대규모 IoT 환경에서의 중복 및 비정상 데이터 처리 기법

김민우[O*], 이태호[*], 이병준[*], 김경태[**], 윤희용[*]
성균관대학교, 정보통신대학 전자전기컴퓨터공학과[O]
성균관대학교, 정보통신대학 전자전기컴퓨터공학과[*]
성균관대학교 소프트웨어대학 소프트웨어학과[**]
e-mail: {kimmw95, leetaeho, byungjun}@skku.edu[O*], kyungtaekim76@gmail.com[**], youn7147@skku.edu[*]

Redundant and Abnormal Data Processing Scheme in Large-scale IoT Environment

Min-Woo Kim[O], Tae-Ho Lee[*], Byung-Jun Lee[*], Kyung-Tae Kim[**], Hee-Yong Youn[*]
Dept. of Electrical and Computer Engineering, Sungkyunkwan University[O]
Dept. of Electrical and Computer Engineering, Sungkyunkwan University[*]
Dept. of Software, Sungkyunkwan University[**]

● 요 약 ●

최근 IoT 환경에서는 고밀도로 노드가 분포되어진다. 이러한 센서 노드들은 데이터 전송 시 혼잡을 초래하는 중복 데이터를 생성하여 데이터의 정확도를 저하시킨다. 이에 따라 본 연구에서는 데이터 집중으로 인해 발생하는 네트워크의 정체 문제를 해결하기 위해 제안 기법은 사 분위(Interquatile, IRQ) 분석과 코사인 유사도 함수를 통해 데이터의 이상치와 중복성을 측정하여 중복 데이터 및 특이치를 제거한다. 본 연구를 통하여 최적의 데이터 전송을 통하여 IoT의 통신 성능을 향상시킬 수 있으며 결과적으로 데이터 감소율, 네트워크 수명 및 에너지의 효율성을 높일 수 있다.

키워드: IoT(Internet of Things), 사 분위 분석(Interquatile Analysis), 대용량 데이터(Big Data), 코사인 유사도 함수(Cosine Similarity Function)

I. Introduction

최근 IoT(Internet of Things)와 Big Data 관리 기술의 급격한 성장으로 인해 수많은 노드의 효율적인 데이터 전송을 가능하게 하며 계산 복잡도를 감소시켜 데이터 처리비용을 줄이기 위한 대용량 데이터의 전처리 단계 기술이 중요한 이슈이다. 이러한 전처리 기술은 수집된 데이터의 정확한 분류와 중복데이터 제거 및 복잡한 데이터 단순화 등의 분석 기술들이 요구된다[1]. 이때, 학습 데이터 세트의 정확성을 높이고 데이터 스케일의 관리를 통한 효율적인 계산 복잡도를 제공해야한다. 이를 위해 본 논문에서는 사 분위(Interquatile) 분석과 코사인 유사도 함수를 통해 CH(Cluster Head)에서 데이터의 이상치와 중복성을 측정하여 중복 데이터 및 특이치를 제거한다. 결과적으로 네트워크의 전체적인 효율성을 증가시킬 수 있으며 전처리를 기반으로 최적화 된 데이터 전송을 통해 IoT 기기의 통신 성능을 향상시킬 수 있다.

II. Preliminaries

1. Related works

수집되어진 파일 시스템들은 흔하게 데이터 중복 현상이 발생한다. 또한 센서 노드를 통해 수집된 데이터는 간혹 이상치 데이터 값을 저장하는 오류를 범하기도 한다. 이는 같은 값을 여러 번 저장하고 전송하며 지우는 과정 등에서 중복되는 데이터와 이상치 값들이 생겨난다. 데이터 중복제거 기법은 크게 두 가지 방법으로 나누어진다 [2]. 먼저, 데이터가 저장된 이후에 중복 데이터를 제거하는 방법이 있으며 실시간 적으로 데이터가 저장소에 저장되기 전에 중복제거가 이루어지는 방법이 있다. 첫 번째 방법은 데이터 검출을 위해 다시 데이터를 읽어야하며 불필요한 저장으로 인해 불필요한 에너지 소모와 여분의 저장 공간이 요구된다는 단점이 있으며 실시간 처리는 중복데이터 제거를 위해 처리 비용 및 성능에 오버헤가 있을 수 있다. 본 연구에서는 오버헤드로 인한 이상치 값을 최소로 하고 높은 전송정확도와 효율을 위해 노드에 저장된 값들을 측정하여 제거한 후 전송하는 방법을 택하였다.

III. The Proposed Scheme

데이터에서 특이치를 검출하고 제거하기 위해 사 분위(IRQ) 분석이 사용되었다. 아래의 그림은 사 분위 기법의 다이어그램을 나타낸다.

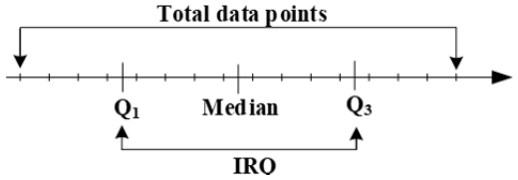

Fig. 1. 사 분위 기법의 다이어그램

아래의 식에서 Q_1과 Q_3는 위 그림과 같이 샘플 데이터의 첫 번째 사 분위와 세 번째 사 분위를 각각 나타낸다.

$$IRQ = Q_3 - Q_1$$

이때 데이터 포인트가 이상 값인지 여부를 결정하기 전에, 먼저 데이터 포인트의 예상수치를 아래식과 같이 식별한다. U_r 및 L_r은 각각 특이치를 식별하는 데이터 포인트의 상한 및 하한 범위를 나타낸다.

$$L_r = Q_1 - (1.5 \times IRQ),\ U_r = Q_3 + (1.5 \times IRQ)$$

가설(Hypotheses):

$$\phi : x_i < L_r \text{ or } x_i > U_r$$

$$\sigma : x_i = y_m$$

가설 ϕ에서 x_i는 이상치 값으로 입력 데이터 포인트 x_i의 중복성을 측정하기 위해 코사인 유사성 함수가 다음과 같이 사용된다.

$$cs(X,Y) = \frac{\sum_{i=1}^{n}\sum_{k=1}^{m}(x_i \cdot y_{i+k})}{\sqrt{\sum_{i=1}^{n}(x_i)^2}\sqrt{\sum_{k=1}^{m}(y_k)^2}}$$

$cs(x_i, y_m) = 1$이면 가설 σ가 받아들여지고 데이터 점 x_i는 중복됨. 결과적으로, 이상치와 중복 데이터는 아래 식에 의해 삭제된다.

$$Y^* = Y - \phi - \sigma$$

IV. Conclusions

본 논문에서는 효율적인 데이터 전송을 위해 중복 및 이상치 데이터를 제거하는 사 분위 분석 기법과 코사인 유사도 함수를 사용한 전처리 기법에 대해 제안하였다. 본 연구를 통해 대규모 노드로 구성된 환경에서의 네트워크 전송 속도와 효율성을 증가시킬 수 있으며 높은 QoS를 얻을 수 있을 것으로 예상된다. 향후 본 연구를 기반으로 성능 평가를 진행할 예정이며, 차후 본 기법을 적용하여 edge computing 환경에 사용되는 모델과 결합하여 더 뛰어난 성능을 기대해 볼 수 있다. 이는 네트워크 트래픽 및 에너지 효율성에 큰 영향을 미치며 데이터의 전송 양을 줄일 수 있다.

ACKNOWLEDGEMENT

본 연구는 과학기술정보통신부 및 정보통신기술진흥센터의 정보통신·방송연구 개발 사업(No. 2016-0-00133, 초연결 IoT 노드의 군집 지능화를 통한 Edge Computing 핵심 기술 연구), SW중심대학지원사업(2015-0-00914), 한국연구재단 기초연구사업(No.2016R1A6A3A11931385, 실시간 공공안전 서비스를 위한 소프트웨어 정의 무선 센서 네트워크 핵심기술 연구, 2017R1A2B2009095, 실시간 스트림 데이터 처리 및 Multi-connectivity를 지원하는 SDN 기반 WSN 핵심 기술 연구, 2019R1I1A1A01058780, 머신러닝 기술을 사용한 SDN기반 무선센서네트워크의 효율적 관리), BK21PLUS 사업의 일환으로 수행되었음.

REFERENCES

[1] Data Deduplication Technolopy Review [Online]. http://www.computerweekly.com/report/Data-Deduplication-technolopy-review

[2] Data Deduplication, [Online]. http://en.wikipedia.org/wiki/Data_deduplication