허프만 코드의 선택적 암호화에 관한 연구

박 상 호*

요 약

네트워크에서 데이터의 보안은 암호화에 의해 제공된다. 최근 영상이나 비디오와 같은 대용량의 데이터 파일의 암호화의 비용 및 복잡성을 줄이기 위한 방안으로 선택적 암호화가 제안되었다. 본 논문에서는 허프만 코딩으로 압축된 데이터의 암호화 방안을 제안하고 그 성능에 대해 논한다. 허프만 코드에 적용 가능한 단순한 선택적 암호화 방법을 제안하고 불안전한 채널에서 암호화 방법의 효율성에 대해 논의한다. 암호화 과정과 데이터 압축 과정이 하나로 결합될 수 있으며 그렇게 함으로서 데이터를 압축하고 암호화하는 과정을 간략화 하고 시간을 줄일 수 있다.

A Study on Selective Encryption of Huffman Codes

Sangho Park*

ABSTRACT

The security of data in network is provided by encryption. Selective encryption is a recent approach to reduce the computational cost and complexity for large file size data such as image and video. This paper describes techniques to encrypt Huffman code and discusses the performance of proposed scheme. We propose a simple encryption technique applicable to the Huffman code and study effectiveness of encryption against insecure channel. Our scheme combine encryption process and compression process, and it can reduce processing time for encryption and compression by combining two processes.

Key words : Selective Encryption, Huffman Coding

* 안동대학교 전자정보산업학부 부교수

1. 서 론

인터넷 기술의 발달로 오디오, 영상, 비디오 등의 멀티미디어 서비스의 요구가 증가하고 있다. 멀티미디어 서비스는 두 가지 문제점을 해결하여야 하는데 하나는 대용량의 데이터를 저장하거나 전송하기에 적절한 크기로 압축하는 것이며, 다른 하나는 네트워크를 통하여 전송할 때 외부의 침입으로부터 데이터를 보호하는 것이다. 따라서 멀티미디어 서비스의 기본적인 기술은 데이터의 압축 기술이며 멀티미디어 데이터 는 다양한 형태로 압축된 후 저장되거나 네트워크를 통하여 전송된다.

데이터의 압축은 무손실 압축과 손실 압축으로 누눌 수 있다[1]. 무손실 압축은 원 데이터와 압축파일을 디코딩 하여 복원된 데이터 간에 오차가 없는 방식이다. 손실 압축은 원데이터와 압축파일을 디코딩한 복원된 데이터 간에 오차가 존재하나 두 데이터 품질의 차이를 느낄 수 없다. 손실 압축은 압축률이 무손실 압축보다 높아 눈, 귀와 같은 인간의 감각기관을 이용하는 형태의 데이터인 오디오, 영상, 그리고 비디오 등을 압축하는데 사용된다. 무손실 압축은 압축률은 낮으나 복원된 데이터의 무결성을 보장해야 하는 응용분야에 사용된다. 대표적인 응용분야는 진단용 의료영상, 텍스트 등의 압축이며, 또한 손실압축에서 압축된 파일을 추가적으로 압축하는데 사용된다.

허프만 코딩 알고리즘[2]은 무손실 압축 방식으로 구조가 간단하고 압축효율이 높아 멀티미디어 데이터 압축에 많이 사용되고 있다. 허프만 코딩은 JPEG, MPEG 등과 같은 표준안에 채택되어 사용되고 있다. 손실 압축된 데이터는 최종적으로 허프만 코딩과 같은 엔트로피 코딩과정을 거치게 되고 무손실 코딩은 엔트로피 코딩과정을 통해 데이터를 압축한다. 무손실 압축된 파일이나 무손실 압축된 파일이나 최종 데이터는 엔트로피 코드로 이루어진다. 따라서 멀티미디어 신호의 전송을 위한 암호과정은 엔트로피 코딩된 데이터의 암호화라고 할 수 있다. 최근에 Maples 등은 영상데이터를 암호화하기 위하여 모든 데이터를 암호화하지 않고 데이터의 일부분만 암호화하는 방법을 제안하였다[3].

본 논문에서는 엔트로피 코드의 하나인 허프만 코드를 저장하거나 네트워크를 통하여 전송할 때 보안을 위하여 암호화하는 과정을 기술한다. 허프만 코드 중 일부 비트만 암호화 하는 방법을 고찰하고 암호화 효과를 알아본다. 본 논문에서 제안하는 암호화는 비트의 값을 반전하여 정당하지 않은 사용자가 허프만 코드를 복호 화할 때 잘못된 데이터를 출력으로 줌으로서 원래의 데이터를 알 수 없게 하거나, 오디오, 영상, 비디오와 같은 인간의 감각기관과 관계되는 데이터는 데이터의 품질이 급격히 낮아져서 상업적으로 가치가 없게 하는 효과가 있다.

2. 허프만 코딩 알고리즘

허프만 코딩 알고리즘은 최적의 접두사 코드(prefix code)에 관한 아래의 두 가지 관찰에 기초를 두고 있다[1].

① 최적의 코드에서 빈도수(또는 발생확률)가 많은 심벌(symbol)은 빈도수가 작은 심벌보다 짧은 코드를 갖는다.
② 최적의 코드에서 가장 빈도수가 작은 두 심벌의 코드의 길이는 같다.

허프만 코딩 알고리즘을 설명하기 위하여, 순서쌍(ordered pair) $H = (S, P)$로 정의되는 소스 H를 가정하자. S는 심벌 $S = \{s_1, s_2, \cdots, s_n\}$이고 각 심벌의 발생확률 $P(s_i) = p_i$, 여기서 p_{i-1}은 p_i 보다 작거나 같다. <표 1>에 주어진 H_1에 대해 허프만 코드를 생성하는 과정을 보였다. H_1는 7개의 심벌 A, B, C, D, E, F, G로 구성되어 있고 각각의 발생

확률은 43%, 25%, 15%, 7%, 6%, 3%, 1%이다.

〈표 1〉 주어진 H_1에 대한 허프만 코드 생성과정

과정	심벌(발생확률)
1	A B C D E F G 0.43 0.25 0.15 0.07 0.06 0.03 0.01
2	A B C D E S1 0.43 0.25 0.15 0.07 0.06 0.04
3	A B C S2 D 0.43 0.25 0.15 0.10 0.07
4	A B S3 C 0.43 0.25 0.17 0.15
5	A S4 B 0.43 0.32 0.25
6	S5 A 0.57 0.43

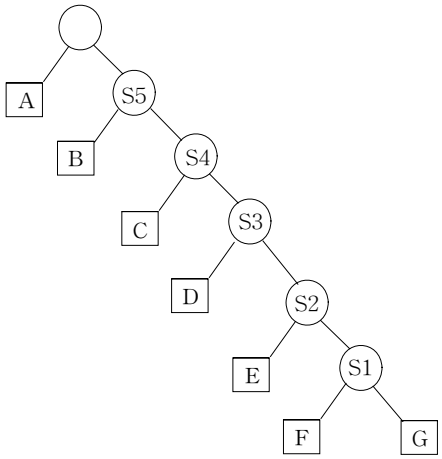
허프만 코드를 생성하기 위한 첫 번째 과정은 심벌을 발생확률에 따라 내림차순으로 정렬하는 것이다. 과정 1에서 정렬된 심벌들 중에서 가장 발생확률이 낮은 두 개의 심벌 F와 G를 묶어 새로운 심벌 S1을 만든다. S1의 발생확률은 F와 G의 확률의 합이다. 과정 2에서 심벌 F와 G 대신 새로운 심벌 S1을 이용하여 발생확률에 따라 새로 심벌을 내림차순으로 정렬한다. 이와 같은 과정은 2개의 심벌만 남을 때 까지 반복한다. 위의 예제에서는 과정 6에 S5와 A의 두 개의 심벌 만 남았다. <표 1>에서 S1과 E를 묶어 S2를, S2와 D를 묶어 S3을 S3과 C를 묶어 S4를, S4와 B가 합하여 S5를 생성하였다.

〈표 2〉 주어진 H_1에 대해 생성될 허프만 코드의 길이

code length	symbols
6	F G
5	S1 E
4	S2 D
3	S3 C
2	S4 B
1	S5 A

<표 1>에서 과정 6의 심벌들은 1-bit 코드, 과정 5의 심벌 중 과정 6에 나타나지 않은 심벌들은 2-bit 코드기 할당된다. 같은 방법으로 모든 과정에 있는 심벌들에 코드가 할당된다. 과정 1의 심벌 중 과정 2에 나타나지 않은 심벌은 6-bit 코드가 할당된다. <표 2>에 생성된 허프만 코드의 길이를 나타내었다.

(그림 1)에 생성된 허프만 코드의 이진 트리를 보였다.



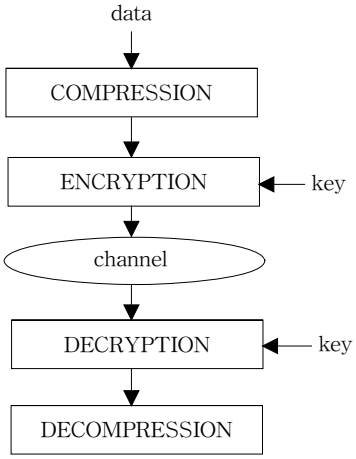
(그림 1) 6레벨 허프만 트리

허프만 트리에서 루트 노드에서 시작하여 각 노드의 좌측은 0, 우측은 1을 할당하면 각 심벌에 할당된 허프만 코드는 A = 0, B = 10, C = 110, D = 1110, E = 11110, F = 111110, G = 111111이 된다.

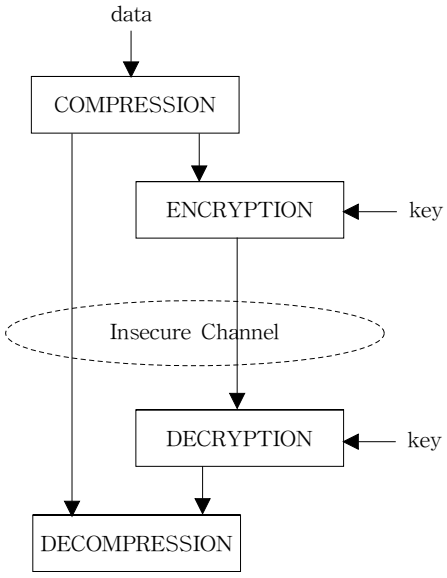
3. 선택적 암호화

안전하지 않은 채널을 통하여 데이터를 전송하기 위하여 메시지의 내용을 암호화하여 보낸다. 멀티미디어 데이터의 암호화 과정은 평문(plaintext)을 암호문(ciphertext)으로 변환하거나[4], 워터마크나 스테가노그래피(steganography)를 사용한다[5].

멀티미디어 데이터는 전송하기 전에 대역폭을 줄이기 위하여 압축과정을 거치게 된다. 손실 압축을 하거나 무손실 압축을 하거나 최종 데이터는 엔트로피 코딩과정을 거친 코드이며 이 코드들은 암호화한 후 전송한다. 이러한 과정이 (그림 2)에 나타나있다.



(그림 2) 데이터의 압축 및 암호화 과정



(그림 3) 데이터 압축 후 선택적 암호화 과정

수신 측에서는 수신된 신호의 암호를 해독한 후 압축된 데이터를 복원한다. 이러한 시스템은 송신 측에서는 모든 전송 데이터를 암호화하여야 하며 수신 측에서는 모든 데이터의 암호를 해독하여야 한다. 따라서 멀티미디어 신호와 같이 데이터의 양이 큰 신호들을 위한 송수신 시스템이 복잡해지고 암호화와 해독에 많은 시간이 소요된다. 이러한 문제들을 해결하기 위하여 최근에 영상 신호를 위한 선택적인 암호화 방법이 제안되었다[3, 6, 7]. 선택적 암호화 방식은 (그림 3)에 나타내었다.

송신 측에서는 원 데이터를 압축한 후 선택적으로 비트 또는 블록을 암호화한다. 수신 측에서는 키를 이용하여 선택적으로 암호화한 비트 또는 블록을 해독하고 원 데이터로 복원한다. 수신 측에서 키를 모르는 경우 암호를 해독하지 않고 압축된 데이터를 복원할 수는 있다. 그러나 복원된 데이터는 압축하기 이전 데이터와는 다른 값을 갖는다. 데이터가 텍스트인 경우에는 부분적으로 다른 내용으로 복원될 것이며, 오디오나 영상 EH는 비디오인 경우 원 데이터가 가진 메시지와 비슷한 메시지를 복원하더라도 데이터의 품질이 그 파일을 이용하기에 충분하지 않아 개인적으로 사용하거나 상업용으로 사용할 수 없어 암호화의 목적을 달성할 수 있다.

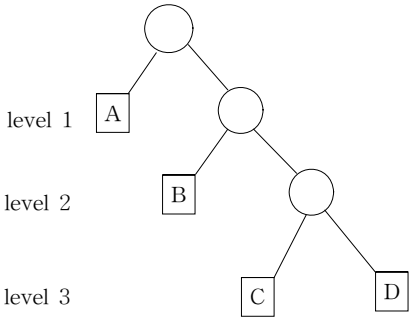
4. 허프만 코드의 선택적 암호화

4.1 허프만 코드의 비트 값의 반전을 이용한 선택적 암호화

허프만 코드는 발생빈도가 높은 심벌에는 짧은 코드가 할당되고 발생빈도가 낮은 심벌에는 긴 코드가 할당된다. 허프만 코드의 할당은 이진 트리로 나타낸다. (그림 4)의 간단한 호프만 트리를 이용하여 허프만 트리에 대해 살펴보자.

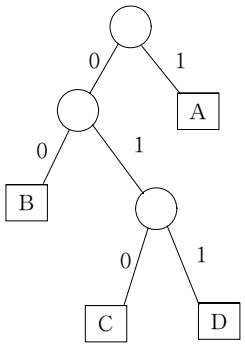
각 심벌에 할당된 코드는 A = 0, B = 10, C = 110,

D = 111이다. 각 코드의 제일 왼쪽에 있는 MSB는 레벨 1에서 할당된 비트이며 왼쪽에서 두 번째 비트는 레벨 2, 세 번째 비트와 네 번째 비트는 레벨 3에서 할당된 비트이다. 각 비트는 할당된 레벨이 같으면 비트의 웨이트가 같다. 레벨의 수를 트리의 높이(height) h라 하면 각 레벨의 L에서 할당된 비트의 웨이트는 2^{h-L}이다.



(그림 4) 3개의 레벨을 갖는 허프만 트리

전송하려는 데이터가 BACBDC라 가정하면 압축 과정을 거친 후 생성된 허프만 코드는 1001101011 1110 이다. 레벨 1에서 할당된 비트의 값을 반전하면 A = 1, B = 00, C = 010, D = 011이 되어 허프만 코드는 001010000011010으로 바뀌게 된다. 비트 반전 후 바뀐 허프만 트리는 (그림 5)와 같다.



(그림 5) 레벨 1비트가 반전된 허프만 트리

수신 측에서 수신된 비트들을 복호 화하여 생성한 데이터는 AACAACB이다. 레벨 1에서 할당된 비트들을 반전함으로서 BACBDC가 수신 측에서는 AACAACB로 해석이 되어 허프만 코드 중 비트를 선택하여 선택적인 비트 반전은 암호화 효과가 있음을 알 수 있다.

레벨 2의 비트를 반전한 경우 허프만 코드는 A = 0, B = 11, C = 100, D = 101로 바뀐다. 따라서 11 0100111011100이 전송되어 수신 측에서는 원래의 메시지가 아니라 CBADACA로 해석한다.

멀티미디어 신호를 엔트로피 코딩한 허프만 코드를 선택적으로 암호화하기 위하여 허프만 코드를 일정한 크기의 블록으로 나누고 각 블록의 비트 열에 일정한 비트패턴을 XOR하여 블록의 비트 열 중 일부분의 비트의 값이 반전되도록 하였다. XOR하는 비트 패턴은 송수신 측에 private key로서 사용된다. 암호화 시스템은 XOR로 수행되므로 암호화 및 암호해독 과정 은 단순하며 송수신측이 동일한 복잡도와 수행시간을 갖는다.

앞의 예제에서 생성된 10011010111110을 7비트의 블록으로 나누면 1001101과 0111110이 된다. 암호화하는 비트 패턴을 0001000이라고 하면 첫 번째 블록은 1000101로 두 번째 블록은 0110110으로 암호화되어 전송되는 비트 열은 10001010110110이다. 송신 측에서 BACBDC라는 메시지를 전송하였으나 암호화한 비트패턴을 알지 못하면 즉 private key가 없으면 수신측은 BAABBCC로 메시지를 해석한다.

4.2 허프만 코드의 선택적 암호화 방안

허프만 코드를 블록으로 묶고 각 블록마다 특정한 비트패턴에 따라 비트의 값을 반전함으로서 암호화할 수 있었다. 본 절에서는 비트 값의 반전을 이용한 다양한 암호화 방안에 대해 고찰한다. 동일한 수 또는 적은 수의 비트를 선택하고 암호화는 더욱 강화하기 위하여 비트들의 블록들을 몇 개씩 묶어 크기가 $M \times N$인 대단위의 블록(대블록)

을 형성한다. 여기서 M은 블록의 크기이고 N은 블록의 개수이다. 각 대블록에 대해 암호화 비트 패턴을 결정하여 허프만 코드로 이루어진 대블록의 비트 열에 XOR하여 암호화 한다. 이때 블록단위로 암호화하는 경우에는 모든 블록에 동일한 비트패턴을 사용하므로 M 비트의 추가적인 정보를 수신 측에 송신해야하며, 대블록을 사용하는 경우 $M \times N$ 비트의 추가적인 정보를 송신하여야 하므로 오버헤드가 증가하게 된다.

5. 실험 및 고찰

허프만 코드의 선택적 암호화 기법의 성능을 평가하기 위하여 심벌의 개수가 9개인 예[8]에서 임의로 데이터를 생성하여 블록을 이용한 암호화와 대블록을 이용한 암호화의 경우 수신 측에서 메시지를 해석할 수 있는 비율을 알아보았다. 실험에 사용한 심벌 및 허프만 코드는 <표 3>과 같다.

〈표 3〉 실험에 사용한 허프만 코드

심벌	발생빈도	허프만 코드
A	48%	0
B	31%	10
C	7%	1100
D	6%	1101
E	5%	1110
F	2%	11110
G	1%	11111

실험을 위하여 <표 3>의 발생빈도 만큼 랜덤하게 심벌을 발생시켰다. 생성된 데이터를 허프만 코드를 이용하여 코딩한 후 다양한 블록 패턴을 이용하여 블록 데이터를 XOR하여 암호화하였다. 대블록을 이용한 암호화의 효과를 평가하기 위하여 $M \times N$ 비트패턴을 이용하여 대블럭을 XOR한 후 암호화 하였다. 블록 비트패턴과 대블럭 비트패턴은 선택하는 비트이 수만 결정하고 비트 패턴은 랜덤하게 발생하였다. 원 데이터와 복원 데이터의 에러의 정도를 나타내기 위하여 원 데이터의 심벌 열과 복원된 데이터의 심벌열 사이에 틀리게 복원된 심벌, 추가로 삽입된 심벌, 또는 제외된 심벌의 개수의 합과 원 데이터의 심벌의 합의 비로 나타내었다. 블록을 이용한 암호화는 10% 정도 비트 값을 반전 하였을 때 50% 정도의 복원 오차가 나타났다. 대블럭을 이용한 암호화는 10% 암호화 하였을 때 70% 정도 복원 오차가 나타났다.

6. 결 론

본 논문에서는 허프만 코드로 압축된 파일의 전송 시 암호화하는 방안을 제안하였다. 암호화시스템의 복잡성과 암호화 시간을 줄이기 위하여 데이터 중 일부만을 암호화 하였다. 허프만 코드의 특성에 따라 비트의 값을 반전하는 것은 암호화 효과가 있음을 보였다. 비트영릉 블록별로 나누어 블록을 암호화 하였고 블록을 묶어 대블럭으로 암호화하는 방안을 실험하였다. 실험결과 10% 정도의 암호화로 70% 정도의 복원오차가 발생하였다.

참 고 문 헌

[1] K. Sayood, Introduction to Data Compression, Third edition, Morgan Kaufmann, 2006.

[2] D. A. Huffman, “A Method for the Construction of Minimum Redundancy Codes”, Proc. IRE, Vol. 40, pp. 1098-1101, Sept. 1951.

[3] T. Maples and G. Spanos, “Performance study of a selective encryption scheme for the security of networked, real-time video”, in Proceedings of the 4th International Con-

ference on Computer Communications and Networks, Las Vegas, Nevada, September 1995.

[4] B. Schneier, Applied cryptography, John Wiley & Sons, Second edition, 1996.

[5] S. Katzenbeisser and F. Petitcolas, Information hiding techniques for steganography and digital watermarking, Artech House, 2000.

[6] M. Van Droogenbroeck and R. Benedett, "Techniques for a Selective Encryption of Uncompressed and Compressed Images", in Proceedings of Advanced Concepts for Intelligent Vision Systems (ACIVS) 2002, Ghent, Belgium, September 2002.

[7] M. Podesser, H-P. Schmidt, and A. Uhl, "Selective Bitplane Encryption for Secure Transmission of Image Data in Mobile Environments", 5th Nordic Signal Processing Symposium, on board Hurtigruten, Norway, October 2002.

[8] R. Hashemian, "Memory Efficient and High-Speed Search Huffman Coding", IEEE Trans., Commun., Vol. 43, No. 10, pp. 2576-2581, 1995.

박 상 호

1979년 경북대학교 전자공학 (공학사)

1981년 영남대학교 전자공학과 (공학석사)

1989년 Syracuse University ECE (MS)

1995년 State University of New York at Buffalo, ECE (Ph.D.)

1996년~현재 안동대학교 전자정보산업학부 부교수