Journal of The Korea Society of Computer and Information
Vol. 19, No. 2, February 2014

www.ksci.re.kr
http://dx.doi.org/10.9708/jksci.2014.19.2.133

수정된 감마 코딩 기반 의료 검진 데이터 압축

구 동 윤*, 박 재 욱*, 이 용 규*

Compression of Medical Examination Data Based on Modified Gamma-Coding

Dong Youn Ku*, Jae Wook Park*, Yong Kyu Lee*

요 약

의료 정보 시스템의 발달로 인해 환자들의 진료 시간이 짧아지고 그에 따라 진료 받는 환자들의 수가 증가하여 진료 데이터 양이 급속도로 증가하고 있으며, 증가하는 환자들의 데이터를 효율적으로 저장, 관리하기 위해 다양한 압축 방법에 대한 연구가 진행 중이다. 그러나 기존 방법들은 측정값을 원시 데이터로 압축하여 저장하기 때문에 압축률이 떨어지는 단점이 있다. 이러한 문제를 해결하기 위해 본 논문에서는 비트 단위로 압축 가능한 감마 코딩 기법을 이용하여 측정값과 정상범위 값과의 편차를 부호화하여 압축하는 방법을 제안한다. 또한 측정값과 편차가 가장 작은 과거 데이터를 기준치로 삼아 그 편차를 부호화하여 압축하는 방법을 제안한다. 제안하는 방법은 매우 간단하며 편차를 압축하기 때문에 기존 방법보다 압축률이 높은 장점이 있다. 성능평가를 통하여 제안한 방법이 기존 압축방법보다 우수하다는 것을 검증한다.

▸ Keywords : 압축, 감마 코딩, 의료 검진 데이터

Abstract

According to the development of medical information systems, shortened examination time per patient could increase the number of treatments, resulting in the rapid growth of the amount of medical data. Studies on how to efficiently compress and store medical text data of increasing patients are in progress. However, previous methods have the shortcoming of compressing medical text data as it is, resulting in low compression rate. This research tries to overcome the problem

•제1저자 : 구동윤 •교신저자 : 이용규
•투고일 : 2013. 06. 18., 심사일 : 2013. 11. 15, 게재확정일 : 2013. 12. 09.
* 동국대학교 컴퓨터공학과-서울(Dept of Computer Science & Engineering, Dongguk University-Seoul)
※ 본 연구는 미래창조과학부 및 정보통신산업진흥원의 산학협력 특성화 지원사업의 연구결과로 수행되었음 (NIPA-2013-ITAH0803130110010001000100100)
※ 이 논문은 2012년도 한국멀티미디어학회 추계학술발표대회에서 '의료 검진 데이터 압축 방법'의 제목으로 발표된 논문을 확장한 것임

by using the gamma coding method which enables compression in bit unit. We propose a new compression scheme which encodes the deviations between measured values and normal range values. Furthermore, we suggest to use the previous value with the least deviation from the measurement as the standard value to encode that deviation. Even though the suggested methods are simple, they have high compression rates. Through performance evaluation, we show that the suggested methods are more efficient than the previous methods.

▸ Keywords : Compression, Gamma-coding, Medical Examination Data

I. 서 론

의료 정보 시스템이 발달하고 진료 받는 환자들의 수가 늘어남에 따라 기록되는 진료 데이터의 양이 급속도로 증가하고 있다. 지속적으로 증가하는 의료 데이터를 저장하기 위해서는 많은 양의 저장 공간이 필요하기 때문에 데이터의 저장 공간을 효율적으로 관리할 수 있는 압축 기술에 대한 연구가 증가하고 있다. 또한 의료기기의 발전으로 인해 원격진료(Telemedicine), 체내이식용 의료기기(Implantable Medical Devices) 등을 이용하여 진료하는 방법이 증가하고 있다. 그러나 원격진료, 체내이식용 의료기기 등은 소용량 저장소 환경이기 때문에 지속적으로 저장되는 환자의 진료 데이터를 저장하기에는 한계점이 있다. 이러한 문제점을 해결하기 위하여 소용량 저장소 환경의 의료기기에서 진료 데이터를 압축하여 효율적으로 저장하기 위한 연구가 활발히 진행중이다[1][2][5][19][20].

데이터의 저장 공간을 줄이는 압축 방법으로는 크게 손실 압축과 무손실 압축 기법으로 나눌 수 있다. 손실 압축은 데이터를 압축하고 복원시킬 때 데이터의 손실이 발생하는 것으로 사진이나 그림, 영상 등 미디어 데이터를 처리할 때 대표적으로 쓰인다. 무손실 압축 기법은 데이터를 압축하고 복원시킬 때 데이터의 손실이 발생하지 않는 것으로 텍스트, 수치 데이터 등에 적용되는 기법이다. 무손실 압축 기법 중 데이터를 부호화 할 때 대표적으로 쓰이는 기법에는 RLE 방법(Run Length Enco-ding), LZW 방법(Lempel Ziv Welch), 옵셋 압축방법(Offset Compression) 등이 있다. 이러한 방법들은 중복되는 데이터를 부호화하는 데 적합하지만 중복되지 않고 계속해서 값이 변화하는 환자들의 데이터의 경우 압축률이 낮아 저장 공간 관리가 비효율적이라는 단점이 있다. 또한 기존 방법들은 알고리즘이 복잡하고 측정값을 그대로 저장하기 때문에 압축률이 떨어지는 단점이 있다. 이러한 문제를 해결하기 위해 중복이 발생하지 않는 데이터를 처리하는 다양한 연구가 진행 중이다[1][2][3][5][15].

중복되지 않는 데이터를 부호화하는 방법에는 감마 코드 인코딩[8][9] 방법이 있다. 감마 코드 인코딩은 비트 단위로 압축하는 방법으로 값이 작은 경우일 때 사용한다. 그러나 부호화할 값이 크면 클수록 감마 코드의 길이가 배 이상으로 증가하여 압축률이 낮아지는 단점이 있다. 본 논문에서는 기존 압축 기법을 개선하고 압축률을 높이기 위하여 두 가지 방법을 제안한다.

첫 번째는 범위가 있는 의료 데이터 중 정상범위에 기준치를 결정하고 환자의 측정값과 기준치와의 편차값을 부호화시켜 저장하는 압축방법을 제안한다.

두 번째는 환자의 측정값과 과거 진료 받았던 데이터와의 편차값을 계산하여 가장 편차가 작은 과거 데이터를 결정하고 그 편차를 압축하여 저장하는 방법을 제안한다.

본 논문에서는 병원의 건강검진 데이터를 이용하여 기존 압축 방법과 제안하는 두 가지 압축 방법에 대한 수검자 수에 따른 평균 비트 수와 각 항목 별 평균 비트수를 비교하여 제안하는 방법의 우수함을 검증한다.

본 논문의 구성은 다음과 같다. 2장에서는 기존의 압축 기법에 대해 설명하고, 3장에서는 제안하는 압축 기법을 이용하여 의료 데이터를 압축하는 방법을 소개한다. 4장에서는 성능평가를 통하여 기존 압축방법보다 제안한 방법의 우수함을 검증하고 5장에서 결론을 맺는다.

II. 관련 연구

병원 내 의료 데이터를 관리하기 위한 기존의 압축 방법에는 환자들의 진료 데이터가 손실되지 않고 정확하게 저장 되어야 하므로 무손실 압축 방법이 많이 쓰인다. 또한 의료 데이터 처리 분야에서 연구가 증가하고 있는 소용량 스토리지 환경에서의 데이터 무손실 압축과 영상 데이터에 대한 무손실 압축이 많이 연구되고 있다.

본 장에서는 무손실 압축 방법에 대한 설명과 의료분야의 소용량 스토리지 환경에서의 텍스트, 수치 데이터를 처리하는 압축 방법과 의료 영상 데이터를 압축하는 기존 방법에 대한 기술과 문제점에 대해 설명한다. 먼저, 무손실 압축 방법에 대한 설명은 다음과 같다.

무손실 압축 방법에는 RLE(Run Length Encoding)[8][11][14][17], LZW(Lempel Ziv Welch)[8][10][13][14][17][18], offset compression[1][3] 등이 있다. 이러한 압축 방법은 수치 데이터와 영상 데이터를 압축할 때 많이 쓰이는 압축 방법들이다.

RLE 방법(Run Length Encoding)[8][11][14][17]은 연속해서 중복되는 문자열을 하나의 문자열로 표현하여 데이터를 줄이는 것으로 압축 기법 중에서 가장 간단한 방법 중의 하나이다. RLE 압축 기법은 데이터를 저장할 때 중복된 데이터가 있을 경우 데이터의 중복된 개수를 카운트하여 저장하는 방식이다. 그러나 병원 내 환자들의 의료 데이터는 측정되는 값이 각각 다르기 때문에 RLE의 중복성을 이용한 압축 방식은 병원 내 서로 다른 의료 데이터를 가진 환경에서는 비효율적이라는 단점이 있다.

LZW 방법(Lempel Ziv Welch)[8][10][13][14][17][18]은 사전 기반 압축 방법으로 자주 중복되어 나오는 값을 테이블에 따로 저장하여 특정값으로 변환한다. 즉, 모든 데이터들을 사전에 미리 저장하여 압축함으로써 중복되는 데이터의 압축률을 높을 수 있다. 그러나 주기적으로 검사받는 환자의 의료 데이터의 경우 사전에 저장되는 데이터의 양이 증가하여 오히려 저장 공간의 관리가 비효율적이라는 단점이 있다.

옵셋 압축방법(offset compression)[1][3]은 측정한 값을 기준 값으로부터의 차이를 구하여 인수로 표현하는 기법이다. 기준점을 하나로 정하여 측정하는 값으로부터 차이를 계산하여 편차만을 저장하면 상대적으로 많은 용량을 줄일 수 있다. 그러나 측정한 값이 기준 값과 차이가 많이 벌어질수록 많은 저장 공간을 소비하게 된다. 그러므로 본 논문에서는 감마 코딩을 이용하여 환자들의 측정값 중 범위가 있는 의료 데이터의 경우 기준점을 늘려 기존의 옵셋 압축 기법보다 우수한 압축방법을 제안한다.

감마 코딩은 어떠한 수치의 간격을 길이와 오프셋 쌍으로 가변 길이 부호화를 하는 것[9]를 말한다. 감마 코딩은 부호화 할 때 비트 수가 두배 이상 증가하는 큰 숫자보다는 작은 숫자를 부호화하기 위한 방법으로 최소한의 비트 수를 가지고 부호화한다. 환자의 의료 검진 데이터의 경우 구하려는 편차 값은 작은 수치에 해당 되므로 비트 압축을 위한 감마 코딩이 적합하다. 감마 코딩은 오프셋과 길이의 쌍과 이것을 구분하기 위한 0이라는 값의 한 비트의 구분점으로 이루어져 있다. 오프셋은 숫자를 이진수로 표현한 것에서 가장 앞의 1을 제거한 값이 표현된 것이고, 길이는 오프셋의 길이만큼을 1의 개수로 표현한 것이다. 본 논문에서는 위의 설명한 감마 코딩 방법을 개선하여 수치 데이터와 영상 데이터를 부호화 할 때 적용한다.

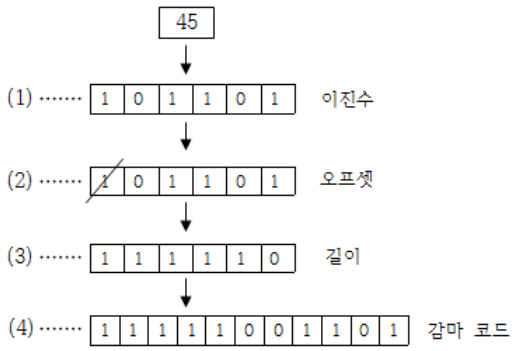


그림 1. 감마 코딩 과정
Fig. 1. Gamma-Coding Processing

다음 그림1은 정수 13을 기존의 감마 코딩으로 부호화하는 과정을 나타낸 것이다.

(1) 정수 13을 이진수로 표현한다.
(2) 이진수의 가장 앞의 1을 제거한 값이 오프셋이 된다.
(3) (2)에서 구해진 오프셋의 길이만큼을 1로 표현한다. 그리고 오프셋과 길이를 구별하기 위한 0값을 마지막 비트에 추가한다.
(4) 구해진 길이와 오프셋을 이어서 감마 코드로 표현한다.

무손실 압축 방법들은 텍스트 데이터, 수치 데이터를 압축하기에 적합한 방법이다. 텍스트, 수치 데이터를 압축하는 방법은 체내이식용 의료기기(Implantable Medical Devices)[19], 원격의료 시스템(Telemedicine)[20] 등과 같은 소용량 스토

리지 환경의 의료기기에서 많은 연구가 진행되고 있다. 이러한 소용량 스토리지를 사용하는 의료기기는 데이터의 전송속도나 저장 공간의 영향을 많이 받기 때문에 전송 속의 개선과 효율적인 저장 공간을 관리하기 위한 요구가 증가하고 있다. 이러한 문제를 해결하기 위해 본 논문에서 제시하는 압축 방법을 응용하여 소용량 스토리지 환경의 의료기기에서 저장 공간을 효율적으로 관리할 수 있다.

또한 기존 압축 방법들은 영상 데이터에도 많이 이용되고 있다. 영상 데이터를 압축하는 대표적으로 기존 방법들 중 대표적으로 가장 많이 쓰이는 방법은 블록 단위 압축 방법[2][6]이다. 블록 단위 압축 방법이란 영상 이미지를 (N × N) 크기의 블록으로 나누어 각 블록이 가지는 화소값을 부호화하는 방법이다. 화소값을 부호화하는 방법에는 영상 데이터에서 중복되는 값을 처리하는 방법과 특정 임계값을 설정하여 처리하는 방법 등이 있다. 중복값을 처리하는 방법은 블록 간의 화소값이 중복될 때 처리하는 방법으로 RLE(Run Length Encoding) 방법을 이용하여 처리한다. 임계치를 설정하여 처리하는 방법은 어떤 특정한 임계값을 정하여 블록의 화소값과 임계값과의 편차를 저장하는 방법이다. 이러한 경우 편차가 작을수록 압축률이 높아지기 때문에 적절한 임계값을 정하는 것이 중요하다.

그러나 기존 영상 압축 방법은 압축하려고 하는 영상데이터와 유사한 과거의 영상데이터가 존재하더라고 이를 고려하지 않고 압축하려고 하는 영상 데이터에 제안된 알고리즘을 적용하여 압축하기 때문에 영상 데이터 간의 압축율은 비슷하다는 단점이 있다. 이러한 문제점을 해결하기 위하여 본 논문에서는 과거 데이터와의 비교를 통해 원본 데이터를 전부 저장하지 않고 편차값 만을 저장하여 압축율을 높하는 방법을 제안한다.

III. 데이터 압축 저장 방법

본 장에서는 급증하는 데이터를 효율적으로 저장하기 위하여 데이터를 압축하여 저장하는 방법을 제안한다.

3.1 기준점을 이용한 압축 기법

저장된 의료 데이터는 크게 정상치의 범위가 있는 데이터와 범위가 없는 데이터로 나눌 수 있다. 범위가 있는 데이터는 혈압과 같이 최저 혈압 정상범위와 최고 혈압 정상범위로 두 개의 정상범위를 가지고 있는 데이터가 있고, 혈당량과 같이 한 개의 범위를 가지고 있는 데이터가 있다. 그리고 범위가 없는 단일 수치 데이터로는 키(cm), 몸무게(kg) 등이 있다. 혈압과 같이 두 개의 정상범위를 가지는 데이터는 각각의 정상범위를 따로 처리할 수 있기 때문에 본 논문에서는 두 개의 정상범위를 가지고 있는 데이터도 각각 하나의 정상범위로 정하여 처리한다. 기준치를 정하는 방법은 정상범위가 그림2와 같을 때 다음과 같이 세 가지로 분류할 수 있다.

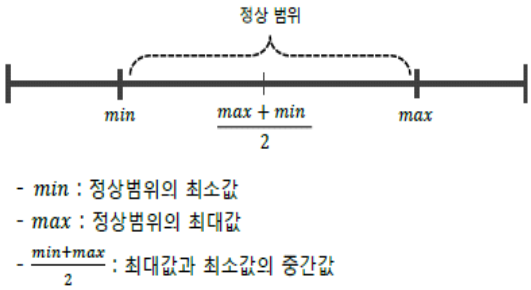


그림 2. 진료데이터의 정상 범위
Fig. 2. Nomal Range of Examination Data

첫 째 측정값이 정상범위의 최소값보다 작은 경우에는 정상범위의 최소값(min)이 기준치가 되고, 둘 째 측정값이 정상범위의 최대값보다 클 경우에는 정상범위의 최대값(max)이 기준치가 된다. 마지막으로 측정값이 정상범위 내에 있는 경우에는 정상범위의 최소값과 최대값의 중간값인 $\frac{\min+\max}{2}$가 기준치가 된다.

기준치가 결정되면 측정값과 비교하여 그 편차만을 저장한다. 편차를 계산하는 방식은 환자의 측정값이 각 정상범위 내에 있는 경우 정상범위의 최대값과 최소값의 중간값(=기준치)보다 크면 중간값으로부터 증가한 값을 저장하고, 작으면 감소한 값을 절대값으로 저장한다.

측정값이 정상범위의 최대값(=기준치)보다 큰 경우에는 최대값으로부터 증가한 값을 저장하고, 정상범위의 최소값(=기준치)보다 작은 경우는 최소값으로부터 감소한 값을 절대값으로 저장한다.

그리고 편차 값을 구하여 인코딩할 때 편차 값이 어느 구간에 속해 있는지를 구별하기 위해 각 구간별로 플래그 비트(flag bit)를 지정한다. 환자의 측정값이 최소값 이하인 경우는 플래그 비트가 00이고, 정상범위 내 중앙값보다 작은 경우는 01, 중앙값보다 큰 경우에는 10, 최대값 이상일 경우는 11으로 지정한다. 또한, 범위가 없는 데이터인 경우에는 기준치보다 작으면 01, 기준치보다 크면 10으로 플래그비트를 지정하고 편차값은 절대값으로 저장한다. 다음 표1은 구간별 플

래그 비트를 설명한 표이다.

표 1. 구간별 플래그 비트
Table 1. Flag-Bit for Each Section

플래그 비트	측정값의 해당 구간	기준치
00	정상범위의 최소값 보다 작은 경우	min : 정상범위의 최소값
01	정상범위 내 중간값 보다 작은 경우	$\frac{\min+\max}{2}$: 정상범위의 최소값과 최대값의 중간값
10	정상범위 내 중간값 보다 큰 경우	
11	정상범위의 최대값 보다 큰 경우	max : 정상범위의 최대값

본 논문에서는 편차가 적을수록 저장하는 비트가 작은 기존의 감마 코딩[9]하는 비트 압축 방식에 플래그 비트를 추가시켜 변형한 방식을 이용한다. 제안하는 감마 코딩 방식은 편차값을 이진수로 변환하고 가장 앞의 값인 1을 제거하여 오프셋을 구성한 후 오프셋의 길이를 유너리 코드로 변환하여 오프셋과 연결하고 가장 앞에 플래그 비트를 추가한다. 다음 그림3은 각 구간별 측정값에 따라 제안한 감마 코딩 방식으로 표현하여 감마 코드 값으로 나타낸 예이다.

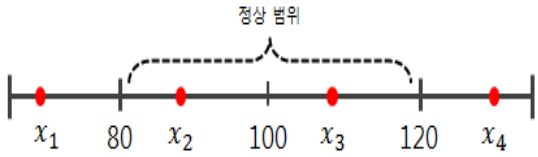
그림 3에서 측정값 x_1의 값이 76인 경우에는 정상범위의 최소값 이하이므로 플래그 비트는 00이 되고, 최소값과의 편차인 4를 감마 코드로 표현한다.

x_2의 값이 91인 경우에는 정상범위 내에 있는 값이고 중앙값보다 작으므로 플래그 비트는 01이 되고, 중앙값을 기준치와의 편차인 9를 감마 코드로 표현한다.

x_3의 값이 113인 경우에는 정상범위 내에 있는 값이고 중앙값보다 크므로 플래그 비트는 10이 되고, 중앙값과의 편차인 13을 감마 코드로 표현한다.

x_4의 값이 122인 경우 정상범위의 최대값 이상이므로 플래그 비트는 11이 되고, 최대값과의 편차인 2를 감마 코드로 표현한다.

편차값을 부호화하여 구해진 감마코드의 비트 수를 구하는 식은 기존의 감마 코딩의 비트 수를 계산하는 식[9]를 변형하여 각 편차값의 비트 수를 계산하며 그 식은 다음의 [수식1]과 같다.



x_n : 측정값 $(n=1,2,3,4)$

측정값	편차값	플래그 비트	유너리 코드	오프셋
x_1 : 76	4	00	110	11
x_2 : 91	9	01	1110	001
x_3 : 113	13	10	1110	101
x_4 : 122	2	11	10	0

그림 3. 구간별 측정값의 감마 코딩
Fig. 3. Gamma-Coding of Measurement Value for each Section

$$len(a,b)=(2\times\lfloor\log_2(|a-b|)\rfloor+3) \quad \text{[수식 1]}$$

(단, $a-b\neq 0$)

-a : 해당 항목의 측정값

-b : 해당 항목의 기준값

위 [수식1]에서 유너리 코드와 오프셋을 구별하기 위해 쓰이는 1비트를 제외하면 유너리 코드와 오프셋은 항상 쌍을 이룬다. 그리고 각 구간을 구별하기 위해 쓰이는 플래그 비트 2비트와 유너리 코드와 오프셋을 구별하기 위해 쓰이는 1비트를 더한 총 3비트를 더한다. 위 식을 이용하여 각 항목에 대한 편차값의 길이를 비교할 수 있다.

3.2 최대 압축 패턴을 이용한 저장

건강검진과 같이 정기적으로 진료를 받는 환자들이 과거에 진료 받았던 데이터는 병원 시스템에 저장된다. 그리고 저장된 환자의 의료 데이터로부터 현재 측정한 값과 비교하여 환자들의 상태를 비교할 수 있다. 그러나 환자들의 건강검진 기록이 계속해서 누적됨으로써 저장 공간을 효율적으로 관리하기 위한 기술이 필요하다. 이러한 문제를 해결하기 위해 두 번째로 제안하는 방법은 현재 측정한 값을 그대로 압축하여 저장하지 않고, 각각의 진료 항목에 대하여 과거 진료 받았던 데이터와 현재 측정값을 비교하는 방법을 이용하여 해결한다. 과거 진료 받았던 데이터와 현재 측정값을 비교하는 방법은 다음과 같다.

먼저, 과거 진료 받았던 데이터와 현재 측정값의 각 항목에 대한 편차를 구한다. 그리고 각 항목에 대해 구해진 편차값을 더한다. 더해진 편차값 중 가장 작은 값에 해당하는 과거 데이터를 기준값으로 선정한다. 마지막으로 선정된 년도의 과거 데이터 항목들과 현재 측정 항목들과의 편차를 구해 편차값을 감마 코딩으로 저장한다.

최대 압축 패턴을 이용한 저장에서는 현재 측정값에서 기준점이 되는 과거 측정값과의 편차값을 구한다. 이 때 편차값은 음수 또는 양수로 표현이 된다. 그러나 각각의 편차값을 더하여 비트 수가 가장 작은 것을 구하는 것이기 때문에 음수값에는 절대값을 적용하여 항상 양수로 계산하여 비교한다.

표 2. 최대 압축 패턴에서의 플래그 비트
Table 2. Flag-Bit in Maximum Compression Parttern

플래그 비트	편차값
0	측정값(현재데이터) 〈 기준값(과거데이터)
1	측정값(현재데이터) 〉 기준값(과거데이터)

위 표2처럼 최대 압축 패턴을 이용한 저장에서는 기준점을 이용한 저장 방식에서 사용한 플래그 비트 방식을 이용하지만 편차값이 두 개의 구간으로 나누어지기 때문에 0 과 1로 표현할 수 있는 최소 1비트를 추가한다. 여기서 음수를 표현하는 플래그 비트는 0이고, 양수를 표현하는 플래그 비트는 1이다.

현재 데이터와 비교했을 때 편차값이 가장 작은 과거 데이터를 결정하기 위해 각 항목의 편차의 합을 구하는 식은 기존의 감마 코드의 비트 수를 계산하는 식[9]를 변형하여 편차값을 감마 코드로 인코딩한 비트 수의 합으로 계산하며 그 식은 다음의 [수식 2]와 같다.

$$len(a,b)=\sum_{i=1}^{n}(2\times \lfloor \log_2(|a_i-b_i|) \rfloor +2) \quad \text{[수식 2]}$$

(단, $a_i-b_i \neq 0$)

- a_i : i번째 검사의 현재 측정값
- b_i : i번째 검사의 과거 측정값
- n : 전체 검사 개수

위 [수식 2]에서 편차가 없는 경우, 즉 과거와 값이 동일한 경우는 편차와 인코딩한 비트의 길이가 0이므로 편차가 0이 아닌 경우에만 계산하고, 제거된 가장 앞의 수를 구하기 위한 1비트와 플래그 비트의 1비트를 합한 2비트를 더한다.

위 수식을 이용하여 편차 값이 가장 작은 즉, 저장해야 할 비트수가 가장 적게 계산되는 과거 데이터를 선택하고 각 검사별 편차값을 제안하는 감마 코드로 부호화하여 저장한다.

제안하는 최대 압축 패턴을 이용한 저장 방법은 환자들의 과거 데이터를 그대로 저장했을 경우 보다 비트 수를 비교했을 때 현저히 줄어들며 편차값이 작으면 작을수록 더욱 효과적인 방법이다.

제안하는 방법은 의료와 관련된 영상 데이터에도 적용 가능하다. 최대 압축 패턴을 이용한 저장 방법을 활용하여 영상 데이터를 압축하는 방법은 과거 측정된 영상 데이터와 현재 측정한 영상 데이터의 각 블록 별 화소값에 대하여 편차값을 계산한다. 그리고 위와 동일한 방식으로 편차값의 합이 가장 작은 과거 영상 데이터를 기준값으로 선정한다. 마지막으로 선정된 과거 영상 데이터와 현재 영상 데이터와의 편차값을 구하여 감마 코딩으로 저장한다.

IV. 성능 평가

기존 진료 데이터 압축 방법과 본 논문에서 제안하는 압축 방법을 비교하기 위하여 성능평가를 실시하였다. 성능평가는 300명의 수검자를 대상으로 한 진료 데이터로 실시하였고 표 3은 본 연구의 실험을 위하여 전체 건강 검진 데이터 중 일부를 나타낸 것이다.

성능평가에서는 전체 건강 검진 데이터 중 일부 항목인 혈압, HDL-C, r-GTP, ALP에 대해서 측정한 데이터만을 추출하였다. 정상범위와의 편차를 부호화하는 압축방법을 평가하기 위하여 표3의 데이터를 이용하여 각각의 측정값과 기준점과의 편차를 구하여 편차값을 부호화 하였고, 최대 압축 패턴 이용방법을 평가하기 위하여 2008년부터 2011년까지의 데이터를 비교하여 측정하였다.

기존방법A는 기존의 감마 코딩을 이용하여 편차값을 구하지 않은 환자들의 측정값을 부호화하여 측정한 것이며, 기존방법B[19]는 RLE압축을 이용하여 의료데이터를 압축하는 방법이다. 제안방법1은 기준점을 이용한 압축 방법이며, 제안방법2는 최대 압축 패턴을 이용한 저장 방법이다. 기존방법A, B와 제안방법을 비교했을 때 제안방법이 우수하다는 것을 증명한다.

표 3. 건강 검진 데이터 일부
Table 3. Part of Medical Examination Data

진료항목 / 년도	**혈압** (최고혈압/최저혈압)	HDL-C	r-GTP	ALP
. . .	. . .	. . .	. . .	. . .
2008년	120/75	90	95	76
2009년	123/77	87	102	77
2010년	118/72	85	117	76
2011년	127/80	101	168	80
현재값	131/84	98	81	74

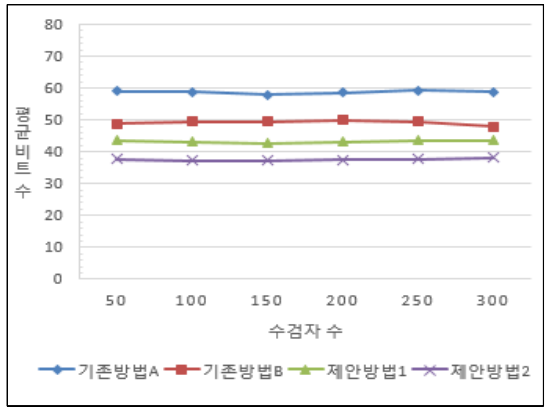
아래 표 4는 제안방법과 기존 압축 방법의 차이점을 요약한 표이다. 제안방법1에서 기준치는 정상범위를 가지는 데이터의 최소값과 최대값 그리고 최소값과 최대값의 중간값으로 총 3개의 기준치를 가지고 측정값은 해당하는 기준치와의 편차값을 구하여 저장한다. 제안방법2는 과거 측정값이 기준치가 되어 현재 측정값과의 편차를 구한 값을 압축한다. 그러나 기존 압축 방법은 원시 데이터를 그대로 감마 코딩하여 저장하기 때문에 많은 비트 수를 저장한다. 제안방법1과 제안방법2는 기준치와의 편차값을 이용하여 저장하는 비트 수를 줄일 수 있는 장점이 있다. 그 결과 같은 값을 압축하여 저장하였을 때 총 저장 비트 수는 플래그 비트가 1비트 더 작은 제안방법2가 제안방법1보다 우수하다는 것을 알 수 있다.

표 4. 기존 압축 방법과 제안방법의 차이
Table 4. Comparison between the Previous Method and the Proposed Methods

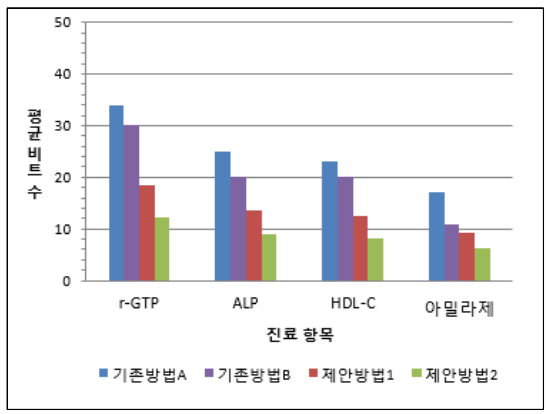
구분	기존방법A	기존방법B	제안방법1	제안방법2
기준치	없음	없음	있음	있음
플래그 비트	없음	1bit	2bit	1bit
*총 비트 수 비교	제안 방법2 〉 제안 방법1 〉 기존 방법			

*비트 수는 같은 값을 기준으로 감마 코딩 하였을 때의 총 저장 비트 수를 말함

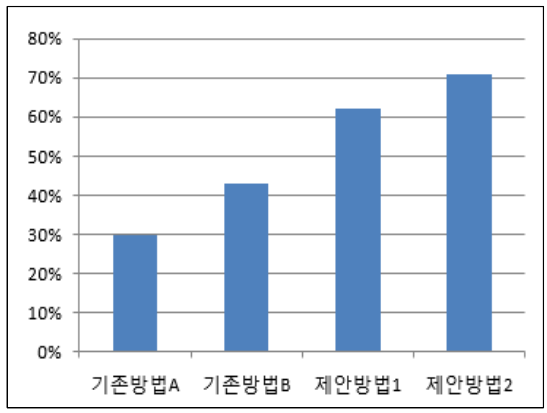
그림4의 (a)는 수검자 수 증가에 따른 일인 당 평균 비트 수를 계산한 결과이고, 그림4의 (b)는 각 진료 항목 마다 저장되는 평균 비트 수를 계산한 결과이다. 그리고 그림4의 (c)는 그림4의 (a)와 (b)에서 보여준 기존방법과 제안하는 압축 방법을 적용했을 때의 압축률을 나타내고 있다.



(a) 수검자 수 증가에 따른 평균 비트 수



(b) 각 진료 항목별 평균 비트 수



(c) 의료 검진 데이터 압축률 비교

그림 4. 압축률 실험 결과
Fig. 4. The Result of Experiments for Compression Rate

위 그림4의 (a)와 (b)에서 제안방법1은 기준값과 측정값의 차이를 저장하는 방법이고. 제안방법2는 최대 압축 패턴을 이용한 방법이다. 성능평가 결과 측정값을 그대로 압축하는 방법인 기존방법A 보다 기존방법B가 더 우수하다는 것을 알 수 있지만 편차값을 사용한 제안 방법1과 2가 기존방법B보다 우수하다는 것을 알 수 있다.

위 그림4의 (c)에서는 보는 바와 같이 제안방법1은 기존의 측정값을 그대로 압축하여 저장하는 기존방법A보다 압축률이 약 32% 개선되었고, 제안방법2는 압축률이 약 43% 개선된 것을 확인하였다. 그리고 기존 진료데이터를 압축하는 방법인 기존방법B보다 제안방법1과 2가 약 18%, 28%가 개선된 것을 확인하였다.

V. 결 론

본 논문에서는 의료 정보 시스템의 발달로 진료 받는 환자들의 수가 급증하면서 환자들의 데이터를 효율적으로 저장하기 위한 압축 방법을 제안하였다.

기준치를 이용한 압축 방법은 범위가 있는 진료 데이터에서 환자의 측정값에 따라 기준치를 선정하여 환자의 측정값과의 편차값을 저장하는 것이다. 기준점으로부터 값의 거리를 구하는 기존 방법에서 기준점을 늘려 더 작은 편차의 값을 저장함으로써 기존의 방식보다 효율적이라는 것을 알 수 있었다.

두 번째 제안 방법은 환자의 현재 측정값과 환자들이 측정했던 과거 데이터와의 각 항목에 대한 편차를 계산하여 편차의 합이 가장 작은 값을 구한다. 편차의 합이 가장 작은 값을 구하면 그에 해당하는 데이터가 기준이 되어 현재 측정값과의 변화량을 저장하였다. 저장 시 인코딩 방식은 비트-압축 방법인 감마코딩을 변형한 식을 이용하여 인코딩 하였다. 측정값을 그대로 저장하지 않고 과거 데이터와의 편차를 이용하여 최소의 비트를 저장함으로써 압축률이 좋아진 것을 알 수 있었다.

기존 방법과 제안하는 방법을 비교하기 위해 환자들의 진료 데이터를 임의로 지정하여 성능 평가를 수행한 결과 기존방법A보다 압축률이 각각 약 32%와 43%가 향상되었다. 또한 기존에 연구되었던 진료 데이터를 압축하는 방법인 기존방법B 보다 압축률이 각각 약 18%, 28% 개선된 것을 알 수 있다. 의료 정보 시스템에서 본 논문이 제안하는 방법은 기존 방법보다 효율적인 것을 확인하였다. 또한 의료 분야의 영상 데이터와 의료 정보 시스템의 데이터와 저장 방식이 유사한 다른 분야에도 적용 가능하다.

참고문헌

[1] Weehyuk Yu, Jongil Jeong, Dongkyoo Shin and Dongil Shin, "Compress transmission of XML-based Clinical Document," Proceedings of Korean Insitute of Information Scientists and Engineers, Vol. 32, No. 2, pp. 250-252, May. 2005.

[2] Myeong-Chan Kim, Yong-Taek Jeong, Tae-Sung Yoon and Young Huh, "New Lossless Compression Method Using Direction Block," Proceedings of Korea Multimedia Society, pp. 206-211, Nov. 1999.

[3] Hyung-Ju Cho and Chin-Wan Chung, "An Efficient Compression Method for Multi-dimensional Index Structures," Journal of Korean Insitute of Information Scientists and Engineers, Vol. 30, No. 5, pp. 429-437, Oct. 2003.

[4] Jeuyoung kim, Yoonhee Kim and Chan Hyun Youn, "Web service based Distributed Medical Data Management," Journal of Korean Insitute of Information Scientists and Engineers, Vol. 34, No. 1(B), pp. 339-343, June. 2007.

[5] Dong Youn Ku, Jae Wook Park and Yong Kye Lee, "Compression Methods for Medical Examination Data," Proceedings of Korea Multimedia Society, Vol. 15, No. 2, pp. 41-44, Nov. 2012.

[6] J.S. Lee, O.S. Kwon, J.Y. Koo, Y.H. Han and S.H. Hong, "A Lossless Medical Image Compression Using Variable Block," Journal of biomedical engineering research, Vol. 19, No. 4, pp. 361-367, 1998.

[7] Jae-Sung Jung and Chang-Hun Lee, "An Efficient Medical Image Compression Considering Brain CT Images with Bilateral Symmetry," Journal of the Webcasting, Internet and Telecommunication, Vol. 12, No. 5, pp. 39-54, 2012.

[8] Guy E. Belloch, "Introduction to Data Compression," Carnegie Mellon University press,

pp. 25-34, Sept. 2010.
[9] Christopher D. Manning and Prabhakar Raghavan, "Introduction to Information Retrieval," Cambridge University Press, pp. 96-100, June. 2008.
[10] Ziv, J. and Lempel A., "Compression of Individual sequences via Variable-Rate Coding," IEEE Transaction on Information Theory, Vol. 24, No. 5, pp. 530-536, Sept. 1978.
[11] Serkan Eryilmaz, "Mean success run length," Journal of the Korean statistical society, Vol. 38, No. 1, pp. 65-71, Mar. 2009.
[12] Lee D. Davisson, "Universal Noiselss Coding," IEEE Transactions on Information Theory, Vol. IT-19, No. 6, pp. 783-795, Nov. 1973.
[13] Ziv, J. and Lempel, A., "A Universal Algorithm for Data Compression," IEEE Transaction on Information Theory, Vol. 23, No. 3, pp. 337-343, May. 1977.
[14] Debra A. Lelewer and Daniel S. Hirschberg, "Data Compression," ACM Computing Surveys, Vol. 19, No. 3, Sept. 1987.
[15] E. Yang, A. Kaltchenko and J.C. Kieffer, "Universal Lossless Data Compression With Side Information by Using a Conditional MPM Grammar Transform," IEEE Transaction on Information Theory, Vol. 47, No. 6, Sept. 2001.
[16] J.C. Kieffer and E. Yang, "Structured Grammar-Based codes for universal lossless data compression," Communication in Information and Systems, Vol. 2, No. 1, pp. 29-52, June. 2002.
[17] Khalid Sayood, "Introduction to Data Compression," Morgan Kaufmann Publishers, Inc. pp. 25-283, 1996.
[18] Welch T., "A Technique for High Performance Data Compression," IEEE Computer, Vol. 17, No. 6, pp. 8-19, 1984.
[19] LA Koyrakh, "Data Compression for Implantable Medical Devices," IEEE Computers in Cardiology, 2008, pp. 417-420, Sept. 2008.
[20] Milanova, M., Kountchev, R., Todorov, V. and Kountcheva R., "New Method for Lossless Compression of Medical Records," IEEE ISSPIT 2008, pp. 23-28, Dec. 2008.
[21] Jung-Yeon Park, "A Study on the Network QoS Management for Telemedicine Service based on Internet," Journal of The Korea Society of Computer and Information, Vol. 7, No. 4, pp. 24-32, Dec. 2002.
[22] Soon-Hyoung Joung and Jong-Ryeol Park, "Study on Telemedicine system in Medical Law," Journal of The Korea Society of Computer and Information, Vol. 17, No. 12, pp. 241-249, Dec. 2012.

저 자 소 개

구 동 윤

2012: 서울호서전문학교
컴퓨터공학과 공학사
현 재: 동국대학교
컴퓨터공학과 석사과정 재학
관심분야: 데이터베이스,
빅데이터 관리, 데이터 압축
Email : hofgee@dongguk.edu

박 재 욱

1999 : 서울과학기술대학교
산업공학과 공학사
2006 : 서울시립대학교 경영대학원
이비지니스학과 경영학석사
2013 : 동국대학교
컴퓨터공학과 공학박사
현 재 : 새마을금고복지회
영업개발팀 과장
관심분야 : 온톨로지, 데이터베이스,
e-비즈니스
E-mail : ssebbok@dongguk.edu

이 용 규

1986 : 동국대학교
전자계산학과 공학사
1988 : 한국과학기술원
전산학과 공학석사
1996 : Syracuse University
전산학 박사
1978년~83년 : 행정직 국가공무원
1988년~93년 : 국방정보체계연구소
선임연구원
1996년~97년 : 한국통신 선임연구원
2002년~03년 : 콜로라도대학교
컴퓨터학과 방문교수
1997년~현재 : 동국대학교
컴퓨터공학과 교수
관심분야 : 데이터베이스,
정보검색, 웹사이언스,
빅데이터 관리
E-mail : yklee@dongguk.edu