부호화 선택기반 시계열 데이터 압축기법

(Encoding selection-based time series data compression)

황 상 호†, 천 승 만†, 석 수 영†, 김 성 호†*
†(재)경북IT융합산업기술원
(Sang-Ho Hwang, SeungMan CHUN, Soo-Young Suk, Sungho Kim)
(†Gyeongbuk Institute of IT Convergence Industry Technology (GITC))

Abstract : In this paper, we implement a lossless compression technique for time series data generated by IoT devices. This technique has improved the Forecasting Module, which is performed to predict time series data in the Sprintz technique, and utilizes the encoding selection technique to reduce the number of bit depths of the predicted time series data. The proposed method reduces the error value through encoding selection and enables higher compression than the existing encoding method in bit packing. In experiments, the proposed method showed a higher compression ratio compared to the Delta and Double Delta methods.

Keywords : IoT, Time series data, Data compression, Lossless

Ⅰ. 서 론

스마트팩토리, 스마트팜, 자율주행 기술이 발전하면서 많은 센서들이 사용되고 있으며 이로 인해 많은 시계열 데이터들이 클라우드로 저장되고 있다. 시계열 데이터는 시간의 흐름에 따라 순차적으로 쌓이는 데이터로 연속된 데이터들 사이에 강한 상관관계를 가지고 있는 특징이 있다. 최근에는 인공지능기술의 발달로 각종 센서들이 다양한 분야에 적용되고, 이로 인하여 생산되는 시계열 데이터가 폭발적으로 늘어나고 있다[1].

단말노드에서 수집되는 시계열 데이터는 데이터 분석을 위해 유/무선 네트워크를 통해 엣지 컴퓨팅 장비 또는 클라우드로 전송이 되며, 데이터 전송에 따라 많은 에너지 소비 및 네트워크 자원을 소비하고 있다. 따라서, 최근 IoT 장비에서 발생하는 에너지 및 자원 소비를 줄이기 위하여 시계열 데이터 압축에 관한 연구가 활발히 이루어지고 있다.[2].

시계열 데이터 압축은 손실 및 무손실 기법이 있으며, 손실 압축은 데이터 복원 시 원본 데이터 정보의 일부가 손실되는 기법이며, 무손실 압축은 원본과 동일하게 복원이 가능한 기법이다. 이러한 특성으로 인하여 일반적으로 손실 압축은 압축율이 높아 자원이 한정적인 분야에서 많이 활용이 되며, 무손실 압축은 원본 데이터가 민감한 정보를 포함하고 있어 손실을 허용하지 않는 분야에 활용할 수 있다. 시계열 데이터에 대한 무손실 압축 기법들 중, Sprintz 알고리즘은 시계열 데이터에 대한 예측을 수행하는 Forecasting, 예측된 값과 실제 값의 차이(error)의 배열에 대한 비트를 줄이는 Bit packing, 0으로 이루어진 값들에 대한 비트를 줄이는 RLE(Run Length Encoding), 마지막으로 엔트로피 부호화로 허프만 부호화를 순차적으로 적용하고 있다[3].

본 논문에서는 IoT 장비에서 생성되는 시계열 데이터에 대한 무손실 압축 기법을 구현한다. 본 기법은 Sprintz 기법에서 시계열 데이터의 예측에 수행하는 Forecasting 모듈에 대한 개선이며, 예측되는 시계열 데이터의 비트 깊이(bit depth)의 수를 줄이기 위해 부호화 선택 기법을 활용하고 있다. 제안하는 기법은 인접한 값에 대한 예측오류의 분산을 줄여 Bit packing에서 기존 부호화 기법에 비해 더 높은 압축이 가능하다.

Ⅱ. 시계열 데이터 압축 기법

시계열 데이터에서 다음 값에 대한 예측에는 인

* 교신저자(Corresponding Author)
황상호, 천승만, 석수영, 김성호 : (재)경북IT융합산업기술원
※ 본 연구는 중소벤처기업부의 규제자유특구혁신사업육성 지원에 의한 연구임 [P0020333]

공지능 등 다양한 알고리즘들이 활용될 수 있으나 일반적으로 단말노드는 전력, 프로세싱 자원 등의 한계로 복잡한 연산이 필요한 기법들보다 Delta 부호화 같은 간단한 수식으로 이루어진 기법들이 적용되어야한다. Delta 부호화는 인접한 데이터와의 차이를 구하여 압축하는 기법으로 연속된 데이터들 사이에 강한 상관관계를 가지고 있는 시계열 데이터에서 상당한 압축효과를 가지고 있어 시계열 데이터를 위한 데이터베이스 등에 많이 활용되고 있다. Double Delta 부호화 기법은 Delta 부호화로 생성된 차이 값에 대하여 Delta 부호화를 적용한 알고리즘으로 꾸준히 증가하거나 감소하는 데이터에서 높은 압축 효과를 가지고 있다.

제안하는 기법은 Delta 부호화와 Double Delta 부호화를 선택적으로 사용하고 있으며 샘플링되는 패킷내 예측에러가 가장 적게 나는 부호화를 선택하여 예측을 수행한다. 그림 1은 제안하는 기법에서 수행하는 Forecasting 절차의 순서를 보여주고 있다.

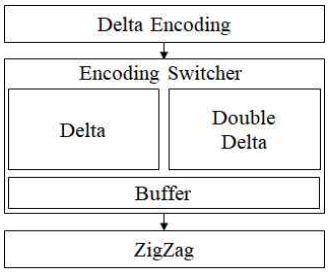

그림 1. Forecasting 절차
Fig. 1. Forecasting procedure

표 1은 그림 1에서 Encoding Switcher의 2가지 부호화에 대한 수식을 보여주고 있다. 표 1의 E_i은 i번째 예측오류를 의미하며 Δx_i는 i번째 차이 값을 의미한다. 상태0은 앞서 Delta 부호화가 이루어진 값을 해당 절차에 그대로 사용하여 Delta 부호화에 해당하며 상태1은 Delta 부호화를 또다시 수행하는 Double Delta 부호화에 해당한다.

표 1. 부호화 수식
Table 1. Encoding formula

0	$E_i = \Delta x_i$
1	$E_i = \Delta x_i - \Delta x_{i-1}$

표 1에 의해 부호화가 완료된 값은 ZigZag 알고리즘을 통해 부호비트 위치를 변경하고 Bit packing이 이루어진다. 샘플링된 패킷의 헤더에는 1bit 크기의 플래그가 추가되며 이 플래그에는 해당 패킷에 적용된 부호화 기법을 표시한다. 클라우드로 보내진 압축 데이터는 수행한 부호화 절차를 역으로 수행하여 복호화된다.

Ⅲ. 실험 및 분석

1. 데이터 및 실험환경

제안한 기법의 성능비교를 위해 본 논문에서는 Gas sensor array temperature modulation 데이터 셋을 활용하였다[4]. 이 데이터 셋은 UCI에 공개된 데이터로 공기 상태에 대한 시계열 데이터를 포함하고 있다.

시계열 데이터 셋의 샘플링 간격은 8개로 설정하였으며 성능비교에는 원본데이터, Delta 및 Double Delta 부호화기법을 활용하였다.

2. 실험결과

그림 2는 각각의 시계열 데이터 압축 후의 파일 크기를 비율로 보여주고 있다.

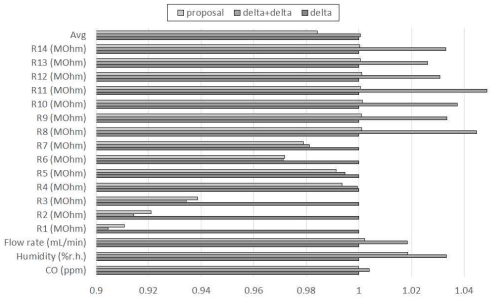

그림 2. 정규화된 파일 크기 비율
Fig. 2. Normalized file size ratio

그림 2에서의 파일크기는 Delta 부호화를 수행한 후의 파일크기를 기준으로 정규화를 한 결과를 보여주고 있다. 제안하는 기법은 Delta, Double Delta기법과 비교하여 평균적으로 약 2%의 성능 향상을 보이고 있다.

Ⅳ. 결 론

본 논문에서는 IoT 장비에서 생성되는 시계열 데이터에 대한 무손실 압축 기법을 구현한다. 본 기법은 Sprintz 기법에서 시계열 데이터의 예측에 수

행하는 Forecasting 모듈에 대한 개선이며, 예측되는 시계열 데이터의 비트 깊이의 수를 줄이기 위해 부호화 선택 기법을 활용하고 있다. 제안하는 기법은 인접한 값에 대한 예측오류의 분산을 줄여 Bit packing에서 기존 부호화 기법에 비해 더 높은 압축이 가능하다. 실험에서 제안하는 기법은 Delta, Double Delta기법과 비교하여 압축 비율이 높음을 보였다. 향후 부호화 선택 알고리즘 내 부호화 기법에 대한 개선을 진행할 예정이다.

참 고 문 헌

[1] Chen, Jianguo, et al. "A periodicity-based parallel time series prediction algorithm in cloud computing environments." Information Sciences, Vol. 496, pp.506-537, 2019.

[2] Azar, Joseph, et al. "Robust IoT time series classification with data compression and deep learning." Neurocomputing, Vol. 398, pp.222-234, 2020.

[3] Blalock, Davis, Samuel Madden, and John Guttag. "Sprintz: Time series compression for the internet of things." Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, Vol. 2, No. 3, pp.1-23, 2018.

[4] J. Burgués, J.M. Jiménez-Soto, and S. Marco, "Estimation of the limit of detection in semiconductor gas sensors through linearized calibration models", Analytica chimica acta, Vol. 1013, pp. 13-25, 2018.