호화까지 예외 부호화를 할 때, 각각의 압축율과 테이블 용량면에서 서로 비교·분석을 통해 차수를 선택하여 부호화 효율이 가장 높은 곳에서 허프만 코드를 생성하도록 한다.

2. 한글코드 종류 및 구조

현재의 한글코드는 두 가지(표준조합형과 표준완성형) 모두를 사용하고 있으며 수시로 변환이 필요한 경우가 많다. 내부코드 처리를 위해서는 비교적 간단한 조합형을 이용하고 외부출력을 위해서는 자체의 변화엔진을 이용하여 완성형을 사용한다. 또한 조합형과 완성형에 관한 영역을 같이 배정하고 있는 유니코드도 사용한다.

· 표준조합형코드(KSC5601-1992)인 경우, 자소에 의미를 부여하고 즉, 초성, 중성, 종성으로써 가능한 한글(11,172자)을 모두 표현할 수 있으며 자소 당 5bit를 할당한다.[10][12] 코드체계는 그림 1과 같다.

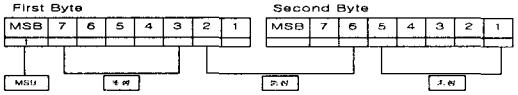

그림 1. 표준조합형코드 체계

· 표준완성형코드(KSC5601-1987)인 경우, 음절에 의미를 부여하고 일반적으로 자주 쓰이는 한글(2350자)만을 표현하며 음절당 2byte를 할당한다.[10][12] 코드체계는 그림 2와 같다.

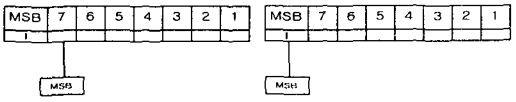

그림 2. 표준완성형코드 체계

· 유니코드(KSC5700)인 경우, 완성형과 같이 음절에 의미를 부여하나 조합형에서 가능한 모든 문자에 대한 코드를 배정하고 있으며 한 음절당 2byte로 표현된다.[10][12] 구조는 그림 3과 같다.

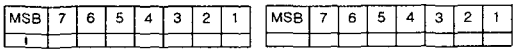

그림 3. 유니코드 체계

3. 텍스트 데이터를 위한 DTV 규격

우리 나라에서는 현재 지상파 DTV 잠정 규격(안)은 미국의 ATSC의 A/65 표준을 따르고 있다. ATSC의 Multiple String Structure의 경우 표 1와 같다.[4]

표 1. Multiple String Structure

Syntax	Bits	Format
multiple_string_structure( ) {		
number_strings	8	uimsbf
for (i=0;i<number_strings;i++) {		
ISO_639_language_code	8 * 3	uimsbf
number_segments	8	uimsbf
for(j=0;j<number_segments;j++) {		
compression_type	8	uimsbf
mode	8	uimsbf
number_bytes	8	uimsbf
for (k=0;k<number_bytes;k++)		
compressed_string_byte[k]	8	bslbf
}		
}		
}		

· 영문데이터의 압축 방식으로는 허프만코딩을 사용하며 압축 종류로는 표 2.에서 보는 바와 같이 두 가지 Table(0x01, 0x02)을 할당되어 있다.[4]

표 2. 압축 종류

compression_type	compression method
0x00	No compression
0x01	Huffman coding using standard encode/decode tables defined in Table C.4 and C.5 in Annex C.
0x02	Huffman coding using standard encode/decode tables defined in Table C.6 and C.7 in Annex C.
0x03 to 0xAF	reserved
0xB0 to 0xFF	user private

4. 허프만 부호

메시지에 의한 디지털 정보원의 평균 정보량(엔트로피)은 다음 식과 같이 정의된다.

$$H(x)=\sum_{i=0}^{n-1} P(x_i)\log_2[1/P(x_i)][bits/symbol]$$

허프만 부호를 만드는 알고리즘의 다음과 같다.

허프만 부호의 알고리즘

1) 각 기호에 대응하는 잎을 만들고 발생확률을 기록한다,

2) 발생확률이 가장 낮은 2개의 잎을 1개의 새 로운 잎으로 결합하여 한쪽에는 "0" 다른 한쪽에는 "1"을 할당하고, 두 발생확률의 합을 새로운 잎에 기록한다.

3) 잎이 1개일 때, 즉 발생확률의 합이 1이 될 때까지 2)를 반복한다.

4) 뿌리에서 각 기호의 잎으로 연결되는 가지에 붙여진 "0"과 "1"의 계열이 그 기호의 부호어가 된다.

허프만 부호는 다음과 식을 만족해야한다.

$$H(x)\le L$$

또한 허프만 부호의 평균길이가 엔트로피보다 어느 선까지 커져야하는지는 크래프트-맥밀란 부등식에 정의되어 있다. 다음 식은 반드시 만족해야한다.

$$H(x)\le L < H(x)+1/n$$

위 식에서 n은 차수 즉, 심벌을 n개씩 블럭화한다는 의미이다.

예의 부호화에 이용된 식으로 부호화 효율은 다음과 같다.[2]

$$\eta = H(x)/L$$

5. 실험결과 및 고찰

디지털 방송에서 발생할 한글데이터는 현재 방송중인 아날로그 방송에서 사용된 데이터와 유사한 것으로 가정하여 방송 3사에서 방송한 뉴스, 드라마, 영화 등 각종 데이터를 수집하여 실험에 이용하였다.

5.1 각각의 분야별 확률분포

각각의 분야별(드라마, 뉴스, 영화)한글의 발생확률을 서로 비교한다.

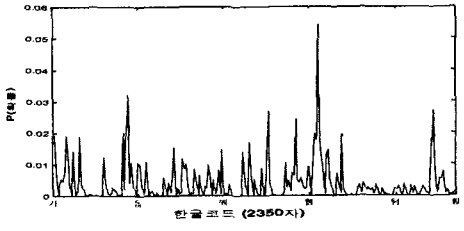

그림 5. 뉴스 발생 분포확률

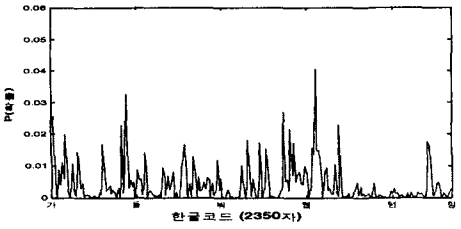

그림 4. 드라마 발생 분포확률

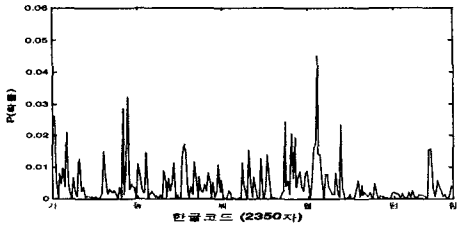

그림 6. 영화 발생 분포확률

각각의 분야별로 발생분포확률을 보면 거의 비슷한 결과를 볼 수가 있다.

5.2 표준완성형과 유니코드 코딩 비교

표준완성형과 유니코드에서 한글 발생분포확률을 비교하면 그림 7와 8와 같다.

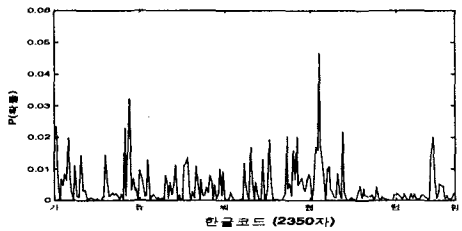

그림 7. 표준완성형코드 발생분포확률

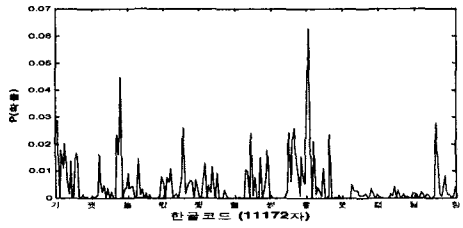

그림 8. 유니코드 발생분포확률

위 실험은 3개의 분야별 집단을 하나로 묶어서 표준완성형은 10자씩, 유니코드는 50자씩 묶어서 비교한 것으로 서로 발생확률이 매우 유사함을 알 수 있다. 그 이유는 표준완성형코드 이외의 한글 문자는 매우 낮은 발생확률을 가지기 때문이다.[12]

표준완성형과 유니코드의 엔트로피(Entroty)와 허프만 평균부호길이(Av_Huf_L)와 평균길이(Av_L)를 비교하면 표 4와 같다.

표 4..표준완성형과 유니코드의 허프만 코딩

구분	표준완성형코드	유니코드
Entropy	7.195822[b/s]	7.203528[b/s]
Av_Huf_L	7.465641[b/s]	7.497881[b/s]
Av_L	7.520426[b/s]	7.532901[b/s]

그림 7, 8와 표 4에서 보는 것과 같이 이 두 코드는 서로 문자코드만 다를 뿐 허프만 코딩에서는 거의 같다고 보면 된다.

따라서 유니코드도 앞으로는 표준완성형코드와 같은 맥락이라고 생각하면 된다.

5.3 표준완성형과 조합형코드 허프만 코딩

표준완성형은 2byte를 하나의 샘플로 하여 허프만 코딩을 하였으나, 표준조합형은 자소 즉, 5bit를 하나의 샘플로 하여 코딩을 한 결과는 표 5.와 같다.

표 5. 표준완성코드와 표준조합형코드의 비교

분야	코드 / 비교내용	조합형코드	완성형코드
뉴스	허프만 평균 부호길이(b/s)	4.9285	7.1129
	압축율(%)	11	47
드라마	허프만 평균 부호길이(b/s)	4.8635	6.9200
	압축율(%)	13	47
영화	허프만 평균 부호길이(b/s)	4.8650	6.9326
	압축율(%)	14	46

위 실험을 통해 표준조합형코드인 경우, 자소당 비트수와 평균부호길이를 비교하면 약간 줄었음을 볼 수 있다. 그러나 표준완성형인 경우 음절 당 비트 수와 평균부호길이를 비교할 때 약 8비트가 감소한 것을 볼 수 있다. 또한 압축율에서도 월등하다는 것을 알 수 있다.

5.4 ECS 코딩 문자 선정

사상의 수가 증가하면 대응하는 허프만 코드의 수와 코드 길이가 증가하게 되며 구현 시 많은 메모리가 필요하게 된다. 따라서 일반적으로 발생확률이 적은 사상에 대해 코드를 부여하지 않고 예외적으로 부호화한다. 이는 전체 평균부호화 길이를 증가시키는데 본 실험에서는 최대의 부호화 효율을 가지는 지점에서의 확률적으로 낮은 심벌을 예외 부호화를 통해 실험을 행하였다.

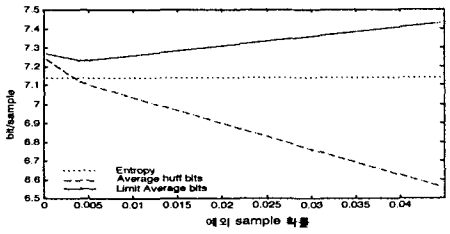

그림 9. 예외 부호화에 따른 평균부호길이

위 실험결과로 가장 높은 부호화 효율을 가지는 지점은 엔트로피와 부호화 평균길이가 차가 가장 적은 지점이 된다. 즉, 그 지점이 이상의 확률을 가지는 사상만 허프만 부호화 적용된다.
전체 허프만 부호화 한 것과 예외 허프만 부호화 한 것을 비교하면 표 6.과 같다.

표 6. 전체와 예외 허프만 부호화 비교

구분 \ data	뉴스	드라마	영화
엔트로피(bits/sample)	7.1129	6.9200	6.9326
전체 코딩 압축율(%)	47	47	46
평균부호길이(b/s)	7.2551	7.0285	7.0773
예외 코딩 압축율(%)	48	49	48
평균부호길이(b/s)	7.2157	6.9828	7.0084

표 6.을 보면 예외 허프만 코딩을 한 경우 평균부호길이가 줄어든 결과를 볼 수 있다. 압축율 또한 감소도 가져왔다.
위 실험결과를 통해 우리는 전체 샘플에 대해 0.43 이하의 확률을 가지는 문자를 허프만 코딩에서 예외코딩을 함으로써 부호화 효율이 가장 높은 허프만 부호를 얻었다. 여기서 허프만 부호화에서 제외되는 문자에 대해서는 허프만 코딩된 ESC코드(특수문자)를 붙여 전송함으로써 압축된 데이터와 구분한다.

5.5 1차와 2차 허프만 코딩의 비교

이 실험에서는 1차와 2차 허프만코딩의 경우 예외 부호화 하지 않은 압축율과 비교한 것이다.

표 7. 차수에 따른 비교

구분 \ 분야	뉴스	드라마	영화
1차의 압축율(%)	42	41	45
1차 디코딩 테이블 용량(kb)	6.8	7.3	6.6
2차의 압축율(%)	72	73	69
2차 디코딩 테이블 용량(kb)	27.5	24.3	42.3

앞에서도 언급되었지만 2차 허프만 부호화로 하면 당연히 압축율이 1차 허프만 부호화보다 뛰어난 것이 보여졌다. 그러나 테이블 용량에서 보면 2차의 경우가 1차보다 더 많은 용량을 차지한다.

6. 결 론

압축방식은 여러 알고리즘이 있으나 압축할 데이터의 특징과 사용목적에 따라 압축 방식이 정해진다고 할 수 있다. 본 논문에서는 다른 알고리즘에 대한 언급은 없었지만 실시간으로 이루어지는 방송용에서는 허프만 알고리즘이외의 다른 알고리즘은 적합하지 않다는 전제하에 한글에 적합한 차수 선택만을 언급되었다.
2차 허프만 부호화로 하면 압축율은 상대적으로 월등히 높으나 인코더와 디코더 테이블의 메모리 용량이 너무 커져 실시간으로 전송이 어렵게 되므로 방송용에서는 적합하지 않다.
따라서 본 논문은 방송용 한글은 1차 허프만 부호화를 통해 디지털 방송용 한글 전용 인코딩 테이블과 디코딩 테이블을 얻어냈다.

참 고 문 헌

[1]황재정, 정동훈, "DTV를 위한 데이터 방송 시스템," 대한전자공학회 하계학술대회, pp. 507-510, 1999. 6.
[2]윤정욱, 박지환 "허프만 블록 부호화의 불규칙성을 효과적으로 구하는 방법". 한국통신학회 하계종합학술발표회 논문집, vol/no. 17/2, pp.1044-1047 1998.
[3]ATSC Interactive services protocols for terrestrial broadcast and cable, Feb. 1999.
[4]ATSC data broadcast spec. Mar. 1999.
[5]Khalid Sayood, Introduction to data compression, Morgan Kaufmann Publishers, 2000.
[6]M.Nelson, J.Gailly,The data compression book, M&T books, 1996.
[7]The Unicode Standard Version 3.0, Addison Wesley Longman, Inc. ISDN 0-201-61633-5, "The Unicode Consortium".
[8]지상파 디지털 TV 실험방송전단반 서브그룹 2 지상파 디지털 텔레비젼 방송 규격(안) v10 2000. 7.
[9]이문호 실용정보이론, 복두출판사, 1998.
[10]한국어정보처리연구소, C로 구현한 한글 코드 시스템 프로그래밍, 도서출판 골-드, 1999.
[11]http://toocan.philabs.research.philips com
[12]http://ncadl.nca.or.kr/HTML/1997/97031/ 97031.htm